# A Continuous-class Queueing Model With Proportional Hazards-based Routing

Neal Master
Department of Electrical Engineering, Stanford University, Stanford, CA 94305, nmaster@stanford.edu

Martin I. Reiman
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, martyreiman@gmail.com

Can Wang
Department of Electrical Engineering, Stanford University, Stanford, CA 94305, canw@stanford.edu

Lawrence M. Wein
Graduate School of Business, Stanford University, Stanford, CA 94305, lwein@stanford.edu

Motivated by jail overcrowding and the U.S. correctional system's widespread use of risk models to aid in inmate release decisions both prior to trial (i.e., pretrial release) and near the end of their jail sentence (i.e., split sentencing), we formulate and analyze a queueing network model with two novel features: there is a continuum of customer classes corresponding to an inmate's continuous risk level $p$, and routing in the network is dictated by a Cox proportional hazards model, where the hazard rate associated with recidivism (i.e., committing another crime) during release is proportional to $e^{\gamma p}$ for some parameter $\gamma$. We perform an exact analysis of a continuous-class $M/M/c/c$ (i.e., Erlang B) model where preemptive priority is awarded according to the risk level $p$, and use it to develop approximate performance measures for a queueing network of a jail with a two-threshold policy, which dictates who is granted pretrial release and who receives a split sentence. For a slightly simplified version of the model, we derive sufficient conditions under which no inmates are offered pretrial release unless all inmates are given a split sentence.

*Key words*: queueing theory, loss systems, proportional hazards models, jails

## 1. Introduction

Due to the expanding availability of data, customers often arrive to service operations (e.g., hospitals, organ transplant waiting lists, matching markets, web pages) with an individualized set of features (i.e., explanatory variables) that is readily observable by the system manager. This leads to the challenge of how to optimally manage – e.g., accept/reject, route, match, prioritize, advertise to, charge – individual customers. As a result, traditional operations research models,

such as the newsvendor model (Ban and Rudin 2016), inventory management (Ban, Gallien and Mersereau 2017), the multi-armed bandit (Auer 2002), dynamic pricing (Cohen, Lobel and Leme 2016, Qiang and Bayati 2016, Ban and Keskin 2017), advertising (Langford and Zhang 2007) and auctions (Amin, Rostamizadeh and Syed 2014), are being endowed with customers (or resources) that possess a vector of observable features.

In this paper, motivated by jail overcrowding, we do the same with a queueing model. This paper builds on Usta and Wein (2015), which develops a simulation model of the Los Angeles (LA) County jail system. Due to severe prison overcrowding, the U.S. Supreme Court (*Brown v. Plata*, 2011) forced the state of California (CA) to reduce its prison population by 25% within two years. As a result, CA passed the Public Safety Realignment Act (Assembly Bill 109), which required low-level felons and state parole violators to be incarcerated in CA county jails rather than state prisons. While successfully reducing the state prison population, this realignment led to significant overcrowding in the CA county jails, particularly in LA County (Petersilia 2014).

CA counties have two primary options for reducing their jail population. They can offer pretrial release, which runs the risk that defendants will either recidivate – i.e., commit another crime – before their case is decided or fail to appear for their court date. They can also offer split sentencing to low-level felons (indeed, Assembly Bill 1468 requires this, unless the court finds that it is not in the interest of justice), which splits their sentence between jail time and mandatory supervision.

To guide these decisions, many correctional facilities throughout the U.S. employ validated risk-based tools, which predict the probability of recidivism and/or failing to appear in court, based on a defendant's criminal history and demographic data (Yang, Wong and Coid 2010). Usta and Wein (2015) estimate risk-based survival models for recidivism and failure to appear using results from the risk-based tools, and embed them into a queueing network model of the jail system. The focus of their study is to find the optimal mix of pretrial release and split sentencing to minimize the amount of recidivism subject to a constraint on the jail population.

Here, we consider a simplified version of the model in Usta and Wein (2015): we ignore the distinction between felony and non-felony charges, the short time delay between arrest and arraignment,

and the possibility that cases lead to acquittal or probation, and we assume that all non-recidivating defendants on pretrial release successfully appear in court for their case disposition. We make one assumption that makes the model more difficult to analyze: whereas Usta and Wein (2015) assumes that LA County rents jail beds from other counties when it reaches capacity, we consider a jail with a fixed number of beds. However, in our model, inmates who are released for lack of space and go on to recidivate do not re-enter the system. While the goal of Usta and Wein (2015) is to present a reasonably accurate model of the LA County jail system for purposes of informing public policy (Wein and Usta 2015), the goals here are to introduce a new class of queueing systems to the management science community, and to present an initial analysis of such a system. We also note that – in contrast to some of the studies of other problems mentioned above – there is no attempt here to optimize the exploration-exploitation tradeoff: the parameters of the risk tool are assumed known (as is appropriate in the jail context, given the previous validation of the risk models), and we focus on a performance analysis of the queueing model.

We construct a model with a continuum of classes spanning the range of inmate risk levels that could be generated by the risk-based tool. In our jail setting, the servers are jail beds and risk is used to prioritize inmates, in that lower-risk inmates are more apt to be granted pretrial release or split sentencing. Recidivism is modeled by a proportional hazards model that explicitly uses each customer's risk level, and inmates who recidivate while on pretrial release or supervision re-enter the jail system. The rest of the model is Markovian: arrivals are Poisson, the time from arrest to case disposition, the post-sentence jail term, and the length of supervision are all exponential. Hence, the model has the structure of a loss queue with delayed random feedback due to recidivism during pretrial release and supervision. We undertake an approximate performance analysis of a two-threshold control policy, where inmates with priority lower than one threshold are granted pretrial release and inmates with priority lower than a second threshold (which may be larger or smaller than the first threshold) are given a split sentence. After the optimal threshold values are found, this analysis allows us to generate tradeoff curves of the crime rate vs. the mean jail population.

Our modeling of a continuum of classes is similar in spirit to a large stream of queueing literature that goes back at least to Mendelson (1985), where arriving customers have a PDF over their valuation of receiving service; in addition, we briefly discuss measure-valued processes in queues in §4 as they relate to possible generalizations of the model. Similarly, very few queueing models incorporate proportional hazards models: going back to Zenios, Chertow and Wein (2000), the lifetime of a recipient with a donated kidney can be modeled by a proportional hazards model and used to allocate kidneys to candidates on the transplant waiting list, but the models in this area typically have a finite number of customer classes.

In §2, we analyze a continuous-class $M/M/c/c$ queue with preemptive priority, which generalizes the classic results for two-class (Helly 1961) and multiclass (Burke 1962) models. This analysis provides the main building blocks for analyzing the jail model, which is formulated and analyzed in §3. Concluding remarks, including a discussion of the potential use of these ideas in modeling hospitals and call centers, appear in §4.

## 2. The Continuous-class $M/M/c/c$ Model

The queueing model is formulated in §2.1 and analyzed in §2.2-2.3. A crime model is superimposed on the queuing model in §2.4.

### 2.1. The Model

As in the classic $M/M/c/c$ (i.e., Erlang B) queueing system, we assume that customers arrive according to a Poisson process with rate $\lambda$, each of $c$ servers has independent exponential service times with rate $\mu$, and there is no waiting room. Arriving customers are assigned a random priority $p$, which is observable by the system manager. For simplicity, we assume that $p \sim U[0,1]$, although we note that risk-based recidivism tools are sometimes constructed so as to generate uniformly distributed risk scores (e.g., §2.7 of Northpointe (2015)), and that eject and reject probabilities are preserved under monotone transformations of the priority levels.

In the traditional $M/M/c/c$ queue, a customer is blocked if he arrives to find all $c$ servers busy. In our model, there are two types of events that can occur when a customer arrives to a system

with no idle servers. He is *rejected* (i.e., leaves the system without entering service) if he has a lower priority than everyone in service; otherwise, the lowest priority customer in service is immediately *ejected* from the system to make room for the new arrival. Within the context of our model, we use the term *blocking* to refer to the sum of rejects and ejects. As in the standard model, an arriving customer in our model immediately enters service when there is an idle server available.

### 2.2. Eject and Reject Analysis

By construction, the total number of customers in the system at each point in time and the total blocking (i.e., rejects plus ejects) are exactly the same as in a $M/M/c/c$ system. Our primary aim is to calculate $P^r(p)$, which is the probability that an arriving customer with priority $p$ is rejected, and $P^e(p)$, which is the probability that a priority $p$ customer is initially accepted into the system and then ejected before completing service.

By the preemptive nature of our queueing discipline, a customer with priority $p$ is not affected by any customer with priority less than $p$. Hence, the rejection probability for a customer with priority $p$ is given by

$$P^r(p) = B(c, a(1-p)), \tag{1}$$

where

$$a = \frac{\lambda}{\mu}$$

is the offered load, and

$$B(x, y) = \frac{y^x/x!}{\sum_{k=0}^{x} y^k/k!} \tag{2}$$

is the Erlang blocking formula with $x$ servers and offered load $y$ (e.g., pp. 5 and 79 of Cooper (1981)).

The total reject plus eject rate (i.e, per unit time) of customers with priority levels in $[p, 1]$ is $\lambda \int_p^1 (P^r(x) + P^e(x))\ dx$. By construction, this quantity must equal the total blocking rate in an Erlang B system with arrival rate $\lambda(1-p)$, which yields

$$\lambda \int_p^1 (P^r(x) + P^e(x))\ dx = \lambda(1-p)B(c, a(1-p)). \tag{3}$$

Let $B_2(x,y) = \frac{\partial}{\partial y} B(x,y)$. Dividing both sides of (3) by $\lambda$ and differentiating with respect to $p$ yields

$$-\left(P^r(p) + P^e(p)\right) = -B(c, a(1-p)) - a(1-p)B_2(c, a(1-p)). \tag{4}$$

Using (1) to substitute for $P^r(p)$ in (4), we obtain

$$P^e(p) = a(1-p)B_2(c, a(1-p)). \tag{5}$$

We can write (5) in terms of $B(x,y)$ rather than $B_2(x,y)$ by leveraging Theorem 15 of Jagerman (1974), which states that

$$B_2(c, a(1-p)) = B(c, a(1-p)) \left( \frac{c}{a(1-p)} - 1 + B(c, a(1-p)) \right). \tag{6}$$

Substituting (6) into (5) gives

$$P^e(p) = B(c, a(1-p))[c - a(1-p)(1 - B(c, a(1-p)))]. \tag{7}$$

### 2.3. Stochastic Ordering of Ejected and Rejected Customers

Let $R$ be the random priority of a rejected customer and $E$ be the random priority of an ejected customer, and let the corresponding probability density functions (PDFs) be $f_r(\cdot)$ and $f_e(\cdot)$. In this subsection, we show that $R$ is smaller than $E$ in the likelihood ratio order by using equations (1) and (7):

$$\frac{f_e(p)}{f_r(p)} = \frac{P^e(p) / \int_0^1 P^e(x)\ dx}{P^r(p) / \int_0^1 P^r(x)\ dx} \propto \frac{P^e(p)}{P^r(p)} = \frac{B(c, a(1-p))[c - a(1-p)(1 - B(c, a(1-p)))]}{B(c, a(1-p))}, \tag{8}$$

$$= c - a(1-p)(1 - B(c, a(1-p))). \tag{9}$$

The derivative of (9) with respect to $p$, after applying (6), is

$$a(1 - P^r(p) - P^e(p)). \tag{10}$$

Because $1 - P^r(p) - P^e(p)$ is the probability that a customer with priority $p$ is successfully served, the derivative in (10) is nonnegative. Therefore, $f_e(p)/f_r(p)$ is nondecreasing in $p$, which gives the

desired result. Theorem 1.C.1 in Shaked and Shanthikumar (2007) implies that $R$ is smaller than $E$ in the hazard rate ordering, reverse hazard rate ordering, and usual stochastic ordering.

Finally, to better understand the relative magnitudes of $f_e(p)$ and $f_r(p)$, we consider a $M/M/c/c$ system with unit service rate and arrival rate $a(1-p)$. Then a fraction $1-B(c,a(1-p))$ of the arriving traffic is served and hence the departure rate of serviced traffic is $a(1-p)(1-B(c,a(1-p)))$. If each server was constantly working, the departure rate of serviced traffic would be $c$. Hence, $a(1-p)(1-B(c,a(1-p)))<c$. The term $(1-B(c,a(1-p)))$ decreases only sublinearly in $a$, and so if we take the limit as $a$ tends to infinity, we get

$$a(1-p)(1-B(c,a(1-p)))\uparrow c \text{ as } a\to\infty.$$

Because $c$ is a finite integer, $a$ must be much larger than $c$ before $a(1-p)(1-B(c,a(1-p)))$ is comparable to $c$. That is, only when $a$ is much larger than $c$ will $f_e(p)$ and $f_r(p)$ be similar in magnitude; otherwise, the quantity in (9) will be large and $f_e(p)$ will be much greater than $f_r(p)$.

### 2.4. Crime Rate

If we now view this $M/M/c/c$ queue as a jail, we are interested in the crime rate due to ejected and rejected customers (i.e., inmates) who recidivate within $\exp(\mu)$ time units, which is the amount of time they would have been jailed had we possessed ample jail capacity. From §2.2, we know that the ejection rate of inmates with priority $p$ is $\lambda P^e(p)$ and the rejection rate of inmates with priority $p$ is $\lambda P^r(p)$. If we assume that inmates recidivate according to a proportional hazards model (Cox 1972) with constant baseline hazard rate $\eta>0$ and regression parameter $\gamma>0$, then upon ejection or rejection an inmate with priority $p$ will commit a crime after an exponential amount of time with rate $\eta e^{\gamma p}$. Therefore, upon ejection or rejection, an inmate with priority $p$ will commit a crime when he could have been in jail with probability $\eta e^{\gamma p}/(\eta e^{\gamma p}+\mu)$. Hence, the crime rate for ejected inmates in the $M/M/c/c$ continuous class model is

$$\lambda\int_0^1 P^e(p)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu}\,dp=\lambda\int_0^1 B(c,a(1-p))[c-a(1-p)(1-B(c,a(1-p)))]\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu}\,dp,$$

and the crime rate for rejected inmates is

$$\lambda\int_0^1 P^r(p)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu}\,dp=\lambda\int_0^1 B(c,a(1-p))\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu}\,dp.$$

# 3. The Jail Model

Our jail model in this section is more complex than the model considered in §2, and we use a nearly completely decomposable Markov chain approximation (Courtois 1977) applied to a two-class priority Erlang loss system to analyze the system performance. The model is formulated in §3.1, analyzed in §3.2-3.3, and a numerical example is carried out in §3.4 to assess the accuracy of the approximations. In §3.5, we derive sufficient conditions for the dominance of split sentencing over pretrial release for a slightly simplified version of the problem that corresponds to an underloaded jail system.

## 3.1. The Model

Inmates arrive according to a Poisson process with rate $\lambda$ (Fig. 1). Each inmate has a random priority $p \sim U[0,1]$, which is observed by the system manager. We consider a family of double threshold policies, where inmates are awarded pretrial release if their priority $p < \theta_R$ and undergo pretrial detention if $p > \theta_R$, and inmates receive a split sentence if $p < \theta_S$ and do not receive a split sentence if $p > \theta_S$ (Fig. 1). In §3.2-§3.3, we analyze the cases $\theta_R \geq \theta_S$ and $\theta_R < \theta_S$, respectively. If an inmate with priority $p$ is out on pretrial release or supervision, he recidivates according to a proportional hazards model with a constant baseline hazard rate; i.e., he commits a crime after an exponential amount of time with rate $\eta e^{\gamma p}$.

Inmates on pretrial release wait an exponential amount of time with rate $r$ until case disposition. If an inmate recidivates during this time, he returns to the start of the process but his priority $p$ does not change; i.e., he undergoes pretrial release again. An inmate experiencing pretrial detention waits in jail for an exponential amount of time with rate $\mu_1$ until case disposition.

We assume that all inmates are found guilty at case disposition. After case disposition – regardless of whether the inmate was released or detained prior to trial – an inmate with priority $p > \theta_S$ receives an exponential post-sentence jail term with rate $\mu_2$ and then exits the system, and an inmate with priority $p < \theta_S$ receives a split sentence, which consists of an exponential post-sentence jail term with rate $\mu_3$ (where $\mu_3 > \mu_2$) followed by an exponential amount of time under mandatory

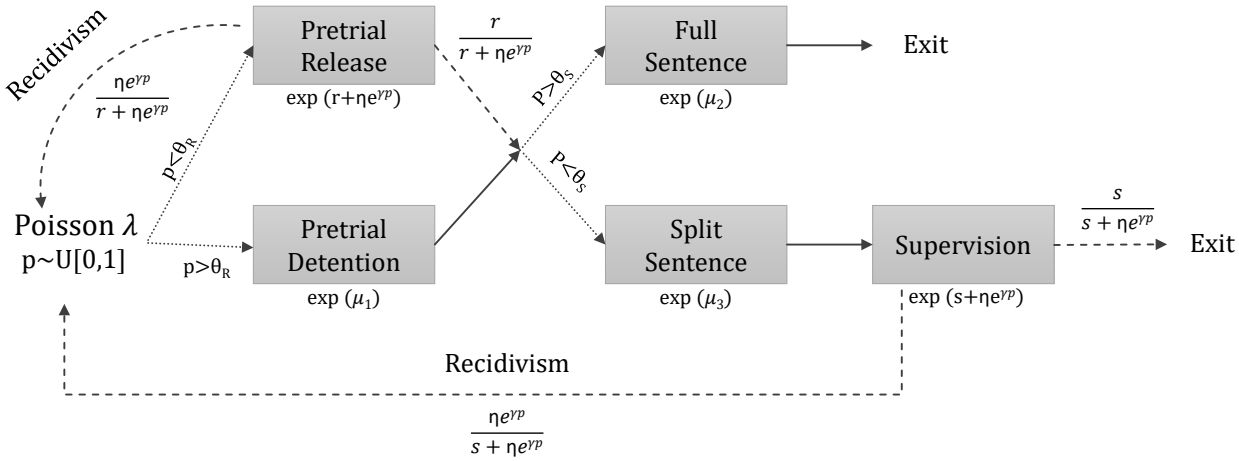


**Figure 1** The jail model. Dotted lines correspond to decisions and dashed lines correspond to probabilistic routing. The parameters are described in Table 1.

supervision with rate $s$. If an inmate does not recidivate while on supervision, then he exits the system. If he does recidivate, he returns to the beginning of the process without changing priority level (so that he receives pretrial release if $p \le \theta_R$ and pretrial detention if $p > \theta_R$).

Note that pretrial release can be modeled as an infinite-server queue with an exponential service rate $r + \eta e^{\gamma p}$ for customers with priority $p$, after which customers are routed back to the beginning of the process with probability $\eta e^{\gamma p}/(r + \eta e^{\gamma p})$ and are routed to post-sentencing with probability $r/(r + \eta e^{\gamma p})$ (Fig. 1). Similarly, supervision can be viewed as an infinite-server queue with exponential service rates $s + \eta e^{\gamma p}$, with departing inmates returning to the beginning of the process with probability $\eta e^{\gamma p}/(s + \eta e^{\gamma p})$ and exiting the system with probability $s/(s + \eta e^{\gamma p})$.

In contrast, the jail is a finite-server queue with no waiting room, which is managed using the continuous priority levels of the inmates. If an inmate arrives to the jail and there is no available space, then he will be rejected if he has lower priority than every inmate currently in jail, and otherwise the lowest priority inmate in jail will be ejected and the arriving inmate will take his place. We assume that ejected and rejected inmates are released from the jail system and are at risk of recidivism for the amount of time that they would have been kept in the jail if it had infinite capacity. For example, consider the case when $\theta_R = \theta_S = 0$ and hence neither pretrial release nor

split sentencing are available to any inmate. In this case (Fig. 1), a rejected inmate will be at risk of recidivism for a random amount of time given by the sum of two exponential random variables with rates $\mu_1$ and $\mu_2$, and an ejected inmate will be at risk of recidivism for an exponential amount of time with rate $\mu_2$ if he is in post-sentencing and a random amount of time given by the sum of two exponential random variables with rates $\mu_1$ and $\mu_2$ if he is in pretrial detention. For simplicity, we assume that ejected or rejected inmates who recidivate do not re-enter the system; the implications of this assumption are discussed in §4.

### 3.2. The Case $\theta_R \geq \theta_S$

We begin by assuming that $\theta_R \geq \theta_S$, which divides the inmates into three different process flows (Fig. 2): inmates with $p \in (\theta_R, 1]$ spend all of their pretrial and post-sentence time in jail, inmates with $p \in [\theta_S, \theta_R]$ receive pretrial release but spend all of their post-sentence time in jail, and inmates with $p \in [0, \theta_S)$ receive both pretrial release and a split sentence. The steady-state number of inmates in jail from these three process flows are denoted by $Q_1$, $Q_2$ and $Q_3$, respectively (Fig. 2). We are interested in two performance measures: the mean steady-state jail population and the crime rate.

**Mean Jail Population.** Inmates with priority $p > \theta_R$ are unaffected by lower priority customers. Hence, $Q_1$ is given by the number of customers in a $M/G/c/c$ system with arrival rate $\lambda(1-\theta_R)$ and service times that are the sum of two independent exponentials with rates $\mu_1$ and $\mu_2$. Let us define $\mu_{12}$ so that

$$\mu_{12}^{-1} = \mu_1^{-1} + \mu_2^{-1}, \tag{11}$$

and let

$$a_1 = \frac{\lambda(1-\theta_R)}{\mu_{12}}. \tag{12}$$

By Little's formula, we have

$$E[Q_1] = a_1 \left(1 - B(c, a_1)\right). \tag{13}$$

To analyze $Q_2$, we use the nearly completely decomposable Markov chain approach (e.g., Courtois (1977)), and assume that, conditioned on $Q_1 = i$, $Q_2$ is the number of customers in an Erlang loss

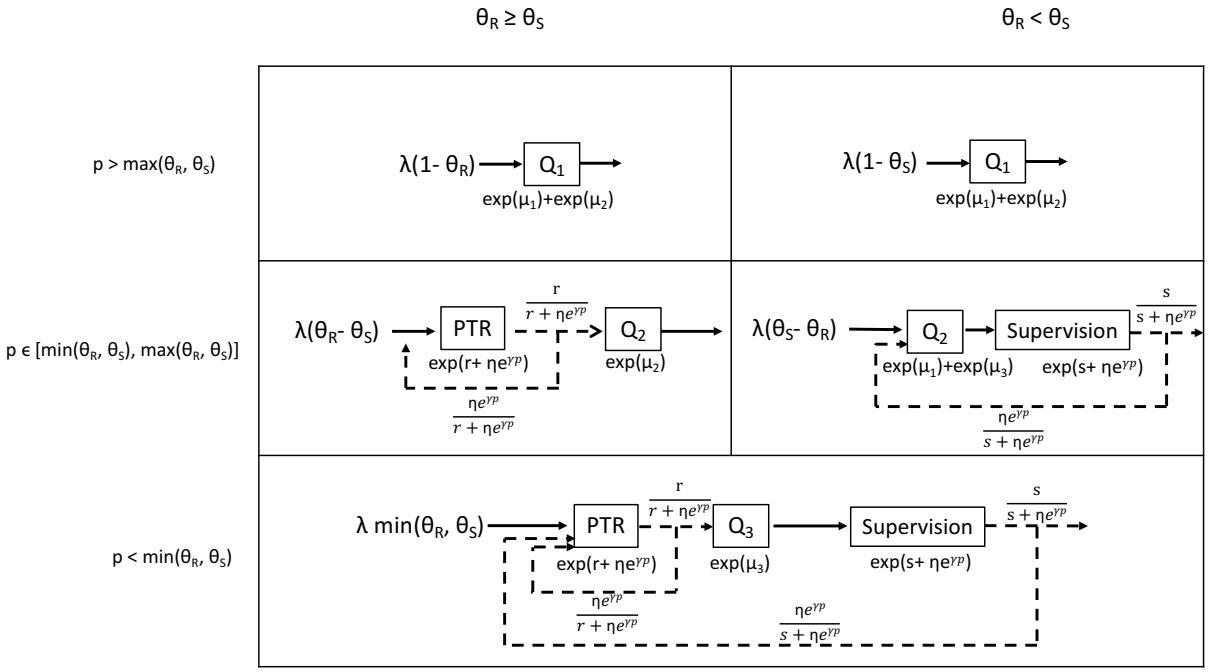


**Figure 2** The three process flows corresponding to $Q_1$, $Q_2$ and $Q_3$ for the cases $\theta_R \geq \theta_S$ and $\theta_R < \theta_S$. The quantities under the boxes are service rates and the quantities near the dashed lines are routing probabilities.

system with arrival rate $\lambda(\theta_R - \theta_S)$, service rate $\mu_2$ and $c-i$ servers (Fischer 1980). In the absence of feedback to pretrial release (Fig. 2), this approximation is asymptotically correct as $\mu_{12}/\mu_2 \to 0$ (Fischer 1980), and for our set of parameter values (Table 1), $\mu_{12}/\mu_2 = 0.842$. The arrival process to $Q_2$ is Poisson because, with feedback to pretrial release, the pretrial release box in Fig. 2 for this case can be viewed as an infinite-server queue with immediate Markovian feedback, which has a Poisson exit process (Corollary 2.6 of Kelly (1979)). If we define

$$a_2 = \frac{\lambda(\theta_R - \theta_S)}{\mu_2},$$

then this approximation yields

$$E[Q_2] \approx a_2 \left(1 - \sum_{i=0}^{c} \frac{a_1^i/i!}{\sum_{k=0}^{c} a_1^k/k!} B(c-i, a_2)\right). \tag{14}$$

In passing, we note that an alternative approach to approximating $E[Q_2]$, which is not pursued here, would be to assume $\mu_1 = \mu_2$, in which case

$$E[Q_2] = E[Q_1 + Q_2] - E[Q_1],$$

$$= \frac{\lambda(1-\theta_S)}{\mu_2}\left(1 - B\Big(c, \frac{\lambda(1-\theta_S)}{\mu_2}\Big)\right) - \frac{\lambda(1-\theta_R)}{\mu_{12}}(1 - B(c, a_1)).$$

We follow a similar approach to analyze $Q_3$ and assume that, conditioned on $Q_1 + Q_2 = i$, $Q_3$ is the number of customers in an Erlang loss system with offered load $a_3$ and $c-i$ servers. To compute the offered load, we first consider only the inmates with priorities in $[p, p+dp]$ for some $p < \theta_S$. These inmates have a hazard rate of approximately $\eta e^{\gamma p}$. Ignoring any ejections or rejections, the rate at which these customers exit the system after supervision is $\lambda dp$. Because inmates leaving supervision exit the system with probability $s/(s+\eta e^{\gamma p})$, it follows that the flow into supervision – and hence through the $Q_3$ queue – is $(\eta e^{\gamma p} + s)\lambda dp/s$ (Fig. 2). Integrating this value for priorities $p \in [0, \theta_S]$ and dividing by $\mu_3$ gives the offered load,

$$\begin{aligned} a_3 &= \frac{1}{\mu_3}\int_0^{\theta_S} \frac{\eta e^{\gamma p} + s}{s}\lambda \; dp, \\ &= \frac{\lambda}{\mu_3}\Big(\frac{\eta}{s\gamma}(e^{\gamma\theta_S} - 1) + \theta_S\Big). \end{aligned} \tag{15}$$

The mean queue length can be approximated by

$$E[Q_3] \approx a_3 \left(1 - \sum_{\substack{i_1=0 \\ i_1+i_2\le c}}^{c}\sum_{i_2=0}^{c} \frac{(a_1^{i_1}/i_1!)(a_2^{i_2}/i_2!)}{\sum_{\substack{j_1=0 \\ j_1+j_2\le c}}^{c}\sum_{j_2=0}^{c}(a_1^{j_1}/j_1!)(a_2^{j_2}/j_2!)} B(c - i_1 - i_2, a_3)\right). \tag{16}$$

With approximations for $E[Q_1]$, $E[Q_2]$, and $E[Q_3]$ in (13), (14) and (16), we can estimate $E[Q]$, the mean jail population, by

$$E[Q] \approx E[Q_1] + E[Q_2] + E[Q_3]. \tag{17}$$

**Crime Rate.** Referring to Fig. 2, we see that inmates with $p \in [\theta_S, \theta_R]$ can recidivate during pretrial release, inmates with $p < \theta_S$ can recidivate during pretrial release and supervision, and all inmates can recidivate if they are ejected or rejected due to all the servers being busy. We compute the crime rate (the number of recidivisms per unit time) for each of the three process flows in Fig. 2.

Inmates with priority $p > \theta_R$ can only recidivate if they are ejected or rejected. Recall that these inmates are not impacted by inmates with priority $p \le \theta_R$. Hence, these inmates have priority levels

that are uniformly distributed between $\theta_R$ and 1, and have cumulative distribution function (CDF) $(p-\theta_R)/(1-\theta_R)$ in this range. Because the eject and reject probabilities are preserved under monotone transformations of the priority levels, we can calculate these probabilities by replacing the priority level $p$ in equations (1) and (7) by the CDF $(p-\theta_R)/(1-\theta_R)$. In addition, we now add the subscripts $c$ and $a$ to equations (1) and (7) to streamline our presentation.

If an inmate with priority $p>\theta_R$ is rejected, then the time that he is exposed to recidivism is the sum of two exponentials with rates $\mu_1$ and $\mu_2$ (Fig. 2), and the probability that he commits a crime during this time is

$$\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}+\left(1-\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}\right)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}.$$

Integrating the product of the rejection probability and the recidivism probability over the PDF of priorities $p>\theta_R$ and multiplying by their arrival rate gives the crime rate due to rejected inmates with priorities $p>\theta_R$,

$$\lambda(1-\theta_R)\int_{\theta_R}^{1}\left[\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}+\left(1-\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}\right)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}\right]P^r_{c,a_1}\left(\frac{p-\theta_R}{1-\theta_R}\right)\,dp. \tag{18}$$

Because the arrival process is Poisson, if one of these inmates is ejected, then with probability $\mu_1^{-1}/\mu_{12}^{-1}$ he is exposed to recidivism for a length of time that is the sum of two exponentials with rates $\mu_1$ and $\mu_2$, and with probability $\mu_2^{-1}/\mu_{12}^{-1}$ he is exposed to recidivism for an exponential amount of time with rate $\mu_2$. Hence, the crime rate due to ejected inmates with priorities $p>\theta_R$ is

$$\lambda(1-\theta_R)\int_{\theta_R}^{1}\left[\frac{\mu_1^{-1}}{\mu_{12}^{-1}}\left(\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}+\left(1-\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}\right)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}\right)+\frac{\mu_2^{-1}}{\mu_{12}^{-1}}\left(\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}\right)\right]P^e_{c,a_1}\left(\frac{p-\theta_R}{1-\theta_R}\right)\,dp. \tag{19}$$

Inmates with priority $p\in[\theta_S,\theta_R]$ can recidivate due to ejection, rejection or pretrial release. These inmates are exposed to possible recidivism for an exponential amount of time with rate $\mu_2$ if there is an ejection or rejection (Fig. 2). Following the steps leading to (18)-(19), we again use the approximation that, conditioned on $Q_1=i$, the $Q_2$ queue behaves as the number of customers in an Erlang loss system with $c-i$ servers and offered load $a_2$. Noting that inmates arriving to $Q_2$

have priority levels that are uniformly distributed on $[\theta_S, \theta_R]$, we find that the crime rate due to ejected and rejected inmates with priority $p \in [\theta_S, \theta_R]$ is approximately

$$\lambda(\theta_R - \theta_S) \sum_{i=0}^{c} \frac{a_1^i/i!}{\sum_{k=0}^{c} a_1^k/k!} \int_{\theta_S}^{\theta_R} \frac{\eta e^{\gamma p}}{\eta e^{\gamma p} + \mu_2} \left( P^e_{c-i,a_2}\left(\frac{p-\theta_S}{\theta_R-\theta_S}\right) + P^r_{c-i,a_2}\left(\frac{p-\theta_S}{\theta_R-\theta_S}\right)\right) \, dp. \tag{20}$$

By Fig. 2, we see that inmates with priority $p \in [\theta_S, \theta_R]$ recidivate while on pretrial release a geometric number of times $(0, 1, 2, \ldots)$ with mean $\eta e^{\gamma p}/r$. Hence, the crime rate due to inmates with priority $p \in [\theta_S, \theta_R]$ on pretrial release is

$$\lambda \int_{\theta_S}^{\theta_R} \frac{\eta e^{\gamma p}}{r} \, dp = \frac{\lambda \eta}{r\gamma}(e^{\gamma \theta_R} - e^{\gamma \theta_S}). \tag{21}$$

Inmates with priority $p < \theta_S$ can recidivate in four ways: ejection, rejection, pretrial release or supervision. Following the same infinitesimal argument preceding equation (15), the CDF of priority levels for inmates with $p < \theta_S$ entering $Q_3$ (Fig. 2) is

$$\begin{aligned} G(p) &= \frac{\int_0^p \left(\frac{\eta}{s} e^{\gamma x} + 1\right) \, dx}{\int_0^{\theta_S} \left(\frac{\eta}{s} e^{\gamma x} + 1\right) \, dx}, \\ &= \frac{\frac{\eta}{s\gamma} e^{\gamma p} + p}{\frac{\eta}{s\gamma} e^{\gamma \theta_S} + \theta_S}. \end{aligned} \tag{22}$$

These inmates are exposed to recidivism for an exponential amount of time with rate $\mu_3$ if they are ejected or rejected. Again making the assumption that, conditioned on $Q_1 + Q_2 = i$, $Q_3$ is the number of customers in an Erlang loss system with $c - i$ servers and offered load $a_3$, we approximate the crime rate due to ejected and rejected inmates with $p < \theta_S$ by

$$a_3\mu_3 \sum_{\substack{i_1=0 \\ i_1+i_2\le c}}^{c} \sum_{i_2=0}^{c} \frac{(a_1^{i_1}/i_1!)(a_2^{i_2}/i_2!)}{\sum_{\substack{j_1=0 \\ j_1+j_2\le c}}^{c} \sum_{j_2=0}^{c} (a_1^{j_1}/j_1!)(a_2^{j_2}/j_2!)} \int_0^{\theta_S} \frac{\eta e^{\gamma p}}{\eta e^{\gamma p} + \mu_3} \left( P^e_{c-i_1-i_2,a_3}(G(p)) + P^r_{c-i_1-i_2,a_3}(G(p))\right) \, dp. \tag{23}$$

The number of recidivisms due to an inmate with priority $p < \theta_S$ on pretrial release is a random sum of independent random variables, where each random variable is geometric $(0, 1, 2, \ldots)$ with mean $\eta e^{\gamma p}/r$ as in the case with inmates with priority $p \in [\theta_S, \theta_R]$, and the number of variables represents the number of times the inmate starts the entire process from the beginning, which is

geometric $(1,2,\ldots)$ with mean $(\eta e^{\gamma p}/s)+1$. Hence, the crime rate due to inmates with priority $p<\theta_S$ on pretrial release is

$$\lambda\int_0^{\theta_S}\left(\frac{\eta e^{\gamma p}}{r}\right)\left(\frac{\eta e^{\gamma p}}{s}+1\right)\,dp=\frac{\lambda\eta}{r\gamma}\left[\frac{\eta}{2s}(e^{2\gamma\theta_S}-1)+(e^{\gamma\theta_S}-1)\right]. \tag{24}$$

Using the same reasoning as in (21), we find that the crime rate due to inmates with priority $p<\theta_S$ on supervision is

$$\lambda\int_0^{\theta_S}\frac{\eta e^{\gamma p}}{s}\,dp=\frac{\lambda\eta}{s\gamma}(e^{\gamma\theta_S}-1). \tag{25}$$

The total crime rate in the case $\theta_R\geq\theta_S$ is approximated by the sum of (18)-(21) and (23)-(25).

### 3.3. The Case $\theta_R<\theta_S$

When $\theta_R<\theta_S$, there are again three process flows (Fig. 2), where inmates with priority $p\in(\theta_S,1]$ spend all of their pretrial and post-sentence time in jail, inmates with $p\in[\theta_R,\theta_S]$ receive a split sentence but spend all of their pretrial time in jail, and inmates with $p\in[0,\theta_R)$ receive both pretrial release and a split sentence.

**Mean Jail Population.** We retain the same notation (e.g., $a_i, Q_i$) as in §3.2. As can be inferred from Fig. 2, the results for $Q_1$ and $Q_3$ are identical to those in (11)-(13) and (15)-(16), except that $\theta_R$ and $\theta_S$ are swapped, and hence we now have

$$a_1=\frac{\lambda(1-\theta_S)}{\mu_{12}} \tag{26}$$

and

$$a_3=\frac{\lambda}{\mu_3}\Big(\frac{\eta}{s\gamma}(e^{\gamma\theta_R}-1)+\theta_R\Big). \tag{27}$$

The analysis of $Q_2$ is different than in §3.2 because these inmates undergo supervision rather than pretrial release (Fig. 2). Define the mean service time at $Q_2$ by

$$\mu_{13}^{-1}=\mu_1^{-1}+\mu_3^{-1}. \tag{28}$$

Repeating our analysis leading to (15), we see that the offered load associated with $Q_2$ is

$$\begin{aligned}a_2&=\frac{1}{\mu_{13}}\int_{\theta_R}^{\theta_S}\frac{\eta e^{\gamma p}+s}{s}\lambda\,dp,\\&=\frac{\lambda}{\mu_{13}}\Big(\frac{\eta}{s\gamma}(e^{\gamma\theta_S}-e^{\gamma\theta_R})+\theta_S-\theta_R\Big).\end{aligned} \tag{29}$$

Once again conditioning on $Q_1 = i$ and assuming that $Q_2$ is the number of customers in an Erlang loss system with offered load $a_2$, we approximate the mean queue length for the second process flow by equation (14).

As before, we can approximate $E[Q]$, the mean total jail population, by equation (17).

**Crime Rate.** The crime rate of inmates with priority $p > \theta_S$ is identical to the crime rate of inmates with priority $p > \theta_R$ in §3.2, but with $\theta_S$ replacing $\theta_R$ in (18)-(19). Hence, the crime rate of inmates with priority $p > \theta_S$ due to rejections is

$$\lambda(1-\theta_S)\int_{\theta_S}^{1}\left[\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}+\left(1-\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}\right)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}\right]P^r_{c,a_1}\left(\frac{p-\theta_S}{1-\theta_S}\right)\,dp, \tag{30}$$

and the crime rate of inmates with priority $p > \theta_S$ due to ejections is

$$\lambda(1-\theta_S)\int_{\theta_S}^{1}\left[\frac{\mu_1^{-1}}{\mu_{12}^{-1}}\left(\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}+\left(1-\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}\right)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}\right)+\frac{\mu_2^{-1}}{\mu_{12}^{-1}}\left(\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}\right)\right]P^e_{c,a_1}\left(\frac{p-\theta_S}{1-\theta_S}\right)\,dp. \tag{31}$$

Inmates with priority $p \in [\theta_R, \theta_S]$ can recidivate due to ejection, rejection or supervision. Following the same arguments as in §3.2, inmates with priority $p < \theta_S$ entering jail have a CDF of $G(\cdot)$ given in (22). If an inmate with priority $p \in [\theta_R, \theta_S]$ is rejected, the time that he is exposed to recidivism is the sum of two exponentials with rates $\mu_1$ and $\mu_3$ (Fig. 2), and hence the crime rate due to rejected inmates with priority $p \in [\theta_R, \theta_S]$ is approximately

$$\lambda(\theta_S-\theta_R)\sum_{i=0}^{c}\frac{a_1^i/i!}{\sum_{k=0}^{c}a_1^k/k!}\int_{\theta_R}^{\theta_S}\left[\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}+\left(1-\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}\right)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_3}\right]P^r_{c-i,a_2}\left(G(p)\right)\,dp. \tag{32}$$

If an inmate with priority $p \in [\theta_R, \theta_S]$ is ejected then with probability $\mu_1^{-1}/\mu_{13}^{-1}$ he is exposed to recidivism for a length of time that is the sum of two exponentials, and with probability $\mu_3^{-1}/\mu_{13}^{-1}$ he is exposed to recidivism for an exponential amount of time with rate $\mu_3$. Hence, the crime rate due to ejected inmates with priority $p \in [\theta_R, \theta_S]$ is approximately

$$\lambda(\theta_S-\theta_R)\sum_{i=0}^{c}\frac{a_1^i/i!}{\sum_{k=0}^{c}a_1^k/k!}\int_{\theta_R}^{\theta_S}\left[\frac{\mu_1^{-1}}{\mu_{13}^{-1}}\left(\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}+\left(1-\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_1}\right)\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_3}\right)+\frac{\mu_3^{-1}}{\mu_{13}^{-1}}\left(\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_3}\right)\right]P^e_{c-i,a_2}\left(G(p)\right)\,dp. \tag{33}$$

We see from Fig. 2 that inmates with priority $p \in [\theta_R, \theta_S]$ recidivate while on supervision a geometric number of times $(0, 1, 2, \ldots)$ with mean $\eta e^{\gamma p/s}$. Hence, the crime rate due to inmates with priority $p \in [\theta_R, \theta_S]$ on supervision is

$$\lambda \int_{\theta_R}^{\theta_S} \frac{\eta e^{\gamma p}}{s}\, dp = \frac{\lambda\eta}{s\gamma}(e^{\gamma\theta_S} - e^{\gamma\theta_R}). \tag{34}$$

Inmates with priority $p < \theta_R$ can recidivate in four ways: ejection, rejection, pretrial release or supervision, and the results are identical to those in (22)-(25), but with $\theta_S$ replaced by $\theta_R$. Hence, the CDF of priority levels for inmates with $p < \theta_R$ is

$$H(p) = \frac{\frac{\eta}{s\gamma}e^{\gamma p} + p}{\frac{\eta}{s\gamma}e^{\gamma\theta_R} + \theta_R},$$

the crime rate due to ejected and rejected inmates with $p < \theta_S$ is approximately

$$a_3\mu_3 \sum_{\substack{i_1=0 \\ i_1+i_2\le c}}^{c}\sum_{i_2=0}^{c} \frac{(a_1^{i_1}/i_1!)(a_2^{i_2}/i_2!)}{\sum_{\substack{j_1=0 \\ j_1+j_2\le c}}^{c}\sum_{j_2=0}^{c}(a_1^{j_1}/j_1!)(a_2^{j_2}/j_2!)} \int_0^{\theta_R} \frac{\eta e^{\gamma p}}{\eta e^{\gamma p} + \mu_3} \left(P^e_{c-i_1-i_2,a_3}(H(p)) + P^r_{c-i_1-i_2,a_3}(H(p))\right) dp, \tag{35}$$

the crime rate due to inmates with priority $p < \theta_S$ on pretrial release is

$$\frac{\lambda\eta}{r\gamma}\left[\frac{\eta}{2s}(e^{2\gamma\theta_R} - 1) + (e^{\gamma\theta_R} - 1)\right], \tag{36}$$

and the crime rate due to inmates with priority $p < \theta_S$ on supervision is

$$\frac{\lambda\eta}{s\gamma}(e^{\gamma\theta_R} - 1). \tag{37}$$

The total crime rate in this case $\theta_R < \theta_S$ is approximated by the sum of (30)-(37).

### 3.4. A Numerical Example

In this subsection, we assess the accuracy of our approach and illustrate how the model can be used. The base-case values of our parameters appear in Table 1 and are derived in §1 of the Online Supplement, based on information in Usta and Wein (2015). We perform a discrete-event simulation to compute the true crime rate and the true mean jail population for 121 different scenarios of

| Parameter | Description | Value |
|---|---|---|
| $\lambda$ | arrival rate | 113.8/day |
| $r^{-1}$ | mean time on pretrial release | 155.0 days |
| $\mu_1^{-1}$ | mean time in pretrial detention | 27.1 days |
| $\mu_2^{-1}$ | mean full post-sentence jail term | 144.3 days |
| $\mu_3^{-1}$ | mean split post-sentence jail term | 72.15 days |
| $s^{-1}$ | mean time on supervision | 72.15 days |
| $c$ | number of jail beds | 19,000 |
| $\eta$ | baseline hazard rate | $3.79 \times 10^{-4}$/day |
| $\gamma$ | regression parameter for priority level | 1.6517 |

**Table 1** The parameters, which are depicted in Fig. 1, and their values, which are derived in §1 of the Online Supplement.

$(\theta_R, \theta_S)$, using the values $(\theta_R, \theta_S) \in \{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}^2$. The simulation details are in §2 of the Online Supplement. For a given scenario, we compute the absolute value of the relative crime rate error,

$$\frac{|\text{simulated crime rate } - \text{approximate crime rate}|}{\text{simulated crime rate}} \times 100\%.$$

We summarize the results by averaging this relative error over all 121 scenarios. We take an analogous approach for the mean jail population.

Before showing the results, we discuss two computational issues related to the analytical results in §3.2-3.3. First, to avoid the need to calculate large factorials, we apply the approximation, motivated by the Central Limit Theorem for a Poisson random variable $A$,

$$P(A=k) \approx \Phi\left(\frac{k-a}{\sqrt{a}} + 0.5\right) - \Phi\left(\frac{k-a}{\sqrt{a}} - 0.5\right) \triangleq g(k;a),$$

to equations (20), (23) and (35). For example, we approximate (20) by

$$\lambda(\theta_R-\theta_S)\sum_{i=1}^{c}\frac{g(i;a_1)}{\sum_{k=0}^{c}g(k;a_1)}\int_{\theta_S}^{\theta_R}\frac{\eta e^{\gamma p}}{\eta e^{\gamma p}+\mu_2}\left(P^e_{c-i,a_2}\left(\frac{p-\theta_S}{\theta_R-\theta_S}\right)+P^r_{c-i,a_2}\left(\frac{p-\theta_S}{\theta_R-\theta_S}\right)\right)\,dp.$$

Second, if we define

$$\beta(p)=\frac{c-a(1-p)}{\sqrt{a(1-p)}}\quad\text{for}\quad p\in[0,1)$$

and $\beta(1)=\infty$, then Brockmeyer, Halstrom and Jensen (1948) proved something that was first observed by Erlang: the blocking probability can be very well approximated by

$$B(c,a(1-p))\approx\frac{\phi(\beta(p))}{\Phi(\beta(p))\sqrt{a(1-p)}} \tag{38}$$

for large values of $c$ (and $a(1-p)$). In our setting, $c=19,000$ jail beds, and so it would be natural to use (38) in our analysis. However, due to the small values in the denominator on the right side of (38) for some values of $p$, we found that it was more numerically stable to directly use equation (2) in our computations. Nonetheless, substituting (38) into (13), (14) and (16), and using the convention that the probability beyond three standard deviations in the standard normal is negligible, we identify ranges of $\theta_R$ and $\theta_S$ where the mean jail population reduces to $E[Q_i]\approx a_i$ for $i=1,2,3$. More specifically, $E[Q_1]\approx a_1$ if $c-a_1>3\sqrt{a_1}$, $E[Q_2]\approx a_2$ if $s\sqrt{a_1}-a_2>3\sqrt{a_2}$, and $E[Q_3]\approx a_3$ if $c-a_1-a_2-a_3>3\sqrt{a_3}$. Substituting in values from Table 1 yields $E[Q_1]\approx a_1$ if $\min\{\theta_R,\theta_S\}>0.047$, $E[Q_2]\approx a_2$ if $\max\{\theta_R,\theta_S\}\in(0.026,0.672)$, and $E[Q_3]\approx a_3$ if $\min\{\theta_R,\theta_S\}<0.672$ or $\max\{\theta_R,\theta_S\}<0.0448$. Because the offered load decreases as $\theta_R$ and $\theta_S$ increase, for values of $\theta_R$ and $\theta_S$ above the upper endpoint of these ranges, the jail system is underloaded and an infinite-server model, where $E[Q_i]=a_i$, becomes more accurate. Consequently, even though the normal approximation may break down outside of these ranges, $E[Q_i]\approx a_i$ not only in these ranges but nearly always, as can be seen in Table 2 of the Online Supplement. Similarly, the crime rates in equations (18)-(19) and (30)-(33) are approximately zero if $\min\{\theta_R,\theta_S\}>0.047$, and results in Table 3 of the Online Supplement show that this is true even when $\min\{\theta_R,\theta_S\}\le 0.047$.

The expressions derived in §3.2-3.3 are very accurate: the average absolute relative error is 0.87% for the total crime rate and 0.17% for the total mean jail population. Moreover, as can

be seen in Tables 2-4 in the Online Supplement, our approximations are quite accurate for all three components of the mean jail population, all nine components of the crime rate, and all 121 scenarios.

We can also use our analysis to generate an optimal crime rate vs. mean jail population tradeoff curve. Let $C$ denote the total crime rate. For a given weight $w \geq 0$, we find the joint pretrial release and split sentencing policy by solving

$$\min_{0 \leq \theta_R, \theta_S \leq 1} \quad C + wE[Q]. \tag{39}$$

The optimal tradeoff curve, which appears in Fig. 3, is derived by solving (39) for many different values of $w$. As expected, the curve generated by simulations and the curve generated by our expressions in §3.2-3.3 are visually indistinguishable. As in Usta and Wein (2015), split-sentencing is more effective than pretrial release at optimizing the crime rate vs. mean jail population tradeoff. More specifically, the optimal solution in both curves in Fig. 3 uses pretrial release only after split sentencing is offered to nearly everyone; i.e., as $w$ increases (i.e., we move up and to the left on the tradeoff curve), the optimal solution $(\theta_R^*, \theta_S^*)$ moves from $(0,0)$ to $\approx (0, 0.95)$ to $\approx (0.05, 1)$ to (1,1). In other words, the optimal solution is usually of the form $(0, \theta_S)$ or $(\theta_R, 1)$ for some $0 \leq \theta_R, \theta_S \leq 1$.

### 3.5. On the Dominance of Split Sentencing Over Pretrial Release

Motivated by the facts that $E[Q_i] \approx a_i$ (Table 2 in the Online Supplement), the crime rates from ejected and rejected customers are approximately zero (Table 3 in the Online Supplement), and the almost complete dominance of split sentencing over pretrial release (Fig. 3), in this subsection we assume $E[Q_i] = a_i$ and the crime rate from ejected and rejected customers are zero, and derive sufficient conditions for the *dominance* of split sentencing over pretrial release (i.e., the inmates who receive pretrial release also receive split sentencing), and the *complete dominance* of split sentencing over pretrial release (i.e., all inmates are given split sentencing before any inmate is given pretrial release). First, we confirm that assuming $E[Q_i] = a_i$ for $i = 1, 2, 3$ and taking the

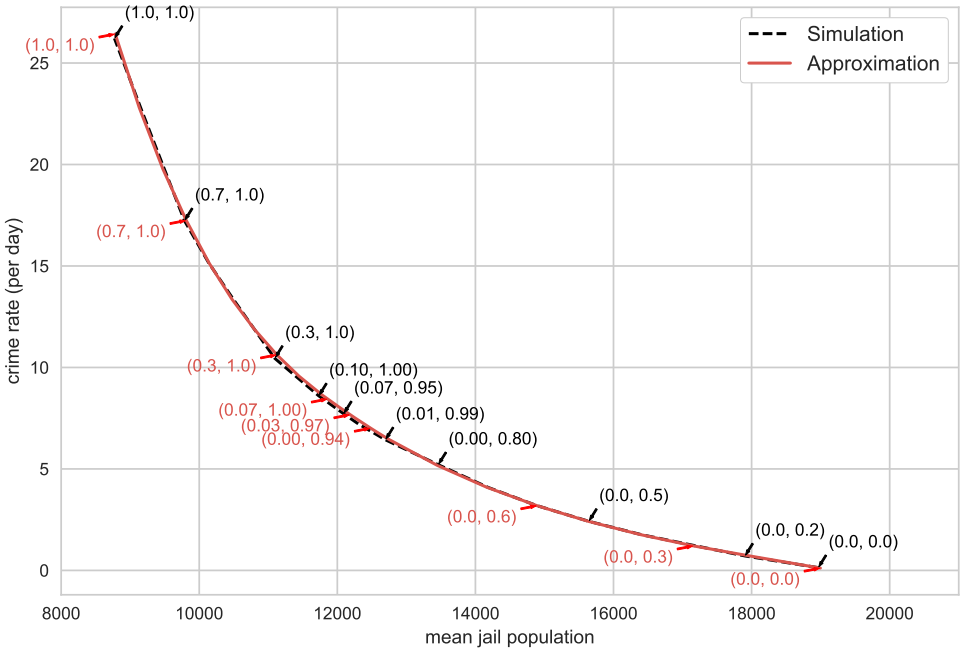


**Figure 3** Tradeoff curves for minimizing the crime rate subject to a constraint on the mean jail population, using both simulation and analytical approximations. The optimal $(\theta_R, \theta_S)$ values appear along various points of the tradeoff curves.

crime rate of ejected and reject inmates to be equal to zero, the optimal tradeoff curve is very similar to the simulated tradeoff curve (Fig. 4 in the Online Supplement).

The following two propositions are proved in §3 of the Online Supplement.

PROPOSITION 1. *(Dominance) Suppose* $E[Q_i] = a_i$ *for* $i = 1, 2, 3$*, so that the crime rates in equations (18)-(20), (23) and (30)-(33) are zero. If*

$$\frac{1}{r} > \frac{1}{s} \tag{40}$$

*and*

$$\frac{1}{\mu_2} > \left(\frac{\eta e^{\gamma}}{s} + 1\right)\left(\frac{1}{\mu_1} + \frac{1}{\mu_3}\right), \tag{41}$$

*then for all* $w > 0$*, the solution to (39) satisfies*

$$\theta_S^* \geq \theta_R^*. \tag{42}$$

PROPOSITION 2. *(Total Dominance) Suppose* $E[Q_i] = a_i$ *for* $i = 1, 2, 3$*, so that the crime rates in equations (18)-(20), (23) and (30)-(33) are zero. If (40), (41) and*

$$\frac{\frac{1}{r}}{\frac{1}{\mu_1}} > e^{\gamma} \frac{\frac{1}{s}}{\frac{1}{\mu_1} + \frac{1}{\mu_2} - (\frac{\eta}{s} + 1)(\frac{1}{\mu_1} + \frac{1}{\mu_3})} \tag{43}$$

*all hold, then for all* $w > 0$*, the solution to (39) satisfies:*

$$\text{if} \quad \theta_S^* < 1 \quad \text{then} \quad \theta_R^* = 0. \tag{44}$$

Solution (42) implies that split sentencing dominates pretrial release: i.e., the inmates who receive pretrial release also receive split sentencing, but the reverse may not hold. Solution (44) implies that split sentencing completely dominates pretrial release; i.e., split sentencing is awarded to all inmates before any inmates receive pretrial release. Condition (40) states that the mean time on pretrial release is longer than the mean time on supervision. The quantity $\frac{\eta e^{\gamma}}{s} + 1$ in (41) is the mean number of times that an inmate with the highest risk level ($p = 1$) that receives split sentencing starts the entire process from the beginning. Thus $\left(\frac{\eta e^{\gamma}}{s} + 1\right)\left(\frac{1}{\mu_1} + \frac{1}{\mu_3}\right)$ is the mean time in jail for a $p = 1$ inmate that receives split sentencing, which is also an upper bound for the mean time in jail of an inmate that receives split sentencing. Therefore, condition (41) states that the mean time in jail for any inmate that receives split sentencing is less than the mean full post-sentence jail term.

The numerators on the left and right side of (43) are the extra time an inmate spends outside of jail (i.e., the extra time we are exposed to a risk of recidivism) when we use pretrial release and split sentencing, respectively. The left denominator in (43) is the mean time in pretrial detention and the right denominator is the difference between the mean time in jail with no pretrial release or split sentencing and the mean time in jail of an inmate with the highest risk level that receives a split sentence but no pretrial release. Hence, the left denominator is the mean time in jail saved by pretrial release and the right denominator is a lower bound on the mean time in jail saved by split sentencing. Taken together, condition (44) states that the recidivism exposure time per jail

time saved for pretrial release is greater than $e^{\gamma}$ times the recidivism exposure time per jail time saved by split sentencing.

Substituting parameter values from Table 1 into (40), (41) and (43) yields $155 > 72.15$, $144.3 > 113.4$ and $5.72 > 6.49$, respectively. That is, conditions (40)-(41) are easily satisfied, but condition (43) is violated (although only modestly), and hence split sentencing dominates pretrial release, but does not completely dominate it. This is consistent with Fig. 3, where only a small portion of the optimal tradeoff curve violates condition (44).

## 4. Concluding Remarks

The main contribution of this paper is to introduce and analyze a queueing model that incorporates customer features, which are now available in many service settings, in a natural way: via a proportional hazards model, where each customer has an individualized hazard rate (based on his features) that influences his movement through the queueing network. This leads to a queueing model with a continuum of classes – one for each possible value of $p = \sum_k \beta_k X_k$, where $X_k$ is a customer's value for feature $k$ and $\beta_k$ is the known regression parameter for feature $k$ – and allows for queue management (in our case, routing decisions) at the individualized level.

Our model is motivated by a jail system (Usta and Wein 2015), where the proportional hazards model measures the time until a released inmate commits a crime and the features include the inmate's demographic data and criminal history (Yang, Wong and Coid 2010). In our jail model in §3, we consider a family of two-parameter policies that sets thresholds for whether inmates are offered pretrial release and/or a split sentence based on their individual features, with the goal of optimizing the tradeoff between public safety and jail congestion.

Our model offers a starting point to incorporate individualized customer management in queueing models of other service systems. The most prominent example is a hospital, where a proportional hazards model could measure the time until death (Bastiampillal, Sharfstein and Allison 2016) or readmission if a patient is released (our continuous-class model could also incorporate the simpler setting where a logistic regression model is used to predict whether or not readmission will occur

(e.g., Anderson, Golden, Jank and Wasil (2012), Bayati, Braverman, Gillam *et al.* (2014)), and the system manager decides when to release each patient depending on his features. A proportional hazard model could also represent the service time (i.e., length of stay) in models that manage the patient flow through an intensive care unit of a hospital (e.g., KC and Terwiescsh (2012), Hu, Chan, Zubizarreta *et al.* (2016)). A key difference between the jail setting and the hospital setting is that patient health changes on a faster time scale than an individual's recidivism risk. Hence, one might want to generalize the model so that the parameter $p$ changes randomly over time for each patient. The analysis of such a queueing system would be much more challenging, perhaps requiring the use of a measure-valued process (e.g., Doytchinov, Lehoczky and Shreve (2001), Gromoll (2004)) that tracks the evolution of each person's risk level.

A second example is a call center, where the proportional hazards model dictates each caller's individualized time until abandonment, and customer features would be used to prioritize callers in the queue. A continuous-class queueing model could also have $p = \sum_k \beta_k X_k$ represent the expected revenue from a customer's call, which could be used for prioritization. For a call center, it would be more natural to consider a many-server queue in the Halfin-Whitt regime (Halfin and Whitt 1981) rather than an Erlang B model; see Mandelbaum and Momcilovic (2017) for a time-varying many-server fluid model in this spirit.

As noted earlier, our jail model in §3 is a simplified version of the simulation model in Usta and Wein (2015). Our goal here is not to analyze the most complicated jail model possible, but rather to introduce a novel continuous-class queueing model as clearly as possible while including only the most salient features. In particular, the model in Usta and Wein (2015) incorporates (i) the time delay between the crime and the arraignment, (ii) a logistic regression model to allow for the risk-based possibility that an inmate fails to appear in court, (iii) the possibility that some cases end in dismissal (i.e., innocence) or probation rather than incarceration, (iv) non-exponential service times, (v) the dependence of the length of the post-sentence jail term on whether an inmate is released or detained prior to case disposition, (vi) two types of inmates, depending upon whether

their current crime is a felony or a non-felony, (vii) non-Poisson arrivals, (viii) a different survival curve that fits the data better than a proportional hazards model, and (ix) the return to jail of ejected and rejected inmates who recidivate.

Extensions (i)-(v) would be tedious, but somewhat straightforward. The time delay in (i) can be addressed by inserting an additional infinite-server queue into the jail model. Regarding (ii), the logistic regression model for the failure to appear in court would be easier to analyze than the proportional hazards model (the former affects only routing, while the latter affects timing and routing), although analyzing them simultaneously would be messy. Case dismissal in (iii) could be incorporated by adding a random exiting branch if the dismissal probability was independent of $p$; otherwise, it could be handled as in (ii). For non-exponential service times in (iv), one could use the method of stages to generalize beyond our sum of two exponentials. Extension (v) could be handled by adding classes and making additional use of the nearly completely decomposable approach (Courtois 1977).

Extensions (vi)-(ix) would be much more challenging. Incorporating felons and non-felons in (vi) would be unwieldy because felons differ from non-felons in their arrival rates, service rates, risk levels and recidivism rates (Usta and Wein 2015). Non-Poisson arrivals in (vii) have been studied in simpler blocking systems by using heavy-traffic (Whitt 1984) and diffusion (Srikant and Whitt 1996) approximations. Regarding (viii), Usta and Wein (2015) find that a split lognormal model with heteroskedasticity provides the best fit to the recidivism data, but this choice would likely make the queueing model intractable.

Although allowing the return of ejected and rejected customers who recidivate in (ix) would be very difficult, we use a back-of-the-envelope calculation for a somewhat simpler model, which suggests that ignoring these retrials has a negligible effect, at least in our jail setting. We focus on the worst case, $(\theta_R, \theta_S) = (0, 0)$, which generates the most ejected and rejected inmates. We also consider the worst case where all ejected and rejected inmates eventually recidivate, albeit after a long exponential delay that is independent of the priority level $p$ This yields a continuous class

$M/G/c/c$ queue with arrival rate $\lambda$, mean service time $\mu_1^{-1}+\mu_2^{-1}$, $c$ servers and priority levels $p \sim U[0,1]$. Then, ignoring the fact that service times are the sum of two exponentials rather than exponential, Cohen's equation (equation (1) in Avram, Janssen, and Van Leeuwaarden (2013)) implies that the re-entry rate is the unique root $\Omega$ to

$$\Omega = (\lambda+\Omega)B\Big(c, \frac{\lambda+\Omega}{\mu_{12}}\Big). \tag{45}$$

Substituting $\lambda$, $c$ and $\mu_{12}$ from Table 1 into (45) and solving yields $\Omega/\lambda = 0.014$, which would have a minor effect in the worst case. For cases where even a few percent of inmates receive pretrial release or a split sentence, the impact of ignoring these retrials vanishes.

Our computational results focus on the parameter values based on the LA County jail system (Usta and Wein 2015), and the system is underloaded for most values of the thresholds, $(\theta_R, \theta_S)$. We have not attempted to assess the accuracy of our approach in §3.2-3.3 under more heavily-loaded conditions.

Another possible generalization is to assume that the regression parameters are unknown and to jointly estimate the regression parameters and manage the queue, which requires addressing the exploration-exploitation tradeoff inherent in such a problem. However, in the jail setting and perhaps the hospital setting, this approach raises ethical issues regarding whether to release the riskiest customers in order to learn their recidivism parameters.

Because our jail model is a simplified version of the simulation model in Usta and Wein (2015), any apparent insights about jail management from the computational results in §3 are overridden by the insights in Usta and Wein (2015) and Wein and Usta (2015). Nonetheless, our analysis does reveal several new insights. By the analysis in §2.3, in a continuous-class Erlang loss model, a customer arriving to a system with all servers busy is much more likely to cause an ejection of a customer currently in queue than to be rejected himself, unless the offered load is much larger than the number of servers. Propositions 1 and 2 give a set of nonobvious but intuitive sufficient conditions for the dominance and complete dominance of split sentencing over pretrial release; these conditions can be viewed as a refinement of the cruder intuition offered in Usta and Wein (2015) and Wein and Usta (2015).

## References

Amin K, Rostamizadeh A, Syed U (2014). Repeated contextual auctions with strategic buyers. *Advances in Neural Information Processing Systems* 27:622-630.

Anderson D, Golden B, Jank W, Wasil E (2012) The impact of hospital utilization on patient readmission rate. *Health Care Management Science* 15:29-36.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learn. Res.* 3:397-422.

Avram, F., Janssen, A. J. E. M., and Van Leeuwaarden, J. S. H. (2013). Loss systems with slow retrials in the HalfinWhitt regime. *Advances in Applied Probability* 45:274-294.

Ban, G-Y, Gallien J, Mersereau A (2017) Dynamic procurement of new products with covariate information: the residual tree method. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2926028.

Ban G-Y, Keskin NB (2017) Personalized dynamic pricing with machine learning. London Business School, London, UK. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972985.

Ban G-Y, Rudin C (2016) The big data newsvendor: practical insights from machine learning. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2559116.

Bastiampillai T, Sharfstein SS, Allison S (2016) Increase in US suicide rates and the critical decline in psychiatric beds. *Journal of the American Medical Association* 316:2591-2592.

Bayati M, Braverman M, Gillam M, Mack KM, Ruiz G, Smith MS, Horvitz E (2014) Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLOS ONE* 9:e109264.

Brockmeyer E, Halstrom HL, Jensen A (1948) *The life and works of A. K. Erlang*, pp. 277 (Academy of Technical Sciences, Copenhagen).

Burke PJ (1962) Priority traffic with at most one queuing class. *Operations Research* 10:567-569.

Cohen MC, Lobel I, Leme RP (2016) Feature-based dynamic pricing. Available at *SSRN*.

Cooper RB (1981) *Introduction to Queueing Theory*, 2nd edition (North Holland, New York).

Courtois PJ (1977) *Decomposability, Queueing and Computer Systems Applications* (Academic Press, New York).

Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* 34:187-220.

Dotchinov B, Lehoczky J, Shreve S (2001) Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Annals of Applied Probability* 11:332-378.

Fischer MJ (1980) Priority loss systems - unequal holding times. *AIIE Transactions* 12:47-53.

Gromoll HC (2004) Diffusion approximation for a processor sharing queue in heavy traffic. *Annals of Applied Probability* 14:555-611.

Halfin S, Whitt W(1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research* 39:567-588.

Helly W (1961) Two doctrines for the handling of two-priority traffic by a group of $N$ servers. *Operations Research* 10:268-269.

Hu W, Chan CW, Zubizarreta JR, Escobar GJ (2016) An examination of early transfers to the ICU based on a physiologic risk score. To appear, *Manufacturing & Services Operations Management.*

Jagerman DL (1974) Some properties of the Erlang Loss function. *Bell System Technical Journal* 53:525-551.

KC DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14:50-65.

Kelly FP (1979) *Reversibility and Stochastic Networks* (Wiley, New York).

Langford J, Zhang T (2007) The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems* 20:1096-1103.

Mandelbaum A, Momčilović P (2017) Personalized queues: the customer view, via a fluid model of serving least-patient first. *Queueing Systems* 87:23-53.

Mendelson H (1985) Pricing computer services: Queueing effects. *Communications of the ACM* 28:312-321.

Northpointe, Inc. (2015) Practitioner's guide to COMPAS Core. Available at http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf.

Petersilia J (2014) California prison downsizing and its impact on local criminal justice systems. *Harvard Law & Policy Review* 8:327-357.

Qiang S, Bayati M (2016) Dynamic pricing with demand covariates. Available at www.ssrn.com/abstract=2765257.

Shaked M, Shanthikumar JG (2007) *Stochastic Orders* (Springer Science & Business Media, New York).

Srikant R, Whitt W (1996) Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 6:7-52.

Usta M, Wein LM (2015) Assessing risk-based policies for pretrial release and split sentencing in Los Angeles County jails. *PLoS ONE* 10(12): e0144967.

Wein LM, Usta M (2015) One way to reduce jail populations. *New York Times*, October 23, page A31.

Whitt W (1984) Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal* 63:689-708.

Yang M, Wong SCP, Coid J (2010) The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin* 136:740-767.

Zenios SA, Chertow GM, Wein LM (2000) Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research* 48:549-569.

# Online Supplement

We estimate the parameter values in §1, provide simulation details in §2 and derive sufficient conditions for the dominance and complete dominance of split sentencing over pretrial release in §3. Fig. 4 is discussed in the main text.

## 1 Parameter Estimation

All of our parameter values are based on information in Usta and Wein (2015). From Table 3 in Usta and Wein (2015), we set $c = 19,000$ jail beds and assume that 55.8% of arriving inmates are non-felons and 44.2% are felons. Referring to the first row of Table 4 in Usta and Wein (2015), we compute the mean time on pretrial release as a weighted average of the values for felons and non-felons:

$$\begin{aligned} r^{-1} &= 0.442e^{5.13+0.47^2/2} + 0.558(1.07)(119.78), \\ &= 155.0 \text{ days.} \end{aligned}$$

The mean time in pretrial custody is also computed from the first row of Table 4 in Usta and Wein (2015),

$$\begin{aligned} \mu_1^{-1} &= 0.442(0.67)(76.81) + 0.558(0.46)(16.80), \\ &= 27.1 \text{ days.} \end{aligned}$$

We estimate the post-sentence jail term from inmates who were in pretrial custody in Table 4 in Usta and Wein (2015), which yields

$$\begin{aligned} \mu_2^{-1} &= 0.442e^{2.064+0.628^2/2}\Big(\frac{365}{12}\Big) + 0.558(0.397)(77.08), \\ &= 144.3 \text{ days,} \end{aligned}$$

where 365/12 is a conversion from months to days. Because post-sentence jail terms are split roughly evenly between jail time and supervision under a split sentence Usta and Wein

(2015), we set

$$\mu_3^{-1} = s^{-1} = \frac{1}{2}\mu_2^{-1} = 72.15 \text{ days.}$$

The lower right portion of Fig. 2(a) in Usta and Wein (2015) implies that the offered load in the absence of pretrial release and supervision is 19,500, which implies

$$\lambda \left( \frac{1}{\mu_1} + \frac{1}{\mu_2} \right) = 19,500,$$

or $\lambda = 113.8/\text{day}$.

We compute the two proportional hazard parameters, $\eta$ and $\gamma$, from two equations: the probability of recidivism (averaged over inmates of all risk levels) within three years is 0.613 (computed from $\sum_{i=1}^{3}\sum_{j=1}^{3} N_{ij}/N$ in §E in the Supporting Material file of Usta and Wein (2015)), and the risk-based tools for predicting recidivism have an area under the curve (AUC) of the receiver operating characteristic curve equal to 0.7 (Yang, Wong and Coid 2010). In terms of our risk model, the first equation can be expressed as

$$\int_0^1 \left(1 - e^{-3\eta e^{\gamma p}}\right) \, dp = 0.613, \tag{1}$$

where $\eta$ is an annual rate.

The second equation implies that the probability that a three-year recidivist has a higher risk score than a three-year non-recidivist is 0.7. Let $i$ index the non-recidivist and $j$ index the recidivist, and define their random priority levels $p_i$ and $p_j$ to be independent $U[0,1]$ random variables. If we let $T_i$ and $T_j$ be exponential random variables with rates $\eta e^{\gamma p_i}$ and $\eta e^{\gamma p_j}$, which are assumed to be conditionally independent given $p_i$ and $p_j$, then (assuming $\gamma > 0$)

$$\text{AUC} = P(p_i < p_j | T_i > 3 \geq T_j).$$

Bayes' rule implies that

$$\text{AUC} = \frac{P(T_i > 3 \geq T_j | p_i < p_j) P(p_i < p_j)}{P(T_i > 3 \geq T_j)}. \tag{2}$$

We now compute the three terms on the right side of (2), beginning with

$$P(p_i < p_j) = \frac{1}{2}. \tag{3}$$

Conditioning on $(p_i, p_j)$ and taking expectations, we get

$$\begin{aligned} P(T_i > 3 \geq T_j | p_i < p_j) &= E[P(T_i > 3 \geq T_j | p_i < p_j, p_i, p_j)], \\ &= E[P(T_i > 3 \geq T_j | p_i, p_j) 2 I_{\{p_i < p_j\}}], \\ &= 2E[e^{-3\eta e^{\gamma p_i}}(1 - e^{-3\eta e^{\gamma p_j}}) I_{\{p_i < p_j\}}], \end{aligned} \tag{4}$$

where $I_{\{x\}}$ is the indicator function of the event $x$. Similarly, we have that

$$P(T_i > 3 \geq T_j) = E[e^{-3\eta e^{\gamma p_i}}(1 - e^{-3\eta\gamma p_j})]. \tag{5}$$

Substituting (3)-(5) into (2) yields

$$\text{AUC} = \frac{E[e^{-3\eta e^{\gamma p_i}}(1 - e^{-3\eta e^{\gamma p_j}}) I_{\{p_i < p_j\}}]}{E[e^{-3\eta e^{\gamma p_i}}(1 - e^{-3\eta e^{\gamma p_j}})]}. \tag{6}$$

Expressing the expectations in (6) as double integrals, we obtain our second equation:

$$\frac{\int_0^1 \int_0^{p_j} e^{-3\eta e^{\gamma p_i}}(1 - e^{-3\eta e^{\gamma p_j}}) \, dp_i \, dp_j}{\int_0^1 \int_0^1 e^{-3\eta e^{\gamma p_i}}(1 - e^{-3\eta e^{\gamma p_j}}) \, dp_i \, dp_j} = 0.7. \tag{7}$$

We solve for the two parameter values that minimize the sum of absolute errors in the solution of equations (1) and (7), which yields $\eta = 3.79 \times 10^{-4}$/day and $\gamma = 1.6517$, with a total absolute error of $5.3 \times 10^{-4}$.

# 2 Simulation Details

For each value of $(\theta_R, \theta_S) \in \{0.0, 0.1, \ldots, 0.9, 1.0\}^2$, we run a simulation for 10 years using the parameter values in Table 1 of the main text. At time 0 of each simulation, no inmates are in pretrial release or on supervision during a split sentence, and the jail has $c$ inmates with iid $U[0, 1]$ priority levels. For all class 1 inmates and for class 2 inmates when $\theta_R < \theta_S$,

the priority level does not uniquely determine the residual service time (the service times are the sum of two exponential random variables). In these situations, each inmate is randomly assigned to be in either pre-sentencing or post-sentencing with equal probability. To reduce the bias introduced by these initial conditions, we discard the first two years of the simulation, and use the remaining eight years to compute the crime rate and mean jail population.

Although we perform this procedure for the 121 scenarios $(\theta_R, \theta_S) \in \{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}^2$, in the interest of comprehensibility and brevity, Tables 2-4 only display the results for the 36 scenarios $(\theta_R, \theta_S) \in \{0.0, 0.2, \ldots, 1.0\}^2$.

# 3 Sufficient Conditions for the Dominance and Complete Dominance of Split Sentencing Over Pretrial Release

We formulate the optimization problem in §3.1, prove Proposition 1 of the main text in §3.2 and prove Proposition 2 of the main text in §3.3.

## 3.1 Problem description

Throughout this section, we assume $E[Q_i] = a_i$ for $i = 1, 2, 3$ and the crime rates in equations in equations (18)-(19) and (30)-(33) in the main text are equal to zero. When $\theta_R \geq \theta_S$, the mean jail population is

$$a_1 + a_2 + a_3 = \frac{\lambda}{\mu_{12}} - \frac{\lambda\eta}{\mu_3 s\gamma} - \frac{\lambda\theta_R}{\mu_1} - \frac{\lambda\theta_S}{\mu_2} + \frac{\lambda\theta_S}{\mu_3} + \frac{\lambda\eta}{\mu_3 s\gamma} e^{\gamma\theta_S},$$

where

$$\mu_{12}^{-1} = \mu_1^{-1} + \mu_2^{-1},$$

and the crime rate is

$$\frac{\lambda\eta}{r\gamma}(e^{\gamma\theta_R}-e^{\gamma\theta_S})+\frac{\lambda\eta}{r\gamma}\left[\frac{\eta}{2s}(e^{2\gamma\theta_S}-1)+(e^{\gamma\theta_S}-1)\right]+\frac{\lambda\eta}{s\gamma}(e^{\gamma\theta_S}-1)$$
$$=\frac{\lambda\eta}{\gamma}(\frac{1}{r}e^{\gamma\theta_R}+\frac{\eta}{2sr}e^{2\gamma\theta_S}+\frac{1}{s}e^{\gamma\theta_S})-\frac{\lambda\eta^2}{2sr\gamma}-\frac{\lambda\eta}{r\gamma}-\frac{\lambda\eta}{s\gamma}.$$

When $\theta_R < \theta_S$, the mean jail population is

$$a_1+a_2+a_3=\frac{\lambda}{\mu_{12}}-\frac{\lambda\eta}{\mu_3 s\gamma}-\frac{\lambda\theta_R}{\mu_1}-\frac{\lambda\theta_S}{\mu_2}+\frac{\lambda\theta_S}{\mu_3}+\frac{\lambda\eta}{\mu_3 s\gamma}e^{\gamma\theta_S}+\frac{\lambda\eta}{\mu_1 s\gamma}(e^{\gamma\theta_S}-e^{\gamma\theta_R}),$$

where

$$\mu_{13}^{-1}=\mu_1^{-1}+\mu_3^{-1},$$

and the crime rate is

$$\frac{\lambda\eta}{s\gamma}(e^{\gamma\theta_S}-e^{\gamma\theta_R})+\frac{\lambda\eta}{r\gamma}\left[\frac{\eta}{2s}(e^{2\gamma\theta_R}-1)+(e^{\gamma\theta_R}-1)\right]+\frac{\lambda\eta}{s\gamma}(e^{\gamma\theta_R}-1)$$
$$=\frac{\lambda\eta}{\gamma}(\frac{1}{r}e^{\gamma\theta_R}+\frac{\eta}{2sr}e^{2\gamma\theta_R}+\frac{1}{s}e^{\gamma\theta_S})-\frac{\lambda\eta^2}{2sr\gamma}-\frac{\lambda\eta}{r\gamma}-\frac{\lambda\eta}{s\gamma}.$$

We consider the constrained version of the Lagrangian relaxation problem in (39) of the main text, where we choose the pair $(\theta_R, \theta_S)$ that minimizes the crime rate when the mean jail population is smaller than a given bound $\tilde{L}$.

Consider the optimization problem

$$\begin{aligned}
\underset{\theta_R,\theta_S}{\text{minimize}} \quad & \frac{\lambda\eta}{\gamma}(\frac{1}{r}e^{\gamma\theta_R}+\frac{\eta}{2sr}e^{2\gamma\theta_S}+\frac{1}{s}e^{\gamma\theta_S})-\frac{\lambda\eta^2}{2sr\gamma}-\frac{\lambda\eta}{r\gamma}-\frac{\lambda\eta}{s\gamma} \\
\text{subject to} \quad & \frac{\lambda}{\mu_{12}}-\frac{\lambda\eta}{\mu_3 s\gamma}-\frac{\lambda\theta_R}{\mu_1}-\frac{\lambda\theta_S}{\mu_2}+\frac{\lambda\theta_S}{\mu_3}+\frac{\lambda\eta}{\mu_3 s\gamma}e^{\gamma\theta_S}\le\tilde{L}, \\
& 0\le\theta_S\le\theta_R\le 1,
\end{aligned} \tag{8}$$

with optimal solution $(\theta_R^{*1}, \theta_S^{*1})$, and another optimization problem

$$\begin{aligned}
\underset{\theta_R,\theta_S}{\text{minimize}} \quad & \frac{\lambda\eta}{\gamma}(\frac{1}{r}e^{\gamma\theta_R}+\frac{\eta}{2sr}e^{2\gamma\theta_R}+\frac{1}{s}e^{\gamma\theta_S})-\frac{\lambda\eta^2}{2sr\gamma}-\frac{\lambda\eta}{r\gamma}-\frac{\lambda\eta}{s\gamma} \\
\text{subject to} \quad & \frac{\lambda}{\mu_{12}}-\frac{\lambda\eta}{\mu_3 s\gamma}-\frac{\lambda\theta_R}{\mu_1}-\frac{\lambda\theta_S}{\mu_2}+\frac{\lambda\theta_S}{\mu_3}+\frac{\lambda\eta}{\mu_3 s\gamma}e^{\gamma\theta_S}+\frac{\lambda\eta}{\mu_1 s\gamma}(e^{\gamma\theta_S}-e^{\gamma\theta_R})\le\tilde{L}, \\
& 0\le\theta_R\le\theta_S\le\tilde{1},
\end{aligned} \tag{9}$$

with optimal solution $(\theta_R^{*2}, \theta_S^{*2})$.

We know that the optimal solution, $(\theta_R^*, \theta_S^*)$, to our constrained problem is $(\theta_R^{*1}, \theta_S^{*1})$ if the optimal value in (8) is smaller than the optimal value in (9), and is $(\theta_R^{*2}, \theta_S^{*2})$ otherwise.

Equivalently, we only need to consider the following two problems:

$$\begin{aligned} \underset{\theta_R,\theta_S}{\text{minimize}} \quad & \frac{1}{r}e^{\gamma\theta_R} + \frac{\eta}{2sr}e^{2\gamma\theta_S} + \frac{1}{s}e^{\gamma\theta_S} \\ \text{subject to} \quad & -\frac{\lambda\theta_R}{\mu_1} - \frac{\lambda\theta_S}{\mu_2} + \frac{\lambda\theta_S}{\mu_3} + \frac{\lambda\eta}{\mu_3 s\gamma}e^{\gamma\theta_S} \leq L, \\ & 0 \leq \theta_S \leq \theta_R \leq 1, \end{aligned} \tag{10}$$

and

$$\begin{aligned} \underset{\theta_R,\theta_S}{\text{minimize}} \quad & \frac{1}{r}e^{\gamma\theta_R} + \frac{\eta}{2sr}e^{2\gamma\theta_R} + \frac{1}{s}e^{\gamma\theta_S} \\ \text{subject to} \quad & -\frac{\lambda\theta_R}{\mu_1} - \frac{\lambda\theta_S}{\mu_2} + \frac{\lambda\theta_S}{\mu_3} + \frac{\lambda\eta}{\mu_3 s\gamma}e^{\gamma\theta_S} + \frac{\lambda\eta}{\mu_1 s\gamma}(e^{\gamma\theta_S} - e^{\gamma\theta_R}) \leq L, \\ & 0 \leq \theta_R \leq \theta_S \leq 1. \end{aligned} \tag{11}$$

When $L = \tilde{L} - \frac{\lambda}{\mu_{12}} + \frac{\lambda\eta}{\mu_3 s\gamma}$, the optimal solution of (10), $(\theta_R^{*3}, \theta_S^{*3})$, is the same as that of (8), $(\theta_R^{*1}, \theta_S^{*1})$, and the optimal solution of (11), $(\theta_R^{*4}, \theta_S^{*4})$, is the same as that of (9), $(\theta_R^{*2}, \theta_S^{*2})$. Also $(\theta_R^*, \theta_S^*) = (\theta_R^{*3}, \theta_S^{*3})$ if the optimal value of (10) is smaller than that of (11), and $(\theta_R^*, \theta_S^*) = (\theta_R^{*4}, \theta_S^{*4})$ otherwise.

## 3.2 Proof of Proposition 1

Define the constants $A = \frac{1}{r}$, $B = \frac{\eta}{2sr}$, $C = \frac{1}{s}$, $D = \frac{\lambda}{\mu_1}$, $E = \frac{\lambda}{\mu_2} - \frac{\lambda}{\mu_3}$, $F = \frac{\lambda\eta}{\mu_3 s\gamma}$, $G = \frac{\lambda\eta}{s\gamma}\left(\frac{1}{\mu_1} + \frac{1}{\mu_3}\right)$ and $H = \frac{\lambda\eta}{\mu_1 s\gamma}$. Note that $F = G - H$ and $A, B, C, D, E, F, G, H > 0$ (by definition of a split sentence, the mean full post-sentence jail term, $\mu_2^{-1}$, is greater than the mean split post-sentence jail term, $\mu_3^{-1}$). Problems (10) and (11) become

$$\begin{aligned} \underset{\theta_R,\theta_S}{\text{minimize}} \quad & Ae^{\gamma\theta_R} + Be^{2\gamma\theta_S} + Ce^{\gamma\theta_S} \\ \text{subject to} \quad & -D\theta_R - E\theta_S + Fe^{\gamma\theta_S} \leq L, \\ & 0 \leq \theta_S \leq \theta_R \leq 1, \end{aligned} \tag{12}$$

and

$$\begin{aligned} \underset{\theta_R,\theta_S}{\text{minimize}} \quad & Ae^{\gamma\theta_R} + Be^{2\gamma\theta_R} + Ce^{\gamma\theta_S} \\ \text{subject to} \quad & -D\theta_R - E\theta_S + Ge^{\gamma\theta_S} - He^{\gamma\theta_R} \leq L, \\ & 0 \leq \theta_R \leq \theta_S \leq 1. \end{aligned} \tag{13}$$

Proposition 1 in the main text is equivalent to: if $A > C$ and $D - E + G\gamma e^{\gamma} < 0$, the optimal value of (13) is less than or equal to the optimal value of (12).

Suppose ($\theta_R = x$, $\theta_S = y$) is the optimal solution of (12). We will show that ($\theta_R = y$, $\theta_S = x$) is a feasible point of (13) and achieves a smaller value of the objective function.

(i) The policy ($\theta_R = y$, $\theta_S = x$) satisfies the second constraint in (13) because, by assumption, ($\theta_R = x, \theta_S = y$) satisfies the second constraint in (12).

(ii) The policy ($\theta_R = y$, $\theta_S = x$) satisfies the first constraint in (13) because

$$\begin{aligned} -Dy - Ex + Ge^{\gamma x} - He^{\gamma y} &= -Dx - D(y-x) - Ey - E(x-y) + Ge^{\gamma x} + Fe^{\gamma y} - (F+H)e^{\gamma y}, \\ &= -Dx - Ey + Fe^{\gamma y} + (D-E)(x-y) + G(e^{\gamma x} - e^{\gamma y}), \\ &\leq -Dx - Ey + Fe^{\gamma y} + (D - E + G\gamma e^{\gamma})(x-y), \qquad (14) \\ &\leq -Dx - Ey + Fe^{\gamma y}, \\ &\leq L. \end{aligned}$$

Inequality (14) holds because $e^{\gamma x} - e^{\gamma y} = (x-y)\gamma e^{\gamma z}$, where $y < z < x$, by the intermediate value theorem, which implies that $e^{\gamma x} - e^{\gamma y} \leq (x-y)\gamma e^{\gamma}$.

(iii) The value of the objective function in (13) is less than or equal to the value of the objective function in (12):

$$\begin{aligned} Ae^{\gamma y} + Be^{2\gamma y} + Ce^{\gamma x} &= Ae^{\gamma x} + Be^{2\gamma y} + Ce^{\gamma y} + Ae^{\gamma y} - Ae^{\gamma x} + Ce^{\gamma x} - Ce^{\gamma y}, \\ &= Ae^{\gamma x} + Be^{2\gamma y} + Ce^{\gamma y} + (C-A)(e^{\gamma x} - e^{\gamma y}), \\ &\leq Ae^{\gamma x} + Be^{2\gamma y} + Ce^{\gamma y}. \end{aligned}$$

Because the optimal value of (13) is no greater than the value under the feasible policy

$(\theta_R = y,\, \theta_S = x)$, the optimal value of (13) will always be less than or equal to the optimal value of (12). Thus, the optimal solution satisfies $\theta_S^* \geq \theta_R^*$.

## 3.3 Proof of Proposition 2

Proposition 1 in the main text implies that we only need to consider problem (13). We start by proving that problem (13) is equivalent to

$$\begin{aligned} \underset{\theta_R,\theta_S}{\text{minimize}} \quad & Ae^{\gamma\theta_R} + Be^{2\gamma\theta_R} + Ce^{\gamma\theta_S} \\ \text{subject to} \quad & -D\theta_R - E\theta_S + Ge^{\gamma\theta_S} - He^{\gamma\theta_R} = L, \\ & 0 \leq \theta_R \leq \theta_S \leq 1. \end{aligned} \tag{15}$$

First we show by contradiction that when $L < G - H$, the optimal solution of problem (13) is always achieved when the inequality constraint is tight. Suppose $(\theta_R = x, \theta_S = y)$ is the optimal solution of problem (13) and $-Dx - Ey + Ge^{\gamma y} - He^{\gamma x} = L^* < L$, $0 \leq x \leq y \leq 1$. Define $h(x) = -Dx - He^{\gamma x}$ and $l(y) = -Ey + Ge^{\gamma y}$. We know $h(x)$ is decreasing in $x \in [0, 1]$ because $h'(x) = -D - H\gamma e^{\gamma x} < 0$ for any $x$. Similarly, $l(y)$ is decreasing in $y \in [0, 1]$ because $l'(y) = -E + G\gamma e^{\gamma y} < 0$ for any $y \leq 1$ due to the assumption that $D - E + G\gamma e^{\gamma} < 0$. Therefore, we will not have $x = y = 0$ because $-Dx - Ey + Ge^{\gamma y} - He^{\gamma x} = L^* < L < G - H$. Thus there exists $(\theta_R = x, \theta_S = y - \epsilon)$ or $(\theta_R = x - \epsilon, \theta_S = y)$ such that $L^* < -D\theta_R - E\theta_S + Ge^{\gamma\theta_S} - He^{\gamma\theta_R} \leq L$, $0 \leq \theta_R \leq \theta_S \leq 1$ and $Ae^{\gamma\theta_R} + Be^{2\gamma\theta_R} + Ce^{\gamma\theta_S} < Ae^{\gamma x} + Be^{2\gamma x} + Ce^{\gamma y}$, which contradicts that $(\theta_R = x, \theta_S = y)$ is the optimal solution. Finally, turning to the case where $L \geq G - H$, the optimal solution of problem (13) is $x = y = 0$, which is the optimal solution of problem (15) when $L = G - H$.

Defining $f(x) = Ae^{\gamma x} + Be^{2\gamma x}$, $g(y) = Ce^{\gamma y}$ and again denoting $\theta_R$ by $x$ and $\theta_S$ by $y$,

we can rewrite problem (15) as

$$\begin{aligned} \underset{x,y}{\text{minimize}} \quad & f(x) + g(y) \\ \text{subject to} \quad & h(x) + l(y) = L, \\ & 0 \le x \le y \le 1. \end{aligned} \tag{16}$$

Recall $h(x)$ and $l(y)$ are both decreasing functions within the feasible range of $x$ and $y$ in [0,1]. We only consider the values of $L$ such that the feasible set is not empty, i.e., $h(1) + l(1) \le L \le h(0) + l(0)$.

First we consider a relaxation of problem (16),

$$\begin{aligned} \underset{x,y}{\text{minimize}} \quad & f(x) + g(y) \\ \text{subject to} \quad & h(x) + l(y) = L, \\ & 0 \le x, y \le 1. \end{aligned} \tag{17}$$

We now derive the conditions for problem (17) to have either $(x^*, 1)$ or $(0, y^*)$ as the optimal solution. In this case, the optimal solution satisfies $x \le y$ and hence is also optimal for problem (16).

Because $h(x)$ and $l(y)$ are both decreasing functions, we know that the constraint $h(x) + l(y) = L$, defines $y$ as a decreasing function of $x$. We restrict our attention to the region $0 \le x, y \le 1$, where the continuous function $y(x)$ has domain being some subset of $[0, 1]$ and codomain also being some subset of $[0, 1]$. By implicit differentiation, within the feasible set of problem (17), we have $y'(x) = -h'(x)/l'(y)$. We denote the objective value as $O(x) = f(x) + g(y(x))$. Then the derivative

$$O'(x) = f'(x) - \frac{g'(y)h'(x)}{l'(y)}, \tag{18}$$

where $(x, y)$ is a feasible point of problem (17).

If we have $O'(x) > 0$ for all feasible points of problem (17) for any $L \in [h(1) + l(1), h(0) + l(0)]$, then for any $L$, the value of $x^*$ will be the minimum value of $x$ in the

feasible set. When $L \in [h(0)+l(1), h(0)+l(0)]$, $(0, l^{-1}(L-h(0)))$ is a feasible point; thus $x^* = 0$ and the optimal solution will be $(0, l^{-1}(L-h(0)))$. When $L \in [h(1)+l(1), h(0)+l(1)]$, the minimum value of $x$ in the feasible set is $h^{-1}(L-l(1))$ and thus the optimal solution is $(h^{-1}(L-l(1)), 1)$.

Because $h(x)$ is a decreasing function, (18) implies that $O'(x) > 0$ if and only if

$$\frac{f'(x)}{h'(x)} < \frac{g'(y)}{l'(y)}.$$

We can see that

$$\frac{f'(x)}{h'(x)} = -\frac{A\gamma e^{\gamma x} + 2B\gamma e^{2\gamma x}}{D + H\gamma e^{\gamma x}} = -\frac{A\gamma + 2B\gamma e^{\gamma x}}{De^{-\gamma x} + H\gamma}$$

decreases as $x$ increases because $A, B, D, H$ and $\gamma$ are all positive, and

$$\frac{g'(y)}{l'(y)} = -\frac{C\gamma e^{\gamma y}}{E - G\gamma e^{\gamma y}} = -\frac{C\gamma}{Ee^{-\gamma y} - G\gamma}$$

decreases as $y$ increases because $C, E, G$ and $\gamma$ are all positive and $Ee^{-\gamma} - G\gamma > De^{-\gamma} > 0$ because we already assume $D - E + G\gamma e^{\gamma} < 0$, which is condition (41) in Proposition 1 of the main text.

Thus $\frac{g'(y)}{l'(y)} \geq \frac{g'(1)}{l'(1)}$ for any $y \in [0,1]$ and $\frac{f'(x)}{h'(x)} \leq \frac{f'(0)}{h'(0)}$ for any $x \in [0,1]$. If we have $\frac{f'(0)}{h'(0)} < \frac{g'(1)}{l'(1)}$, i.e.,

$$\frac{A\gamma + 2B\gamma}{D + H\gamma} > \frac{C\gamma e^{\gamma}}{E - G\gamma e^{\gamma}},$$

which is equivalent to condition (43) in Proposition 2 of the main text, then $O'(x) > 0$ holds for all $(x, y) \in [0,1]^2$ and thus also holds for all feasible points of problem (17), thereby completing the proof.

# References

Usta M, Wein LM (2015) Assessing risk-based policies for pretrial release and split sentencing in Los Angeles County jails. *PLoS ONE* 10(12): e0144967.

Yang M, Wong SCP, Coid J (2010) The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin* 136:740-767.

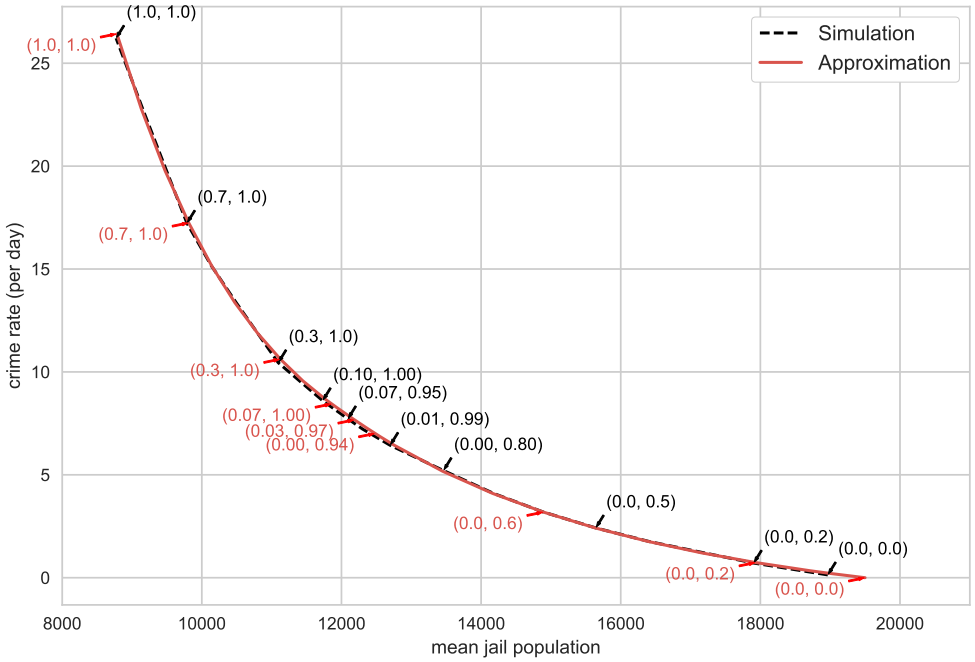


**Figure 4** Tradeoff curves for minimizing the crime rate subject to a constraint on the mean jail population, using both simulation and the cruder analytical approximation introduced in §3.5 of the main text, which assumes that $E[Q_i] = a_i$ and ignores the crimes from ejected and rejected inmates. The optimal $(\theta_R, \theta_S)$ values appear along various points of the tradeoff curves.

| $\theta_R$ | $\theta_S$ | $C_{\text{sim}}$ | $C_{\text{app}}$ | $E[Q_1]_{\text{sim}}$ | $E[Q_1]_{\text{app}}$ | $a_1$ | $E[Q_2]_{\text{sim}}$ | $E[Q_2]_{\text{app}}$ | $a_2$ | $E[Q_3]_{\text{sim}}$ | $E[Q_3]_{\text{app}}$ | $a_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.20 | 0.17 | 18967.59 | 18966.42 | 19500.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.2 | 0.71 | 0.74 | 15583.08 | 15600.00 | 15600.00 | 2321.47 | 2332.05 | 2332.05 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.4 | 1.75 | 1.76 | 11740.64 | 11700.00 | 11700.00 | 4671.95 | 4692.73 | 4692.73 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.6 | 3.20 | 3.19 | 7798.02 | 7800.00 | 7800.00 | 7089.27 | 7093.24 | 7093.24 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.8 | 5.21 | 5.17 | 3884.76 | 3900.00 | 3900.00 | 9576.75 | 9549.16 | 9549.16 | 0.00 | 0.00 | 0.00 |
| 0.0 | 1.0 | 7.90 | 7.93 | 0.00 | 0.00 | 0.00 | 12086.27 | 12082.20 | 12082.20 | 0.00 | 0.00 | 0.00 |
| 0.2 | 0.0 | 1.62 | 1.58 | 15615.97 | 15600.00 | 15600.00 | 3216.61 | 3261.15 | 3284.27 | 0.00 | 0.00 | 0.00 |
| 0.2 | 0.2 | 2.37 | 2.37 | 15585.85 | 15600.00 | 15600.00 | 0.00 | 0.00 | 0.00 | 1683.06 | 1695.29 | 1695.29 |
| 0.2 | 0.4 | 3.47 | 3.40 | 11738.31 | 11700.00 | 11700.00 | 2366.21 | 2360.68 | 2360.68 | 1683.06 | 1695.29 | 1695.29 |
| 0.2 | 0.6 | 4.81 | 4.82 | 7816.31 | 7800.00 | 7800.00 | 4757.55 | 4761.18 | 4761.18 | 1689.77 | 1695.29 | 1695.29 |
| 0.2 | 0.8 | 6.79 | 6.81 | 3924.67 | 3900.00 | 3900.00 | 7198.81 | 7217.11 | 7217.11 | 1693.37 | 1695.29 | 1695.29 |
| 0.2 | 1.0 | 9.58 | 9.57 | 0.00 | 0.00 | 0.00 | 9744.35 | 9750.14 | 9750.14 | 1699.52 | 1695.29 | 1695.29 |
| 0.4 | 0.0 | 3.80 | 3.79 | 11693.38 | 11700.00 | 11700.00 | 6549.67 | 6568.54 | 6568.54 | 0.00 | 0.00 | 0.00 |
| 0.4 | 0.2 | 4.56 | 4.57 | 11706.87 | 11700.00 | 11700.00 | 3277.32 | 3284.27 | 3284.27 | 1701.35 | 1695.29 | 1695.29 |
| 0.4 | 0.4 | 5.76 | 5.70 | 11711.04 | 11700.00 | 11700.00 | 0.00 | 0.00 | 0.00 | 3415.77 | 3411.39 | 3411.39 |
| 0.4 | 0.6 | 7.13 | 7.13 | 7824.00 | 7800.00 | 7800.00 | 2409.02 | 2400.51 | 2400.51 | 3405.78 | 3411.39 | 3411.39 |
| 0.4 | 0.8 | 9.14 | 9.11 | 3888.94 | 3900.00 | 3900.00 | 4851.27 | 4856.43 | 4856.43 | 3413.98 | 3411.39 | 3411.39 |
| 0.4 | 1.0 | 11.93 | 11.87 | 0.00 | 0.00 | 0.00 | 7384.14 | 7389.47 | 7389.47 | 3409.47 | 3411.39 | 3411.39 |
| 0.6 | 0.0 | 6.86 | 6.85 | 7846.75 | 7800.00 | 7800.00 | 9816.73 | 9852.80 | 9852.80 | 0.00 | 0.00 | 0.00 |
| 0.6 | 0.2 | 7.69 | 7.64 | 7777.88 | 7800.00 | 7800.00 | 6505.66 | 6568.54 | 6568.54 | 1698.69 | 1695.29 | 1695.29 |
| 0.6 | 0.4 | 8.80 | 8.76 | 7789.64 | 7800.00 | 7800.00 | 3265.59 | 3284.27 | 3284.27 | 3386.66 | 3411.39 | 3411.39 |
| 0.6 | 0.6 | 10.33 | 10.38 | 7803.89 | 7800.00 | 7800.00 | 0.00 | 0.00 | 0.00 | 5150.08 | 5156.44 | 5156.44 |
| 0.6 | 0.8 | 12.39 | 12.37 | 3910.94 | 3900.00 | 3900.00 | 2451.40 | 2455.92 | 2455.92 | 5176.76 | 5156.44 | 5156.44 |
| 0.6 | 1.0 | 15.17 | 15.13 | 0.00 | 0.00 | 0.00 | 4980.10 | 4988.96 | 4988.96 | 5154.48 | 5156.44 | 5156.44 |
| 0.8 | 0.0 | 11.08 | 11.11 | 3898.65 | 3900.00 | 3900.00 | 13048.63 | 13137.07 | 13137.07 | 0.00 | 0.00 | 0.00 |
| 0.8 | 0.2 | 11.97 | 11.90 | 3911.15 | 3900.00 | 3900.00 | 9813.32 | 9852.80 | 9852.80 | 1688.77 | 1695.29 | 1695.29 |
| 0.8 | 0.4 | 12.96 | 13.03 | 3891.96 | 3900.00 | 3900.00 | 6595.07 | 6568.54 | 6568.54 | 3396.63 | 3411.39 | 3411.39 |
| 0.8 | 0.6 | 14.63 | 14.65 | 3914.46 | 3900.00 | 3900.00 | 3281.22 | 3284.27 | 3284.27 | 5155.39 | 5156.44 | 5156.44 |
| 0.8 | 0.8 | 16.93 | 17.01 | 3892.33 | 3900.00 | 3900.00 | 0.00 | 0.00 | 0.00 | 6945.21 | 6941.78 | 6941.78 |
| 0.8 | 1.0 | 19.83 | 19.77 | 0.00 | 0.00 | 0.00 | 2539.75 | 2533.04 | 2533.04 | 6948.84 | 6941.78 | 6941.78 |
| 1.0 | 0.0 | 16.98 | 17.05 | 0.00 | 0.00 | 0.00 | 16451.02 | 16421.34 | 16421.34 | 0.00 | 0.00 | 0.00 |
| 1.0 | 0.2 | 17.83 | 17.84 | 0.00 | 0.00 | 0.00 | 13118.12 | 13137.07 | 13137.07 | 1685.34 | 1695.29 | 1695.29 |
| 1.0 | 0.4 | 18.87 | 18.96 | 0.00 | 0.00 | 0.00 | 9841.01 | 9852.80 | 9852.80 | 3429.48 | 3411.39 | 3411.39 |
| 1.0 | 0.6 | 20.46 | 20.58 | 0.00 | 0.00 | 0.00 | 6542.85 | 6568.54 | 6568.54 | 5171.61 | 5156.44 | 5156.44 |
| 1.0 | 0.8 | 22.91 | 22.94 | 0.00 | 0.00 | 0.00 | 3313.55 | 3284.27 | 3284.27 | 6938.88 | 6941.78 | 6941.78 |
| 1.0 | 1.0 | 26.22 | 26.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 8772.27 | 8783.18 | 8783.18 |

**Table 2** For 36 $(\theta_R, \theta_S)$ scenarios, simulation vs. approximate results for the total crime rate $C$ and the mean jail population $E[Q_i]$, as well as $a_i$, for each inmate class $i = 1, 2, 3$. The subscripts "sim" and "app" represent simulation and approximate results.

| $\theta_R$ | $\theta_S$ | $E_{1,\text{sim}}$ | $E_{1,\text{app}}$ | $E_{2,\text{sim}}$ | $E_{2,\text{app}}$ | $E_{3,\text{sim}}$ | $E_{3,\text{app}}$ | $R_{1,\text{sim}}$ | $R_{1,\text{app}}$ | $R_{2,\text{sim}}$ | $R_{2,\text{app}}$ | $R_{3,\text{sim}}$ | $R_{3,\text{app}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.1929 | 0.1676 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0036 | 0.0032 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.0 | 0.2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.0 | 0.4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.0 | 0.6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.0 | 0.8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.0 | 1.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.2 | 0.0 | 0.0000 | 0.0000 | 0.0132 | 0.0016 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.2 | 0.2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.2 | 0.4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.2 | 0.6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.2 | 0.8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.2 | 1.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.4 | 0.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.4 | 0.2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.4 | 0.4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.4 | 0.6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.4 | 0.8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.4 | 1.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.6 | 0.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.6 | 0.2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.6 | 0.4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.6 | 0.6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.6 | 0.8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.6 | 1.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.8 | 0.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.8 | 0.2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.8 | 0.4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.8 | 0.6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.8 | 0.8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.8 | 1.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1.0 | 0.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1.0 | 0.2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1.0 | 0.4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1.0 | 0.6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1.0 | 0.8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1.0 | 1.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

**Table 3** For 36 $(\theta_R, \theta_S)$ scenarios, simulation vs. approximate results for the crime rates of ejected $(E_i)$ and rejected $(R_i)$ inmates of class $i = 1, 2, 3$. The subscripts "sim" and "app" represent simulation and approximate results.

| $\theta_R$ | $\theta_S$ | $SS_{2,sim}$ | $SS_{2,app}$ | $SS_{3,sim}$ | $SS_{3,app}$ | $PTR_{2,sim}$ | $PTR_{2,app}$ | $PTR_{3,sim}$ | $PTR_{3,app}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.2 | 0.71 | 0.74 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.4 | 1.75 | 1.76 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.6 | 3.20 | 3.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.8 | 5.21 | 5.17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.0 | 1.0 | 7.90 | 7.93 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.2 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.61 | 1.58 | 0.00 | 0.00 |
| 0.2 | 0.2 | 0.00 | 0.00 | 0.74 | 0.74 | 0.00 | 0.00 | 1.63 | 1.63 |
| 0.2 | 0.4 | 1.08 | 1.03 | 0.74 | 0.74 | 0.00 | 0.00 | 1.65 | 1.63 |
| 0.2 | 0.6 | 2.41 | 2.45 | 0.73 | 0.74 | 0.00 | 0.00 | 1.67 | 1.63 |
| 0.2 | 0.8 | 4.43 | 4.44 | 0.74 | 0.74 | 0.00 | 0.00 | 1.63 | 1.63 |
| 0.2 | 1.0 | 7.23 | 7.20 | 0.75 | 0.74 | 0.00 | 0.00 | 1.60 | 1.63 |
| 0.4 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.80 | 3.79 | 0.00 | 0.00 |
| 0.4 | 0.2 | 0.00 | 0.00 | 0.73 | 0.74 | 2.21 | 2.20 | 1.63 | 1.63 |
| 0.4 | 0.4 | 0.00 | 0.00 | 1.77 | 1.76 | 0.00 | 0.00 | 3.99 | 3.94 |
| 0.4 | 0.6 | 1.43 | 1.43 | 1.76 | 1.76 | 0.00 | 0.00 | 3.94 | 3.94 |
| 0.4 | 0.8 | 3.44 | 3.41 | 1.74 | 1.76 | 0.00 | 0.00 | 3.96 | 3.94 |
| 0.4 | 1.0 | 6.25 | 6.17 | 1.74 | 1.76 | 0.00 | 0.00 | 3.93 | 3.94 |
| 0.6 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 6.86 | 6.85 | 0.00 | 0.00 |
| 0.6 | 0.2 | 0.00 | 0.00 | 0.74 | 0.74 | 5.33 | 5.27 | 1.62 | 1.63 |
| 0.6 | 0.4 | 0.00 | 0.00 | 1.78 | 1.76 | 3.08 | 3.06 | 3.94 | 3.94 |
| 0.6 | 0.6 | 0.00 | 0.00 | 3.22 | 3.19 | 0.00 | 0.00 | 7.11 | 7.20 |
| 0.6 | 0.8 | 1.98 | 1.98 | 3.22 | 3.19 | 0.00 | 0.00 | 7.19 | 7.20 |
| 0.6 | 1.0 | 4.78 | 4.75 | 3.18 | 3.19 | 0.00 | 0.00 | 7.21 | 7.20 |
| 0.8 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 11.08 | 11.11 | 0.00 | 0.00 |
| 0.8 | 0.2 | 0.00 | 0.00 | 0.74 | 0.74 | 9.56 | 9.53 | 1.67 | 1.63 |
| 0.8 | 0.4 | 0.00 | 0.00 | 1.74 | 1.76 | 7.32 | 7.33 | 3.89 | 3.94 |
| 0.8 | 0.6 | 0.00 | 0.00 | 3.22 | 3.19 | 4.21 | 4.26 | 7.20 | 7.20 |
| 0.8 | 0.8 | 0.00 | 0.00 | 5.12 | 5.17 | 0.00 | 0.00 | 11.82 | 11.83 |
| 0.8 | 1.0 | 2.77 | 2.76 | 5.20 | 5.17 | 0.00 | 0.00 | 11.86 | 11.83 |
| 1.0 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.98 | 17.05 | 0.00 | 0.00 |
| 1.0 | 0.2 | 0.00 | 0.00 | 0.75 | 0.74 | 15.43 | 15.46 | 1.65 | 1.63 |
| 1.0 | 0.4 | 0.00 | 0.00 | 1.76 | 1.76 | 13.17 | 13.26 | 3.95 | 3.94 |
| 1.0 | 0.6 | 0.00 | 0.00 | 3.11 | 3.19 | 10.21 | 10.20 | 7.14 | 7.20 |
| 1.0 | 0.8 | 0.00 | 0.00 | 5.18 | 5.17 | 5.93 | 5.93 | 11.80 | 11.83 |
| 1.0 | 1.0 | 0.00 | 0.00 | 7.90 | 7.93 | 0.00 | 0.00 | 18.32 | 18.49 |

**Table 4** For 36 $(\theta_R, \theta_S)$ scenarios, simulation vs. approximate results for the crime rates of inmates of class $i = 2, 3$ on pretrial release ($PTR_i$) and split-sentence supervision ($SS_i$). The subscripts "sim" and "app" represent simulation and approximate results.