Stopping the Revolving Door: MDP-Based Decision Support for Community Corrections Placement

Xiaoquan Gao
School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, gao568@purdue.edu

Pengyi Shi
Mitchell E. Daniels, Jr. School of Business, Purdue University, Lafayette, IN 47907, shi178@purdue.edu

Nan Kong
Weldon School of Biomedical Engineering, Purdue University, Lafayette, IN 47907, nkong@purdue.edu

Community corrections (CC) offer an alternative to incarceration that can reduce jail overcrowding and recidivism rates. The aim is to address the root causes behind criminal behavior, ultimately breaking the cycle of reincarceration. However, placing all eligible individuals in CC may can strain case managers, resulting in reduced supervision, increased violations, and higher recidivism rates, which undermines the intended benefits for all participants in the programs. We take the first step to build a comprehensive analytical framework based on a queueing system that incorporates various system- and individual-level factors. We develop a Markov Decision Process (MDP) to systematically study the intricate tradoffs among individual recidivism risks of individuals and the negative effects of overcrowded jail and CC programs. Unlike conventional queueing routing problems, our model incorporates salient features in the criminal justice setting, including deterministic service times (sentenced length) and convex holding and violation costs that vary with program occupancy, which present significant theoretical and algorithmic challenges. To tackle the theoretical challenge, we develop a unified approach with system coupling and policy deviation bounding to compare value functions, establishing the superconvexity of the value function. This provides a theoretical basis to innovate an efficient algorithm based on a separation of time scales. We showcase the effectiveness of our algorithm in solving the originally intractable real-world problems through a case study using data from our community partner. Our study provides valuable policy insights, highlighting the significance of our approach in breaking the cycle of recidivism and offering guidance for capacity planning.

Key words: Criminal Justice, Social Good, Two-time-scale, Actor-critic Algorithm
History:

1. Introduction

Jail crowding is a prominent challenge to the current criminal justice systems in the United States (U.S.), causing various negative consequences on individuals and their communities (Specter 2010). The number of incarcerated people per capita has more than quadrupled in the U.S. over the past 40 years. In recent years, there has been a record high of 2.3 million individuals incarcerated in prisons and jails, with the jail population reaching 0.75 million (Minton and Golinelli 2013).

What makes the overcrowding issue worse is the well-known "revolving door" phenomenon in the criminal justice system. Over 75% of individuals released from incarceration are re-arrested within five years (Alper et al. 2018). These released individuals often have difficulties in accessing housing, education, job and rehabilitation opportunities, reducing their chances of successful reintegration into society. Consequently, this increases the likelihood of recidivism, i.e., relapsing into criminal behavior shortly after release (Spohn and Holleran 2002). This creates a *vicious cycle*, in which the elevated rates of repeat offenses result in an increased number of people being incarcerated, exacerbating the already critical problem of jail overcrowding.

Given the serious negative effects as above, as well as the often staggering expense of jails and prisons to governments and taxpayers, addressing this vicious cycle between jail overcrowding and recidivism has become a top priority for many policymakers (MIT News 2017). The Community Corrections Act in the U.S. was established in 1979 for the purpose of encouraging counties to develop a coordinated local criminal justice system in order to divert from traditional incarceration (Pitts et al. 2014). According to the Indiana Department of Correction, community corrections (CC) "is a community-based supervision agency used for the primary purpose of providing sentencing alternatives for felony offenders in lieu of incarceration." CC programs go beyond mere diversion; they play a pivotal role in enhancing the long-term health and well-being of participants by addressing the underlying root causes behind criminal behavior. For example, CC programs provide individuals with access to services such as substance use disorder (SUD) treatment, mental health services, and vocational training. These programs help participants seek jobs, improve life skills, and ultimately increase their chances of success on reintegration into the society (Prins et al. 2009). Studies show that CC programs help improve long-term community safety and reduce crime (Duwe and McNeeley 2021); this reduces the need for costly new prison construction and the burden on overcrowded jail facilities, saving taxpayer money.

While CC programs could help decrease long-term recidivism rates, they are not a universal remedy suitable for all individuals. For example, since CC programs do not impose strict incarceration, individuals placed in CC may still face the risk of reoffending while participating in these programs (Hawken 2016). Moreover, the effectiveness of reintegration is closely tied to the caseloads of CC case managers. Current practices primarily follow a risk-based approach, often utilizing acceptance criteria based on risk factors (Desmarais and Lowder 2020). For cases that are not a clear cut, additional factors including jail or CC occupancies are taken into consideration beyond the risk assessment (Garrett et al. 2019), sometimes resulting case-by-case outcomes and not necessarily the same for individuals who appear to be similar. However, even with all these more nuanced considerations, placement decisions tend to be isolated in current practice, focusing on one individual at a time. They often do not take into account the overall status of the jail and

CC programs, as well as the potential impact of placing this one individual on all other existing participants. This can cause unintended consequences. For example, indiscriminately placing all eligible individuals to CC can shift crowding from jail to CC, overburdening case managers and compromising supervision, thereby increasing technical violations and reoffenses. This can undermine the benefits of CC programs and even exacerbate the cycle of incarceration (see discussions in Section 3). Therefore, the decision on whom to divert to CC is far more nuanced than an individual assessment or following a rigid rule – it must strike a balance among various tradeoffs and avoid unintended consequences caused by overlooking the system effect. In essence, it necessitates proactively considering how the placement decision for one individual impacts themselves *and* all other existing participants in CC or jail, especially for those cases that fall into a "gray area," the incarceration or diversion decision for whom is not clear. While adding jail capacity or increasing CC staffing whenever necessary may seem like the obvious solution to avoid any of the tough decisions, this option is not always feasible due to limited state funding and various other constraints. It necessatiates a sophisticated, system-wide view to understand the various tradeoffs comprehensively and develop decision support to effectively use existing resources for the benefit of both individuals and the broad society in a dynamic environment characterized by stochastic fluctuations. Furthermore, gaining a thorough understanding of placement decisions serves as a crucial stepping stone for capacity planning at a strategic level.

1.1. Paper Overview and Contribution

We take an important first step to develop an analytics-based approach to study placement decisions for individuals in the incarceration setting. Our approach offers a holistic, system-wide perspective that differs from current siloed and experience-based decision-making practices. Our analytical framework considers the complex interplay between individual and system-level factors that have not been systematically studied before, presenting a pioneering step and filling a critical gap in this important public sector area that has lacked OR analytics support. Our research is also practically grounded based on close collaboration with the community partner – Tippecanoe County Community Corrections (TCCC) in Indiana, USA, with ongoing implementation and research translation from this work. We provide an overview of our work and summarize both our theoretical and practical contributions as follows. We stress that criminal justice is a very complicated system, and we take an initial stride in this paper to understand the incarceration diversion from a system's view. We make a few important remarks on how to interpret results from our analytical model at the end in Section 1.1.2.

1.1.1. Technical contributions.

I. New Modeling Framework. We build a modeling framework that comprehensively examines the tradeoffs involved in community correctional placement. These tradeoffs consider both individual-level (recidivism and violation risks in the short- and long-term) and system-level factors (staffing needs and the negative impact of overcrowding on all participants). Prior works primarily focus on reducing jail overcrowding and often overlook the CC perspective (Usta and Wein 2015), or solely consider individual risks and ignore the *"externalities"* brought by placing the current individual on other existing participants in the jail or CC programs (Brogan et al. 2015). In Section 3, we specify the important individual- and system-level factors we consider and the tradeoffs. Our MDP model, as described in Section 4, effectively integrates these intricate trade-offs through a novel cost modeling approach. Specifically, to connect individual-level considerations and system-level impacts within the decision making, we transform the costs associated with individual recidivism and violations into a convex cost structure based on program occupancy (in addition to the traditional holding cost), where the occupancy is risk-dependent. This approach allows us to capture the increasing negative impact (the externalities) when assigning an individual to an already crowded program, while still differentiating based on individual risk levels. In addition, we account for salient features in this application context, including (a) pre-set and deterministic sentence lengths, which render the system non-memoryless and necessitate the tracking of the length of stay (LOS) for each participant; and (b) different recidivism windows, as individuals can re-offend during CC programs but not in jail. These features bring further complication due to their interplay with the occupancy and resulted tradeoff. These new features not only set our MDP problem apart from classic queueing routing problems but also establish a contextual foundation for future research in this important area of public policy.

II. Unified Approach for Value Function Comparison. We investigate the interplay among different tradeoffs via structural analyses of the MDP policy, including when CC placement is preferred under different system congestion levels, individual risk levels, and recidivism windows. Most importantly, we establish the superconvexity property of the value function of the "right" state variables (see Section 5.1 for counterexamples using insufficient state variables). These analyses are much more complicated than conventional structrual analyses due to the unique features of our model (deterministic service times and convex costs). We develop a unified approach that leverages sample path coupling and policy deviation bounding techniques to decompose and compare the optimal value functions starting at different states. This technical innovation avoids the tediousness of using the traditional operator methods and directly analyzes the value function via coupling and monotonicity results, which has independent technical interest for analyzing similar systems beyond our application. This structural analysis result is significant since it provides not just valuable insights but also establishes a theoretical foundation for our algorithmic development.

III. Two-Timescale Algorithmic Solution. To solve the placement decision problem in realistic settings, we need an algorithmic solution. However, the non-memoryless and long LOS significantly enlarge the state space and make the problem computationally intractable. To overcome this challenge, we propose a scalable actor-critic policy gradient algorithm that leverages a separation of timescales uniquely to our domain application. Specifically, on the slow time-scale, we transform the original MDP into a form that (i) considers batched decisions, for which global optimality is guaranteed by the established superconvexity results; (ii) allows us to recover the transition dynamics accurately without tracking individual LOS, significantly reducing the state space. We demonstrate the effectiveness of our algorithm in solving the originally intractable real-world problems via a comprehensive case study using data from our community partner. By comparing against benchmark algorithms, we also show that neither a pure risk-based policy nor a static threshold policy is sufficient to address the complex tradeoffs involved in the placement decision problem. A careful balance among the factors as captured by our decision framework is necessary to achieve the desired benefits of CC programs – stopping the revolving door of incarceration. We elaborate more on the implication of our decision support and our practical contributions next.

1.1.2. Results Interpretation and Practical Contributions While this paper presents an important step forward in understanding and optimizing placement decisions in the community corrections setting, we recognize that our analytical framework is a simplified abstraction of a highly complex and multifaceted reality. The criminal justice system is inherently complex and involves various legal, social, and individual factors that cannot be fully captured in any mathematical model. Thus, before getting into the details in the rest of the paper, we want to be upfront and make a few important remarks here on how to interpret our analytical results and how our decision-support should be viewed. This also highlights the broader practical contribution of this work beyond the modeling and algorithm development.

First and foremost, our framework prescribes state-dependent placement decision recommendation, which may raise concerns about consistency and fairness in the placement process: how can an individual be placed to jail instead of CC just because CC is crowded? The answer is far more multifaceted than what it seemed to be. It is crucial to emphasize that for clear-cut cases such as those involving serious violent offenses or first-time low-level drug offenses, human (judge) experiences are likely to be sufficient. In these cases, our placement recommendations are mainly driven by individual cost balance and will most likely be consistent with the practice. However, in cases where an individual falls into the "gray area," a lot more factors need to be taken into consideration, hence, the decisions are not necessarily consistent for individuals alike even in practice. Indeed, system congestion is a considering factor as current prosecuting guidelines recommend

considering the availability of CC programs (Indiana Prosecuting Attorneys Council 2019) and the growing emphasis on diverting more individuals due to jail overcrowding (Davis 2019, Gaines and Miller 2021). Despite having these guidelines, the criminal justice system is known for its siloed nature, where various stakeholders operate independently (Hamilton 2010). Hence, the detrimental effects of system congestion tend to be overlooked (e.g., judges may not fully grasp the extensive workload of probation officers) until severe overcrowding becomes evident, amplifying the vicious cycle. Our approach brings the important *forward-looking connected view* into the existing practice for assisting these "gray-area" cases, where the system state represented via occupancy effectively encapsulates a much broader spectrum of considerations through our cost modeling. In other words, we present an inclusive trade-off analysis involving individual risk and the ripple effects on others, thereby adding an extra layer of consideration that can aid various stakeholders in the decision-making process, including judges, CC case managers, prosecutors, and more.

Second, our model should be viewed as a versatile analytical framework rather than a definitive prescription for placement decisions. While our model provides a valuable interconnected perspective to assist decision-making, the ultimate decisions still rest with the judges. Nevertheless, our practical contributions extend beyond mere placement recommendations as our model can serve various additional purposes. One important application is its usage as a counterfactual tool, allowing decision-makers to explore "what-if" scenarios and understand the potential outcomes of different placement decisions for the individual in question and the effect it brings to others. Additionally, it can assist in making decisions regarding temporary placements when specific community corrections programs reach capacity, or even in the context of early release and parole from jail or prison; see more in Future Work in Section 8.

Lastly, we emphasize that the dynamics of the system and capacity constraints make trading off between individual considerations with broader societal benefits certainly unavoidable, so as variations in the placement recommendations for similar individuals, particularly those in the gray area. This is not unique to criminal justice but probably universal in any system with resource constraints. For example, in healthcare, patients may need to be discharged earlier when hospital capacity becomes constrained (Shi et al. 2021) or denied admission to the most suitable units (Zhalechian et al. 2023). Similarly, the insufficient community corrections program capacity or jail space necessitates making choices that considers the overall system's performance beyond solely individual considerations. Our approach cannot eliminate the inherent challenges of managing constrained resources in a dynamic environment and tough choices must sometimes be made. On the other hand, our findings in the case study shed lights to staffing and capacity management: (1) without capacity increase, our recommended decisions could help avoid case manager burnout and reduce their workload volatility (see Section 7.2), reducing potential turnover and maintaining

the workforce; (2) we show that investing in more case managers can significantly reduce recidivism and alleviate jail overcrowding, which contrasts the traditional approach of solely expanding jail capacities and contributes to changing the stigma surrounding incarceration.

2. Literature Review

We review two streams of literature: one on applying operations tools in criminal justice work; the other on the methodology side: state-dependent admission/routing decisions and solving high-dimensional MDP.

OR in criminal justice. The application of OR tools in criminal justice systems is relatively understudied. Among the few recent papers applying operations or system engineering to analyze jail overcrowding and related issues, Hughes (2013) develops a simulation modeling based on simple Markovian transitions for planning jail diversion programs for individuals with mental illness. The author shows that jail diversion programs can be an effective incarceration alternative for these individuals but highlights that "a key missing ingredient in jail diversion programs to date is a knowledge of the requirements for capacity and staffing of systems." Usta and Wein (2015) develop a comprehensive simulation model to analyze the impact of different interventions on jail population levels. They use data from the Los Angeles County Jail and show that split sentencing could not only reduce the jail population by 20% but also cut recidivism by 7%. Master et al. (2018) build upon this paper and develop queueing theory to provide more analytical insights.

Another related line of work studies infectious disease or public health issues among the incarcerated population (Franco-Paredes et al. 2021). Malta et al. (2019) provide a systematic review of opioid-related interventions delivered before, during, and after incarceration. They find that individuals treated with opioid agonist treatment in corrections have higher adherence to addiction treatment, lower rates of relapse, and other improved outcomes. In the context of hepatitis C treatment in U.S. prisons, Ayer et al. (2019) develop a restless bandit modeling framework to support treatment prioritization decisions. The authors derive a closed-form index policy that is adjusted for capacity and provide valuable insights into several contentious health policy matters regarding hepatitis treatment, including considerations beyond liver health status and the influence of remaining sentence lengths. Araz et al. (2022) present an analytical framework using a computational epidemiology model for demonstrating the potential impact of jail-based screen-treat programs on the prevention of sexually transmitted diseases.

Methodology. The placement decision considered in this paper is related to a vast body of literature on state-dependent admission and routing decisions in queueing systems. Various studies have analyzed the performance and optimization of such systems using different methodologies and policies. Mandelbaum and Pats (1998) examine state-dependent queueing networks, considering

queue-length-dependent arrival and service rates, routing probabilities, and they derive analytical results based on diffusion approximations. Movaghar (1997) establishes the optimality of "join the shortest queue" (JSQ) policy for identical single-server stations with Markovian dynamics. Ansell et al. (2003a) extend the JSQ policy to include heterogeneous stations and develop routing policies considering congestion using DP-based policy improvement. Glazebrook et al. (2009) propose a Markovian model for optimal control of admissions and routing in a system with heterogeneous stations, providing an explicit construction of an index policy based on Whittle's index. More specifically in the healthcare context, Patrick et al. (2008) use an MDP approach with hard capacity constraints to model the scheduling of diagnostic systems and apply approximate dynamic programming (ADP) for solution. Samiedaluie et al. (2017) formulate an infinite-horizon average-cost MDP to study patient admission policies in a fixed-capacity neurology ward with multiple types of patients. Huang et al. (2015) analyze a multiclass queueing model with feedback to minimize congestion costs in emergency departments.

Our work differs from this line of literature in that our model incorporates salient features of the criminal justice system, which differs from conventional queueing systems in other service operations. As a result, new theories and algorithms have to be developed to address these specific challenges. One key distinction is the unconventional assumption of deterministic service times. This feature, along with the convex cost structure to capture the negative consequences of congestion, results in the curse of dimensionality and presents significant challenges for structural analysis. Furthermore, our model takes into account the impact of service lengths on the recidivism window and the eventual probability of recidivism. This introduces more complex tradeoffs compared to typical routing decisions in conventional queues.

Our research also contributes to the literature on developing scalable algorithms to address the curse of dimensionality in MDPs. Various techniques, including value function approximation and policy gradients, have been proposed to alleviate the computational burden of solving large-scale MDPs; see review in Powell (2007). For queueing networks, Dai and Gluzman (2022) recently demonstrate that a class of deep reinforcement learning algorithms, known as proximal policy optimization, can produce control policies that consistently outperform state-of-the-art queueing control policies. Several studies have also explored leveraging problem structure for dimensionality reduction, as shown in the works of George et al. (2008), Nadar et al. (2018), Ulmer (2020), among others. In our paper, we introduce a novel technique to address the curse of dimensionality in MDPs. Specifically, we propose a two-timescale approximation scheme that takes advantage of the long service times and relatively stable placement policies observed in the criminal justice setting. This technique not only addresses the challenges specific to our application but also has the potential for broader applications in dimensionality reduction in other domains.

3. Background

In this section, we provide an overview of community corrections operations. We use our partner, the Tippecanoe County Community Corrections (TCCC), as the main motivational example to provide detailed context while drawing upon relevant criminal justice literature to make the overview general. We start from describing the process of intake and placement to CC programs in Section 3.1. We then discuss the various tradeoffs involved in the placement decision process, particularly with regard to the impact of system congestion, in Section 3.2, highlighting the complexity of the placement decision.

3.1. Processes of Jail Diversion and Community Corrections Intake

CC programs serve as an alternative to traditional forms of incarceration such as jail. The process of diverting individuals from jail to CC typically consists of five stages: identification, assessment, referral, participation, and societal reintegration. Figure 1 shows the process flow for the jail diversion and CC intake in our community partner, which we elaborate on below. While there may be some variations among community correction centers, such as differences in eligibility criteria for intake, this illustration provides a typical example of the intake and placement procedures for community correction programs in other states (Magaña et al. 2022).

The diversion process begins with individual identification by law enforcement and court officials, followed by an assessment using the Indiana Risk Assessment System (IRAS). Eligible individuals who are willing to participate in the diversion programs may be sentenced[1] to one of the CC programs or incarcerated in jail, depending on the decision made by the judge with inputs from prosecutors and CC managers. After completing the designated sentence length in their assigned location, they exit the system with the goal of reintegrating into society. However, it is important to note that some individuals may reoffend and get rearrested either after leaving the system or while participating in the CC programs – known as *recidivism*; see details in Section 3.2.

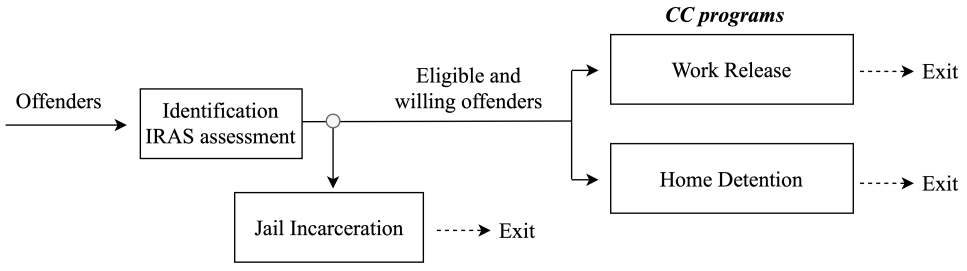


Figure 1 An illustration of the jail diversion process incorporating TCCC's practices.

[1] We use sentence and place interchangeably in this paper.

The main programs offered at TCCC are *Home Detention* and *Work Release*, which are common programs at community correction centers under possibly different names (Gaines and Miller 2021). Home Detention (HD) allows individuals to serve their sentence from their own homes through electronic monitoring and regular reporting to CC case managers. Work Release (WR), on the other hand, permits participants to leave to work during the daytime and return to stay on-site overnight, with access to education and treatment programs. Our main analysis in this paper focuses on the HD program, for several reasons. First, HD is the most prevalent and widely accessible program in various community corrections, which is also known as electronic monitoring, house arrest, or home confinement (Aloyan 2020). At TCCC, 77% of its participants are placed in the HD program. Second, HD captures the most salient feature in community programs that differ from traditional incarceration, i.e., loosened monitoring and supervision with the involvement of case managers. The nature of this setup creates complex tradeoffs in placement decisions (jail vs. CC), which we discuss in the following section. For tractability, we use HD to represent CC programs in our technical analysis unless stated otherwise; we differentiate between HD and WR in the case study in Section 7 where we calibrate a realistic simulation platform with TCCC data.

3.2. Key Considerations in Placement Decisions

The decision to place an individual in jail or CC is primarily made by the court (judge), with input from the CC director. We classify these factors into two main categories: individual-level risk and system-level congestion. While current practices tend to focus on the individual level, the system-level effects (particularly workload on case managers) are often overlooked, potentially leading to unintended consequences. We summarize these factors in Table 1 and elaborate on each of them as follows. We emphasize that the placement process is a complex and multifaceted task that requires careful consideration of numerous factors. Our intention is not to override the decisions made by judges, but rather to provide a comprehensive and interconnected perspective on this intricate process, laying the foundation to support decision-making for all stakeholders.

Table 1 Tradeoffs behind the criminal placement decisions between jail and CC.

	Individual Level			**System Level**	
	Violation	In-service Recidiv	Long-term Recidiv	Holding Cost	Congestion Effect
Jail	No	No	Risk increase	Expensive	Risk increase via mixing
CC	Yes	Yes	Risk reduce	Cheap	Loosened monitoring

Individual level. The risk of recidivism is a primarily used factor when deciding whether to sentence an individual to CC or incarceration in practice. Individuals undergo the IRAS assessment (Figure 1), which evaluates various factors (e.g., criminal history, substance use) to determine their reoffending/recidivism risk and criminogenic needs. TCCC has specific eligibility criteria on offense

types and risk assessments (e.g., prioritizing low-risk individuals with mental health or SUD issues); also see similar practice used in New Jersey (Russell 2022). Considering the risk of recidivism upon entry is important because CC programs differ from strict incarceration settings like jail, which serve as a barrier between individuals and the community. A high-risk individual re-offending during a CC program poses a greater risk to society. In addition, non-criminal technical violations can also occur during CC program participation, such as drug test failures, missed appointments or treatment programs, and non-compliance with case manager or probation officer reporting (Gould et al. 2011). These violations can cause negative societal impact and increase CC program costs due to stricter supervision.

What complicates the placement decision is that not only the short-term risk level at the entry point should be considered, but also the long-term recidivism risk after completing the sentence. In other words, the decision should account for the "treatment effect" in reducing risk post-sentencing. Incarceration diversion aims to address the root cause to reduce long-term recidivism rates (Brogan et al. 2015). Community correctional programs, including substance abuse treatment, cognitive behavioral programs, and education and employment programs, have been found to contribute to significant reductions in recidivism rates (McNiel and Binder 2007). On the other hand, individuals in jail may lack access to necessary treatment and support programs, making reintegration and long-term risk reduction more challenging. While jail can prevent short-term reoffending, it may lead to a higher long-term recidivism risk. In contrast, CC programs reduce long-term risk but may not prevent short-term technical violations and reoffenses. Therefore, the placement decision should strike a balance, considering both short- and long-term recidivism risks for each individual.

System level. While the current focus is primarily on individual-level risk consideration, the impact on the system level is often disregarded due to the siloed decision-making structure prevalent in the criminal justice system (a primary issue that prompted this study). It is important to recognize that system overcrowding affects both jails and CC programs in various aspects.

The direct impact of overcrowded jails is increased operational and maintenance costs, such as the need for additional staff, supplies, and infrastructure. This "holding cost" is known for being expensive (Henrichson 2015). Overcrowding also limits access to essential medical care and rehabilitation services, resulting in inadequate care and a higher risk of illness or injury for incarcerated individuals. CC programs, which receive partial subsidies from participants, have lower holding costs compared to jails. While CC programs were designed to alleviate jail overcrowding and avoid the high costs of incarceration, an excessive number of participants in CC programs can still overload case managers and have other negative effects as elaborated below.

The impact of system overcrowding on the risk of recidivism and violation is more nuanced, as it affects jail and CC differently. Jail overcrowding is known to be a contributing factor to a

higher likelihood of recidivism after release. In overcrowded jails, shared living spaces result in increased violence, aggression, and detrimental effects on the well-being of inmates (Spohn and Holleran 2002). On the other hand, heavy caseloads in CC put strain on case managers and hinder the effectiveness of these programs. In particular, staff shortages often make it challenging for overburdened case managers to provide adequate supervision and treatment to individuals in need. Consequently, increased probation caseloads are associated with higher rates of violations and reoffenses (Worrall et al. 2004, Camp and Camp 1999).

The effects of overloading CC programs are often neglected, in comparison to the widely discussed issue of jail overcrowding. Studies showcasing the benefits of CC programs are typically conducted in controlled trial settings with ample resources. However, real-world implementation often encounters resource constraints that diminish the program's effectiveness. While existing guidelines advocate weighing CC program availability amid overcrowding, they fall short of providing a holistic understanding of the interconnected challenges and broad impact of congestion in the criminal justice system. In the following numerical illustration, we highlight the necessity of considering these system-level factors.

Nuanced Decisions. Placement decisions in the current practice are commonly guided by a risk-based approach. We refer to this current practice as the *static policy* since it focuses solely on individual risk, which may result in either jail or CC becoming overcrowded due to stochastic fluctuations. Table 2 illustrates this pitfall of the static policy using a simulation platform calibrated with data from TCCC (see Section 7.1). The results show that the static policy performs similarly in a system with no congestion impact and in a system with sufficient capacity. However, when capacity is limited due to insufficient case managers, the static policy leads to dramatically higher total costs. Specifically, recidivism and violation costs more than double, exacerbating the vicious cycle. This highlights how the static policy fails to account for resource constraints and negative impacts of system congestion. Ignoring externalities and focusing solely on individual risk can be detrimental under crowded conditions, which undermines the original goal of CC programs to break the cycle of re-incarceration and places additional strain. As discussed in Section 1.1.2, the stochastic fluctuations and capacity constraints necessitate making tradeoffs that considers the overall system's performance beyond solely individual considerations. A one-size-fits-all solution or reliance on ad-hoc experiences is insufficient for this multifaceted placement process, calling for sophisticated decision support that considers individual risk and can adaptively adjust to the system state.

4. Modeling Framework

We formulate an infinite-horizon, discounted-cost discrete-time MDP for the placement decision problem. We introduce the system model in Section 4.1 and specify the MDP in Section 4.2.

Table 2 Performance comparison of the static policy under different scenarios. The unit cost of recidivism is set to 1 and all other costs are scaled proportionally. Each number is averaged over 100 simulation replications with each replication corresponding to a three-year time horizon.

Scenario	Recidivism Cost	Violation Cost	Holding cost	Total Cost
No Congest-impact	2117	441	171	2729
High Staff (15 CC case managers)	2507	541	185	3233
Low Staff (9 CC case managers)	4580	1721	210	6511

4.1. System Modeling

We consider a two-station queueing network that is abstracted from the process flow of the placement decision to incarceration and CC, as introduced in Section 3.1. Figure 2 illustrates this network model, where one station corresponds to jail, and the other station corresponds to CC (with the two CC programs combined for analytical tractability). In the rest of the paper, we refer to the individual going through the network as "customer" and use it interchangeably with "individual" or "client."

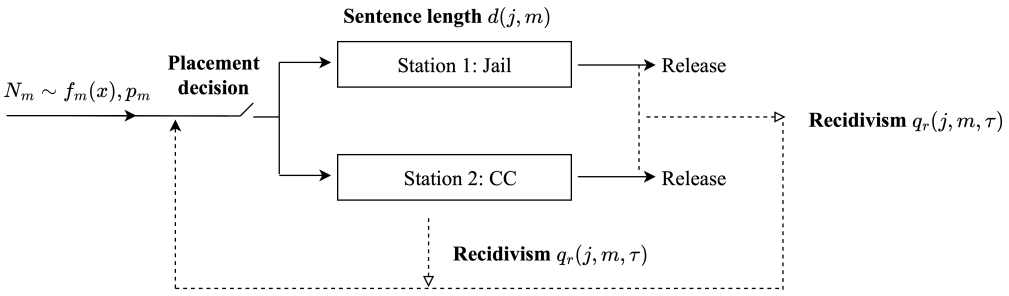


Figure 2 Two-station network.

Customers arrive to the system from M classes, where the classes correspond to different risk categories determined during the intake assessment. We assume that the risk-class distribution is characterized by $\{p_m\}_{m=1}^M$, where each p_m is the probability of a customer belonging to class m and $\sum_{m=1}^M p_m = 1$. The system evolves continuously, but we observe the system state at discrete time epochs. Without loss of generality, we assume the observation is made daily and use day as the time unit. Under this discrete-time setup, we assume that the daily number of new arrivals from class m, denoted as N_m, is independently and identically distributed (i.i.d). It follows a general discrete distribution with support on $\mathbb{Z}_+$ and probability mass function $f_m(\cdot)$. Upon arrival, a placement decision is made, determining which station $j \in \mathcal{J} = \{\texttt{jail, CC}\}$ to assign the customer to. After the placement, the customer starts the service process – corresponding to serving the sentence length decided by the court. Let $d(j,m)$ denote the sentence length for class m customers placed in station j. In contrast to the prevalent assumption of memoryless service times in the literature,

we assume a *deterministic* value for $d(j,m)$ for each risk class and station combination. This is a salient feature in the context of criminal justice since the sentence length is predetermined.

A customer may recidivate after completing their service or even during it, depending on whether they are placed in jail (with no recidivism during incarceration) or CC (with a possibility of recidivism during service). We use $q_r(j,m,\tau)$ to denote the probability that a class m customer placed in station j will recidivate on the τ^{th} day since "leaving the incarceration" – the time of service completion if placed in jail and the time of service start if placed in CC. Therefore, the cumulative recidivism probability in a given time-window T equals $q_r(j,m) = \sum_{\tau=0}^{T} q_r(j,m,\tau)$. Note that $q_r(j,m,\tau)$ allows us to capture *time-dependent* recidivism risk, which is another salient feature in the criminal justice system: the recidivism risk is highest in the months immediately following release and gradually decreases over time (Greenberg 1978, Schmidt and Witte 2012).

4.2. MDP Model

Our primary decision is to select a station j to assign an incoming customer. To model this decision problem, we formulate an infinite-horizon, discounted-cost discrete-time MDP, with decision epochs occurring daily ($t = 0, 1, \dots$). We introduce the system state, action, transition dynamics, costs, and objective function in sequence. We omit the explicit dependency on the epoch index t in the description for simplicity. A summary of notations is provided in Appendix EC.1 for reference.

State. At the beginning of each day (decision epoch), we observe the pre-action system state $S \in \mathcal{S}$,

$$S = \{X_{j,m,0}, X_{j,m,1}, ..., X_{j,m,d(j,m)}\}_{j\in\mathcal{J};m=1,...,M},$$

where $X_{j,m,l}$ is the number of class m customers who have been in station j for l days. Tracking the length-of-stay (LOS) already spent, l, is necessary since the total sentence length $d(j,m)$ is pre-determined. Thus, the history-dependent nature is integral to our analysis. We have $l \in \{0, 1, \dots, d(j,m)\}$, with $l = 0$ denoting customers just admitted yesterday and $l = d(j,m)$ for customers to be released today. We refer to $d(j,m) - l$ as the remaining LOS.

Action. The decision maker determines the placement of new arrivals based on the observed state. In the discrete-time setting, at the start of each day, a batch of arrivals may be waiting to be placed. Let n_m be a realization of N_m, denoting the number of new customers of class m who arrived on the current day. The action is a vector $A_m = \{A_{m,k}\}$ with

$$A_{m,k} \in \mathcal{J},\ k = 1, 2, ..., n_m.$$

The realized action $a_{m,k} = j \in \mathcal{J}$ means that the kth new arrival from class m is placed in station j. The decisions are made sequentially: (i) we start with the highest-risk group and then proceed to the lowest-risk group to prioritize public safety; (ii) within a risk group, placements are made

sequentially based on the order of arrival times (i.e., first-come-first-serve). The decision for each individual k depends on the current system state and can be influenced by prior placements. A comprehensive discussion on validity of state-dependent action can be found in Section 1.1.

One-epoch cost. For a given state $s = \{x_{j,m,l}\}$'s and action $a = \{a_{m,k}\}$'s, we calculate the pre-action, pre-transition cost $C(s)$ for the current day (epoch). This cost comprises three components: holding cost $c_h(s)$, recidivism cost $c_r(s)$, and violation cost $c_v(s)$. These costs capture the key tradeoffs discussed in Section 3.2. Mathematically, we have:

$$C(s) = c_h(s) + c_r(s) + c_v(s).$$

First, the holding cost $c_h(s)$ represents the cost associated with maintaining individuals in different stations, reflecting the consumption of resources such as beds in jail and case managers in CC. We assume a convex relationship between the holding cost and the occupancy of each station:

$$c_h(s) = \sum_{j \in \mathcal{J}} c_h^j(s_j), \tag{1}$$

where $c_h^j(\cdot)$ is a convex function, and $s_j(s) = \sum_{m=1}^{M} \sum_{l=0}^{d(j,m)} x_{j,m,l}$ is the occupancy (total number of individuals) in station j. A commonly used convex function is the quadratic function, i.e., $c_h^j(s_j) = h_j s_j^2$, where h_j is the unit holding cost in station j. The convex form of the holding cost function captures the key aspect of escalating negative consequences as congestion levels deteriorate in jail or CC. This can have detrimental effects on participants' well-being and potentially overwhelm case managers (Ansell et al. 2003b).

Second, the recidivism cost $c_r(s) = \sum_{j \in \mathcal{J}} c_r^j(s)$, where $c_r^j(s)$ denotes the recidivism cost from station j and follows

$$c_r^j(s) = h_r \sum_{m=1}^{M} s_{j,m}(s) \tilde{q}_r(j,m). \tag{2}$$

Here, h_r is the unit recidivism cost, $s_{j,m}(s) = \sum_{\ell=0}^{d(j,m)} x_{j,m,\ell}$ is the occupancy of class m customers in station j, and $\tilde{q}_r(j,m)$ is an adjusted recidivism probability for cost accounting purposes. The reason for this adjustment is that, unlike holding costs, recidivism costs are incurred both during service (for CC) and after service completion (for both jail and CC). The adjustment allows us to appropriately distribute the future recidivism cost incurred after service completion across each period during the service duration $d(j,m)$. This ensures a unified form that connects each type of cost to the system's occupancy, effectively linking them to the congestion state. The specifics of this adjustment for a given recidivism window T are provided in Appendix EC.4.1.

Lastly, the violation cost captures the fact that CC participants may engage in non-criminal technical violations that do not necessarily amount to recidivism. Different from the recidivism

cost, this violation cost is specific to CC participants and occurs only during their service, not after service completion. Specifically, $c_v(s) = c_v^{\texttt{CC}}(s)$ with

$$c_v^j(s) = h_v \sum_{m=1}^{M} s_{j,m}(s) q_v(s_j, m), \quad j = \texttt{CC}. \tag{3}$$

Here, h_v is the unit violation cost, $s_{j,m}(s)$ is defined as before, and $q_v(s_j, m)$ is the probability of violation for a class m customer in each epoch. The summation part corresponds to the expected number of technical violations that will happen in the current epoch, and we assume that $q_v(s_j, m)$ is increasing with the occupancy level s_j. As a result, $c_v^j(s)$ is increasing and convex with the occupancy level s_j. This convex form reflects the increasing violation risk under loosened supervision when CC gets more overcrowded, as discussed in Section 3.2.

The assumptions on the cost functions are summarized in Assumption 1. Both recidivism and violation costs are computed by summing over costs from all risk classes. This differentiation between risk classes, as indicated by risks $\tilde{q}_r(j, m)$ and $q_v(s_j, m)$, allows us to account for individual-level effects. The system-level effects are encapsulated by the convexity of costs $c_h^j(s)$ and $c_v^j(s)$ in the class-LOS specific occupancy $X_{j,m,l}$, given that the occupancy $s_j = \sum_{m=1}^{M} s_{j,m}$ and $s_{j,m} = \sum_{l=0}^{d(j,m)} X_{j,m,l}$. This convexity captures the increasing marginal effects that each new customer has on the risk levels of all existing customers. Unlike traditional congestion-based routing in queueing control literature, our model, centered around occupancy, covers a much broader spectrum of consideration beyond the typical holding cost. This unified cost formulation also facilitates the theoretical analysis of value function comparison; see details in Section 5.

Assumption 1. *The holding cost $c_h(s)$ and violation cost $c_v(s)$ are convex in each $X_{j,m,l}$, and the recidivism cost $c_r(s)$ is linear in $X_{j,m,l}$.*

We remark that the linear recidivism cost primarily serves the purpose of interpreting the recidivism risk. All our technical analyses can accommodate a convex form of recidivism cost, similar to what we have done for the violation risk $q_v(s_j, m)$. However, as recidivism can occur after the completion of service when computing the adjusted $\tilde{q}_r(j, m)$ to distribute future recidivism costs across the service duration, there is ambiguity in selecting which day's occupancy s_j to use if we adopt the convex form. Because this occupancy can change during the service duration. To maintain interpretability, we assume that the risk of recidivism depends solely on the class and the number of days since release, rather than occupancy, leading to the linear cost assumption.

Transition dynamics. Denote the current state as $S = \{X_{j,m,0}, X_{j,m,1}, ..., X_{j,m,d(j,m)}\}_{j;m}$, and the state on the next day as $S' = \{X'_{j,m,0}, X'_{j,m,1}, ..., X'_{j,m,d(j,m)}\}_{j;m}$. The state evolution follows:

$$\begin{aligned} X'_{j,m,0} &= \sum_{k=1}^{N_m} \mathbb{1}(A_{m,k} = j), \quad \forall j, m; \\ X'_{j,m,l+1} &= X_{j,m,l}, \quad \forall j, m, \text{ and } l = 0, ..., d(j,m) - 1. \end{aligned} \tag{4}$$

Equation (4) states that the number of new participants in each station is determined by the placement decisions, where N_m is the total number of new class-m arrivals. For existing participants who have stayed $l \le d(j,m)-1$ days, they become participants who spent $l+1$ days on the next day; those who have stayed $d(j,m)$ days will be released today.

Remark 1. Because the recidivism window is long (measured in years), it induces a smoothing effect on the impact of returning arrivals (resulting from recidivism) on the daily arrival rate (Greenberg 1989), as long as the placement decisions remain relatively stable. For the sake of maintaining the Markov property in theoretical analysis, we assume that the stream of recidivism arrivals is independent of past actions and is considered part of the exogenous arrivals included in N_m. In Appendix EC.6.1, we demonstrate how these recidivism arrivals can be integrated into the exogenous arrival process and how to calculate the rate of N_m when the system is in a steady state. Importantly, we relax this assumption in the case study presented in Section 7, where we explicitly model the recidivated customers as a separate stream of returning flow that can be influenced by prior placement decisions.

Objective. We define the placement policy $\pi : \mathcal{S} \to \mathcal{J}^M$ as a mapping from state S to action $\{A_m\}_{m=1,..,M}$. We focus on considering the class of stationary policies for π, s.t., $\pi_t = \pi, \forall t$. For a given state s, $\pi(s)$ is a vector with $\pi(s) = \{\pi_m(s)\}_{m=1,...,M}$, where $\pi_m(s) \in \mathcal{J}$ prescribes the placement action for each class m. For any optimal policy in a discounted MDP, the corresponding value function $V(\cdot)$ must satisfy the following Bellman optimality equation (Puterman 2014):

$$V(s) = \min_{\pi \in \mathcal{J}^{M\cdot|\mathcal{S}|}} \; C(s) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,\pi) V(s'), \quad \forall s \in \mathcal{S}, \tag{5}$$

where $\gamma \in (0,1)$ is the discounting factor, $C(s)$ is the one-epoch cost function, and $p(s'|s,\pi)$ represents the transition probability under policy π. Given the finiteness of the action space, there exists an optimal deterministic stationary policy; see Theorem 6.2.10 in Puterman (2014).

5. Structure Property

In this section, we study the structure of the optimal policies, which offer important insights into the interplay among different tradeoffs and desired analytical properties for algorithm design in Section 6. In Section 5.1, we show the preference of sending a new arrival to jail versus CC with respect to (i) the number of existing individuals with varying LOS and (ii) different risk classes. We also showcase, through counterexamples, that there is *no* monotonicity when solely considering total occupancy or risk classes. This intricate nature prompts a thorough investigation into the properties of the value function in Section 5.2. We establish the superconvexity of the value function with respect to appropriately defined occupancy categories, the main technical result of this paper. This superconvexity plays a pivotal role in our algorithm design in Section 6 to

provide justification for the existence of global optimality. In Section 5.3, we explore the influence of the recidivism window in individual's placement decision, which yields valuable insights into the long-term consequences of placing individuals in CC versus jail.

5.1. Policy Monotonicity on LOS and Risk Class

In this section, we show in Proposition 1 that the optimal placement policy is monotone with respect to the number of each class of participants in each LOS category in each station. Through this result, we establish two corollaries to show the impact of LOS and risk classes. We denote the unit occupancy of class m in LOS category l in station j as $e_{j,m,l}$. Denote $\pi^*_{m_0}(s)$ as the optimal policy for customers of class m_0 at state s.

PROPOSITION 1. *Under Assumption 1, the optimal policy π^* is monotone with the number of participants of each class in each LOS category. For any given state s and class m_0, $\pi^*_{m_0}(s) = \texttt{CC}$ implies $\pi^*_{m_0}(s+e_{\texttt{Jail},m,l}) = \texttt{CC}$, and $\pi^*_{m_0}(s) = \texttt{jail}$ implies $\pi^*_{m_0}(s+e_{\texttt{CC},m,l}) = \texttt{jail}$ for all $m = 1, ..., M$ and $l = 0, ..., d(j,m)$.*

This analysis connects the placement of individuals of each class-LOS combination to its impact on existing customers. Specifically, when jail gets one more class m customers with LOS of l epochs (add $e_{\texttt{Jail},m,l}$ to jail occupancy), it becomes less preferred to place a new customer in jail and more preferred to CC; vice versa for CC. Although this may seem like a congestion-based policy, the underlying reason is that as programs become more crowded, adding new customers raises more risk for participant, making the placement less desirable.

Based on Proposition 1, we establish two corollaries, with a focus on individual-level considerations, and then discuss their policy implications. The first corollary considers the impact of existing participants with different LOS, highlighting the sensitivity of the optimal policy to the LOS distribution among existing customers.

COROLLARY 1. *Under Assumption 1, for any given state s, class m_0, and two LOS $l_1 < l_2$, $\pi^*_{m_0}(s + e_{\texttt{Jail},m,l_1}) = \texttt{jail}$ implies $\pi^*_{m_0}(s + e_{\texttt{Jail},m,l_2}) = \texttt{jail}$, and $\pi^*_{m_0}(s + e_{\texttt{CC},m,l_1}) = \texttt{CC}$ implies $\pi^*_{m_0}(s+e_{\texttt{CC},m,l_2}) = \texttt{CC}$, for all $m = 1, ..., M$.*

Corollary 2 establishes the placement prioritization between different risk types. This prioritization is based on the expected cost reduction $C_\Delta(m,s)$, when placing a class-m individual in CC instead of jail at a given state s. This cost reduction comes from the difference in the three parts of costs mentioned in Section 4.2 and its detailed formulation is delegated to Appendix EC.5.2 (see Equation (EC.19) there).

COROLLARY 2. *Under Assumption 1, for any given state s, and two customer classes m_1 and m_2, if $C_\Delta(m_1,s) \le C_\Delta(m_2,s)$, then $\pi^*_{m_1}(s) = \texttt{CC}$ implies $\pi^*_{m_2}(s) = \texttt{CC}$; if $C_\Delta(m_1,s) \ge C_\Delta(m_2,s)$, then $\pi^*_{m_1}(s) = \texttt{jail}$ implies $\pi^*_{m_2}(s) = \texttt{jail}$.*

The proofs of Proposition 1 and Corollaries 1 and 2 are in Appendix EC.3 and EC.5.2.

Policy implications. For the results on LOS, intuitively, a "newer" participant has longer remaining time in the system. This implies longer-lasting impacts on other customers' risks, thereby reducing the tendency to place future customers in the station. Corollary 1 establishes this intuition formally.

The result on the risk classes is more nuanced. Specifically, the value of $C_\Delta(m,s)$ is determined by comparing the reduction in long-term recidivism risk and jail occupancy cost (benefits from placing the individual in CC) and the increase in short-term recidivism risk, violation risk, and CC occupancy cost (drawback of placing in CC). For example, the increase of violation cost can be expressed as $h_v \sum_{t=0}^{d(\mathtt{CC},m)} \gamma^t \mathbb{E}(c_v^{\mathtt{CC}}(s_{t,\mathtt{CC}} + e_{\mathtt{CC},m,t}) - c_v^{\mathtt{CC}}(s_{t,\mathtt{CC}}))$, where the expectation is taken over future occupancy with $s_{t,\mathtt{CC}}$ denoting the occupancy of CC under the optimal policy after t epochs of seeing state s. Therefore the increase of violation cost is determined by both the new customer's risk class m and the future occupancy level $s_{t,\mathtt{CC}}$. When two customers belong to classes with a large difference in risks, their personal risk difference will dominate the cost difference, as well as the placement prioritization. This is similar to the clear-cut case discussed in Section 1.1. However, when two customers belong to similar classes, the cost difference will be mainly driven by the future system occupancy due to cost convexity. In other words, there is no simple rule-of-thumb priority among the risk classes, especially for individuals whose risk levels fall into the "gray area." The placement decision necessitates a careful balance between individual risks and system-wide effects.

Counterexamples for non-monotonicity. Proposition 1 establishes policy monotonicity with respect to the number of participants in each class-LOS category. While this result is intuitive, we have counterexamples showing that the policy is not monotone (i) in the total occupancy or (ii) considering risk classes alone. (ii) has been discussed in the preceding paragraph, and for (i), we show that when the jail occupancy is higher, it is possible that the optimal decision still places new participants in jail, e.g., when the remaining LOS of existing individuals in jail is short and they are close to being released. This reaffirms our earlier discussion that there is no fixed, universal preference ordering, rendering the static policy suboptimal. This motivates us to take a deep dive into examining the superconvexity property of the value function in Section 5.2.

5.2. Superconvexity of Value Function

In this section, we establish the main technical result of this paper, the superconvexity of the value function, in Theorem 1. This result highlights how the system effect from one customer is intertwined with the characteristics of other customers in the system. We provide a sketch of proof for Theorem 1, which is based on our novel, unified approach developed through value function decomposition, coupling, and policy deviation bounding. We discuss the important implications of

Theorem 1 at the end of this section. For ease of exposition, we focus on the single-class setting ($M = 1$) and omit the subscript m, so $e_{j,l}$ denotes the unit occupancy for LOS l and station j, and d_j denotes the sentence length in station j. The applicability of the unified approach extends naturally to multi-class models. We start by defining the superconvexity, denoted as $SuperC(e_i, e_j)$ property below.

DEFINITION 1. A function $f : \mathcal{S} \to \mathbb{R}$ satisfies the $SuperC(e_i, e_j)$ property if for all $s \in \mathcal{S}$, $f(s + 2e_i) - f(s + e_i) \geq f(s + e_i + e_j) - f(s + e_j)$ and $f(s + 2e_j) - f(s + e_j) \geq f(s + e_i + e_j) - f(s + e_i)$.

The superconvexity is a property closely related to both supermodularity and convexity (Koole 1998). Now, we are ready to state the main theorem that shows the superconvexity property of the value function.

THEOREM 1. *Under Assumption 1 and some mild technical condition, the optimal value function V^* satisfies $SuperC(e_{jail,l}, e_{CC,l})$, i.e., for all $s \in \mathcal{S}$ and $l = 0, 1, ..., \min\{d_{jail}, d_{CC}\}$,*

$$V^*(s + 2e_{CC,l}) - V^*(s + e_{CC,l}) \geq V^*(s + e_{jail,l} + e_{CC,l}) - V^*(s + e_{jail,l}), \tag{6}$$

$$V^*(s + 2e_{jail,l}) - V^*(s + e_{jail,l}) \geq V^*(s + e_{jail,l} + e_{CC,l}) - V^*(s + e_{CC,l}). \tag{7}$$

In the interest of space, we refer to Appendix EC.5.1 for a detailed explanation of the mild technical condition and its verification. Inequality (6) compares the increase in the value function by placing a customer to CC ($e_{CC,l}$) at two current states, $s + e_{\mathtt{CC},l}$ and $s + e_{\mathtt{jail},l}$, respectively, which indicates that the increase is larger when this CC placement is done to the former state with an additional customer of $e_{CC,l}$. Inequality (7) shows a similar result for jail placement, indicating that assigning a customer to jail in a state with an additional customer of $e_{jail,l}$ leads to a more substantial increase in the value function. We outline the challenges in the proof before providing the proof sketch and discussing the policy implications.

Technical challenges. Unlike most existing literature in proving superconvexity, the biggest challenge in proving Theorem 1 comes from the salient feature of our model: the non-memoryless service time. Specifically, the common approach to proving the superconvexity is to use the operator-based method (Koole 1998). However, due to the non-memoryless service time, using this method requires us to include a new operator to account for the deterministic transition in the remaining service time (LOS). This necessitates proving that this operator preserves all desired properties such as increasing and superconvexity for each shift in the LOS, i.e., going from $e_{j,l}$ to $e_{j,l+1}$, which makes the proof extremely tedious and requires imposing strong conditions for the property preserving. Furthermore, the non-memoryless LOS affects the proof for other operators as well due to the need of tracking LOS. For example, when evaluating the operator of placement decisions, we need to evaluate inequalities involving four dimensions, including $e_{\mathtt{jail},0}, e_{\mathtt{CC},0}$ for the impact on

new arrivals and $e_{\texttt{jail},l}, e_{\texttt{cc},l}$ for property preservation on existing customers. This goes beyond the two-dimensional properties typically considered in the operator-based method.

To address these limitations, we propose a unified approach through value function decomposition, leveraging the coupling techniques and policy deviation bounding based on policy monotonicity. We provide a sketch of this framework below and delegate the complete details to Appendix EC.2.

Sketch of proof. Our proof framework includes four steps. First, we decompose the value function into *marginal contribution* from each individual. This starts from decomposing the convex one-epoch cost function into the sum of marginal contributions:

$$C(s) = \sum_j C_j(s_j) = \sum_j \sum_{k=1}^{s_j} (C_j(k) - C_j(k-1)), \tag{8}$$

where $C_j(\cdot)$ is the cost function for station j, s_j is the occupancy level of station j at state s, and $C_j(0) = 0$. Then the value function $V(s) = \min_\pi \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t C(s_t) | s_0 = s, \pi]$ can be decomposed into the discounted sum of marginal costs contributed by individuals. This first step lays out the most important foundation for our following steps in decomposing the value function difference into the marginal cost from "one" additional customer after coupling the arrivals and placement decisions.

The optimal policies π_a and π_b are the same when viewed as state-action mappings. We differentiate between π_a and π_b to highlight the different sets of decisions prescribed by the optimal policy in the two systems with different starting states.

The key difficulty in analyzing the value functions in (6) or (7) is that each inequality necessitates a comparison of costs across four systems under the optimal policy π^* starting from four different states. Despite following the same state-action mapping $\pi^* = \{\pi(s)\}_s$, these varied starting states result in four distinct sets of "unknown" decisions. We overcome this challenge via the *coupling technique* and *policy deviation bounding*. Taking Inequality (6) as an example, we perform the following three steps.

(i) For each side of (6), we couple the arrivals and placement decisions of future customers using the decisions prescribed by the optimal policy starting from one of the two states. By leveraging suboptimality for the other state, we bound the difference in value functions and reduce the number of sets of decisions to be analyzed from four in total to two (one on each side of the inequality).

(ii) Under coupling and the decomposition (8), the value function difference on each side simplifies to the marginal contribution from one extra customer. For example, the left-hand side of (6) can be expressed via $\sum_{t=0}^{d_{\texttt{cc}}-l} \gamma^t \big(C_{\texttt{cc}}(s^1_{t,\texttt{cc}}) - C_{\texttt{cc}}(s^1_{t,\texttt{cc}} - 1)\big)$ after some algebra, where $s^1_{t,j}$ denotes the occupancy at station j after t epochs when starting from $s + 2e_{2,l}$. While this marginal contribution is trivial to analyze under the linear cost setting, our setting of general convex cost functions greatly complicates the analysis. That is, the marginal cost varies with the occupancy levels of the system.

(iii) Even though the same optimal policy π^* is used for both systems, the systems will see different states at each epoch if they start from different states. This can lead to different actions and different future states. This variability makes it difficult to compare the occupancy levels across systems. To address this, we propose a policy deviation analysis approach. By coupling the two systems to have identical arrivals, we assess for each epoch if π^* would prescribe different decisions in the two systems. By exploiting the optimal policy's inherent monotonicity (as detailed in Section 5.1), we can then effectively characterize instances with decision discrepancy and bound the policy deviation. This analysis then provides the occupancy deviation.

Combining all steps gives us the superconvexity result. We conclude this section by discussing two implications from Theorem 1.

Policy implications. The first implication of Theorem 1 is that placement decisions should be state-dependent, which aligns with the implication of Proposition 1. Additionally, when we go from the baseline sequential decisions for each individual to considering batched decisions, which essentially leads to a proportion-based decision, this superconvexity provides a very useful result: the value function with respect to the number of new participants that should be placed to jail (vs CC) could only have one local optimum, which is also globally optimal. This serves as an important stepping stone for our algorithmic solution developed in Section 6, which we will further illustrate there.

The second important implication is that the superconvexity result covers placement decisions for both new arrivals (with $l = 0$) and individuals that come with some served LOS (with $l > 0$). The latter situation corresponds to an important feature in criminal justice: **plea bargains**. In practice, more than 90% of criminal convictions are obtained through plea bargains (Abrams 2011), where a common offer in plea bargains is *time served*. This offer allows the time that a defendant has served in pretrial custody to be credited towards their sentence. Therefore, for customers who take a time-served offer and have been in pretrial detention for l epochs, their first epoch is at jail/CC is equivalent to the lth epoch. Thus, the placement decision for such a participant is $\min\{C(s)+\gamma V^*(s+e_{\texttt{jail},l}), C(s)+\gamma V^*(s+e_{\texttt{CC},l})\}$. Our superconvexity results apply to these individuals with plea bargains. This lays a foundation for our future work on extending the placement decisions to different time points in the criminal justice process; see discussions in Section 8.

5.3. Policy Monotonicity on Recidivism Window

Proposition 2 examines the impact of recidivism window on the optimal policy under Assumption 2. This assumption captures essential aspects of the recidivism risk discussed in Sections 3.2 and 4.1, including long-term recidivism rate reduction through CC programs and the time-dependency of recidivism risk.

ASSUMPTION 2. *For each $m = 1, ..., M$, we assume that (i) $q_r(\texttt{CC}, m, \tau) \leq q_r(\texttt{jail}, m, \tau)$ for each τ; (ii) there exists a $\tilde{\tau}_m > 0$, such that the recidivism window T satisfies $T - d(\texttt{jail}, m) > \tilde{\tau}_m$ and the recidivism risk $q_r(j, m, \tau)$ decreases in τ for all $\tau > \tilde{\tau}_m$ and $j \in \mathcal{J}$.*

PROPOSITION 2. *Consider the first incoming customer after state s as the target customer, and let $a_m^1(s)$ and $a_m^2(s)$ be the optimal decisions for the target customer of class m under recidivism windows T_1 and T_2 (where $T_2 > T_1$), respectively. Customers except for the target customer are subject to a same recidivism window T. Suppose that Assumptions 1 and 2 hold, then for any given state s and class m, $a_m^1(s) = \texttt{CC}$ implies $a_m^2(s) = \texttt{CC}$.*

We prove Proposition 2 by contradiction via a coupling argument, which converts the comparison of the difference in value functions to the recidivism cost of the target customer at the given state s_0. See the complete proof in Appendix EC.4.2.

Policy implication. Proposition 2 says that if a customer should be placed in CC under a shorter T_1, then the same customer should also be placed in CC if their recidivism window is changed to a longer T_2. In other words, CC is a more preferred choice with a longer recidivism window; and vice versa. This result highlights one of the most important tradeoffs in our setting: traditional incarceration (jail) prevents customers from reoffending while they are in service, yet the alternative programs (CC) suffer from the risk that customers can recidivate while in service and impose societal issues – which has been a primary concern for many of the alternative programs. This benefit of preventing the risk of "recidivating during service" by putting people in jail is substantial under a short recidivism window; however, it decreases when we look at a longer recidivism window. Therefore, when we consider the long-term effect, the benefit of CC may outline the risk of recidivism. Decision-makers should account for this factor when considering the pros and cons of non-traditional alternatives for incarceration. Note that Assumption 2 is consistent with findings in the criminology field, i.e., the risk of recidivism peaks fairly quickly and then diminishes as the time goes (Schmidt and Witte 2012). The peak time $\tilde{\tau}$ is several months (Durose et al. 2014), while the sentence length of jail is usually less than one year; hence $T_1 - d(\texttt{jail}, m) > \tilde{\tau}_m, \forall m$ is satisfied when we consider the typical 3-year recidivism window.

6. Efficient Algorithm based on Separation of Timescales

The structural insights derived in Section 5 are valuable for understanding the system dynamics. However, when it comes to practical decision-making, these insights alone are insufficient. To make informed placement decisions (requires calculating the values of V^*), we need an algorithmic solution. The primary hurdle in developing this solution still lies in dealing with the lengthy and non-memoryless sentence length (LOS), as it leads to the state space growing exponentially in the sentence length and renders conventional MDP methods intractable. We develop a novel algorithm

that addresses this curse of dimensionality. The main idea of our algorithm is to leverage the long LOS (relative to the arrival rate) and the natural separation between two timescales. That is, new arrivals happen on a daily scale while the sentence lengths are in the order of weeks and months. Hence, the relative change in the system state is small when looking at the daily level. This allows us to formulate a "stable" placement policy on a weekly basis, transforming the original MDP problem into one using weekly batch arrivals with a *proportional placement decision*, specified in Section 6.1. The key to this transformation is the elimination of the need to track LOS, as we can approximate departures under the stable policy, which significantly reduces the state space's dimensionality. Combining this approximation with a policy-gradient optimization technique, we introduce the final algorithm in Section 6.2.

6.1. Transformed MDP: Batched Decisions

In the transformed MDP, we consider each week as a decision epoch. At the beginning of each week, we observe the system state along with the batch of new arrivals that have accumulated from the previous week and need to be placed. We make two important adaptions compared to the original MDP model. First, instead of determining the placement for each arrival sequentially, we use a placement *proportion*, $\theta_{j,m}$, to determine the number of class m customers, out of the current batch of arrivals, to be placed in station j. This placement proportion allows us to apply a gradient-based method to efficiently solve this transformed MDP, a key for designing the final algorithm that is scalable. Specifically, the proportion $\theta_{j,m}$ is continuous on $[0,1]$ and allows for the search of gradient, contrasting the discrete action space in the original problem. Second, we no longer track the LOS in the state. Instead, we use a novel approximation that leverages the separation of the timescales for the state transition. We elaborate on the two adaptions in sequence.

6.1.1. Batched decisions and implementation. Based on the observed state s in this epoch, the placement policy for the batched arrivals follows

$$\Theta(s) = \{\theta_{j,m}\}_{j\in\mathcal{J},m=1,\dots,M},$$

where $\theta_{j,m} \in [0,1]$ is the proportion of class m customers out of this batch to placed in station j. It is worth noting that this batched decision is not directly employed for decision-making in the original MDP. Instead, the value function solved from the transformed MDP is used to derive a deterministic policy for sequential arrivals in the original problem. This is achieved through a one-step optimization of the Bellman equation, as detailed in Section 7.1. In Appendix EC.6.2, we detail the system state, transition, cost, and objective for this transformed MDP.

For the batched arrival and proportion-based decision, Theorem 1 indicates an important property: the value function with respect to the number of new participants placed in jail versus CC

should only have one local optimum, which is also globally optimal. We formalize this in the following corollary. For notational simplicity, we use the vector $e(\texttt{JL}_{m,0}, \texttt{CC}_{m,0})$ to represent the vector of unit occupancy in the two stations, namely $(e_{\texttt{jail},m,0}, e_{\texttt{cc},m,0})$.

COROLLARY 3 **(Superconvexity for global optimality)**. *Given n new class-m arrivals, the amount to place in jail should only have one local minimum $k \in \{0, 1, ..., n\}$, which satisfies*

$$\begin{aligned} V^*\big(s + (k, n-k)\cdot e(\mathit{JL}_{m,0}, \mathit{CC}_{m,0})\big) &\le V^*\big(s + (k+1, n-k-1)\cdot e(\mathit{JL}_{m,0}, \mathit{CC}_{m,0})\big); \\ V^*\big(s + (k, n-k)\cdot e(\mathit{JL}_{m,0}, \mathit{CC}_{m,0})\big) &\le V^*\big(s + (k-1, n-k+1)\cdot e(\mathit{JL}_{m,0}, \mathit{CC}_{m,0})\big) \end{aligned} \tag{9}$$

for any state $s \in \mathcal{S}$. The equality is attained on the boundary $k = 0$ or n.

Corollary 3 says that only one combination $(k, n-k)$ is optimal for the placement decisions to jail and CC, respectively, out of any given n arriving class m individuals. This indicates that the proportion $\theta_{j,m}$ should also only have one optimum point, which is important for our algorithmic development: if we can find the local optimum $\theta_{j,m}$ (typical for many algorithms including gradient-based ones), this guarantees that we also find the global optimum.

Proof. We prove Corollary 3 by contradiction. Let $\Delta V_n(k) = V^*\big(s + (k, n-k)\cdot e(\texttt{JL}_{m,0}, \texttt{CC}_{m,0})\big) - V^*\big(s + (k-1, n-k+1)\cdot e(\texttt{JL}_{m,0}, \texttt{CC}_{m,0})\big)$. Then (9) is equivalent to $\Delta V_n(k+1) \ge 0$ and $\Delta V_n(k) \le 0$. Suppose that for the n new arrivals, the value function V^* has two local optima, i.e., (9) holds for two values of k_1 and k_2, with $k_1 < k_2$. Then we have

$$\begin{aligned} &\Delta V_n(k_1 + 1) \ge 0 \text{ and } \Delta V_n(k_1) \le 0; \\ &\Delta V_n(k_2 + 1) \ge 0 \text{ and } \Delta V_n(k_2) \le 0. \end{aligned} \tag{10}$$

However, the superconvexity of V^* proved in Theorem 1 implies that $\forall k \in \{0, ..., n-1\}$, $\Delta V_n(k+1) \ge \Delta V_n(k)$. Thus, for $k_2 \ge k_1 + 1$, we have $\Delta V_n(k_2) \ge \Delta V_n(k_1 + 1)$, contradicting (10). □

6.1.2. State transition approximation The second important adaption we make for the transformed MDP is to remove the LOS index and only track the total number of customers of each type in each station in the state, i.e., the system state is captured by $S = \{X_{j,m}\}_{j=1,...,J; m=1,...,M}$. This adaption greatly reduces the state dimension. Meanwhile, under the batched decision setting and Poisson arrival assumption, we are able to approximately characterize the state transitions. The key is to use the *ordered statistics* property for the Poisson process. In the rest of the algorithm development, we assume that the new arrivals from each class follow a Poisson process with rate λ_m, and that $\theta_{j,m}$ is stable over the past weeks under the separation of timescales as discussed earlier. With the Poisson thinning property, the arrivals routed to station j also follow a Poisson process with rate $\lambda_{j,m} = \lambda_m \theta_{j,m}$. Correspondingly, the total number of customers that are routed (placed) to station j by time t, $N_j(t)$, is a Poisson random variable. The ordered statistics property says

that, given $N_j(t) = k$, each of the k customers can be considered as arrived and placed uniformly over $[0, t]$.

To elaborate, suppose the sentence length is three weeks, and a total of six arrivals were placed in one of the stations in the first three weeks. To approximate the state transition calculation, it is sufficient to get the number of customers (out of the six) to be released in the fourth week, that is, the number of participants who arrived in the first week and reached the three-week sentenced length. Conditioning on $N_j(t) = 6$, the arrival times of these six customers, $t_1, \ldots, t_6$, are distributed as ordered statistics of six independent uniform variables. In other words, $\{t_1, \ldots, t_6\} \stackrel{d}{=} \{u_1, \ldots, u_6\}$, where u_i's are independent and identically distributed (i.i.d) random variables. They follow the uniform distribution $U\big(0, d(j,m)\big)$ and are ordered chronically, with $d(j,m) = 3$ weeks in this example. Correspondingly, the probability that each t_i's falls in the first week is $1/3$, and the number of released customers in the fourth week follows a binomial$(6, 1/3)$ distribution.

Using this ordered-statistics argument, we can write the general form for the number of participants to be released in the next epoch as a binomial random variable $B^r_{j,m} \sim \text{Binom}\left(x_{j,m}, \frac{\Delta}{d(j,m)}\right)$, where Δ is the length between two decision epochs (e.g., $\Delta = 1$ week in the example), and $\{x_{j,m}\}$ is the reduced state (class m customers in station j). The state transition can be characterized with this approximate number of releases and the number of new placements $B^a_{j,m} \sim Pois(\lambda_{j,m})$:

$$X'_{j,m} = B^a_{j,m} + (x_{j,m} - B^r_{j,m}).$$

See details of the transition probability specification in Appendix EC.6.3.

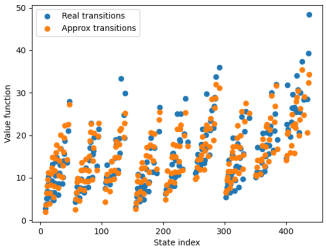
Approximation accuracy. We numerically evaluate the accuracy of this approximation by comparing the value function solved from the original MDP (tracking LOS in the state) and that from the approximate state transition. Even if we utilize Monte Carlo Policy Evaluation (Singh and Sutton 1996) in the numerical comparison, the original MDP can only be solved for small-sized and medium-sized problems due to the large state space. Figure 3(a) shows the quantile distribution of relative estimation errors. Figures 3(b) and 3(c) visually compare the value functions with respect to each state. This comparison result suggests that the approximation is sufficiently accurate.

6.2. Policy Gradient Actor-Critic Algorithm

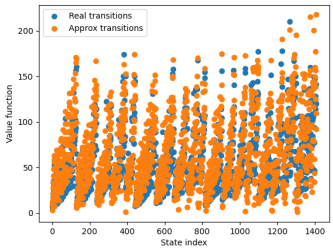
We use a policy-gradient-based, Actor-Critic algorithm to solve the transformed MDP. The idea of actor-critic is based on policy iteration (Puterman 2014). Starting from an initial $\{V(s)\}_{s\in\mathcal{S}}$, the algorithm proceeds between (i) the *policy improvement* step by finding a greedy policy $\Theta(s) = \{\theta_{j,m}(s)\}$ from the current value function $V(\cdot)$ by minimizing $C(s) + \gamma\mathbb{E}_\Theta[V(s')]$ for all $s \in \mathcal{S}$; and (ii) the *policy evaluation* step by evaluating the value function under the obtained policy $\Theta(s)$ via solving the Bellman equation. The algorithm iterates between the two steps until the policy converges and finds the optimal policy. To facilitate the computation without enumerating the entire

	25%	50%	75%
Small model	0.02	0.05	0.06
Medium	0.03	0.06	0.09

(a) Quantiles of relative estimation errors



(b) Small model



(c) Medium model

Figure 3 **Monte Carlo policy evaluation of the value functions under the original MDP model and the transformed model. The numerical setting is detailed in Appendix EC.7.1.**

state space, we use a simulation-based method, which leads to the tabular actor-critic algorithm specified in Algorithm 1. In this algorithm, we denote the probability of making placement decisions $a = \{a_{j,m}\}_{j,m}$ at state s under policy Θ as $\pi_\Theta(a|s)$, where $a_{j,m} \sim Pois\big(\lambda_m \theta_{j,m}(s)\big)$ represents the number of class m individuals placed in station j.

Algorithm 1: Tabular batched actor-critic policy gradient

Input : Step sizes $\alpha_\theta, \alpha_\omega$. Batch size N. Number of iterations T.

Output: Policy $\{\theta_{j,m}(s)\}_{j,m}, s \in \mathcal{S}$ and value function $V(s), s \in \mathcal{S}$.

Initialize $\{\theta_{j,m}(s)\}_{j,m}, s \in \mathcal{S}$ at random, $V(s) = 0, s \in \mathcal{S}$. Initialize state s_1 at random.

for $t = 1, 2, ..., T$ **do**

  for $n = 1, 2, ..., N$ **do**

    Set current state s_t.

    Sample and store the placement of new arrivals $a_n \sim \pi_\Theta(a|s_t)$; sample and store next state $s'_n \sim P(s'|s_t, a_n)$.

  end

  Update the policy parameters:

$$\theta_{j,m} \leftarrow \theta_{j,m} - \alpha_\theta V(s_t) \cdot \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta_{j,m}} \ln \pi_\Theta(a_n|s_t). \tag{11}$$

  Update the value function with TD(0):

$$V(s_t) \leftarrow V(s_t) + \alpha_\omega \left(\frac{1}{N} \sum_{n=1}^{N} \big(C(s_t) + \gamma \cdot V(s'_n)\big) - V(s_t) \right). \tag{12}$$

  Sample the next states $s_{t+1} \sim P(s'|s_t, a_N)$.

end

Policy Improvement. Equation (11) corresponds to the policy improvement step in PI, i.e., updates the policy parameters $\{\Theta\}$'s based on the current value function approximation. The continuous property of placement proportion enables us to learn the policy through policy gradient-based methods. Specifically, we improve the actor via the gradient method, updating parameters in the direction of the estimated policy gradient. According to the policy gradient theorem (Sutton et al. 1999), the policy gradient $\nabla_\Theta V^\pi(s) \propto \mathbf{E}[V^\pi(s)\nabla_\Theta \ln \pi_\Theta(a|s)]$. Denote the stations by $j_1, ..., j_{|\mathcal{J}|}$. In the transformed MDP, $\nabla_{\theta_{j,m}}$ in (11) can be written as:

$$\nabla_{\theta_{j,m}} \ln \pi_{\theta(a|s)} = \frac{a_{j,m}}{\theta_{j,m}} - \frac{a_{j_{|\mathcal{J}|},m}}{1 - \sum_{k=1}^{|\mathcal{J}|-1} \theta_{j_k,m}},\ j = j_1, ..., j_{|\mathcal{J}|-1},\ m = 1, ..., M.$$

See its derivation in Appendix EC.6.4. The policy gradient method is guaranteed to converge to a local minimum, where the norm of the gradient equals zero, i.e., $||\nabla_{\theta_{j,m}}|| = 0$ (Konda and Tsitsiklis 1999). Corollary 3 provides a justification that the local optima is also globally optimal.

Policy Evaluation. Equation (12) corresponds to the policy evaluation step, which uses simulated samples to evaluate the value function $V(s)$ under the policy in this iteration. Specifically, the update in (12) uses temporal difference (TD) learning, which updates the value function via the TD error $\frac{1}{N}\sum_{n=1}^{N}(c + \gamma \cdot V(s'_n)) - V(s_t)$ in the sample-average form (Tesauro 1995).

7. Case Study and Practical Implications

In this section, we perform a case study with data from our community partner, the Tippecanoe County Community Corrections (TCCC), to demonstrate the performance of our algorithmic solution and generate important practical insights. We first introduce our simulation platform and discuss the sequential decision-making process in Section 7.1. Our simulation platform removes several assumptions made for analytical tractability in the structural analysis, allowing us to substantiate the theoretical results. In Section 7.2, we demonstrate the superior performance of our algorithm against benchmark policies over a variety of system parameters. In Section 7.3, we illustrate how our approach can provide valuable insights for capacity planning.

7.1. Simulation Evaluation Platform

As shown in the data pipeline (Figure 4), our simulation platform acts as the "oracle" environment, generating simulation samples for both value function learning and performance evaluation. Compared with the model described in Section 4.2, we incorporate additional two major realistic features:

1. Instead of viewing CC as a single station as in theoretical analysis, we differentiate CC into two programs in the simulation, corresponding to the actual Work Release and Home Detention programs in TCCC.

2. We explicitly model the recidivism through the returning flow of recidivated individuals (Remark 1) and the proportion of technical violations returning to jail from CC.

In addition to these two major additions, we also incorporate the congestion-based recidivism risk and risk- and need-based customer classification with an additional dimension of treatment needs to measure the potential benefits of different CC programs.

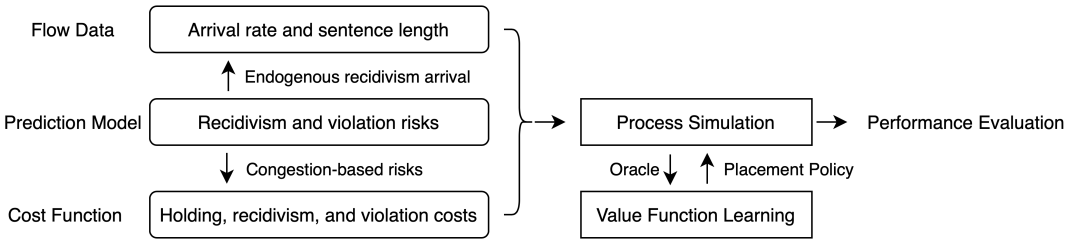


Figure 4 Data pipeline of the process simulation, value function learning, and performance evaluation.

For practical implementation, we do not directly use the solved policy $\theta_{j,m}$ from the transformed MDP in Section 6.2. Instead, we "de-randomize" this proportion-based policy through a one-step optimization and prescribe individual-specific decisions. Specifically, after solving the value functions through the transformed MDP, we obtain the placement decision for each individual by solving a one-step optimization of the Bellman equation (5). The prescribed placement decision for the kth arrival from class m is given by:

$$a^*_{m,k} = \underset{a_{m,k}\in\mathcal{J}}{\arg\min} \big(C(s) + \gamma \cdot \frac{1}{N} \sum_{s'\in\mathcal{S}} V(s')\big), \tag{13}$$

where s denotes the system state prior to the decision, and s' is the system state in the next epoch after taking action $a_{m,k}$. We use the average of N samples to approximate the transition $p(s'|s,\pi)$ in the Bellman equation, where s' is generated by our simulation platform, and $V(s')$ is the value function estimated from Algorithm 1. Note that the value function learned via Algorithm 1 uses the reduced state $s = \{x_{j,m}\}_{j,m}$, which can be computed as $x_{j,m} = \sum_{l=0}^{d(j,m)} x_{j,m,l}$ based on the state we see in the simulation. We leverage the individualized cost functions $C(\cdot)$ and aggregated value functions $V(\cdot)$ in Equation (13) to accommodate customer heterogeneity while keeping the algorithm remains scalable. Specifically, we first solve a single-class-multiple-program model for $V(\cdot)$ using Algorithm 1. Then in decision-making, we employ a cost function $C(\cdot)$ that captures the risk and treatment needs of each individual. More details about the algorithm adaptation for the real-sized case studies and model calibration via real data are provided in Appendix EC.6.5 and EC.7.3, respectively.

Benchmark Policies. We refer to the policy solved from Algorithm 1 (with the above adaptions) as the RL-based `dynamic policy`. We consider two benchmarks for performance comparison. The first one is the `static policy` that only considers individual risk but ignores the system effect. The second benchmark is the `threshold-based policy`, where predefined thresholds for occupancy levels are chosen and optimized offline for each station. Under this policy, individuals are assigned to "available" stations that have occupancy levels below the respective thresholds. In cases where multiple stations are available, the preference is guided by individual-level cost minimization. When no stations are available (below the threshold), individuals are assigned to the station that minimizes their individual cost, ignoring congestion. The handling of batch arrivals adheres to the FIFO order. This threshold-based policy improves over the static policy by accounting for congestion effects. However, the thresholds are static and lack the state-dependent adaptability featured in our RL-based dynamic policy.

7.2. Value of the Dynamic Approach

As explained in Section 1.1.2, constrained capacity necessitates difficult tradeoffs between individual and societal considerations, especially for individuals in the gray area. In this section, we highlight the value of dynamically incorporating the necessary system consideration in breaking the cycle of recidivism by comparing it to the static policy, which can lead to an exacerbated vicious cycle, as discussed in Section 3.2. We then discuss other benefits brought by the dynamic policy including reducing the workload volatility and providing the flexibility to balance across various tradeoffs.

Table 3 compares several performance metrics of the static policy and our RL-based dynamic policy under two staffing scenarios: 15 case managers in the high-staff scenario and 9 case managers in the low-staff scenario. For cost-related metrics, we set the baseline unit holding costs to be \$54 and \$15 for jail and CC, respectively (see Vera Institute of Justice) and the unit costs of recidivism and violation to be \$151,662 and \$50,000, respectively (First 2018, CSG Justice Center 2019). The staffing level affects the recidivism and violation rates among CC participants with the functional form specified in Appendix EC.7.3. We note that WR is a residential program and its availability is limited by the physical space; hence, all policies perform similarly on WR placement and occupancy (Table 3). Consequently, we focus on examining the impact of placement policies on jail and HD occupancy levels, and the resulted recidivism and violation costs in the rest of the analysis.

Avoiding the feedback loop. The first two rows in Table 3 are from simulating the static policy in an environment without any negative consequence from overcrowded jail or CC. Thus, the outcomes are the same under the high- and low-staffing scenarios. The third and fifth rows are from the setting where overcrowded jail or CC increases individual recidivism and violation

Table 3 Performance metrics of the static and dynamic policy: recidivism cost (RecidC), violation cost (VioC), occupancy (average number of individuals) in jail, Work Release (WR), Home Detention (HD), and the average caseload of each case manager in HD.

	Policy	RecidC	VioC	Jail Ocp.	WR Ocp.	HD Ocp.	HD Caseloads
No Congestion	Static (high-staff)	2117	1324	236	147	540	36
	Static (low-staff)	2117	1324	236	147	540	60
With Congestion	Static (high-staff)	2507	1623	252	149	515	34
	Dynamic (high-staff)	2249	1578	223	148	567	38
	Static (low-staff)	4580	5162	292	149	675	75
	Dynamic (low-staff)	3716	3065	328	148	549	61

risks. We can see that the three-year recidivism and violation costs increase by 18% and 23% in the high-staff scenario and by 116% and 289% in the low-staff scenario. These significant increases result from a feedback-loop effect. That is, a higher jail occupancy leads to a higher recidivism rate, which results in more individuals staying in the criminal system and further crowding both the jail and CC. More importantly, the static policy ignores the congestion effect in the HD program, and assigns more-than-optimal number of individuals to HD. This results in a heavy workload for each case manager, particularly in the low-staff scenario (caseload increases from 60 to 75). Consequently, the number of violations rises significantly due to the loosened supervision. Even worse, a portion of these violations end up as jail incarceration, further congesting the jail and triggering more subsequent recidivism events, which increases the number of individuals remaining in the system over the long term. In essence, the static policy, which relies solely on individual risk without proper consideration of system congestion, triggers a *cascading effect* due to congestion and results in a downward spiral.

Our RL-based dynamic policy can effectively avoid this downward spiral, as demonstrated in the fourth and sixth rows in Table 3. The dynamic policy significantly reduces the recidivism and violation costs compared to the static policy: 10% and 3% in the high-staff scenario, and more prominently, 19% and 41% in the low-staff scenario. This superior performance is achieved by properly balancing the system congestion level among the different stations (jail vs CC programs) in a real-time fashion. Because of this load balancing, the overall recidivism and violations are much smaller, leading to fewer individuals in the system in the long run (safer community) and avoiding case manager burnout (the caseload remains around 60 in the low-staff scenario).

Reducing workload volatility. The dynamic policy offers an additional benefit – preventing excessive workload in either station – thereby reducing overall workload volatility. To showcase this, Table 4 reports the standard deviations of workload in jail and HD under different policies. The RL-based dynamic policy demonstrates its ability to significantly reduce workload volatility, particularly in the HD program, which is crucial to case managers. Workload volatility often contributes to a stressful working environment and can lead to burnout. Through the reduction of

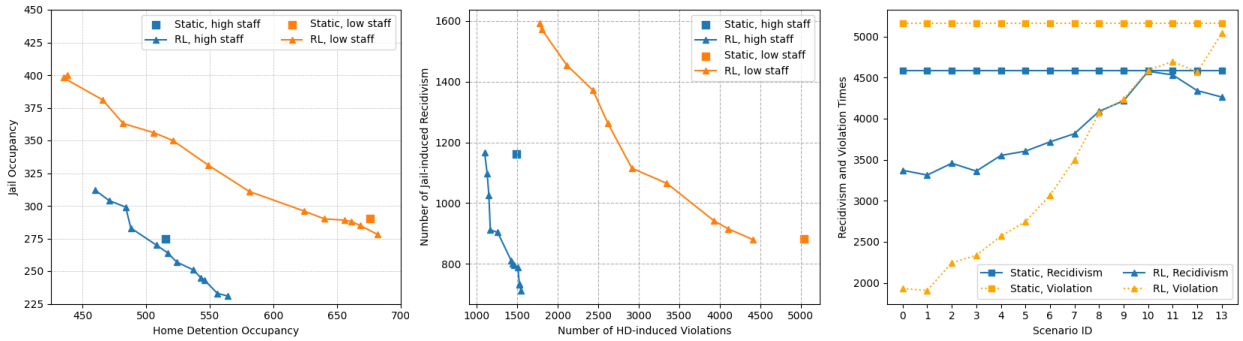
workload volatility, our dynamic policy provides case managers with a more predictable and manageable workload within existing budget and capacity constraints. This, in turn, has the potential to mitigate stress and burnout, enhance supervision efficiency, and improve individual reintegration (Liljegren et al. 2021, Bergman et al. 2022).

Table 4 Standard deviations of workload in jail and CC under different policies. The workload values are normalized across 100 simulations. The number following the ± sign denotes the half-width of the 95% confidence interval for the standard deviation.

	Dynamic Policy	Threshold-based Policy	Static Policy
Jail	$5.00\% \pm 0.03\%$	$5.86\% \pm 0.04\%$	$4.92\% \pm 0.03\%$
Home Detention	$4.00\% \pm 0.03\%$	$5.26\% \pm 0.03\%$	$6.31\% \pm 0.04\%$

Tradeoff curves. The estimates of the unit holding, recidivism, and violation costs vary in a large range, e.g., the jail holding cost (per customer per day) spans from \$21 to \$299 in the state of Indiana (Vera Institute of Justice). Our dynamic policy has the flexibility to balance among key tradeoffs given different cost combinations, i.e., the dynamic policy has the flexibility of achieving various performance outcomes by using the cost parameters as a "tuning knob." We show in Figures 5(a) and 5(b) the tradeoff curves under the high- and low-staff scenarios when we vary the unit holding cost ratio between the jail and CC programs. Specifically, Figure 5(a) plots the tradeoff curve between jail congestion and HD congestion, and Figure 5(b) plots the tradeoff curve between jail-induced recidivism and HD-induced violations. We observe that, if the dynamic policy results in a comparable level of jail-induced recidivism as the static policy, it significantly reduces the number of HD-induced violations; and vice versa. Moreover, Figure 5(c) compares the numbers of recidivism and violations under both policies in the low-staff scenario, with the scenario ID on the x-axis arranged in ascending order of HD occupancy under static policy. As expected, the number of violations increases as the HD occupancy gets higher. The RL-based policy can keep recidivism and violation rates lower than the static policy.

Figure 6 plots a different set of tradeoffs from the cost perspective. We plot the total recidivism and violation costs in million dollars (M\$) when we vary the unit holding cost for the jail or CC programs. Figure 6(a) shows that our dynamic policy results in the lowest (and relatively consistent) recidivism and violation costs compared to the other two benchmark policies. In contrast, the threshold-based and static policies exhibit a rapidly increasing trend in recidivism and violation costs as the unit jail holding cost rises, especially for violation. This trend can be explained by Figure 6(b), which shows that as the unit jail holding cost increases, HD becomes a more favorable option due to its lower holding cost. Consequently, the static and threshold-based policies tend to place more-than-optimal customers in HD and overburden HD case managers, exacerbating the

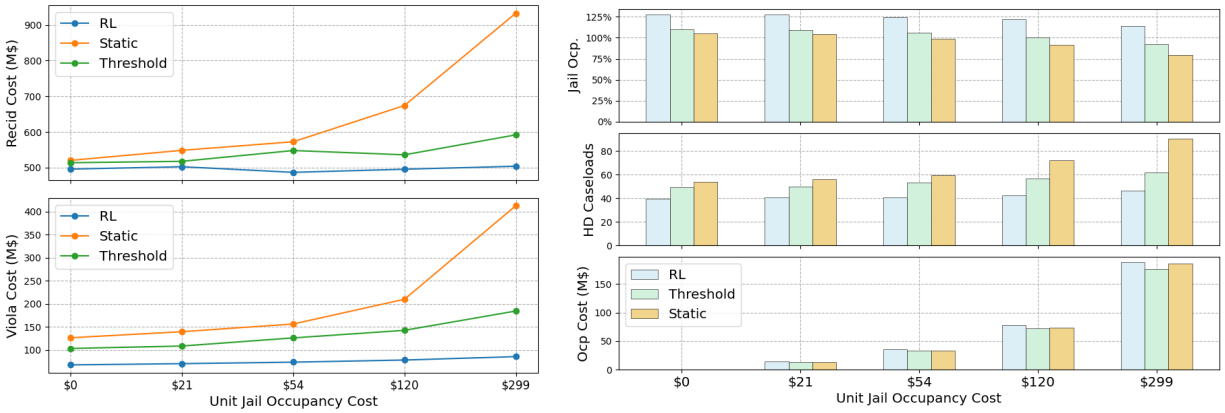


(a) Occupancy level tradeoff (b) Performance tradeoff (c) Performance (low-staff)

Figure 5 Tradeoff curves for the dynamic policy with different congestion sensitivity.

feedback loop and resulting in more recidivism and violations. In particular, the static policy is prone to severe congestion in the HD station. Additional sensitivity analysis results on customer arrival rates and lengths of recidivism window can be found in Appendix EC.7.5.

Importantly, these tradeoff curves and performance analyses can offer stakeholders (such as judges, prosecutors and CC managers) a comprehensive array of potential options by visualizing combinations of jail and HD occupancy levels, along with other essential performance metrics. This approach avoids enforcing a specific "optimization" objective or prescribing single definitive placement decision. In addition, the curves for scenarios with high and low staffing imply performance enhancement through increased staffing, a topic we discuss in the following section.



(a) Individual-level cost (b) System-level cost

Figure 6 Performance comparison under different unit jail holding costs.

7.3. Insights on Capacity Planning

By incorporating system effects, our framework can analyze outcomes under various capacities, thus guiding capacity planning. In this section, we explore outcomes under different placement

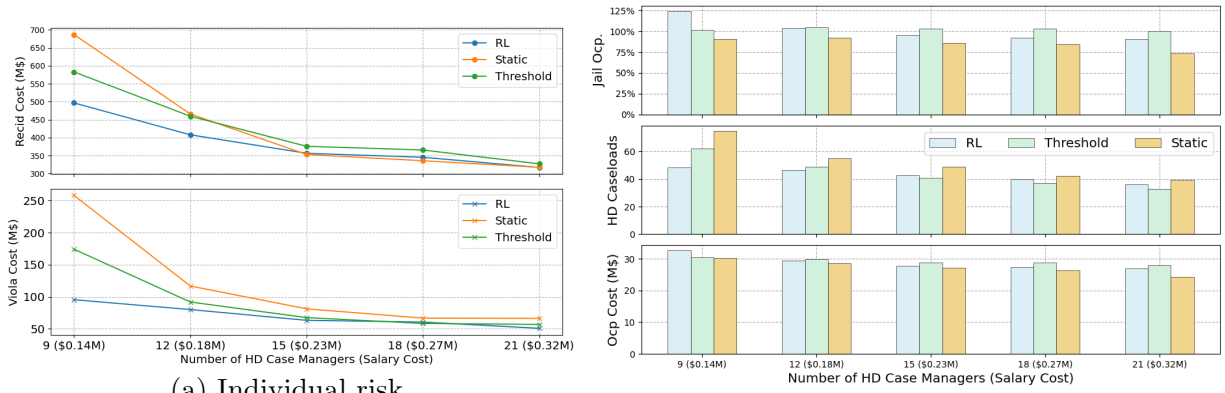
policies and capacities by adjusting capacity parameters in the congestion-dependent risk functions; see details in Appendix EC.7.3. Comparing the two tradeoff curves in Figure 5(a) suggests that having a larger number of HD staff (case managers) can alleviate congestion in both the jail and HD by avoiding the cascading effect of congestion. In particular, note that the tradeoff curve slope is around -0.5 in the low-staff scenario and is around -1 in the high-staff scenario. This is because under low staffing, the HD program is already congested, and placing any extra individuals there would exacerbate the congestion effect. Correspondingly, increasing the number of case managers can bring substantial benefits. Enhanced supervision leads to reduced recidivism and violation rates, fostering system-wide improvement by reducing the number of individuals staying in the criminal system. Ultimately, this reduces the long-term incarceration population.

We further examine the benefit of capacity expansion from the cost-effectiveness perspective, with a focus on increasing the HD staffing level. Unlike jail or WR, which has limited capacities determined by physical space (cells, residential dorms), the capacity of the HD program is more flexible and depends on the number of case managers available. Figure 7(a) shows that adding a few case managers to the HD program can result in a substantial reduction in recidivism and violation costs. Considering a three-year window, increasing the number of case managers from 11 to 15 requires an additional monetary cost of around $2,373,900 (estimated based on the input from TCCC), while the cost saving from avoided recidivism would be as high as $40,657,524. A detailed breakdown of these estimated numbers is in Appendix EC.7.4. Figure 7(b) shows the corresponding occupancy levels in each program and total occupancy costs. We can see that increasing the number of case managers can reduce jail overcrowding by (i) placing more individuals in CC and (ii) decreasing the number of returning individuals due to recidivism and violations. The marginal benefit of staffing additional case managers diminishes though.

The capacity analysis above offers valuable insights to aid community corrections in deciding the ideal number of case managers to hire within their budget limitations, along with the objective of decreasing recidivism, violations, and the strain on criminal justice systems due to overcrowding. It is important to highlight that we have shared our cost-effectiveness analysis with the director of our community partner, who presented our results during a county council meeting to support their upcoming budget justifications.

8. Conclusions

Jail diversion programs, such as community corrections, offer a promising solution to address jail overcrowding and reduce crime rates. However, the current practice tends to focus primarily on individual risks when deciding whom to place into CC programs, without adequately considering the potential impact on system congestion. As CC capacity may not be expanded in real time,



(a) Individual risk

(b) System occupancy

Figure 7 **Performance comparison under different numbers of Home Detention case managers**

sending all eligible individuals to CC can cause extreme crowding under stochastic fluctuation, which shifts the overcrowding issue from jail to CC programs, overwhelming case managers and resulting in increased violations and recidivism rates. This can set off a chain reaction, causing more individuals to be re-sentenced to incarceration and exacerbating crowding within the criminal justice system. To capture all the intricate tradeoffs in this placement decision process, we develop a novel decision-support tool based on an MDP framework. Our approach goes beyond conventional queueing routing models by incorporating key features specific to this context. This includes accounting for predetermined sentence terms (deterministic service time), which requires tracking length of stay (LOS), and explicitly modeling congestion-dependent costs, departing from the typical memoryless and linear holding cost assumptions. We overcome technical challenges posed by these features, establishing the superconvexity of the value function and policy monotonicity, and proposing a novel algorithmic solution to handle the large state space.Through a comprehensive evaluation using data provided by our community partner, we demonstrate the effectiveness of our dynamic approach in mitigating the cascading effect and improving both public safety and individual outcomes. Additionally, we conduct cost-effective analyses to inform staffing planning and prevent case manager burnout. Overall, our research provides a comprehensive framework and practical tools to guide decision-making in jail diversion programs, considering the complex interplay between individual and system-level factors.

It is important to stress that the criminal justice system is inherently complex, and no mathematical model can fully encapsulate its intricacies. Hence, our analytical results should be interpreted with a nuanced perspective that recognizes the limitations of our approach and the broader context of the real-world criminal justice system. As discussed in Section 1.1, our aim is not to provide a definitive solution but rather to provide a connected system view. We also believe our work will

spark future analytical interest and contribute to a growing body of research focused on improving decision-making in this critical area.

We conclude by discussing potential directions for future work. First, our modeling framework, based on grasping the key tradeoffs involved, could extend beyond initial placement decisions to a series of criminal justice decisions. These include placement choices for (i) individuals pleaing guilty with time served, (ii) temporary placements when certain community corrections programs reach capacity, and (iii) early release decisions for the incarcerated population. By exploring placement decisions at different points in time, we can further examine the interplay between these choices, individual risk factors, individual (remaining) sentence length, and system effects. Furthermore, our findings provide valuable insights for budget planning under soft capacity constraints. A natural extension would be to investigate a bilevel problem that considers both capacity planning and placement optimization under hard capacity constraints. Moving forward, a crucial direction is to refine jail diversion decisions to address biases linked to race, ethnicity, and socioeconomic status. We assume an accurate and unbiased risk classification and focus on the placement decision making in this paper. Moving forward, we are interested in developing robust decision support tools that account for potential prediction errors. Beyond this, a more comprehensive approach could also encompass the integration of constraints on the fairness and consistency of placement decisions and outcomes within decision modeling.

References

David S Abrams. Is pleading really a bargain? *Journal of Empirical Legal Studies*, 8:200–221, 2011.

Markus Aloyan. Home confinement in the United States: The evolution of progressive criminal justice reform. *Trento Student Law Review*, 2, 2020.

Mariel Alper, Matthew R Durose, and Joshua Markman. *2018 update on prisoner recidivism: A 9-year follow-up period (2005-2014)*. US Department of Justice, Office of Justice Programs, Bureau of Justice, 2018.

Phil Ansell, Kevin D Glazebrook, and Christopher Kirkbride. Generalised 'join the shortest queue' policies for the dynamic routing of jobs to multi-class queues. *Journal of the Operational Research Society*, 54 (4):379–389, 2003a.

Phil Ansell, Kevin D Glazebrook, José Nino-Mora, and M O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57:21–39, 2003b.

Ozgur M Araz, Mayteé Cruz-Aponte, Fernando A Wilson, Brock W Hanisch, and Ruth S Margalit. An analytic framework for effective public health program design using correctional facilities. *INFORMS Journal on Computing*, 34(1):113–128, 2022.

Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C Spaulding, and Jagpreet Chhatwal. Prioritizing hepatitis C treatment in US prisons. *Operations Research*, 67(3):853–873, 2019.

Alon Bergman, Hummy Song, Guy David, Joanne Spetz, and Molly Candon. The role of schedule volatility in home health nursing turnover. *Medical Care Research and Review*, 79(3):382–393, 2022.

Leah Brogan, Emily Haney-Caron, Amanda NeMoyer, and David DeMatteo. Applying the risk-needs-responsivity (RNR) model to juvenile justice. *Criminal Justice Review*, 40(3):277–302, 2015.

C.G. Camp and G.M. Camp. The corrections yearbook: Adult corrections. *Criminal Justice Institute*, pages 1994–1995, 1999.

CSG Justice Center. Confined costly: How supervision violations are filling prisons and burdening budgets, 2019. URL `https://csgjusticecenter.org/publications/confined-costly/`. Accessed: May 2, 2023.

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

Wendy Davis. Diversion sentencing may offer an alternative path to justice—but how fair is it?, 2019. URL `https://www.aclukansas.org/en/news/diversion-sentencing-may-offer-alternative-path-justice-how-fair-it`.

Sarah L Desmarais and Evan M Lowder. Principles and practices of risk assessment in mental health jail diversion programs. *CNS spectrums*, 25(5):593–603, 2020.

Matthew R Durose, Alexia D Cooper, and Howard N Snyder. *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010*, volume 28. US Department of Justice, Office of Justice Programs, Bureau of Justice, 2014.

Grant Duwe and Susan McNeeley. The effects of intensive postrelease correctional supervision on recidivism: A natural experiment. *Criminal Justice Policy Review*, 32(7):740–763, 2021.

Eugene A Feinberg and Mark E Lewis. On the convergence of optimal actions for Markov decision processes and the optimality of (s, S) inventory policies. *Naval Research Logistics (NRL)*, 65(8):619–637, 2018.

Illinois Result First. The high cost of recidivism, 2018. URL `https://spac.icjia-api.cloud/uploads/Illinois_Result_First-The_High_Cost_of_Recidivism_2018-20191106T18123262.pdf`. Accessed: May 2, 2023.

Carlos Franco-Paredes, Nazgol Ghandnoosh, Hassan Latif, Martin Krsak, Andres F Henao-Martinez, Megan Robins, Lilian Vargas Barahona, and Eric M Poeschla. Decarceration and community re-entry in the COVID-19 era. *The Lancet Infectious Diseases*, 21(1):e11–e16, 2021.

Larry K Gaines and Roger LeRoy Miller. *Criminal justice in action*. Cengage Learning, 2021.

Brandon L Garrett, Alexander Jakubow, and John Monahan. Judicial reliance on risk assessment in sentencing drug and property offenders: A test of the treatment resource hypothesis. *Criminal Justice and Behavior*, 46(6):799–810, 2019.

Abraham George, Warren B Powell, and Sanjeev R Kualkarni. Value function approximation using multiple aggregation for multiattribute resource management. *Journal of Machine Learning Research*, 9(10), 2008.

Kevin D Glazebrook, Christopher Kirkbride, and Jamal Ouenniche. Index policies for the admission control and routing of impatient customers to heterogeneous service stations. *Operations Research*, 57(4): 975–989, 2009.

Laurie A Gould, Matthew Pate, and Mary Sarver. Risk and revocation in community corrections: The role of gender. *Probation Journal*, 58(3):250–264, 2011.

Betsy S Greenberg. M/G/1 queueing systems with returning customers. *Journal of applied probability*, 26 (1):152–163, 1989.

David F Greenberg. Recidivism as radioactive decay. *Journal of Research in Crime and Delinquency*, 15: 124, 1978.

Margaret Hamilton. People with complex needs and the criminal justice system. *Current Issues in Criminal Justice*, 22(2):307–324, 2010.

Angela Hawken. All implementation is local. *Criminology & Public Policy*, 15:1229, 2016.

Christian Henrichson. The price of jails: Measuring the taxpayer cost of local incarceration. 2015.

Junfei Huang, Boaz Carmeli, and Avishai Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

David Hughes. A simulation modeling approach for planning and costing jail diversion programs. *Simulation Strategies to Reduce Recidivism: Risk Need Responsivity (RNR) Modeling for the Criminal Justice System*, page 223, 2013.

Indiana Prosecuting Attorneys Council. Diversion and deferral guidelines, 2019. URL `https://www.in.gov/ipac/files/Diversion-and-Deferral-Guidelines-approved-February-22,-2019.pdf`.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Ger Koole. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing systems*, 30(3-4):323–339, 1998.

Andreas Liljegren, Johan Berlin, Stefan Szücs, and Staffan Höjer. The police and 'the balance'—managing the workload within swedish investigation units. *Journal of Professions and Organization*, 8(1):70–85, 2021.

Erica Jovanna Magaña, Dina Perrone, and Aili Malm. A process evaluation of San Francisco's law enforcement assisted diversion program. *Criminal Justice Policy Review*, 33(2):148–176, 2022.

Monica Malta, Thepikaa Varatharajan, Cayley Russell, Michelle Pang, Sarah Bonato, and Benedikt Fischer. Opioid-related treatment, interventions, and outcomes among incarcerated persons: A systematic review. *PLoS medicine*, 16(12):e1003002, 2019.

Avi Mandelbaum and Gennady Pats. State-dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits. *The Annals of Applied Probability*, 8(2):569–646, 1998.

Neal Master, Marty I Reiman, Can Wang, and Lawrence M Wein. A continuous-class queueing model with proportional hazards-based routing. *Available at SSRN 3390476*, 2018.

Dale E McNiel and Renee L Binder. Effectiveness of a mental health court in reducing criminal recidivism and violence. *American Journal of Psychiatry*, 164(9):1395–1403, 2007.

Todd D Minton and Daniela Golinelli. Jail inmates at midyear 2012–statistical tables. *Non-Criminal Justice*, 241264, 2013.

MIT News. Stopping the revolving prison door: Reducing recidivism. *MIT News*, 2017. URL `https://news.mit.edu/2017/stopping-revolving-prison-door-reducing-recidivism-mit-jpal-0510`.

Ali Movaghar. Optimal assignment of impatient customers to parallel queues with blocking. *Scientia Iranica*, 3(4), 1997.

Emre Nadar, Alp Akcay, Mustafa Akan, and Alan Scheller-Wolf. The benefits of state aggregation with extreme-point weighting for assemble-to-order systems. *Operations Research*, 66(4):1040–1057, 2018.

Jonathan Patrick, Martin L Puterman, and Maurice Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 56(6):1507–1525, 2008.

James MA Pitts, O Hayden Griffin III, and W Wesley Johnson. Contemporary prison overcrowding: Short-term fixes to a perpetual problem. *Contemporary Justice Review*, 17(1):124–139, 2014.

Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

Seth Jacob Prins, Laura Draper, Justice Center, and D John. *Improving outcomes for people with mental illnesses under community corrections supervision: A guide to research-informed policy and practice.* Justice Center, Council of State Governments New York, 2009.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Suzanne Russell. Grant awarded to help NJ parolees kick drug addiction. https://www.mycentraljersey.com/story/news/health/addiction/2022/04/26/3-2-million-grant-awarded-help-nj-parolees-kick-drug-addiction/7443322001/, April 26 2022. Accessed: April 4, 2023.

Saied Samiedaluie, Beste Kucukyazici, Vedat Verter, and Dan Zhang. Managing patient admissions in a neurology ward. *Operations Research*, 65(3):635–656, 2017.

Peter Schmidt and Ann D Witte. *Predicting recidivism using survival models*. Springer Science & Business Media, 2012.

Pengyi Shi, Jonathan E Helm, Jivan Deglise-Hawkinson, and Julian Pan. Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Operations Research*, 69(6):1842–1865, 2021.

Satinder P Singh and Richard S Sutton. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1-3):123–158, 1996.

Donald Specter. Everything revolves around overcrowding: the state of California's prisons. *Federal Sentencing Reporter*, 22(3):194–199, 2010.

Cassia Spohn and David Holleran. The effect of imprisonment on recidivism rates of felony offenders: A focus on drug offenders. *Criminology*, 40(2):329–358, 2002.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Gerald Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3):58–68, 1995.

Marlin W Ulmer. Dynamic pricing and routing for same-day delivery. *Transportation Science*, 54(4):1016–1033, 2020.

Mericcan Usta and Lawrence M Wein. Assessing risk-based policies for pretrial release and split sentencing in Los Angeles county jails. *PloS One*, 10(12):e0144967, 2015.

Vera Institute of Justice. What jails cost statewide in Indiana. URL `https://staging.vera.org/publications/what-jails-cost-statewide/indiana`.

John L Worrall, Pamela Schram, Eric Hays, and Matthew Newman. An analysis of the relationship between probation caseloads and property crime rates in California counties. *Journal of Criminal Justice*, 32 (3):231–241, 2004.

Mohammad Zhalechian, Esmaeil Keyvanshokooh, Cong Shi, and Mark P Van Oyen. Data-driven hospital admission control: A learning approach. *Operations Research*, 2023.

Electronic Companion

EC.1. Notation Table

Table EC.1 summarizes the main notations used in the paper.

Table EC.1 Notation Table

Parameters and Variables in System Description	
p_m	Probability of a customer from class m
N_m	Daily number of new arrivals from class m
$\mathcal{J}$	Set of stations: $\mathcal{J}=\{\texttt{jail},\texttt{CC}\}$
$d(j,m)$	Sentenced length for customers of class m in station j
$q_r(j,m,\tau)$	Probability that a class m customer served at station j will recidivate during the τ^{th} epoch after leaving the incarceration
$T_{j,m}$	Time window that recidivism may happen for class m in station j
$X_{j,m,l}$	State variable: number of class m customers who have been in station j for l epochs
S	System state: $S=\{X_{j,m,0},X_{j,m,1},...,X_{j,m,d(j,m)}\}_{j\in\mathcal{J};m=1,...M}$
$s_j(s)$	Total number of customers in station j given state $S=s$: $s_j(s)=\sum_{m=1}^{M}\sum_{l=0}^{d(j,m)}x_{j,m,l}$
$s_{j,m}(s)$	Total number of class m customers in station j given state $S=s$: $s_{j,m}(s)=\sum_{l=0}^{d(j,m)}x_{j,m,l}$
$e_{j,m,l}$	Unit occupancy of class m customers who have been in station j for l epochs
Parameters and Variables in Cost Functions	
$C(s)$	Pre-action, pre-transition one-period cost given state s
$C_j(s)$	Pre-action, pre-transition one-period cost attributed from station j given state s
h_j	Unit holding cost of station j
h_r	Unit recidivism cost
h_v	Unit violation cost
$c_h(s)$	One-period holding cost
$c_h^j(s_j)$	One-period holding cost from station j
$c_r(s)$	One-period recidivism cost
$c_r^j(s)$	One-period recidivism cost from station j
$c_v(s)$	One-period violation cost
$q_v(s_{\texttt{CC}},m)$	Cumulative probability of violations from a class m customers placed in CC
γ	Discounting factor: $\gamma\in(0,1)$
π	Policy $\pi:\mathcal{S}\to\mathcal{J}^M$ is a mapping from state S to action $\{A_m\}_{m=1,...,M}$ with $\pi(s)=\{\pi_m(s)\}_{m=1,...,M}$ and $\pi_m(s)\in\mathcal{J}$
Parameters and Variables in Transformed MDP	
λ_m	Arrival rate for customers of class m
$\theta_{j,m}$	Proportion of class m customers placed in station j
$\Theta(s)$	Placement policy of the batched decisions $\Theta(s)=\{\theta_{j,m}\}_{j\in\mathcal{J},m=1,...,M}$
$X_{j,m}$	State variable: number of class m customers in station j
S	System state: $S=\{X_{j,m}\}_{j=1,...,J;m=1,...,M}$
$\pi_\Theta(a\|s)$	Probability of making placement decisions $a=\{a_{j,m}\}_{j,m}$ at state s under policy Θ

EC.2. Proof of Theorem 1

We focus on proving (6), re-stated below:

$$V^*(s+2e_{\texttt{CC},l})-V^*(s+e_{\texttt{CC},l})\geq V^*(s+e_{\texttt{jail},l}+e_{\texttt{CC},l})-V^*(s+e_{\texttt{jail},l}).$$

Note that following (6), (7) can be proved by interchanging `jail` and `CC`, since our proof does not rely on the distinction between jail and CC.

EC.2.1. Proof Roadmap

Our proof framework, as discussed in the main text, consists of the following four steps. Figure EC.1 provides a schematic flow chart.

(1) Decompose the convex cost function into the summation of marginal contributions from individuals within the system as given in (8), re-stated below:

$$C(s) = \sum_{j\in\mathcal{J}} C_j(s_j) = \sum_{j\in\mathcal{J}} \sum_{k=1}^{s_j} (C_j(k) - C_j(k-1)),$$

where $C_j(\cdot)$ is the cost function for station j, s_j is the occupancy level of station j at state s, and $C_j(0) = 0$. Then the value function $V(s) = \min_\pi \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t C(s_t)|s_0 = s, \pi]$ can be decomposed into the discounted sum of marginal costs contributed by individuals within the system.

(2) Bound the value function difference in each side of (6) through coupling. This coupling is necessary because when analyzing each side of (6), it necessitates a comparison of the incurred costs under four "unknown" sets of decisions prescribed by the optimal placement policy for future customers given four different starting states. Through coupling, we reduce the number of sets of unknown decisions to compare from four to two.

(3) Decompose the value function difference, after coupling, for future customers into the marginal contribution from one additional customer using (8). While this marginal contribution is trivial under conventional linear cost setting, our setting of general convex cost functions greatly complicates the analysis. This is because the marginal cost varies with the occupancy level, which requires our further analysis on the occupancy difference in the next step.

(4) Bound the marginal cost via bounding the occupancy level deviations of the two systems (Lemma EC.1). This is achieved by comparing instances and states where the optimal policy prescribe different decisions in the two systems (Lemmas EC.2 and EC.3), which further depend on monotonicity results established in Proposition 1 and Corollary 1.

EC.2.2. Main Proof

Proof. For notational simplicity, we refer to jail as Station 1 and CC as Station 2 in the proof. To prove (6), we denote the left-hand side $LHS = V^*(s + 2e_{2,l}) - V^*(s + e_{2,l})$ and the right-hand side $RHS = V^*(s + e_{1,l} + e_{2,l}) - V^*(s + e_{1,l})$. We desire to show $LHS \geq RHS$.

Step (1) The cost decomposition is given in (8). In the single-class model, the total cost attributed from station j, denoted as C_j, is a convex function that only depends on the occupancy level of station j, i.e., s_j. Consequently, in Step (3), when the value function difference is derived to be

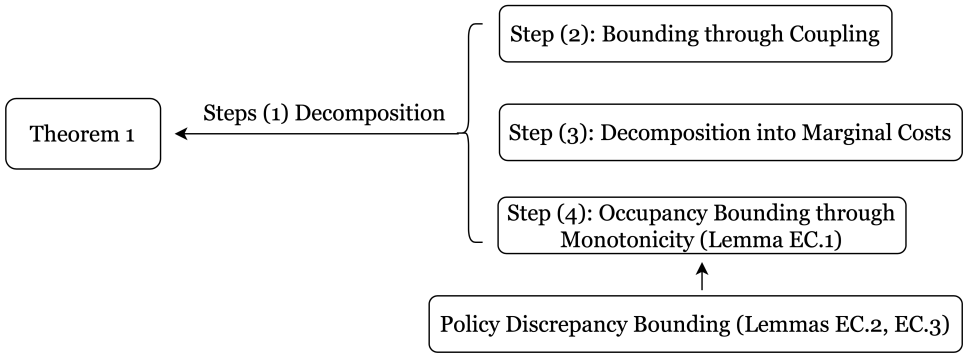


Figure EC.1 Roadmap for Proving Theorem 1.

the marginal cost from one customer, it suffices to analyze the system's occupancy level when the customer is in service.

Step (2) In this step, we bound both LHS and RHS by coupling. Denote the optimal policy after seeing the starting state $s_a = s + 2e_{2,l}$ by π_a. Then for LHS, we have

$$LHS = V^{\pi_a}(s + 2e_{2,l}) - V^*(s + e_{2,l}) \geq V^{\pi_a}(s + 2e_{2,l}) - V^{\pi_a}(s + e_{2,l}), \tag{EC.1}$$

where the inequality comes from the suboptimality of using policy π_a for the system starting from the different state $s + e_{2,l}$, i.e., $V^*(s + e_{2,l}) \leq V^{\pi_a}(s + e_{2,l})$. Note that when we apply policy π_a to the system starting from $s + e_{2,l}$, we couple Systems 1 and 2 to have identical new arrivals and sequentially apply the same decisions prescribed by π_a, resulting in the exactly same placement for new arrivals in the two systems.

Similarly, denote the optimal policy after seeing the starting state $s_b = s + e_{1,l}$ by π_b. Then for RHS, we have,

$$RHS = V^*(s + e_{1,l} + e_{2,l}) - V^{\pi_b}(s + e_{1,l}) \leq V^{\pi_b}(s + e_{1,l} + e_{2,l}) - V^{\pi_b}(s + e_{1,l}), \tag{EC.2}$$

where the inequality comes from the suboptimality of π_b for the system starting from $s + e_{1,l} + e_{2,l}$, i.e., $V^*(s + e_{1,l} + e_{2,l}) \leq V^{\pi_b}(s + e_{1,l} + e_{2,l})$. Therefore, it is sufficient to prove

$$V^{\pi_a}(s + 2e_{2,l}) - V^{\pi_a}(s + e_{2,l}) \geq V^{\pi_b}(s + e_{1,l} + e_{2,l}) - V^{\pi_b}(s + e_{1,l}). \tag{EC.3}$$

Step (3): Decomposition into Marginal Costs In this step, we bound $V^{\pi_a}(s + 2e_{2,l}) - V^{\pi_a}(s + e_{2,l})$ and $V^{\pi_b}(s + e_{1,l} + e_{2,l}) - V^{\pi_b}(s + e_{1,l})$ by decomposing the cost into the marginal contribution from one extra customer.

For LHS, recall that $V^{\pi_a}(s + 2e_{2,l})$ and $V^{\pi_a}(s + e_{2,l})$ are the total discounted costs from the systems starting at states $s + 2e_{2,l}$ and $s + e_{2,l}$ under π_a, respectively. We couple the two systems,

starting from these two states, by having identical arriving times of new arrivals and identical placement decisions prescribed by π_a for each of the new arrivals. Consequently, the departure times of future customers become identical, as do the occupancy levels from future customers in the two systems. Given that the total one-step cost from station j depends solely on the occupancy level of station j, we have the following results on total costs contributed from the jail and CC sides.

- Total costs contributed from the jail side are identical in these two systems. Because the starting state only differs in occupancy in CC (station 2), and the jail occupancy remains the same in both systems at any given time.
- For total costs contributed from the CC side, note that all future customers have identical arrival times and departure times. Therefore, the difference only comes from the one additional customer when starting from state $s+2e_{2,l}$ versus starting from $s+e_{2,l}$. In other words, we just need to focus on bounding the cost contributed by this additional customer at each time epoch $t=0,...,d_2-l$, until this customer leaves the system.

Let $s^a_{t,j}$ denote station j's occupancy level after t epochs when starting from state $s_a=s+2e_{2,l}$ under policy π_a. Using the decomposition given in (8), we have

$$V^{\pi_a}(s+2e_{2,l})-V^{\pi_a}(s+e_{2,l})=\sum_{t=0}^{d_2-l}\gamma^t\big(C_2(s^a_{t,2})-C_2(s^a_{t,2}-1)\big), \tag{EC.4}$$

where C_j contains all three costs from station j and depends on occupancy via a convex form.

Similarly, for RHS, using similar arguments, we have

$$V^{\pi_b}(s+e_{1,l}+e_{2,l})-V^{\pi_b}(s+e_{1,l})=\sum_{t=0}^{d_2-l}\gamma^t\big(C_2(s^b_{t,2}+1)-C_2(s^b_{t,2})\big), \tag{EC.5}$$

where $s^b_{t,j}$ denotes station j's occupancy level after t epochs when starting from $s_b=s+e_{1,l}$ under policy π_b.

Combining (EC.4) and (EC.5) with (EC.3), it is sufficient for us to prove

$$\sum_{t=0}^{d_2-l}\gamma^t\big(C_2(s^a_{t,2})-C_2(s^a_{t,2}-1)\big)\geq\sum_{t=0}^{d_2-l}\gamma^t\big(C_2(s^b_{t,2}+1)-C_2(s^b_{t,2})\big). \tag{EC.6}$$

While these marginal contributions are trivial under conventional linear cost functions, in our context of general convex cost functions (incorporating convex holding and violation costs), these marginal costs vary with the occupancy levels. Therefore, our remaining task is to bound the difference between the occupancy levels in the two systems after coupling: the first system uses policy π_a, the optimal policy, with occupancy $s^a_{t,2}$ at each t; the second system uses policy π_b, also the optimal policy, with occupancy $s^b_{t,2}$ at each t.

Step (4): Occupancy Bounding through Monotonicity In this final step, we establish (EC.6) by comparing the occupancy levels, $s^a_{t,2}$ and $s^b_{t,2}$ with the help of Lemma EC.1.

LEMMA EC.1. *(Bounding the Occupancy Deviation) Under Assumptions 1 and EC.1, for* $t = 0, 1, ..., d_2 - l$*, the CC and jail occupancy satisfy:*

$$\textit{jail: } 0 \le s^b_{t,1} - s^a_{t,1} \le 1. \tag{EC.7}$$

$$\textit{CC: } 1 \le s^a_{t,2} - s^b_{t,2} \le 2; \tag{EC.8}$$

Recall that $s^b_{t,1} - s^a_{t,1} = 1$ and $s^a_{0,2} - s^b_{0,2} = 2$ for the starting states. Lemma EC.1 implies that under the optimal policies π_a and π_b for each of the two systems, respectively, the disparity in the resulting occupancy at any given t would not exceed the disparity in the starting states. The proof of Lemma EC.1 is given in Section EC.2.3.

With Lemma EC.1, we have $s^a_{t,2} \ge s^b_{t,2} + 1$, which gives $C_2(s^a_{t,2}) - C_2(s^a_{t,2} - 1) \ge C_2(s^b_{t,2} + 1) - C_2(s^b_{t,2})$ under the convexity assumption of cost function $C_2(\cdot)$. Consequently, we have:

$$\sum_{t=0}^{d_2-l} \gamma^t (C_2(s^a_{t,2}) - C_2(s^a_{t,2} - 1)) \ge \sum_{t=0}^{d_2-l} \gamma^t (C_2(s^b_{t,2} + 1) - C_2(s^b_{t,2})),$$

i.e., the desired result in (EC.6). This completes the proof. □

EC.2.3. Proof of Lemma EC.1

Lemma EC.1 establishes the relationship between the occupancy levels $s^a_{t,2}$ and $s^b_{t,2}$. To compare these occupancy levels, we analyze the discrepancies in placement decisions between the policies π_a and π_b through Lemmas EC.2 and EC.3. The optimal policies π_a and π_b are the same when viewed as state-action mappings. We differentiate between π_a and π_b to highlight the different sets of decisions prescribed by the optimal policy in the two systems with different starting states.

Proof of Lemma EC.1. We first show in Lemma EC.2 that π_a will not place more customers in CC than π_b; and π_b will not place more customers in jail than π_a. Combining $s^a_{0,2} - s^b_{0,2} = 2$ and $s^b_{t,1} - s^a_{t,1} = 1$, we have $s^a_{t,2} - s^b_{t,2} \le 2$ and $s^b_{t,1} - s^a_{t,1} \le 1$. Then, we show in Lemma EC.3 that π_a will place at most one more customer in jail than π_b; and π_b will place at most one more customer in CC than π_a. Combining $s^a_{0,2} - s^b_{0,2} = 2$ and $s^b_{t,1} - s^a_{t,1} = 1$, we have $s^a_{t,2} - s^b_{t,2} \ge 1$ and $s^b_{t,1} - s^a_{t,1} \ge 0$. This completes the proof for Lemma EC.1. □

Before proving Lemmas EC.2 and EC.3, we first define the policy *discrepancy times*. For $k = 1, 2, ...,$ we denote

- t^+_k: the epoch index when $\pi_a = \texttt{jail}$ while $\pi_b = \texttt{CC}$ for the kth time;
- t^-_k: the epoch index when $\pi_a = \texttt{CC}$ while $\pi_b = \texttt{jail}$ for the kth time.

We prove $t^+_k \le t^-_k \le t^+_{k+1}$, $\forall k = 1, 2, ...$ in Lemmas EC.2 and EC.3. We give a proof outline first before showing the details.

To prove $t_k^+ \le t_k^-, \forall k = 1, 2, \ldots$ in Lemma EC.2, we take an inductive approach with contradiction. In both the base case and inductive step, we first describe the system state, which comprises three components: (1) s_τ: the customers in the initial state s; (2) $\mathcal{A}_\tau$: the customers placed when π_a and π_b make the same decisions; (3) $e_{j,\tau-t^-/t^+}$: the customers placed when π_a and π_b make different decisions. By utilizing the policy discrepancy times, we can express the system states under policies π_a and π_b at any given time with these three components. Subsequently, by leveraging the inductive hypothesis, we can compare the number of customers and their LOS prior to the decision-making epoch on a given policy discrepancy time. This allows us to construct a contradiction that violates the policy monotonicity given in Proposition 1 and Corollary 1. Lemma EC.3 can be proven using a similar approach, with the exception that the contradiction arises from a tie-breaking rule given in Assumption EC.1.

LEMMA EC.2. *(Bounding the Policy Discrepancy 1) Under Assumption 1, for $t = 0, 1, ..., d_2 - l$, the policy discrepancy times satisfy*

$$t_k^+ \le t_k^-, \forall k = 1, 2, \ldots.$$

Proof of Lemma EC.2. We prove by induction.

Base case. For $k = 1$, suppose the first policy discrepancy time occurs at epoch τ. Then the system states right before the decision-making point can be written as

$$\begin{aligned} s_\tau^a &= s_\tau + \mathcal{A}_\tau + 2e_{2,l+\tau}, \\ s_\tau^b &= s_\tau + \mathcal{A}_\tau + e_{1,l+\tau}. \end{aligned}$$

Then by the policy monotonicity on occupancy (Proposition 1), at time τ, $\pi_a = \texttt{CC}$ implies $\pi_b = \texttt{CC}$. Thus if there is policy discrepancy, it could only be $\pi_a = \texttt{jail}$ implies $\pi_b = \texttt{CC}$. In other words, it could only be $\tau = t_1^+$. So we have $t_1^+ \le t_1^-$.

Inductive hypothesis and induction. Suppose Lemma EC.2 holds for $k = 1, ..., n-1$. This inductive hypothesis states that $t_1^+ \le t_1^-$, ..., $t_{n-1}^+ \le t_{n-1}^-$. This means that right before the next policy discrepancy time, denoted as τ, both π_a and π_b places $n-1$ customers in CC during the policy discrepancy times. Then at τ, right before the placement decision is made, the system states are:

$$\begin{aligned} s_\tau^a = s_\tau + \mathcal{A}_\tau &+ e_{1,\tau-t_1^+} + \cdots + e_{1,\tau-t_{n-1}^+} + \\ &+ 2e_{2,l+\tau} + e_{2,\tau-t_1^-} + \cdots + e_{2,\tau-t_{n-1}^-}, \\ s_\tau^b = s_\tau + \mathcal{A}_\tau &+ e_{1,l+\tau} + e_{1,\tau-t_1^-} + \cdots + e_{1,\tau-t_{n-1}^-} \\ &+ e_{2,\tau-t_1^+} + \cdots + e_{1,\tau-t_{n-1}^+}. \end{aligned} \tag{EC.9}$$

Since $t_1^+ \le t_1^-$, ..., $t_{n-1}^+ \le t_{n-1}^-$, we have $\tau - t_1^+ \ge \tau - t_1^-, \cdots, \tau - t_{n-1}^+ \ge \tau - t_{n-1}^-$. That is, before the decision made at τ, s_t^a contains $n-1$ customers who arrived more earlier jail and $k+1$ customers who arrived more recently at CC; s_t^b contains n customers who arrived more recently at jail and $n-1$ customers who arrived earlier at CC. Thus by the policy monotonicity on occupancy and LOS (Proposition 1 and Corollary 1), at τ, $\pi_a = \texttt{CC}$ implies $\pi_b = \texttt{CC}$. In other words, it could only be $\tau = t_n^+$. So we have $t_n^+ \le t_n^-$.

Based on the principle of mathematical induction, $t_k^+ \le t_k^-$ holds true for all values of k. □

LEMMA EC.3. *(Bounding the Policy Discrepancy 2) Under Assumptions 1 and EC.1, for* $t = 0, 1, ..., d_2 - l$*, the policy discrepancy times satisfy*

$$t_k^- \le t_{k+1}^+, \forall k = 1, 2, \ldots.$$

The proof of Lemma EC.3 follows a similar induction proof used in proving Lemma EC.2 and is omitted here.

EC.3. Proof for Proposition 1 and Corollary 1

In this section, we prove policy monotonicity on occupancy (Proposition 1) and LOS (Corollary 1) using induction and coupling. For ease of exposition, we present the results using the single-class model, but the same logic applies to the multi-class model. Let $e_{j,l}$ be the unit customer who has been in station j for l epochs.

EC.3.1. Proof Roadmap

Before proving Proposition 1 and Corollary 1, we provide an outline of the proof structure. We first prove the desired monotonicity results in a finite-horizon MDP with H horizons, where the terminal cost $V_H(s)$ represents the value function from state s in a clearance system (with no new arrivals). After establishing the results within this finite-horizon MDP framework, we then apply Corollary 3.5 (v) in Feinberg and Lewis (2018), which guarantees convergence of the value function from finite-horizon to infinite-horizon MDP in the total discounted cost setting as $H \to \infty$. This allows us to extend the results to the infinite-horizon case.

In the rest of the proof, we focus on proving the results for the finite-horizon MDP, where we prove Proposition 1 and Corollary 1 together in each step through an inductive approach. The proof consists of the following steps:

(1) Base case. Establish the validity of Proposition 1 and Corollary 1 for decisions made at horizon $H-1$ based on the convexity of one-step costs.

(2) Inductive hypothesis and induction. Assuming Proposition 1 and Corollary 1 hold for decisions made at horizons $t+1, ..., H-1$, we proceed to prove their applicability at horizon t using the coupling method.

(2.1) Construct two systems for each state of s_0 and $s_0 + e_{\texttt{jail},l}$: one following the optimal policy; the other following a suboptimal policy that prescribes the same actions as the optimal policy except for the first incoming customer.

(2.2) For the two systems constructed in Step (2.1) starting from each state of s_0 and $s_0 + e_{\texttt{jail},l}$, we couple the sample paths to reduce the difference between their value functions to the marginal contribution from one single customer.

(2.3) Bound the marginal cost via bounding the occupancy level deviations of the two systems. This is achieved by comparing instances and states where the optimal policy prescribes different decisions in the two systems, which further depends on the inductive hypothesis.

EC.3.2. Main Proof of Proposition 1 and Corollary 1

Proof. Denote the optimal policy at state s in horizon t as $\pi_t(s)$. For notational simplicity, we call $j = 1, 2$ to denote $j = \texttt{jail}, \texttt{CC}$, respectively. In this section, we establish the results that in the $H-$horizon MDP, for all $l_2 < l_1 < d_1$, $t = 1, ..., H-1$,

(i) policy monotonicity on occupancy: $\pi_t(s) = \texttt{CC}$ *implies* $\pi_t(s + e_{\texttt{jail},l}) = \texttt{CC}$;

(ii) policy monotonicity on LOS: $\pi_t(s + e_{\texttt{jail},l_1}) = \texttt{CC}$ *implies* $\pi_t(s + e_{\texttt{jail},l_2}) = \texttt{CC}$.

Then the results $\pi_t(s) = \texttt{jail}$ *implies* $\pi_t(s + e_{\texttt{CC},l}) = \texttt{jail}$ (policy monotonicity on occupancy) and $\pi_t(s + e_{\texttt{CC},l_1}) = \texttt{jail}$ *implies* $\pi_t(s + e_{\texttt{CC},l_2}) = \texttt{jail}$ (policy monotonicity on LOS) can be proven using the same reasoning, given the symmetry between jail and CC.

Step (1) Decisions at Horizon $H-1$. In this step, we prove the policy monotonicity on occupancy and LOS for decisions made at Horizon $H-1$ by utilizing the convexity of terminal costs.

Policy monotonicity on occupancy. Result to show: "$\pi_{H-1}(s) = \texttt{CC}$ implies $\pi_{H-1}(s + e_{\texttt{jail},l}) = \texttt{CC}$." If $\pi_{H-1}(s) = \texttt{CC}$, then

$$c(s) + \gamma V_H(s' + e_{2,0}) < c(s) + \gamma V_H(s' + e_{1,0}),$$

where s' is the next-epoch state of s after customers' departure and LOS transition. Thus $V_H(s' + e_{2,0}) < V_H(s' + e_{1,0})$. From the definition of the terminal cost and the convexity of one-step function, we have

$$V_H(s' + e_{1,l} + e_{1,0}) - V_H(s' + e_{1,l} + e_{2,0}) > V_H(s' + e_{1,0}) - V_H(s' + e_{2,0}) > 0.$$

Thus we have

$$c(s) + \gamma V_H(s' + e_{1,l} + e_{2,0}) < c(s) + \gamma V_H(s' + e_{1,l} + e_{1,0}),$$

which leads to $\pi_{H-1}(s+e_{\texttt{jail},l})=\texttt{CC}$.

Policy monotonicity on LOS. Result to show: "for $l_2<l_1<d_1$, $\pi_{H-1}(s+e_{\texttt{jail},l_1})=\texttt{CC}$ implies $\pi_{H-1}(s+e_{\texttt{jail},l_2})=\texttt{CC}$," which can be shown through the same argument for policy monotonicity on occupancy.

Step (2) Inductive Step Assuming the policy monotonicity on occupancy and LOS holds for decisions made at Horizons $t+1,...,H-1$, we proceed to prove their applicability at Horizon t using the coupling method. We present the detailed proof of the policy monotonicity on occupancy at Horizon t. The policy monotonicity on LOS at Horizon t can be proven via the same approach. For notational simplicity, we omit the policy and value function's dependence on horizon index t in the remaining of this section.

We prove the policy monotonicity on occupancy at Horizon t by contradiction. Suppose there exists s_0, s.t., $\pi^*(s_0)=\texttt{CC}$ and $\pi^*(s_0+e_{\texttt{jail},l})=\texttt{jail}$.

Step (2.1) System Construction.

We first construct two systems, both starting from s_0. Consider the first incoming customer after the starting state s_0 as our target customer.

1. The first system follows the optimal policy, denoted as π_a, i.e., place the target customer in CC, i.e., $\pi_a(s_0)=\texttt{CC}$, and so forth.
2. The second system follows a suboptimal policy, denoted as π_1, specified as follows: (1) placing the target customer in jail, i.e., $\pi_1(s_0)=\texttt{jail}$. (2) placing all the other subsequent customers in the same station as prescribed in System 1.

Denote the value functions of Systems 1 and 2 as $V^{\pi_a}(s_0)$ and $V^{\pi_1}(s_0)$, respectively. By the policy suboptimality in System 2, we have $V^{\pi_a}(s_0)-V^{\pi_1}(s_0)\leq 0$.

We then construct two systems, both starting from $s_0+e_{\texttt{jail},l}$. Consider the first incoming customer after the starting state $s_0+e_{\texttt{jail},l}$ as our target customer.

3. The third system follows the optimal policy π_b, i.e., place the target customer in jail, i.e., $\pi_b(s_0+e_{\texttt{jail},l})=\texttt{jail}$, and so forth.
4. The fourth system follows a suboptimal policy, denoted as π_2, specified as follows: (1) placing the target customer in CC, i.e., $\pi_2(s_0+e_{\texttt{jail},l})=\texttt{CC}$. (2) placing all the other subsequent customers in the same station as prescribed in System 3.

Denote the value functions of Systems 3 and 4 as $V^{\pi_b}(s_0+e_{\texttt{jail},l})$ and $V^{\pi_2}(s_0+e_{\texttt{jail},l})$, respectively. By the policy suboptimality in System 4, we have $V^{\pi_b}(s_0+e_{\texttt{jail},l})-V^{\pi_2}(s_0+e_{\texttt{jail},l})\leq 0$.

Step (2.2) Sample Path Coupling and Value Difference Computation.

In this step, we compare value function differences $V^{\pi_a}(s_0)-V^{\pi_1}(s_0)$ and $V^{\pi_b}(s_0+e_{\texttt{jail},l})-V^{\pi_2}(s_0+e_{\texttt{jail},l})$ through coupling, and derive them to be the marginal contribution from the target customers.

We compare $V^{\pi_a}(s_0)$ and $V^{\pi_1}(s_0)$ by coupling Systems 1 and 2 with identical arrivals. In this way, according to the definition of policy π_1, the two systems will have identical subsequent customer placements except for the target customer. Thus, the difference between occupancy levels (and the total costs) under π_a and π_1 is only from different placement of the target customer. Let $s^a_{k,j}$ be the occupancy of station j at k^{th} epoch after s under optimal placement policy π_a. Then the value function difference equals

$$V^{\pi_a}(s_0) - V^{\pi_1}(s_0) = \sum_{k=0}^{d_2} \gamma^k (C_2(s^a_{k,2}) - C_2(s^a_{k,2} - 1)) - \sum_{k=0}^{d_1} \gamma^k (C_1(s^a_{k,1} + 1) - C_1(s^a_{k,1})), \tag{EC.10}$$

where $\sum_{k=0}^{d_2} \gamma^k (C_2(s^a_{k,2}) - C_2(s^a_{k,2} - 1))$ represents the marginal cost contributed by the target customer in System 1, where the CC occupancy changes from $s^a_{k,2} - 1$ to $s^a_{k,2}$ because of the target customer after k epochs. Similarly, $\sum_{k=0}^{d_1} \gamma^k (C_1(s^a_{k,1} + 1) - C_1(s^a_{k,1}))$ represents the marginal cost contributed by the target customer in System 2, where the jail occupancy changes from $s^a_{k,1}$ to $s^a_{k,1} + 1$ because of the target customer after k epochs.

We compare $V^{\pi_b}(s_0 + e_{\texttt{jail},1})$ and $V^{\pi_2}(s_0 + e_{\texttt{jail},1})$ by coupling Systems 3 and 4 with identical arrivals. Similarly, the difference between occupancy levels (and the total costs) under π_b and π_2 is only from different placement of the target customer. Let $s^b_{k,j}$ be the occupancy of station j at k^{th} epoch after $s + e_{\texttt{jail},l}$ under the optimal policy π_b. Then the value function difference equals

$$V^{\pi_2}(s_0 + e_{\texttt{jail},1}) - V^{\pi_b}(s_0 + e_{\texttt{jail},1}) = \sum_{k=0}^{d_2} \gamma^k (C_2(s^b_{k,2} + 1) - C_2(s^b_{k,2})) - \sum_{k=0}^{d_1} \gamma^k (C_1(s^b_{k,1}) - C_1(s^b_{k,1} - 1)), \tag{EC.11}$$

where $\sum_{k=0}^{d_2} \gamma^k (C_2(s^b_{k,2} + 1) - C_2(s^b_{k,2}))$ represents the marginal cost contributed by the target customer in System 4, where the CC occupancy changes from $s^b_{k,2}$ to $s^b_{k,2} + 1$ because of the target customer after k epochs. Similarly, $\sum_{k=0}^{d_1} \gamma^k (C_1(s^b_{k,1}) - C_1(s^b_{k,1} - 1))$ represents the marginal cost contributed by the target customer in System 3, where the jail occupancy changes from $s^b_{k,1} - 1$ to $s^b_{k,1}$ because of the target customer after k epochs.

Combining Equations (EC.10) and (EC.11), we have

$$\begin{aligned}
&(V^{\pi_2}(s_0 + e_{\texttt{jail},1}) - V^{\pi_b}(s_0 + e_{\texttt{jail},1})) - (V^{\pi_a}(s_0) - V^{\pi_1}(s_0)) \\
&= \sum_{k=0}^{d_2} \gamma^k \left((C_2(s^b_{k,2} + 1) - C_2(s^b_{k,2})) - (C_2(s^a_{k,2}) - C_2(s^a_{k,2} - 1)) \right) \\
&\quad - \sum_{k=0}^{d_1} \gamma^k \left((C_1(s^b_{k,1}) - C_1(s^b_{k,1} - 1)) - (C_1(s^a_{k,1} + 1) - C_1(s^a_{k,1})) \right).
\end{aligned} \tag{EC.12}$$

Therefore, our remaining task is to analyze the difference between the occupancy in the two systems after coupling: the first system uses policy π_a, the optimal policy, starting at state s_0; the second system uses policy π_b, also the optimal policy, starting at state $s_0 + e_{\texttt{jail},l}$.

Step (2.3) Occupancy Bounding.

In this step, we establish $(V^{\pi_2}(s_0 + e_{\texttt{jail},1}) - V^{\pi_b}(s_0 + e_{\texttt{jail},1})) - (V^{\pi_a}(s_0) - V^{\pi_1}(s_0)) \leq 0$ by comparing the occupancy levels, $s^a_{k,j}$ and $s^b_{k,j}$ via Lemma EC.4.

LEMMA EC.4. *(Bounding the Occupancy Deviation) Under Assumption 1, for* $k = 0, 1, ..., \max\{d_1, d_2\}$*, the CC and jail occupancy satisfy:*

$$\textit{jail: } s^b_{k,1} - s^a_{k,1} \geq 1 \tag{EC.13}$$

$$\textit{CC: } s^a_{k,2} - s^b_{k,2} \geq 1 \tag{EC.14}$$

By Lemma EC.4, we have $(V^{\pi_2}(s_0+e_{\texttt{jail},1})-V^{\pi_b}(s_0+e_{\texttt{jail},1}))-(V^{\pi_a}(s_0)-V^{\pi_1}(s_0)) \leq 0$, based on the convexity of one-step cost functions. This contradicts with $V^{\pi_a}(s_0)-V^{\pi_1}(s_0) \leq 0$ and $V^{\pi_2}(s_0 + e_{\texttt{jail},1}) - V^{\pi_b}(s_0 + e_{\texttt{jail},1}) \geq 0$, and therefore the original assumption that $\pi^*(s_0) = \texttt{CC}$ and $\pi^*(s_0 + e_{\texttt{jail},l}) = \texttt{jail}$ must be false. Thus the statement $\pi_t(s_0) = \textit{jail}$ *implies* $\pi_t(s_0 + e_{\textit{CC},l}) = \textit{jail}$ holds true for all s_0 at horizon t.

Based on the principle of mathematical induction, $\pi_t(s_0) = \textit{jail}$ *implies* $\pi_t(s_0 + e_{\textit{CC},l}) = \textit{jail}$ for all s_0 at all horizons $t = 1, ..., H-1$. □

Proof of Lemma EC.4 We prove Lemma EC.4 by analyzing the discrepancies in placement decisions between policies π_a and π_b. At the starting states ($k = 0$), we have $s^b_{0,1} - s^a_{0,1} = 1$ and $s^a_{0,2} - s^b_{0,2} = 0$. After the placement of the target customer at time τ in both systems, we have $s^b_{\tau,1} - s^a_{\tau,1} = 2$ and $s^a_{\tau,2} - s^b_{\tau,2} = 1$. Therefore, Lemma EC.4 is equivalent to the condition that after the placement of the target customer, π_a will place at most one more customers in jail than π_b and π_b will not place more customers in CC than π_a. This can be proved similarly to the proofs of Lemma EC.2 and EC.3. □

EC.4. Recidivism Cost Formulation and Proof of Proposition 2

For notational consistency, we use the adjusted recidivism risk in the MDP formulation (Section 4.2), where the cumulative recidivism risk within the recidivism window is distributed as a constant risk during each epoch of service. In this section, we first introduce the formulation of the adjusted recidivism probability $\tilde{q}_r(j, m)$ with respect to the original recidivism risk $q_r(j, m, \tau)$ and the recidivism window T. Then in Section EC.4.2, we provide the proof for Proposition 2 via coupling and contradiction, which establishes the impact of recidivism window on the optimal placement policy.

EC.4.1. Adjusted Recidivism Risk Formulation

We use the *time window* $T_{j,m}$, during which a recidivism may happen to highlight the difference between jail and CC in preventing short-term recidivism. Given a recidivism window T, this time window is calculated as $T_{j,m} = T - d(j,m)\mathbb{1}_{j=\texttt{jail}}$. [2] Denote the total recidivism cost incurred by each type m customer who is placed in station j by $\Psi_r(j,m)$. Then

$$\Psi_r(\texttt{jail},m) = h_r \sum_{\tau=0}^{T_{\texttt{jail},m}} \gamma^{\tau+d(\texttt{jail},m)} q_r(j,m,\tau), \tag{EC.15}$$

$$\Psi_r(\texttt{CC},m) = h_r \sum_{\tau=0}^{T_{\texttt{CC},m}} \gamma^{\tau} q_r(j,m,\tau), \tag{EC.16}$$

where h_r is the unit recidivism cost. Since individuals could recidivate only after being released from incarceration, for individuals placed in jail, the recidivism cost is incurred after $d(\texttt{jail},m)$ epochs after the placement decision (Equation (EC.15)). However, individuals placed in CC may recidivate both during the service and after release (Equation (EC.16)). For the sake of formulation uniformity, we define an adjusted recidivism risk for the class m individual during each in-service epoch if placed in station j as $\tilde{q}_r(j,m)$. To ensure the total cost incurred by this individual remains the same, we have

$$h_r \sum_{\ell=0}^{d(j,m)} \gamma^{\ell} \tilde{q}_r(j,m) = \Psi_r(j,m),$$

which leads to

$$\tilde{q}_r(j,m) = \frac{1}{h_r} \frac{1-\gamma}{1-\gamma^{d(j,m)+1}} \Psi_r(j,m).$$

EC.4.2. Proof for Proposition 2

Proof. We prove by coupling and contradiction. Suppose there exists state s_0 and class m, such that optimal decision for the target customer under T_1 is $a^1_m(s_0) = \texttt{CC}$, while optimal decision for the target customer under T_2 (where $T_2 > T_1$) is $a^2_m(s_0) = \texttt{jail}$. Denote the adjusted recidivism probability under T_1 and T_2 as $\tilde{q}^1_r(j,m)$ and $\tilde{q}^2_r(j,m)$, respectively.

Step (1). System Construction. We construct two systems, both starting from s_0 and having recidivism window T_1:

1. The first system follows the optimal policy, denoted as π^*_1, i.e., placing the target customer in CC, i.e., $\pi^*_{1,m}(s_0) = \texttt{CC}$, and so forth.
2. The second system follows a suboptimal policy, denoted as π_1, specified as follows: (1) placing the target customer in jail, i.e., $\pi_{1,m}(s_0) = \texttt{jail}$. (2) placing all the other subsequent customers according to the system's optimal policy.

[2] For example, if we consider three-year recidivism from the placement decision, i.e., $T = 3 \times 365$, the window for those placed in jail is $T_{\texttt{jail},m} = 3 \times 365 - d(\text{jail},m)$; for those placed in CC, it is $T_{\texttt{CC},m} = 3 \times 365$. If the sentence lengths in both the jail and CC are one year, the recidivism window is 2 years in jail (excluding the sentenced length) but is 3 years in CC.

Denote the value functions of Systems 1 and 2 as $V_1^*(s_0)$ and $V_1(s_0)$, respectively. By the policy suboptimality in System 2, we have $V_1^*(s_0) - V_1(s_0) \leq 0$.

Similarly, we construct another two systems, both starting from s_0 and having recidivism window T_1 except for the target customer, whose recidivism window is changed to T_2:

3. The third system follows the optimal policy, denoted as π_2^*, i.e., placing the target customer in jail with $\pi_{2,m}^*(s_0) = \texttt{jail}$, and so forth.
4. The fourth system follows a suboptimal policy, denoted as π_2, specified as follows: (1) placing the target customer in CC, i.e., $\pi_{2,m}(s_0) = \texttt{CC}$. (2) placing all the other subsequent customers according to the system's optimal policy.

Denote the value functions of Systems 3 and 4 as $V_2^*(s_0)$ and $V_2(s_0)$, respectively. By the policy suboptimality in System 4, we have $V_2^*(s_0) - V_2(s_0) \leq 0$.

Step (2). Sample Path Coupling. We couple Systems 1 and 4 to have identical arrivals. Then they both place the target customer in CC. After this placement, Systems 1 and 4 start from the identical occupancy levels of each station, differing only in the recidivism window of the target customer. However, this difference does not impact the state variables of the systems. Consequently, the optimal decisions for subsequent customers after the target customer remain identical, as they consistently encounter the same system state. Therefore, the only difference between Systems 1 and 4 is the recidivism window of the target customer. In other words, the difference between $V_1^*(s_0)$ and $V_2(s_0)$ is only from the target customer who brings different costs due to different recidivism windows. For the target customer, the holding and violation costs cancel out because the subsequent jail's and CC's occupancy are the same for Systems 1 and 4. Therefore, the cost difference equals

$$\begin{aligned} V_2(s_0) - V_1^*(s_0) &= h_r \sum_{\tau=0}^{d(\texttt{CC},m)} \gamma^\tau \tilde{q}_r^2(\texttt{CC}, m) - h_r \sum_{\tau=0}^{d(\texttt{CC},m)} \gamma^\tau \tilde{q}_r^1(\texttt{CC}, m) \\ &= h_r \sum_{\tau=0}^{T_2} \gamma^\tau q_r(\texttt{CC}, m, \tau) - h_r \sum_{\tau=0}^{T_1} \gamma^\tau q_r(\texttt{CC}, m, \tau), \end{aligned} \tag{EC.17}$$

where $\tilde{q}_r^1(j, m)$ and $q_r(\texttt{CC}, m, \tau)$ are the adjusted recidivism probability and the original recidivism risk, respectively; (see Sections 4.2 and EC.4.1).

We couple Systems 2 and 3 to have identical arrivals. Then they both place the target customer in jail. After this placement, Systems 2 and 3 start from the identical occupancy levels of each station, differing only in the recidivism window of the target customer. However, this difference does not impact the state variables of the systems. Following the same argument for the coupling of

Systems 1 and 4, the only difference between Systems 2 and 3 comes from the recidivism windows of the target customer. Therefore, the cost difference equals

$$\begin{aligned} V_2^*(s_0) - V_1(s_0) &= h_r \sum_{\tau=0}^{d(\texttt{jail},m)} \gamma^\tau \tilde{q}_r^2(\texttt{jail},m) - h_r \sum_{\tau=0}^{d(\texttt{jail},m)} \gamma^\tau \tilde{q}_r^1(\texttt{jail},m) \\ &= h_r \sum_{\tau=0}^{T_2-d(\texttt{jail},m)} \gamma^\tau q_r(\texttt{jail},m,\tau) - h_r \sum_{\tau=0}^{T_1-d(\texttt{jail},m)} \gamma^\tau q_r(\texttt{jail},m,\tau). \end{aligned} \tag{EC.18}$$

Step (3). Contradiction by Value Comparison. From Equations (EC.17) and (EC.18), we have

$$\begin{aligned} &(V_2^*(s_0) - V_1(s_0)) - (V_2(s_0) - V_1^*(s_0)) \\ =&h_r \left(\sum_{\tau=0}^{T_2-d(\texttt{jail},m)} \gamma^\tau q_r(\texttt{jail},m,\tau) - \sum_{\tau=0}^{T_1-d(\texttt{jail},m)} \gamma^\tau q_r(\texttt{jail},m,\tau) \right) - h_r \left(\sum_{\tau=0}^{T_2} \gamma^\tau q_r(\texttt{CC},m,\tau) - \sum_{\tau=0}^{T_1} \gamma^\tau q_r(\texttt{CC},m,\tau) \right) \\ =&h_r \left(\sum_{\tau=T_1-d_{\texttt{jail}}+1}^{T_2-d_{\texttt{jail}}} \gamma^{\tau+d_{\texttt{jail}}} q_r(\texttt{jail},m,\tau) - \sum_{\tau=T_1+1}^{T_2} \gamma^\tau q_r(\texttt{CC},m,\tau) \right) \\ =&h_r \sum_{\tau=T_1+1}^{T_2} \gamma^\tau \big(q_r(\texttt{jail},m,\tau-d_{\texttt{jail}}) - q_r(\texttt{CC},m,\tau) \big) \\ >&0. \end{aligned}$$

The inequality comes from $q_r(\texttt{jail},m,\tau-d_{\texttt{jail}}) \geq q_r(\texttt{jail},m,\tau) \geq q_r(\texttt{CC},m,\tau)$, according to Assumption 2. Thus, $V_2^*(s_0) - V_2(s_0) > V_1(s_0) - V_1^*(s_0)$, which contradicts $V_1(s_0) - V_1^*(s_0) \geq 0$ and $V_2^*(s_0) - V_2(s_0) \leq 0$ (policy suboptimality). Therefore the original assumption that $a_m^1(s_0) = \texttt{CC}$ under T_1, but the optimal action $a_m^2(s_0) = \texttt{jail}$ under T_2 must be false. □

EC.5. Other Proofs

In this section, we present supplemental proofs and technical details. We first describe and verify Assumption EC.1, which is used in the proof of Lemma EC.3. We then provide a detailed formulation of the expected cost reduction achieved by placing new customers in CC instead of jail, as introduced in Proposition EC.5.2 in the main text. Finally, we conclude the section with the proof of the increasingness of the value function.

EC.5.1. Assumption on State Ties

In this section, we introduce Assumption EC.1, which is used in the proof of Lemma EC.1. It is a compliment to the policy monotonicity on occupancy (Proposition 1) and LOS (Corollary 1).

Definition EC.1. (State Tie) We say that there is a tie of states when there are two states s_1 and s_2 that have the same set of existing customers with exactly the same placement and arrival time, and the remaining existing customers satisfy the following conditions:

1. In state s_1, there are k customers who arrived more recently at station 1 $(e_{1,l_{1,1}}, \ldots, e_{1,l_{1,k}})$ and $k+1$ customers who arrived earlier at station 2 $(e_{2,l_{2,1}}, \ldots, e_{2,l_{2,k+1}})$ with $k > 1$.
2. In state s_2, there are k customers who arrived earlier at station 1 $(e_{1,l_{2,1}}, \ldots, e_{1,l_{2,k}})$ and k customers who arrived more recently at station 2 $(e_{2,l_{1,1}}, \ldots, e_{2,l_{1,k}})$.
3. Index ordering for each state: $l_{1,1} \le l_{1,2} \le \cdots \le l_{1,k}$ and $l_{2,1} \le l_{2,2} \le \cdots \le l_{2,k} \le l_{2,k+1}$.
4. Index ordering between two states: $l_{1,1} \le l_{2,1}$, $l_{1,2} \le l_{2,2}$, ..., $l_{1,k} \le l_{2,k}$.

Station 1 and 2 refer to either jail and CC or CC and jail, respectively.

Assumption EC.1. *(Tie Breaking Rule) In case of a state tie defined in Definition EC.1, the optimal policies for s_1 and s_2 satisfy: if $\pi^*(s_1) = 1$, then $\pi^*(s_2) = 1$.*

A state tie arises when two states share the same set of customer states and have clear LOS rankings for the remaining distinct customers. However, encountering a state tie renders policy comparison ambiguous. To address this issue, we can eliminate the customer with the longest LOS from station 2, denoted as $e_{2,l_{2,k+1}}$ in s_1. Subsequently, relying on the policy monotonicity on LOS (Corollary 1), if the optimal decisions differ, they can only involve placing the new customer at station 2 for s_1 and placing the new customer at station 1 for s_2. Nevertheless, due to the addition of $e_{2,l_{2,k+1}}$, the relationship between the optimal policies of the two states cannot be determined solely based on the policy monotonicity. Hence, in situations where an additional customer with the longest LOS is present, we introduce a tie-breaking rule in Assumption EC.1 to supplement the policy monotonicity. Assumption EC.1 outlines the decision rule employed to resolve the state tie. Even when a state tie occurs and s_1 includes a customer with a longer LOS than any other customer in s_1 and s_2, the policy monotonicity on LOS still holds.

Note that tie-breaking rule can be demonstrated to hold in the finite-horizon MDP, employing a similar inductive approach to that used in Section EC.3 for proving the policy monotonicity on occupancy. Suppose the terminal cost $V_H(s)$ is the value function of a system starting at s in an clearance system (with no new arrivals). To validate Assumption EC.1 at Horizon $H-1$, the one-step cost function $C(s)$ needs to satisfy: the difference in total cost resulting from placing a new customer in station 1 compared to station 2 is greater at s_2 than at s_1 in a clearance system. In the inductive step, to verify that the tie-breaking rule holds at Horizon t, a similar argument as in Section EC.3 can be applied.

EC.5.2. Proof for Corollary 2

In this section we prove the Corollary 2 where we establish the relationship between risk type prioritization and expected cost reduction by placing the customer in CC rather than jail.

Proof. We define $C_\Delta(m, s)$ as the expected cost reduction resulting from this placement choice, given by:

$$C_\Delta(m, s) = V(s + e_{\texttt{jail},m,0}) - V(s + e_{\texttt{cc},m,0}).$$

Let π^* denote the optimal policy. According to the Bellman equation, if $V(s+e_{\texttt{jail},m,0}) \geq V(s+e_{\texttt{cc},m,0})$, then $\pi^*_m(s) = \texttt{cc}$. Consequently, for any given state s_0 and two customer classes m_1 and m_2, if $C_\Delta(m_1,s_0) \geq C_\Delta(m_2,s_0)$, it follows that $C_\Delta(m_2,s_0) \geq 0$ implies $C_\Delta(m_1,s_0) \geq 0$. In other words, if customers of class m_2 are to be placed in CC, the same applies to customers of class m_1. □

Formulation of $C_\Delta(m,s)$. Let $s_{t,j}$ denote the occupancy of station j after t epochs. We denote the expected reduction of holding, recidivism, and violation costs by placing class m customers in CC instead of jail at state s as $C^h_\Delta(m,s)$, $C^v_\Delta(m,s)$, and $C^r_\Delta(m,s)$, respectively. Leveraging the decomposition scheme in (8), $C_\Delta(m,s)$ can be expressed by:

$$
\begin{aligned}
&C_\Delta(m,s) = C^h_\Delta(m,s) + C^v_\Delta(m,s) + C^r_\Delta(m,s).\\
&C^h_\Delta(m,s) = \sum_{t=0}^{d(\texttt{jail},m)} \gamma^t \mathbb{E}(c_h^{\texttt{jail}}(s_{t,\texttt{jail}}+1) - c_h^{\texttt{jail}}(s_{t,\texttt{jail}})) - \sum_{t=0}^{d(\texttt{cc},m)} \gamma^t \mathbb{E}(c_h^{\texttt{cc}}(s_{t,\texttt{cc}}+1) - c_h^{\texttt{cc}}(s_{t,\texttt{cc}})).\\
&C^v_\Delta(m,s) = -h_v \sum_{t=0}^{d(\texttt{cc},m)} \gamma^t \mathbb{E}(c_v^{\texttt{cc}}(s_{t,\texttt{cc}}+1) - c_v^{\texttt{cc}}(s_{t,\texttt{cc}})).\\
&C^r_\Delta(m,s) = h_r\Big(\sum_{t=0}^{T_{\texttt{jail},m}} \gamma^t q_r(\texttt{jail},m,t) - \sum_{t=0}^{T_{\texttt{cc},m}} \gamma^t q_r(\texttt{cc},m,t)\Big).
\end{aligned}
\tag{EC.19}
$$

Here the expectations are taken over future occupancy $s_{t,j}$, which depends on the placement policy.

EC.5.3. Value Function Increasingness

In this section, we formally state and prove the increasingness of value functions on state variables.

Lemma EC.5. *Under Assumption 1, the optimal value function V^* is increasing with state variables, i.e., for all $s \in \mathcal{S}, j \in \mathcal{J}, m = 1,\dots,M, l = 0,1,\dots,d(m,j)$, $V^*(s+e_{j,m,l}) \geq V^*(s)$.*

Proof. Denote the optimal policy following $s+e_{j,m,l}$ and s by π_1 and π_2, respectively. Then we have

$$
\begin{aligned}
V^*(s+e_{j,m,l}) &= \mathbb{E}_{\pi_1}[\sum_{t=0}^{\infty} \gamma^t C(s_t) | s_0 = s+e_{j,m,l}]\\
&\geq \mathbb{E}_{\pi_1}[\sum_{t=0}^{\infty} \gamma^t C(s_t) | s_0 = s]\\
&\geq \mathbb{E}_{\pi_2}[\sum_{t=0}^{\infty} \gamma^t C(s_t) | s_0 = s]\\
&= V^*(s),
\end{aligned}
$$

where the first inequality is due to the increasingness of one-step function $C(\cdot)$, and the second inequality is due to the suboptimality of policy π_1 for state s, which leads to $V^{\pi_1}(s) \geq V^{\pi_2}(s) = V^*(s)$. □

EC.6. Additional Details of the Modeling Framework, Transformed Model and Algorithmic Solution

EC.6.1. Exogenous Recidivism Arrival Rate

To maintain the Markov property, we assume that the stream of recidivism arrivals is independent of the policy and treat them as part of the exogenous arrivals (Remark 1). In this section, we present the formulation of the rate of recidivism arrival in steady state. We denote the (cumulative) probability of recidivism within the time window as

$$r(j,m) = \sum_{\tau=0}^{T_{j,m}} q_r(j,m,\tau).$$

Suppose the system enters the steady state. Let $x(j,m) = \mathbb{E}[\sum_{l=0}^{d(j,m)} X_\infty(j,m,l)]$ be the throughput of class m customers in station j, i.e., the expected number of class m customers entering/leaving station j in the steady state. Let λ'_m denote the expected total number of class m customers recidivating within window $T_{j,m}$. Then

$$\lambda'_m = \sum_{m=1}^{M} \sum_{j \in \mathcal{J}} x(j,m) r(j,m) = \sum_{m=1}^{M} \sum_{j \in \mathcal{J}} x(j,m) \sum_{\tau=0}^{T_{j,m}} q_r(j,m,\tau).$$

EC.6.2. Formulation of the Transformed MDP

In this section, we present the formulation of the transformed MDP based on the timescale separation described in Section 6.1. With a slight abuse of notation, we use the same notation for the state (S) and one-step cost function ($C(s)$) as in the original MDP model.

State. The system state is captured by $S = \{X_{j,m}\}_{j=1,\dots,J;m=1,\dots,M}$, where $X_{j,m}$ denotes the total number of class m customers in station j at current decision epoch. Thus, $X_{j,m} = \sum_{l=0}^{d(j,m)} X_{j,m,l}$, where $X_{j,m,l}$ is the state variable of the original MDP.

Action. All the customers arriving within the same week will be placed following the same placement policy, which depends on the system state in the beginning of the epoch. The placement policy is specified by: $\Theta(s) = \{\theta_{j,m}\}_{j \in \mathcal{J}, m=1,\dots,M}$, where $\theta_{j,m} \in [0,1]$ represents the proportion of customers of class m that are to be placed in station j in current epoch.

Cost and objective. We specify the transition dynamics in the next section after introducing the LOS recovery mechanism. Similar with the original MDP, for a given state $s = \{x_{j,m}\}$, the pre-action cost per epoch is composed of holding cost $c_h(s)$, recidivism cost $c_r(s)$, and violation cost $c_v(s)$: $C(s) = c_h(s) + c_r(s) + c_v(s)$. The objective is still minimizing the infinite-horizon discounted cost, with the Bellman equation similar to Equation (5).

EC.6.3. Transition Dynamics in Transformed MDP with Truncated State Space

To keep the state space finite, the arrival process is truncated for each epoch. Let H be the cutoff of the Poisson arrival in each day. For station j, the number of new customers follows a Poisson distribution with rate $\lambda_{j,m} = \lambda_m \theta_{j,m}$. By using the approximation scheme described in Section 6.1.2, we can model the number of participants released at the next slow-timescale epoch as a Hypergeometric distribution Hypergeo$(x_{j,m}, H\Delta, Hd(j,m))$. Consequently, the transition matrix is the convolution of Poisson distributions for new arrivals and hypergeometric distributions for remaining participants:

$$\begin{aligned}
&Pr\left\{X'_{j,m} = x'_{j,m}, m = 1, ..., M, j = 1, ..., J | s = \{x_{j,m}\}_{j,m}\right\} \\
=&\prod_{j=1}^{J}\prod_{m=1}^{M}\sum_{a=0}^{x'_{j,m}} Pr\left\{B^r_{j,m} = k, B^a_{j,m} = x'_{j,m} - k\right\} \\
=&\prod_{j=1}^{J}\prod_{m=1}^{M}\sum_{k=0}^{x'_{j,m}} \left(\frac{\binom{H\Delta}{k}\binom{x_{j,m}-H\Delta}{x_{j,m}-k}}{\binom{Hd(j,m)}{x_{j,m}}} \cdot \frac{(\lambda_{j,m})^{\left(x'_{j,m}-(x_{j,m}-k)\right)} e^{-\lambda_{j,m}}}{\left(x'_{j,m}-(x_{j,m}-k)\right)!} \right).
\end{aligned}$$

Here $B^a_{j,m} \sim Pois(\lambda_{j,m})$ is the number of new customers of class m in station j; and $B^r_{j,m} \sim$ Hypergeo$(X_{j,m}, H\Delta, Hd(j,m))$ is the number of released customers of class m in station j at the next slow-timescale epoch.

For the illustrating example in the second paragraph of Section 6.1.2, suppose the arrival process is truncated at five for each epoch. Then out of the maximum possible 15 arrivals over the past three epochs, six customers actually arrived in the system. Thus, in the fourth epoch, there can be at most five customers to be released who arrived within the first week. Consequently, the number of customers to be released follows a hypergeometric distribution with parameters $K = 6$ (the number of successes in the population), $n = 5$ (the sample size), and $N = 15$ (the population size).

EC.6.4. Derivation of Policy Gradient

To derive the policy gradient in (11), we first write out the randomized policy $\pi_{\theta_{j,m}}(a|s)$ according to the Poison thinning property. Denote realized placement action as $a = \{a_{j,m}\}_{j,m}$. Then the distribution of placement action is

$$\begin{aligned}
\pi_{\theta_{j,m}}(a|s) &= \prod_{m=1}^{M}\prod_{j\in\mathcal{J}} Pr\{A_{j,m} = a_{j,m}|\theta\} \\
&= \prod_{m=1}^{M}\prod_{j\in\mathcal{J}} \frac{e^{-\lambda_m\theta_{j,m}}(-\lambda_m\theta_{j,m})^{a_{j,m}}}{a_{j,m}!},
\end{aligned} \tag{EC.20}$$

where the second equality is because the number of class m customers placed in station j, $A_{j,m}$, follows a Poisson distribution with rate $\lambda_m\theta_{j,m}$.

For any class $m=1,...,M$, the placement probabilities $\theta_{j,m}$ satisfy: $\sum_{j\in\mathcal{J}}\theta_{j,m}=1$. Thus,

$$\theta_{j_{|\mathcal{J}|},m}=1-\sum_{k=1}^{|\mathcal{J}|-1}\theta_{j_k,m},\ m=1,...,M. \tag{EC.21}$$

Substituting (EC.21) into (EC.20) and taking a partial derivative on $\theta_{j,m}$, we have

$$\begin{aligned}\nabla_{\theta_{j,m}}\ln\pi_{\theta_{j,m}}(a|s)&=\nabla\theta_{j,m}\big(a_{j,m}\ln\theta_{j,m}+a_{j_{|\mathcal{J}|},m}\ln(1-\sum_{k=1}^{|\mathcal{J}|-1}\theta_{j_k,m})\big)\\&=\frac{a_{j,m}}{\theta_{j,m}}-\frac{a_{j_{|\mathcal{J}|},m}}{1-\sum_{k=1}^{|\mathcal{J}|-1}\theta_{j_k,m}}.\end{aligned}$$

EC.6.5. Adaptations for real-sized case studies

In this section, we detail adaptations of Algorithm 1 for real-sized case studies: the heuristic algorithm addressing a large number of customer classes and the regression technique for rarely visited states.

To handle the increased complexity caused by the detailed customer classification and separated CC programs, we adapt Algorithm 1 developed for the analytical setting and obtain placement decisions using the one-step optimization of Equation (13). We begin by solving a single-class-multiple-program model to obtain the value function $V(\cdot)$. We then make the placement decision on-the-go, considering two cost components: an immediate individual-level cost (which depends on risk level, medical need, and class), and the value function output from the algorithm (which depends on the total occupancy of each station). Specifically, the optimal placement decision for the kth new customer from class m is determined by:

$$a_{m,k}=\arg\min_{a_{m,k}\in\mathcal{J}}\big(C(s)+\gamma\cdot\frac{1}{N}\sum_{s'\in\mathcal{S}}V(s')\big).$$

The equation has a same formula as (13). However, here s denotes the immediate state after the action $a_{m,k}$, reflecting the class-specific costs associated with new arrivals' placement decisions; s' is the system state in the next epoch where $s'=\{s_j\}_j$ with s_j representing the total occupancy in station j. $V(s')$ is the value function estimated by Algorithm 1 for the one-class-multiple-program model. By switching to the post-decision state s, we are able to leverage $C(s)$ to accommodate the heterogeneity brought by the increased customer classes. Meanwhile, keeping a single-class version of value function $V(s')$ help maintain the tractability of the MDP model.

To estimate the value function of states that are infrequently visited or on the truncation boundary, we start by identifying states that are visited frequently enough, such as those visited over 500 times. Next, we apply a LASSO regression to fit the value functions using transformed state variables as features. Specifically, we consider the occupancy of each station-class combination as a feature. After fitting with LASSO, we replace the value functions of lesser-visited states or those on the truncation boundary with the value functions predicted by the regression.

EC.7. Supplement of Experimental Settings and Results

EC.7.1. Specification of Approximation Accuracy Evaluation in Section 6.1.2

To compare the optimal value functions between the original and transformed MDP, we need to first solve both models. In order to make the MDP solvable to the optimality, we estimate the approximation accuracy on the small and medium models. In both models, the length of slow timescale is two days, i.e., $\Delta = 2$. The parameters of the two models are presented in Table EC.2.

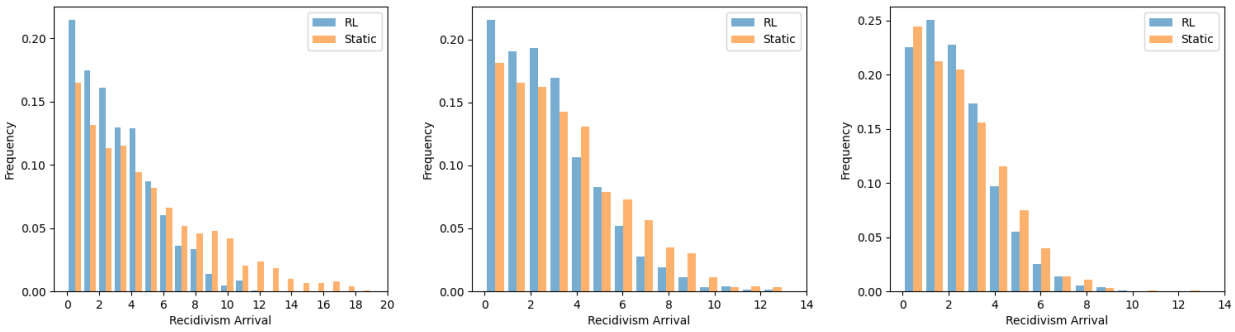
Table EC.2 Model parameters for approximation accuracy estimation.

	Small Model	Medium Model
Risk class M	2	2
Arrival rate λ_m	(0.5, 0.8)	(3, 5)
Jail term	(6, 2)	(12, 6)
CC term	(12, 4)	(24, 12)

EC.7.2. Recidivism Arrival Distributions

During a customer's participation in a CC program and or a customer completes their sentence and is ready to be released from jail or CC, we predict the likelihood and timing of a future recidivism event based on their risk class. If they are predicted to recidivate, we add them to the system's customer record with a future arrival time. This enables us to simulate the impact of returning customers on the system over time.

Figure EC.2 shows the distributions of daily recidivism arrivals from simulation experiments. Two policies are used: the RL-based dynamic policy and the static policy. We compare the distribution of recidivism arrivals under three HD staffing settings. The results indicate that the static policy produced higher recidivism arrivals, particularly when the HD was understaffed.



(a) Low staff (b) Baseline (c) High staff

Figure EC.2 Comparison of recidivism arrival distributions under different HD staffing settings

EC.7.3. Parameter Specification of Case Study in Section 7

Table EC.3: Parameter Specification in Case Study

Parameter	Value and Calibration
Discounting factor	$\gamma = 0.9999$
Time horizon	Simulation length: 3 years. Warm start period: 2 months.
Arrival rate	$\lambda_{\text{base}} = 84$ per week. Estimated using historical County jail and TCCC arrival data: $\lambda_{\texttt{jail}} = 37$ and $\lambda_{\texttt{CC}} = 47$
Risk class proportion	$\lambda_{\texttt{severe}} : \lambda_{\texttt{mild}} = 0.3 : 0.7$. Estimated based on the proportion of each risk class of customers in historical data.
Sentence length	The baseline sentence lengths are 12 weeks for jail, 5 weeks for WR, and 15 weeks for HD, based on the average historical program length. For severe and mild types of customers, sentence lengths are incremented by +1 week and -1 week, respectively, from the baseline.
Base recidivism risk	We fit a logistic regression model to estimate the recidivism and violation risks. We use a 70% cutoff for the high-risk class to estimate the risk distribution p_m's. The base recidivism risk is sampled from the recidivism risk distribution estimated using a logistic regression model with an AUC score of 0.71. Recidivism risk decrease every day exponentially with rate 0.997.
Base violation risk	In WR, the violation risks are 0.001 and 0.0006 for severe and mild customers, respectively; In HD, the violation risks are 0.003 and 0.0018 for severe and mild customers, respectively.
Congestion's impact	Recidivism/violation risk = Base recidivism/violation risk × congestion-adjusted factor f, where f is a piecewise linear function of total occupancy in each station: $$1 + \frac{\max(0, \text{Occupancy}(j) - \text{Capacity}(j))}{\text{Capacity}(j)},$$ where the capacities of jail, WR and HD are set as 300, 120, and 300 in baseline case.
Holding cost	Quadratic function of total occupancy in each station
Risk type transition	$P_{m_1 \to m_2}$ denotes the probability of changing to type m_2 from type m_1 every day. In jail, $P_{0\to1} = 0.005$ and $P_{1\to0} = 0.01$. In WR and HD, $P_{0\to1} = 0.01$ and $P_{1\to0} = 0.005$.

EC.7.4. Cost-effectiveness Analysis of Adding Case Managers in Section 7.3

Holding Cost Holding cost estimation follows the Fiscal Impact Analyses of Illinois Sentencing Policy Advisor Council [3], including:

1. variable costs, which directly relate to services (laundry, food, etc.) – Monitoring cost in our case. Currently, in most Indiana counties, house arrest costs a minimum of $50 to set up and a minimum of $10-$15 per day in community correction monitoring fee [4];
2. personnel costs, which change when staffing levels change – Case manager salary in our case;

[3] https://spac.icjia-api.cloud/uploads/2023%20Update%20-%20Marginal%20Costs%20for%20Fiscal%20Impacts-20230210T16344761.pdf

[4] https://goldmanlegalhelp.com/what-is-house-arrest-in-indiana/

3. fixed costs, which relate to physical space that vary only with large service changes – Not applicable in our case.

Recidivism Cost Recidivism cost includes cost of rearrest, prosecution, and incarceration; economic costs on communities, such as lost productivity [5].

Violation Cost A technical probation violation is when someone violates a rule placed on their probation. Technical probation violations can be things like missing curfew or losing a job. These types of violations are noncriminal violations. Technical violations accounted for 26% of the population in the study's group who went to jail or prison [6]. The Court Services and Offender Supervision Agency's (CSOSA) results show that people arrested for probation violations show a roughly 30% chance of rearrest [7]. Thus we estimate the cost per technical violation as 1/3 of the recidivism, which is $50,000.

EC.7.5. Performance under Different Arrival Rates and Recidivism Window

In this section, we present more numerical results demonstrating the performance robustness of our proposed policy. We evaluate the policy's performance under different customer arrival rates and lengths of recidivism windows to demonstrate its performance robustness in varying crime rates and holding costs.

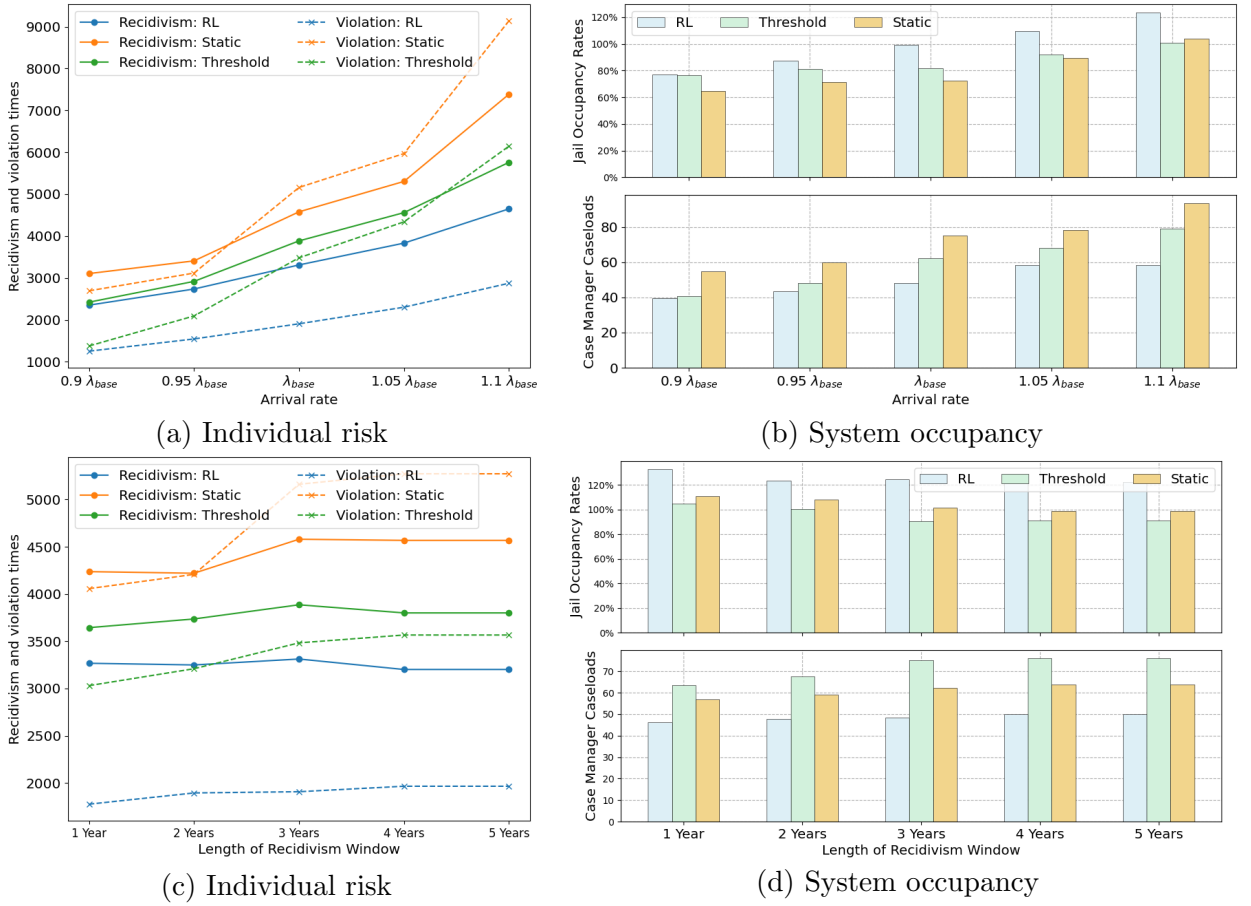
Figure EC.3 illustrates the performance gap across various customer arrival rates. The baseline model assumes an arrival rate of 84 customers per week, estimated through data from the TCCC and the Indiana Department of Correction (IDOC). Simulations are conducted over a three-year period, and we report the average value from 100 replications. The 95% confidence intervals are tight, and we omit them in the plots.

As shown in Figures EC.3(a) and (c), our RL-based policy consistently outperforms the static and threshold-based policies, resulting in few instances of recidivism and violation across all tested scenarios. Notably, as the customer arrival rate increases, the recidivism and violation rates under the RL-based policy increase at a slower rate than those of the other policies. This effect can be attributed to the effective congestion-mitigation strategies of the RL-based policy. Additionally, Figures EC.3(a) and (c) suggest that the threshold-based policy can outperform the static policy with consideration of system congestion, thereby highlighting the benefits of a congestion-adjusted placement policy. Furthermore, the performance gap among these three policies are relatively small under lower arrival rates, where system congestion is of lesser concern. In such scenarios, the static policy, which optimizes station assignments based on individual risks, can be beneficial.

[5] https://spac.icjia-api.cloud/uploads/Illinois_Result_First-The_High_Cost_of_Recidivism_2018-20191106T18123262.pdf

[6] https://ijrd.csw.fsu.edu/sites/g/files/upcbnu1766/files/media/images/publication_pdfs/Going_Back_to_Jail.pdf

[7] https://www.csosa.gov/wp-content/uploads/bsk-pdf-manager/2020/02/CSP-FY2021-Congressional-Budget-Justification-02062020.pdf#page=22



(a) Individual risk (b) System occupancy

(c) Individual risk (d) System occupancy

Figure EC.3 Performance gap under different customer arrival rates and recidivism windows

Figures EC.3(b) and (d) present the three-year average occupancy rates of jail and caseloads in HD under the three policies. Work Release occupancy rates remain consistently high, ranging from 96% to 99%, due to capacity constraints and the benefits of training programs, thus is omitted from the figure. As shown in Figure EC.3(b), the RL-based policy could maintain similar congestion levels at jail and HD and avoid extreme congestion in either station. However, Figure EC.3(b) also indicates that all stations are inevitably more congested under higher customer arrival rates, highlighting the importance of accurate crime rate estimations during capacity and staffing planning. In addition, Figure EC.3(d) demonstrates that, as the length of the recidivism window increases, the HD program becomes a more preferable option, aligning with the policy structure discussed in Section 5.3.