Routing for Fairness and Efficiency in a Queueing Model with Reentry and Continuous Customer Classes

Zhiqiang Zhang,[1] Pengyi Shi,[2] and Amy R. Ward[3]

Abstract— **This paper presents a new framework to study customer routing fairness and efficiency in managing queueing systems with reentry; this framework is motivated by applications in healthcare and criminal justice. Each customer's reentry probability is captured by a proportional hazard model that depends on the type of service offered and the customer's continuous risk score, which is drawn from group-dependent distributions. Unlike the machine learning literature that focuses on fairness in classification with group-dependent risk distributions, we account for fairness directly in the resource allocation captured via the routing decisions. We derive an explicit analytical characterization of the system's efficiency and fairness for two policies: group-unaware and group-aware routing policies. Through the comparison, we quantify the fairness-efficiency trade-off. This trade-off generates useful insights for decision makers into how the cost of fairness depends on different system parameters and when caution should be taken for fairness considerations.**

I. INTRODUCTION

Customers arriving at service systems typically have different attributes and service needs. It is of central importance to allocate proper services that best match with the customers' needs when managing service systems, particularly in resource-limited settings. With recent developments in machine learning tools, accurate prediction of customers' service needs has become much easier. These predictions facilitate decision makers to better target limited resources to subgroups who benefit the most.

At the same time, increasing attention has been brought to the potential bias and the resulting unfairness caused by using such prediction algorithms in resource allocation [1]. That is, significant disparities could have existed in how different population groups accessed and used different services; hence, data collected from the past could be intrinsically biased. Prediction algorithms trained on such biased data will produce biased results and could further exacerbate a disparity when those results are used in resource allocation, particularly in systems with a "feedback loop."

We use two examples to illustrate. The first example comes from healthcare, where decision makers (hospital managers) often use predicted risk scores to target resources for higher risk patients in their postdischarge follow-ups, e.g., using phone calls and home visits, instead of the standard text messages, to increase patients' medical adherence and to reduce the 30-day hospital readmission rate. However, Obermyer et al. [2] find that disparities exist in healthcare utilization among different racial groups. Using an algorithm trained on such utilization data, patients from one group could be much sicker than those from another even though their predicted scores could be the same. In other words, patients from the latter group would be effectively assigned more resources than the former when such biased prediction is used, further enlarging the existing disparity.

The second example comes from criminal justice, where machine learning tools are used to predict recidivism so judges can make probational and jail-diversion decisions. Angwin et al. [3] find that a widely used recidivism prediction tool disproportionally labeled more people from the Black community as high risk, with higher false-positive rates than other groups. Consequently, Black people would get fewer opportunities to receive probation or jail-alternatives and be more likely to be incarcerated. Disparities in incarceration further limit their opportunities to access treatment and support services and reduce chances of social reintegration, leading to the vicious "revolving prison door."

The machine learning literature has developed various ways to address bias and achieve fairness in prediction. Methods include data preprocessing, in-processing, and post-processing [4], with equal-opportunity and equal-odds being the most commonly used criteria. However, their focus is to achieve similar false-negative/-positive rates among different groups in classifications when dealing with group-dependent risk/covariate distributions [5]. Even with an (ideally) debiased prediction, decision makers still need to make resource allocation decisions as a separate step. In this paper, we take a completely different approach to addressing fairness in managing service systems with machine learning–based predictions. We directly work with the resource allocation decision and study the trade-off between efficiency and fairness under different decisions. In other words, we eliminate the two-step "predict-then-decide" procedure and directly account for fairness in the decision.

Motivated by various systems with customer reentry (e.g., patient readmission and recidivism in the two examples), we consider a stylized queueing model with customer reentry. The model has two groups of customers and two stages of services. Each customer has a continuous risk score that is drawn from a risk distribution; the risk distributions may differ between the two groups. Each customer goes through the same first stage, e.g., hospitalization, initial jail sentencing. But in the second stage, there is a decision between two

[1]Zhiqiang Zhang is a Masters student in the Computational and Applied Mathematics Program at the University of Chicago; zqzhang1729@uchicago.edu.

[2]Pengyi Shi is on the faculty of the Krannert School of Management at Purdue University; shi178@purdue.edu.

[3]Amy R. Ward is on the faculty of the Booth School of Business at the University of Chicago; amyward@chicagobooth.edu.

options (special versus standard), which leads to different reentry probabilities. We refer to this stage as the monitoring stage, corresponding to patients being discharged or offenders being placed on probation. Special monitoring includes enhanced follow-up services and extra community-support services. We use a proportional hazard model to capture the reentry probability, where the hazard rate parameters depend on both the customer's risk score and the services they received. Proportional hazard models are found to be suitable for predicting reentry probabilities in service, healthcare, and criminal justice areas [6]. The decision concerns which type of service to assign an incoming customer to, i.e., the *routing* decision. Our aim in this study is to quantify the trade-off between fairness and efficiency in routing, where we measure *fairness* in terms of the proportions of customers assigned to special monitoring from the two groups, and *efficiency* in terms of minimized return probabilities. To do this, we consider a class of state independent randomized routing policies, whose analytic tractability allows us to characterize efficiency and fairness when two types of policies are used: a *group-unaware* policy and a *group-aware* policy. Since the group-aware policy uses the group information in routing while the group-unaware policy does not, we characterize the cost of fairness when the routing probabilities in the group-aware policy are chosen to achieve maximum efficiency. We leave for future work to see if more complicated policies can improve efficiency and reduce unfairness.

This paper makes two major contributions. First, we develop a new queueing-control framework to address fairness in routing for systems with reentry. Unlike typical queues with feedback loops [7], the probability of reentry depends on (i) which service was received in the second stage, and (ii) the risk level of the customer, which is drawn from a continuous distribution that depends on the population group, as opposed to the typical homogeneous or finite-class assumption in the literature. Second, we analytically characterize the trade-off between efficiency and fairness by comparing the group-unaware and the group-aware policies. Our novel modeling framework and our focus on return probabilities differentiate us from the literature of balancing fairness and efficiency, as in the recent work of Mulvany et al. [8] on the customer side, or in the work of Ward et al. [9] on the server side (see also references therein, and Wierman et al. [10] for a discussion on fairness). Our analytical development explicitly allows quantification of the cost of fairness (i.e., the efficiency loss when maintaining fairness) and understanding of how it depends on various system parameters, from which one could identify strategies to reduce this cost. We also generate insights into what parameter regimes result in higher costs of fairness, suggesting caution should be taken when allocating resources in these regimes and more sophisticated routing policies would be warranted to design fair and efficient policies.

II. MODEL DESCRIPTION

We consider a queueing system with two groups of customers, denoted M and F (mnemonic, for males and females, for example). Each customer has a risk level $p \in [0,1]$, drawn from the risk distribution of their corresponding group. We denote $p_M \sim G_M$ and $p_F \sim G_F$, and assume both have density functions. Customers from each group arrive to the system following Poisson processes, at arrival rates λ_M and λ_F, respectively. The system has two stages, modeled as two infinite-server queues. For ease of exposition, we refer to the first stage as the service stage and the second as the monitoring stage based on our motivating examples.

Upon arrival, each customer first enters the service stage, where the service time follows an exponential distribution with rate μ_1. Then, the customer enters the monitoring stage, during which the customer may reenter the first stage. The reentry probability depends on risk p and which type of monitoring they receive in the monitoring stage: special versus standard monitoring. Specifically, the time a customer spends in the monitoring stage is the minimum of two exponential random variables, one with rate μ_2 (the nominal monitoring time) and the other with rate $\eta_i e^{\gamma_i p}, i \in \{1,2\}$, which follows the proportional hazard model to reflect the rate of reentry. Without loss of generality, we use $i = 1$ to denote the special monitoring, and $i = 2$ for the standard monitoring, and $\eta_i, \gamma_i > 0$ are parameters for the hazard model. We also assume that special monitoring reduces reentry, i.e., $\eta_1 e^{\gamma_1 p} \leq \eta_2 e^{\gamma_2 p}, \forall p \in [0,1]$. Therefore, a customer with risk p leaves the monitoring stage following an exponential time with rate $\mu_2 + \eta_i e^{\gamma_i p}$, and depending on which monitoring they receive, this customer reenters the first stage with probability

$$h_i(p) := \frac{\eta_i e^{\gamma_i p}}{\mu_2 + \eta_i e^{\gamma_i p}}, i \in \{1,2\}, \tag{1}$$

and leaves the system with probability $1 - h_i(p)$. We assume that the reentries are memoryless and the customer is randomly assigned a new risk level while reentering the system. So the risk distributions do not change as time passes. Figure 1 illustrates the flow chart of the queueing system.

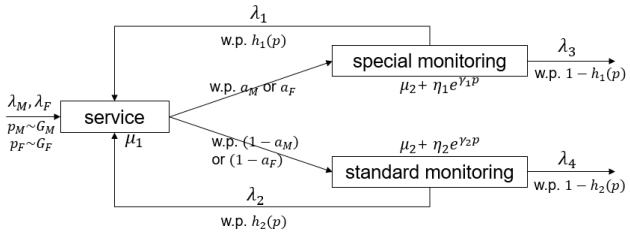


Fig. 1: The queueing model. All the service times are exponential.

We focus on randomized routing policies in this paper. That is, we route customers in group $j \in \{M,F\}$ to receive special monitoring with probability $a_j \in [0,1]$ and standard monitoring with probability $1 - a_j$. We refer to a_M and a_F as the routing probabilities for the two groups. We say a routing policy is *group-aware* if $a_M \neq a_F$ and *group-unaware* if $a_M = a_F$.

To facilitate later analysis, we define the following flow rates, which are also labeled in Figure 1; λ_1 and λ_2 cor-

respond to the rates of reentry from special and standard monitoring, respectively; λ_3 and λ_4 correspond to the departure rates from special and standard monitoring, respectively. These rates are the sum from the two groups, i.e., $\lambda_i = \lambda_{iM} + \lambda_{iF}, i = 1, \cdots, 4$. Note that these flow rates depend on the routing probabilities a_M and a_F, although we have suppressed that dependence for notational simplicity.

A. System-Level Fairness and Efficiency Metrics

To quantify the efficiency and fairness from different routing policies, we define two performance metrics on the system level. For the *fairness metric*, we use the percentage of special monitoring each group receives, since special monitoring leads to a more desirable outcome in our motivating examples (lower reentry probability). For the *efficiency metric*, we use reentry percentage. The two metrics are defined formally as follows.

Definition 1: (Fairness) A routing policy (a_M, a_F) is *fair* if the percentages of special monitoring are equal in two groups; that is, $a_M = a_F$. Otherwise, the policy is *unfair*.

Definition 2: (Efficiency) The *efficiency* metric of a routing policy (a_M, a_F) is the percentage of departure, which is

$$\overline{R}(a_M, a_F) = 1 - R(a_M, a_F) := 1 - \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_M + \lambda_F},$$

where $R(a_M, a_F)$ is the percentage of reentry.

Note that under our definition of efficiency, maximizing efficiency $\overline{R}(a_M, a_F)$ is equivalent to minimizing the percentage of reentry $R(a_M, a_F)$. We use $R(a_M, a_F)$ in later analysis.

B. Group-Level Efficiency

A routing policy with a high efficiency overall does not mean it is highly efficient for both groups. For group-level efficiency, we consider the percentage of reentry in each group. Note that the probability of reentry for a customer with risk level p is

$$h(p, a) := a h_1(p) + (1 - a) h_2(p),$$

with a being the probability of receiving special monitoring. Thus, the percentages of reentry in groups M and F are the expectations of $h(p, a)$ with respect to the distribution of p under $a = a_M$ and $a = a_F$, which can be given as,

$$R_M(a_M) := \mathbb{E}_{p \sim G_M}[h(p, a_M)], \quad R_F(a_F) := \mathbb{E}_{p \sim G_F}[h(p, a_F)].$$

Then the overall efficiency can be interpreted as a weighted average of group-level efficiencies, which has the following form

$$R(a_M, a_F) = \frac{R_M(a_M)(\lambda_{1M} + \lambda_{2M} + \lambda_M) + R_F(a_F)(\lambda_{1F} + \lambda_{2F} + \lambda_F)}{\lambda_{1M} + \lambda_{2M} + \lambda_M + \lambda_{1F} + \lambda_{2F} + \lambda_F}.$$

III. Group-Unaware Routing

We start by analyzing the group-unaware policy, which does not use group information in routing decisions. As a result, the routing probability is the same for each customer regardless group, i.e., $a_M = a_F = a$, and the policy is fair according to our definition of fairness. In the following analysis, we study group-level efficiency when G_M and G_F have a convex ordering relationship. To start, we examine the convexity/concavity of $h(p, a)$.

Lemma 1: Fix $a \in [0, 1]$ and denote $h(p, a)$ as $h(p)$ for $p \in [0, 1]$.

1) If $\mu_2 \geq \max\{\eta_1 e^{\gamma_1}, \eta_2 e^{\gamma_2}\}$, then $h(p)$ is convex in p.
2) If $\mu_2 \leq \min\{\eta_1, \eta_2\}$, then $h(p)$ is concave in p.

The proof of this lemma is straightforward by studying the second-order derivative of $h_i(p), i \in \{1, 2\}$, and thus is omitted here. Based on Lemma 1, we leverage the convex ordering (see, for example, [11]) of the distributions G_M and G_F to study the ordering in the group-level efficiency.

Definition 3: (Convex Ordering) For random variables X and Y, X is said to be smaller than Y in *convex ordering* (denoted as $X \leq_{\text{cx}} Y$) if $\mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(Y)]$, for all convex functions ϕ.

Theorem 1: Fix $a \in [0, 1]$ and denote $h(p, a)$ as $h(p)$ for $p \in [0, 1]$. Suppose $p_M \leq_{\text{cx}} p_F$,

1) If $h(p)$ is convex, then $\mathbb{E}_{p \sim G_M}[h(p)] \leq \mathbb{E}_{p \sim G_F}[h(p)]$.
2) If $h(p)$ is concave, then $\mathbb{E}_{p \sim G_M}[h(p)] \geq \mathbb{E}_{p \sim G_F}[h(p)]$.

Proof: The proof follows directly from Definition 3, noting that if $h(p)$ is concave, then $-h(p)$ is convex. ■

Theorem 1 implies that the group-unaware policy does not produce the same reentry percentage (efficiency) for the two groups in general if the risk distributions are different. Which group has a higher reentry percentage depends on the convexity or concavity of the function $h(p)$, which further depends on the hazard model parameters η_i, γ_i and the nominal service rate μ_2 of the monitoring stage. As a result, the system-level efficiency can be improved by using group information to make the routing decisions, and setting a_M and a_F differently (resulting in a policy that is no longer fair).

IV. Group-Aware Routing

Group-aware routing has an additional degree of freedom by choosing different routing probabilities for the two groups to achieve a higher system-level efficiency. Given this degree of freedom, we formulate an optimization problem to choose routing probabilities for each group to minimize the system-level percentage of reentry $R(a_M, a_F)$ (i.e., maximize efficiency). We then present the analytical solution of this optimization problem and discuss whether the results are fair under the corresponding group-aware policy.

A. Routing Optimization Problem

Since we assume that special monitoring leads to a lower reentry probabilities, the most efficient policy without any constraint is to offer every customer special monitoring in the second stage of service, then the problem becomes trivial. To make this problem practical, we fix the overall percentage of special monitoring offered, denoted by $P(a_M, a_F)$, which reflects the limited budget in reality. The overall percentage of special monitoring $P(a_M, a_F)$ is a weighted average of a_M and a_F,

$$P(a_M, a_F) := \frac{a_M(\lambda_{1M} + \lambda_{2M} + \lambda_M) + a_F(\lambda_{1F} + \lambda_{2F} + \lambda_F)}{\lambda_{1M} + \lambda_{2M} + \lambda_M + \lambda_{1F} + \lambda_{2F} + \lambda_F}.$$

Given a fixed level of $P(a_M, a_F)$, we formulate an optimization problem to minimize the system-level reentry probability:

$$\begin{array}{ll} \min_{a_M, a_F} & R(a_M, a_F), \\ \text{subject to} & P(a_M, a_F) = \bar{a}, \\ & 0 \le a_M \le 1, \\ & 0 \le a_F \le 1, \end{array} \tag{2}$$

where $\bar{a} \in [0,1]$ is the fixed percentage of special monitoring. If we apply Little's law to the service stage, the optimization problem above is equivalent to minimizing the mean population in service stage in Figure 1.

The optimization problem (2) must have a solution. Because the objective function $R(a_M, a_F)$ is continuous in a_M, a_F and the feasible set is closed in the 2-dimensional space for a_M, a_F, the minimum of $R(a_M, a_F)$ can be attained in the feasible set.

B. Traffic Equations

Solving the optimization problem (2) requires characterizing the flow rates $\lambda_{1M}, \lambda_{2M}, \lambda_{1F}, \lambda_{2F}$ as functions of a_M and a_F. The traffic equations can be set up and solved for each group independently, since we have assumed infinite-server queues for both stages in the system. We focus on solving these flow rates for group M below; the rates for group F can be solved for with the same procedure. For group M, we denote the probability density function for its risk distribution G_M as $g_M(p)$. Using (1), we define two important quantities that are the key for our analysis:

$$\alpha_{1M} := \int_0^1 h_1(p) g_M(p) \mathrm{d}p, \quad \alpha_{2M} := \int_0^1 h_2(p) g_M(p) \mathrm{d}p.$$

These two quantities are expectations of the reentry probabilities if all customers in group M are assigned to special monitoring or standard monitoring, respectively. We have $\alpha_{1M}, \alpha_{2M} \in [0,1]$. Note that under randomized routing, a customer, regardless of risk p, is assigned to special monitoring with probability a_M. Thus, the flow rate, λ_{1M}, for those who received special monitoring and reenter the system equals the proportion $a_M \alpha_{1M}$ multiplied by the total flow rate that comes into the system, $\lambda_M + \lambda_{1M} + \lambda_{2M}$. Following similar steps, we set up the following flow-balance equations:

$$\begin{cases} \lambda_{1M} = (\lambda_M + \lambda_{1M} + \lambda_{2M}) a_M \alpha_{1M} \\ \lambda_{2M} = (\lambda_M + \lambda_{1M} + \lambda_{2M})(1 - a_M) \alpha_{2M} \\ \lambda_{1M} + \lambda_{3M} = (\lambda_M + \lambda_{1M} + \lambda_{2M}) a_M \\ \lambda_{2M} + \lambda_{4M} = (\lambda_M + \lambda_{1M} + \lambda_{2M})(1 - a_M) \end{cases}.$$

Solving these equations, we get

$$\begin{cases} \lambda_{1M} = \dfrac{a_M \alpha_{1M}}{1 - R_M(a_M)} \lambda_M, \quad \lambda_{2M} = \dfrac{(1 - a_M)\alpha_{2M}}{1 - R_M(a_M)} \lambda_M \\ \lambda_{3M} = \dfrac{a_M (1 - \alpha_{1M})}{1 - R_M(a_M)} \lambda_M, \quad \lambda_{4M} = \dfrac{(1 - a_M)(1 - \alpha_{2M})}{1 - R_M(a_M)} \lambda_M \end{cases}, \tag{3}$$

where $R_M(a_M) = \mathbb{E}[h(p_M, a_M)] = a_M \alpha_{1M} + (1 - a_M)\alpha_{2M}$.

C. Optimal Group-Aware Policy

With (3), we can solve the optimization problem (2) analytically; the solution will be stated in Theorem 2 below. We will show that optimal solutions are either a strict priority policy or any mixture policy when there is no differentiation. The intuition behind the strict priority policy is simple: if one group benefits more from receiving special monitoring than the other in terms of reentry probability reduction, this group should be assigned to receive special monitoring with higher priority since our objective is to minimize system-level reentry proportions. Thus, the priority group should be given as much special monitoring as possible and the other group none when $\bar{a}$ is small. When $\bar{a}$ is large, everyone in the priority group should be routed to special monitoring and the other group will be routed with the remaining available budget.

As discussed, which group gets prioritized depends on the reentry probability reduction from receiving special monitoring, i.e., the *treatment effect* if we think of special monitoring as an intervention. To facilitate the analysis, we quantify the group-level treatment effect as the ratio of average departure probabilities, $1 - \alpha_{ij}$, with standard and special monitoring ($i = 1, 2$) in each group $j \in \{M, F\}$. We work with the departure probabilities instead of α_{ij} because (i) a higher departure probability means reduced reentry in terms of measuring the treatment effect, and (ii) this method avoids the hassle of accounting for reentry customers in the flow rate back to the system and so simplifies the solution. Formally, we define the treatment effect as follows.

Definition 4: (Group-Level Treatment Effect) The *treatment effect* for group $j \in \{M, F\}$ is the ratio of average departure probabilities with special and standard monitoring in group j:

$$E_j := \frac{1 - \alpha_{1j}}{1 - \alpha_{2j}}, \quad j \in \{M, F\}.$$

We have $E_j \ge 1$ under the assumption that special monitoring reduces the reentry probability and thus increases the departure probability. A larger treatment effect E_j indicates a larger benefit for customers in group j who receive special monitoring. Now we state the optimal solution for (2).

Theorem 2: Fix $\bar{a} \in [0,1]$.

1) If $E_M \ne E_F$, without loss of generality, we assume $E_M > E_F$, then group M is strictly prioritized. The unique optimal solution of (2) $(a_M^\star, a_F^\star)$ satisfies $a_F^\star = 0, P(a_M^\star, 0) = \bar{a}$ if $\bar{a} \le P(1,0)$, or $a_M^\star = 1, P(1, a_F^\star) = \bar{a}$ otherwise. The explicit solution is

$$\begin{cases} \left(\dfrac{(1 - \alpha_{2F})\bar{a}\lambda_M + (1 - \alpha_{2M})\bar{a}\lambda_F}{(1 - \alpha_{2F})\lambda_M - (\alpha_{2M} - \alpha_{1M})\bar{a}\lambda_F}, 0 \right), & \text{if } \bar{a} \le P(1,0), \\ \left(1, \dfrac{-(1 - \alpha_{2F})(1 - \bar{a})\lambda_M + (1 - \alpha_{2M})\bar{a}\lambda_F}{(\alpha_{2F} - \alpha_{1F})(1 - \bar{a})\lambda_M + (1 - \alpha_{1M})\lambda_F} \right), & \text{if } \bar{a} > P(1,0). \end{cases}$$

2) If $E_M = E_F$, then every feasible point for (a_M, a_F) is an optimal solution.

Proof: First we eliminate the equation constraint in optimization problem (2). Treating a_M as given, we can solve

for a_F as a function of a_M from $P(a_M, a_F) = \bar{a}$:

$$a_F(a_M) = \frac{[1-R_M(a_M)]\bar{a}\lambda_F - (1-\alpha_{2F})(a_M-\bar{a})\lambda_M}{[1-R_M(a_M)]\lambda_F + (\alpha_{2F}-\alpha_{1F})(a_M-\bar{a})\lambda_M}. \quad (4)$$

After substituting (4) into optimization problem (2), we can then take the derivative of $R(a_M, a_F(a_M))$ with respect to a_M,

$$\frac{\mathrm{d}R(a_M,a_F(a_M))}{\mathrm{d}a_M} = \frac{[(1-\alpha_{1F})(1-\alpha_{2M})-(1-\alpha_{1M})(1-\alpha_{2F})][1-R_F(\bar{a})]\lambda_M(\lambda_M+\lambda_F)}{\{[1-R_F(a_M)]\lambda_M+[1-R_M(a_M)]\lambda_F\}^2}.$$

Now consider the two cases:

1) If $E_M > E_F$, we have $(1-\alpha_{1M})(1-\alpha_{2F}) > (1-\alpha_{1F})(1-\alpha_{2M})$, which implies $\frac{\mathrm{d}R(a_M,a_F(a_M)}{\mathrm{d}a_M} < 0$. So $a_M^\star$ should be the maximum in its feasible set, and the solution is unique.
   If $\bar{a} \le P(1,0)$, the maximum of a_M can be attained when $a_F = 0$. In this case all customers receiving special monitoring are from group M, i.e., $a_F^\star = 0$, and $a_M^\star$ satisfies $P(a_M^\star, 0) = \bar{a}$. This gives us
   $$a_M^\star = \frac{(1-\alpha_{2F})\bar{a}\lambda_M + (1-\alpha_{2M})\bar{a}\lambda_F}{(1-\alpha_{2F})\lambda_M - (\alpha_{2M}-\alpha_{1M})\bar{a}\lambda_F}.$$
   If $\bar{a} > P(1,0)$, a_M can reach its natural upper bound 1. Thus, $a_M^\star = 1$, and $a_F^\star$ can be computed from $P(1, a_F^\star) = \bar{a}$, which is given by
   $$a_F^\star = \frac{-(1-\alpha_{2F})(1-\bar{a})\lambda_M + (1-\alpha_{2M})\bar{a}\lambda_F}{(\alpha_{2F}-\alpha_{1F})(1-\bar{a})\lambda_M + (1-\alpha_{1M})\lambda_F}.$$
2) If $E_M = E_F$, then $\frac{\mathrm{d}R(a_M,a_F(a_M))}{\mathrm{d}a_M} = 0$. So $R(a_M, a_F(a_M))$ is a constant in the feasible set, which implies that every feasible point for (a_M, a_F) is a optimal solution. ■

Clearly, the optimal group-aware policy is unfair as it strictly prioritizes one group over another group when the treatment effects are different. In other words, the optimal group-aware policy improves efficiency at the cost of fairness in comparison to the group-unaware policy. In the next section, we explicitly quantify this cost of fairness.

V. Fairness-Efficiency Trade-Off

Comparing the group-unaware and the group-aware policy, we can clearly see a trade-off between fairness and efficiency: the group-aware policy is more efficient than the group-unaware policy, but not as fair. To study this fairness-efficiency trade-off, we define the *cost of fairness* as the efficiency loss of the fair, group-unaware routing policy compared with the optimal policy solved from (2), while fixing the overall percentage of special monitoring $\bar{a}$.

Definition 5: (Cost of Fairness) The cost of fairness $C(\bar{a})$ is defined as the difference of system-level efficiency between the group-unaware policy $(\bar{a}, \bar{a})$ and the optimal group-aware policy $(a_M^\star, a_F^\star)$, while fixing the percentage of special monitoring to $\bar{a}$. That is,

$$C(\bar{a}) := R(\bar{a}, \bar{a}) - R(a_M^\star, a_F^\star),$$

where $(a_M^\star, a_F^\star)$ is the optimal solution of (2).

We focus our discussion on when the treatment effects are different to avoid triviality. In particular, if $E_M = E_F$, then the cost of fairness is 0; that is, the group-unaware policy retains the same efficiency if the treatment effects of the two groups are the same.

In the case where $E_M \neq E_F$, we assume $E_M > E_F$ without loss of generality, then the cost of fairness is

$$C(\bar{a}) = \begin{cases} \frac{\widetilde{\lambda}_F(\bar{a})}{\widetilde{\lambda}_M(\bar{a})+\widetilde{\lambda}_F(\bar{a})} \frac{\bar{a}(E_M-E_F)(\lambda_M+\lambda_F)}{\lambda_M/(1-\alpha_{2M})+\lambda_F/(1-\alpha_{2F})}, & \text{if } \bar{a} \le P(1,0), \\ \frac{\widetilde{\lambda}_M(\bar{a})}{\widetilde{\lambda}_M(\bar{a})+\widetilde{\lambda}_F(\bar{a})} \frac{(1-\bar{a})(1/E_F-1/E_M)(\lambda_M+\lambda_F)}{\lambda_M/(1-\alpha_{1M})+\lambda_F/(1-\alpha_{1F})}, & \text{if } \bar{a} > P(1,0), \end{cases}$$

where $\widetilde{\lambda}_j(a) = \lambda_{1j} + \lambda_{2j} + \lambda_j = \frac{1}{1-a\alpha_{1j}-(1-a)\alpha_{2j}}\lambda_j, j = M, F$ is the total entry rate for group j.

We can see that the cost of fairness depends on two factors: (i) the proportion of two group populations in the group-unaware policy $\frac{\widetilde{\lambda}_j(\bar{a})}{\widetilde{\lambda}_M(\bar{a})+\widetilde{\lambda}_F(\bar{a})}, j = M, F$, and (ii) the difference between treatment effects $E_M - E_F$ and $1/E_M - 1/E_F$. If there is a large discrepancy between the treatment effects E_M and E_F, the cost of fairness $C(\bar{a})$ is expected to be large.

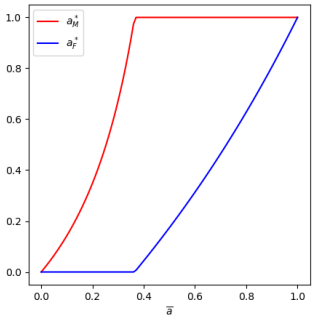
This explicit characterization allows us to understand when the cost of fairness could be large, which suggests more caution should be taken when designing routing policies. Meanwhile, it also allows us to identify special cases in which the cost of fairness is zero, i.e., cases in which we can achieve both fairness and optimal efficiency. Specifically, there exists a optimal solution of (2) $(a_M^\star, a_F^\star)$ that satisfies $a_M^\star = a_F^\star$ under the following two scenarios:

1) If $\bar{a} = 0$, we have no budget for special monitoring. Hence, the only feasible solution $(a_M^\star, a_N^\star) = (0,0)$ is efficiency optimal. It is also fair because $a_M^\star = a_F^\star = 0$.
2) If $\bar{a} = 1$, we have enough budget to offer special monitoring to all customers. Then the unique feasible policy is $(a_M^\star, a_N^\star) = (1,1)$, fair and efficiency optimal.

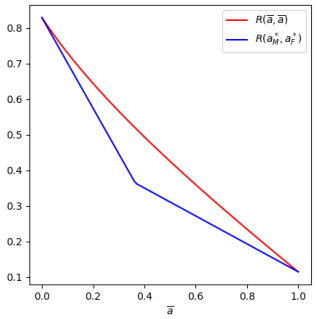
These special cases are not typical in reality. To help generate more insights into the analytical form of the cost of fairness $C(\bar{a})$, we next present numerical studies to visualize how this cost depends on the budget level $\bar{a}$ and the difference in the risk distributions.

A. Numerical Results: Impact of $\bar{a}$

Figure 2 plots the routing probabilities and the resulting system-level efficiencies of the group-aware and group-unaware policies. The parameters for the numerical experiments are given in the caption of Figure 2. We choose the Beta distributions as the risk level distribution, because (1) it has finite support on [0,1], and (2) it is easy to adjust the skewness by varying the parameters. In this parameter setting, group M has a larger treatment effect, and thus, is prioritized. This can be observed in plot (a) in Figure 2. Customers in group F will receive special monitoring only if all customers in group M receive special monitoring, i.e., when $\bar{a} \ge 0.37$. This plot illustrates the extreme unfairness between the two groups under this optimal group-aware policy. Plot (b) in Figure 2 compares the percentage of reentry $R(a_M, a_F)$ under the two policies. We can see that,



(a) The optimal policy routing probabilities.



(b) The percentages of reentry in fair and optimal policies.

Fig. 2: Model parameters are selected as follows: $\lambda_M = \lambda_F = 1$, $\mu_1 = \mu_2 = 0.1$, $\eta_1 e^{\gamma_1 p} = 0.01 \times e^{0.5p}$, $\eta_2 e^{\gamma_2 p} = 0.03 \times e^{5p}$, and $G_M = \text{Beta}(3,1), G_F = \text{Beta}(1,3)$.

except for the extreme cases $\overline{a} = 0, 1$ where all or none of customers receive the special monitoring, the optimal group-aware policy produces lower percentages of reentry than the group-unaware policy. This gap between the two lines is the cost of fairness $C(\overline{a})$. In this example, the largest cost of fairness is about 15.3%, which occurs when $\overline{a}$ is near 0.37. Note that the curve for $R(a_M^\star, a_F^\star)$ is also nonsmooth near 0.37 because that is the point where special monitoring resources are just enough for group M and the optimal allocation from (2) switches at this point.

B. Numerical Results: Impact of Risk Distribution

The risk level distribution greatly affects the cost of fairness as can be seen from the expression of $C(\overline{a})$. To illustrate this impact, we fix the risk level distribution of group M, G_M, and the overall percentage of special monitoring $\overline{a}$, then change the distribution of group F, G_F from left skewed to right skewed. Figure 3 shows the corresponding change in the cost of fairness $C(\overline{a})$. Other model parameters remain the same as in the numerical results in Figure 2.

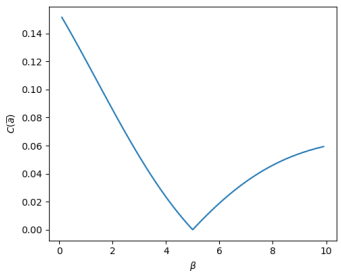


Fig. 3: The change of cost of fairness with risk distribution. Fix $G_M = \text{Beta}(5,5)$, $\overline{a} = 0.5$. Change $G_F = \text{Beta}(\beta, 10-\beta), \beta \in (0,10)$. Other model parameters are the same as in Figure 2.

We can see from Figure 3 that if $\beta = 5$, in which case $G_M = G_F$ in the distribution, then the cost of fairness is zero. When β deviates from 5, the cost of fairness increases as the deviation increases. Note that this deviation $|\beta - 5|$ reflects the discrepancy between the distributions of risk levels in the two groups. Thus, the larger discrepancy we have in the two group distributions, the larger the cost of fairness. Furthermore, we observe that the cost of fairness is larger when G_F is left skewed than right skewed because the effect of special monitoring is not significant for low-risk customers. If G_F is left skewed, most customers in group F are low risk. Therefore, the group-unaware policy is much less efficient since it offers special monitoring to half of F customers.

C. Discussion

In the example above, we obtain an intuitive understanding of how the discrepancy between distributions G_M and G_F affects the cost of fairness, and we see that in general they are positively correlated. Decision makers can more easily impose a fairness restriction when the cost of fairness is low. However, the trade-off between efficiency and fairness must be considered more carefully when the cost is high. This observation motivates quantifying the discrepancy between G_M and G_F, for example by using a stochastic distance metric like the K-L divergence or Wasserstein metric, and then developing the relationship between the distributional discrepancy and the cost of fairness. In cases where the discrepancy and resulting cost of fairness are high, decision makers would benefit from developing more complicated routing policies that also use state information, in an attempt to create more efficient routing policies that are also fair.

REFERENCES

[1] J. Morgenstern and A. Roth, Fairness in prediction and allocation, 2021.

[2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science, 2019, 366(6464): 447–453.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, (2016, May 23), Machine bias, Retrieved from ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR), 2021, 54(6): 1–35.

[5] M. Hardt, E. Price, and N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems, 2016, 29: 3315–3323.

[6] N. Master, M. I. Reiman, C. Wang, and L. M. Wein, A continuous-class queueing model with proportional hazards-based routing, Available at SSRN 3390476, 2018.

[7] T. Phung-Duc, Retrial queueing models: a survey on theory and applications, Available at arXiv:1906.09560, 2021.

[8] J. Mulvany and R. S. Randhawa, Fair scheduling of heterogeneous customer populations, Available at SSRN 3803016, 2021.

[9] A. R. Ward and M. Armony, Blind fair routing in large-scale service systems with heterogeneous customers and servers, Operations Research, 2013, 61(1): 228–243.

[10] A. Wierman, Fairness and scheduling in single server queues, Surveys in OR and Management Science, 2011, 16(1): 39–48.

[11] M. Shaked and G. Shanthikumar, Stochastic ordering, Springer-Verlag New York, 2007.