Feature-Driven Priority Queuing

Simrita Singh[†], Itai Gurvich[‡]*, Jan A. Van Mieghem[‡]

[†]Leavey School of Business at Santa Clara University, [‡]Kellogg School of Management at Northwestern University

Thursday 18th January, 2024

Traditional queuing theory assumes that customer types are known or perfectly observed, and each customer is placed in its type-specific priority queue; we call this type-driven priority queueing. We study feature-driven priority queuing where types are not perfectly observed but are inferred from observed features using a classifier. A practically appealing approach first takes an off-the-shelf classifier that predicts the type posterior and then optimizes priority queueing based on the classification probabilities. We propose instead a direct approach that optimizes the classifier to directly predict the priority queue from features.

The explicit modeling of the classifier in the queueing-system design is the novel contribution of this paper. We present an analytic model to study the optimal queue classification that minimizes queuing delay costs. We study how the optimal number of priority queues and the assignment of features to queues changes with the classifier's accuracy. In a numerical study on a data set of medical images used in digital radiology triage, optimal feature-driven priority queuing improves delay costs by up to 30-60% relative to type classification and queue optimization using state-of-the-art image classifiers, under high system utilization.

Key words: data-driven operations, priority queues, digital triage, feature-driven classification

1. Introduction

Priority queues are used in resource-constrained settings to stratify waiting time (and possibly service speed) by job priority. To ensure that critical medical images experience minimal waiting time, hospitals, imaging centers, and teleradiology companies assign each incoming medical image a priority level (e.g., routine, important, urgent, or critical) and sequence it accordingly into the work queue of the radiologist. Similarly, ecommerce websites assign priorities to customers for live (human) chat agent assignment to minimize lost sales from high-value customers with limited patience.

We consider the problem of assigning a priority level to each job (e.g., patients, customers, or X-ray images) that arrives at a service facility. Each job is of a certain type and may have to wait in a type-specific queue if the service facility is busy at its time of arrival; customers are served in order of their priority in a non-preemptive manner. A job of type-i has a type-specific service time distribution (e.g., exponential with rate μ_i) and a penalty c_i per unit of time spent waiting in queue

* The work of the second author was supported by NSF grant DIS-1935809

for service. Traditional queuing theory assumes that types are perfectly observed and prescribes some version of the celebrated $c\mu$ rule to minimize delay costs (Cox and Smith 1961). The $c\mu$ rule assigns priority to customers in decreasing order of $c_i\mu_i$. We call this type-driven priority queueing.

In practice, the type of a job may not be perfectly observed on arrival but instead must be inferred from observed "features." E.g., round, white shadows on a chest X-ray indicate the presence of lung nodule, which is the type. To minimize waiting costs, one must specify the number of priority queues N and a mapping from features to queues. The probability P_{ij} that a type-i customer is placed in priority queue j depends on this mapping. These probabilities are collected in a $T \times N$ *queue-classification matrix* P where T is the number of types.

A practically appealing "type-first" approach towards building this mapping first takes an off-the-shelf type classifier that predicts the type-probabilities and then optimizes the prioritization of jobs based on the classification probabilities. This is common practice in the setting of X-ray triage that we consider in our numerical study. We propose a "direct approach" to feature-driven priority queuing, that optimizes the classifier to directly predict the priority queue from features. In this direct approach, the matrix P, including the number of priority queues N, is endogenous to the queuing-system design. The queues and the classifier are optimized simultaneously to minimize the average waiting cost.

To partially characterize the impact of feature distributions on the priority queueing design, we construct a lower bound on average waiting cost in terms of the statistical distance between the feature distributions for the two types, and the relative values of their delay costs, service times, and arrival rates. If the statistical distance between feature distributions is small, then the value of a feature-driven classifier for prioritization is limited. We also find the lower bound when the classifier is required to be unbiased. The difference in the two lower bounds captures the loss from using a plug and apply approach with an unbiased type classifier.

We prove that the Bayes type-posterior for a feature vector is statistically sufficient. Hence, any prioritization decision based on features can be equivalently made using their type posteriors. In practice classifiers are imperfect: they do not perfectly predict the Bayes posterior, so that basing the prioritization decision solely on the classifier output (and not features) can be sub-optimal. We demonstrate how the type-first approach may be sub-optimal even when the type classifier is unbiased with expected error equal to zero. In contrast, the direct approach can recover the optimal solution, even with a noisy mis-specified model, by optimizing over the weights of the feature inputs. We demonstrate the analytical tractability of the direct approach for 2 types by showing that the optimization of a direct queue classifier for waiting cost can be formulated as a linear programming problem.

For $T = 3$ types, (1) we investigate the structure of an idealized classifier that can achieve any value satisfying constraints imposed by the statistical distances between the feature distributions, and (2) we compare the waiting costs under three stylized classifiers for 3 and 2 queues to uncover conditions when fewer queues can be better, and when is it better to assign a high (or low) priority to types with medium cost reduction per unit time ($c\mu$). Both the study designs for $T = 3$ types analytically demonstrate the mechanisms by which a classifier is optimized for priority queuing: 1) Statistical pooling: if there is high overlap in feature distributions, it is difficult to distinguish between types from features, the optimal classifier may utilize two queues instead of three; 2) Over and under prioritization of the medium type: when there are two queues, the optimal classifier chooses between grouping more of the medium type-2 with type-1 or with type-3, depending on the relative costs of the three types; 3) Sensitivity over specificity: The optimal classifier chooses to have high sensitivity (the probability that type 1 is classified as queue 1) even if the specificity (the probability that type 3 is classified as queue 2) is low.

To *quantify* the gains from feature-driven queueing and to understand the mechanisms underlying these gains, we present a combination of theoretical results, numerical examples, and computational experiments with real data. Our simulation experiments are based on 100,000 anonymized 2D chest X-ray images labeled with fourteen disease findings and made available by the National Institutes of Health (Wang et al. 2017). We collaborated with a radiologist at a large university hospital to rank different disease findings on chest X-rays and their delay costs, and used state-of-the-art deep learning-based image classifiers (Garyfallos (2019)).

We find that, relative to the type-first approach, the direct approach can reduce delay costs by $30\% - 60\%$ using state-of-the-art image classifiers under high system utilization. We observe that these gains result from the mechanisms identified by our theoretical analysis. The direct approach utilizes three queues instead of the four queues available. The direct approach chooses more over-triage of the medium delay cost type than the type-first approach. The direct approach chooses sensitivity over specificity — it correctly classifies 83% of the highest delay cost type as queue 1. However, the accuracy of correctly classifying the lowest delay cost type as the last queue is only 53%. In contrast, the type-first approach has approximately the same accuracy of correctly classifying the higest delay cost type (69%) as queue 1 and the lowest delay cost type as the last queue (71%).

Our work advocates for a change of practice for managers and clinicians: using plug-and-play AI/ML based classifiers and designing prioritization policies based only on the distribution of their output can be significantly sub-optimal. Instead, the optimization of these classifiers should be “in-house” and calibrated to a service metric like the average waiting cost. Importantly, our study of X-ray triage shows that this is a feasible undertaking even in more complex settings: significant waiting gains are possible with minimal adaptation to off-the-shelf classifiers.

2. Literature Review

Priority queues have been studied extensively since the work of Cobham (1954). Motivated by AI/ML-assisted triage in digital radiology, we focus here on prioritization when types are imperfectly observed.

Zee and Theil (1961) investigate the priority assignment of two customer types–high and low–when the type is not observable for a fraction of customers. For these customers a classifier outputs a probability that the arriving customer is of the high type. They show that it is optimal to have three priority levels (high, medium, and low) and and assign the medium priority level when the classifier probability is in a certain interval.

Beja and Sid (1975) investigate the priority assignment problem when the delay cost c and service rate μ of an arriving customer are random and drawn from a known continuous distribution. They explore how to partition the space of delay-cost and service rate into N queues to minimize the average waiting cost. When full types (both c and μ) are observed, a customer is assigned high priority if and only if the product $c\mu$ lies in a given interval. If only μ is observed but not c, then the optimal assignment is based on the product of expected cost and service rate. An analogous result holds when only cost is observed. In our model, both the delay cost and service time (the type) are unobserved but inferred through a feature-driven classifier that is endogenous to the queuing system.

Bren and Saghafian (2019) consider priority assignment in a queuing system where the type of an arriving customer is known but the service rate for each type is not. They propose a data-driven percentile optimization approach that dynamically utilizes incoming data to learn the service rate. The priority of a class might change as more data is collected about its own service rate and that of others. In our study, the cost and service rates of each type (class) are known yet the type itself is not observed, but predicted from features.

Argon and Ziya (2009) explicitly model the use of customer information (or features) for the priority assignment. Each arriving customer has a label/signal that corresponds to the probability that it is a high-type customer. With linear delay costs, the optimal prioritization policy is highest signal first. If the number of priority queues is restricted to two, then it is optimal to assign high priority to customers with signals above a threshold and assign low priority to all others. The optimal threshold decreases as the system load increases (placing more customers in the high priority queue) and converges to zero as the utilization approaches 100%. The properties of the signal (in our case this would be the output of the classifier) matter: signals that are larger in convex ordering, and thus have higher variance, result in lower waiting costs. Our novelty over their work lies in characterizing the waiting cost based on the statistical distance between the feature distributions of the two types and quantifying the impact of restricting the queue classifier to be unbiased. We compare our direct approach with prioritization based on the distribution of the classifier signal (Section 5, Argon and Ziya (2009)), and prove that the two approaches are identical when the classifier outputs the Bayes posterior that a feature is of type-1. However, when the classifier predicts the Bayes posterior with an error, basing the prioritization threshold on the distribution of the classifier output is sub-optimal.

Sun et al. (2019) study the dynamic state-dependent decision to triage customers or not, to optimize the information-delay trade-off when triage times and triage errors are non-negligible. They consider two types of customers and exogenous triage errors.

The contribution of our study is to endogenize the classifier and jointly optimize the classifier and the design of the priority queue. Earlier work considers the design of the priority queues for a given classifier that outputs a signal such as the probability of the job being of a high type. The queue design then corresponds to choosing a signal threshold above which the job is placed in the high priority queue. We propose direct queue prediction where not only the design of the queue depends on the classifier properties but also the classifier depends on the queues; it is the joint design that promises best priority prediction.

In being explicit about the classifier and its properties, our work speaks also to the growing literature on data-driven operations where historical data is used to predict the distribution of random variables central to the decision model. Much of this literature focuses on inventory management where the random variable is the demand for the product. Kleywegt et al. (2001) take a non-parameteric approach to the newsvendor problem and minimize the sample average approximation (SAA) of the cost of inventory, where the average is computed over an empirical distribution realized by a sample of past demand realizations. In similar spirit to our direct approach, Operational Statistics (Liyanage and Shanthikumar 2005) takes a parametric approach that integrates parameter estimation and optimization of the order quantity.

Features that are correlated with demand can be used for prediction. See and Sim (2010) incorporate various features such as market outlook, oil prices, trend, seasonality, cyclic variation for a multi-period inventory management problem and propose a robust optimization approach. Ban and Rudin (2019) propose distribution-free, machine-learning algorithms for predicting the optimal order quantity from demand features and show that these algorithms fare better than sequential algorithms that first estimate a feature-dependent demand distribution and then optimize the order quantity.

The data-driven management of queuing systems (Chan et al. 2021), as the one we study here, expands the intellectual landscape of data-driven operations. In our study, we predict the optimal priority from features, given a training dataset of features (e.g., chest X-rays) and corresponding true types (e.g., disease labels). We optimize the classifier to minimize the empirical cost of waiting (the 'queuing loss'). Classification to priority queues is conceptually related to the use of ranking loss functions (Cortes and Mohri (2003)) for classifier ranking tasks. But, importantly, queuing loss is a non-linear function of misrankings (misclassifications): the average wait time in a queue in a multi-class queuing system is a non-linear function of the total arrival rate into that queue (the linear sum-product of the arrival rates of all classes and the (mis)classification probabilities of the classes into that queue), and the arrival rates into the higher priority queues.

Our proposed feature-driven queueing approach drops disease prediction as an intermediary and directly "prescribes" the decisions: queue placement based on features. Direct optimization of decisions is obviously superior to separate (or silo-ed) two step optimization. While this general idea is well understood, the mechanisms through which it improves performance relative to the sequential optimization is context dependent; see Bertsimas and Kallus (2020) and (Elmachtoub and Grigas 2021) for general (context independent) frameworks for joint prediction and optimization.

In a queuing context, the joint problem has two important characteristics. The first is the aforementioned non-linearity of the waiting cost in the unknown parameters. The second is that queuing externalities prohibit local optimization for a feature and its "neighborhood": the prioritization of a job with a feature vector X can impact the wait times of all the jobs in the system - including those with feature vectors far away from the neighborhood of X. At the same time, two types with very different feature presentations may join the same priority queue. Priority queues is an unexplored context for the study of joint prediction and optimization. The *non-linear externalities*—the priority assignment of one job impacts the waiting times of other jobs—give rise to new questions about the interaction between classification/prediction and workflow design that warrant a detailed study.

3. A Model of T Noisily Observed Types put into N Priority Queues

We consider the problem of assigning a priority level for service to each job (e.g., patients, customers, or X-ray images) that arrives over time at a service facility. Each job is of a certain type[1] and may have to wait in a queue if the service facility is busy at the time of arrival. We thus consider non-preemptive service of priority queues. We assume the simple setting where type-$i \in \mathcal{T} = \{1, 2, 3, \ldots, T\}$ jobs arrive according to a Poisson process with rate λ_i for processing by a single server[2] whose processing time for type-i requests is exponentially distributed with rate μ_i.

[1] For example, if a chest X-ray shows pneumothorax (a collapsed lung), then pneumothorax is the type of chest X-ray. Similarly, a chest X-ray may be of type pneumonia, effusion, or any other disease diagnosis. If more than one disease is present then the type can be defined as the most serious disease present.

[2] In large medical centers, it is common to have a single worklist (sequence) of chest X-Rays and a dedicated radiologist to review them in that order manually. Also, in the context of radiology (especially teleradiology), it is common for images from a distributed network of hospitals or urgent care centers to be reviewed at a centralized location. A single server non-preemptive priority queuing model with independent Poisson arrivals fits rather well.

We impose natural independence assumptions between the arrival processes and service times of different types. Denoting the type-i load by $\rho_i = \lambda_i/\mu_i$, the server utilization is $\rho = \sum_{i\in\mathcal{T}} \rho_i$. We assume $\rho < 1$ to guarantee stability. A type-i job incurs a *delay cost* c_i per unit of time waiting for service. The relative values of the delay costs serve as a proxy for relative urgency. We label types, without loss of generality, so that

$$c_1\mu_1 \geq c_2\mu_2 \geq c_3\mu_3 \geq ... \geq c_T\mu_T.$$

Let N be the number of priority queues and let queue 1 have the highest priority and queue N have the lowest priority. The assignment of jobs to priority queues is evaluated on its success in stratifying speed by priority and reducing the waiting time for service for the truly urgent jobs. In this study the objective of priority assignment is minimizing the long-run waiting cost averaged over all jobs.

3.1. Noisy Type Observations and the Classification Matrix P

The premise of our paper is that the type of a job is not perfectly observable but must be inferred from imperfect signals of the type. Each arriving job is characterized by a set of observable *features* (for e.g., in radiology, each arriving image can be characterized by its geometric and textural features). We use X to denote a job's observed feature vector, and $\mathcal{X}$ to denote the feature space; for most of our analysis we take $\mathcal{X} = \mathbb{R}^p$. The features for type-$i$ jobs are drawn from a distribution F_i with density f_i, i.e., $X \sim F_i(\cdot)$ for type-$i \in \mathcal{T}$. In feature-driven prioritization, each observed feature $X \in \mathcal{X}$ is deterministically mapped to a priority queue $Q(X) \in \mathcal{N} = \{1, 2, 3 \ldots, N\}$.

To evaluate waiting costs, we must know the delay cost rate (which is type dependent) and the waiting time, which is queue specific. Therefore, the waiting cost depends on how the types are classified into priority queues. This depends on the mapping $Q(\cdot)$ from features X to queues and on the feature density $f_i(\cdot)$ for each type-i. The total effect of classification on the arrival rate of jobs to the different queues is captured by the $T \times N$ classification matrix P, where P_{ij} denotes the probability that a type-i job is placed in queue j: $\forall i \in \mathcal{T}, j \in \mathcal{N}$:

$$P_{ij} = \mathbb{P}(\text{type-}i \text{ job joins queue } j) = \int_{X\in\mathbb{R}^p} f_i(X)\mathbb{1}\{Q(X) = j\}dX.^3 \tag{1}$$

This is a stochastic matrix: $\sum_{j\in\mathcal{N}} P_{ij} = 1, \forall i \in \mathcal{T}$.

The arrivals into priority queue $j \in \mathcal{N}$ follow a Poisson process with total rate $\lambda_j(P) = \sum_{i\in\mathcal{T}} \lambda_i P_{ij}$. If $s_i = 1/\mu_i$ denotes the mean service time of type-i, then the service time of jobs in queue j is hyper-exponentially distributed with mean $\sum_{i\in\mathcal{T}} \lambda_i P_{ij} s_i/\lambda_j(P)$. The jobs wait in a

[3] We henceforth use dX to denote $dX_1 \ldots dX_p$, and use $\mathbb{1}(\mathcal{J})$ to denote a binary variable that takes values 1 and 0 when condition $\mathcal{J}$ is true or false, respectively.

$N-$class $M/G/1$ queue. The stationary waiting time, $W_j(P)$, in queue j under classification matrix P has expectation:

$$\mathbb{E}[W_j(P)] = \frac{\mathbb{E}[\mathcal{S}]}{(1-\sum_{i<j}\rho(i,P))(1-\sum_{i\leq j}\rho(i,P))} \tag{2}$$

where $\mathbb{E}[\mathcal{S}] = \sum_{i\in\mathcal{T}} \lambda_i/\mu_i^2$ is the expected residual service time and

$$\rho_j(P) = \sum_{i\in\mathcal{T}} \rho_i P_{ij} = \sum_{i\in\mathcal{T}} \lambda_i P_{ij}/\mu_i$$

is the net traffic intensity (server utilization) of priority level $j \in \mathcal{N}$; see e.g. Ross (1996, Chapter 5). The mean stationary delay cost (the average waiting cost) depends on the classification matrix P through

$$\mathcal{C}(P) = \sum_{i\in\mathcal{T}} \lambda_i c_i \sum_{j\in\mathcal{N}} P_{ij}\mathbb{E}[W_j(P)] \tag{3}$$

The average waiting cost $\mathcal{C}(P)$ is the average urgency (delay cost) weighted waiting time and captures the speed stratification by priority.

3.2. Queueing Externalities

The average waiting cost $\mathcal{C}(P)$ reflects the externalities present in queueing systems:

1. The misclassification of a single job impacts the waiting time of all the jobs present (and arriving later) in its and lower priority queues in that busy period.
2. The waiting time for a misclassified job not only depends on that job but also on the characteristics of other jobs that are already ahead or may get ahead of that job in queue.

Example 1 (Higher Accuracy does not imply Lower Waiting Cost) *Consider three types such that $c_1 = 25, c_2 = 2, c_3 = 1, \lambda_1 = \lambda_2 = \lambda_3 = 1, \mu_1 = \mu_2 = \mu_3 = 3.2$. Suppose there are three priority queues ($N = 3$), and two classifiers with classification matrices:*

$$P_1 = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } P_2 = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.25 & 0.50 & 0.25 \\ 0 & 0 & 1 \end{bmatrix}.$$

While P_1 predicts types better, P_1 has an average waiting cost of 75.46, exceeding the average waiting cost of 69.91 with P_2.

Under P_1, the average lengths of queues 1,2, and 3 are 0.22, 0.99, and 12.86, respectively. Whereas, under P_2, the average lengths of queues 1,2, and 3 are 0.34, 0.54, and 13.86, respectively. The 20% of the very costly type 1 images that end up in queue 3 have fewer images ahead of them (in queues 1 and 2) under P_2 (0.88) as compared to P_1 (1.21).

Problem Statement: To find the optimal number of queues N^* and the optimal classification matrix P^* that together minimize the average waiting cost $\mathcal{C}(P)$. We consider two approaches to priority queue classification—*direct queue classification* ("direct"), where the classifier predicts the priority queue directly from features and is optimized to minimize the waiting cost, and *type classification and queue optimization* ("type-first") where the classifier is optimized to predict type probabilities and the allocation of predicted type probabilities to queues is optimized to minimize the waiting cost. In the remainder of the paper, we

- Provide performance bounds on the average waiting cost for two types by using the total variation distance between the feature distributions. We demonstrate that well-separable feature distributions across types result in lower waiting costs.
- Show that the unbiasedness of a queue classifier is sub-optimal. Thus, employing off-the-shelf type classifiers directly for priority queueing may not be the most effective strategy.
- Demonstrate that the direct and type-first approaches are equivalent when the type classifier is "perfect," i.e., when it outputs the Bayes' probability of a type given a feature. However, if the classifier is imperfect, the direct approach outperforms the type-first approach.
- Show the analytical tractability of the direct approach by formulating the optimization of a linear classifier for minimizing average waiting cost for two types as a linear programming problem.
- Analytically optimize the classification matrix in a three-type setting by considering the total variation distances between the feature distributions.
- Analyze the performance of the direct and type-first approaches using a real-life Chest X-ray 2D image dataset.

4. Analysis, Results, and Bounds

When types are perfectly observed, priority queues are assigned based on observed types rather than features and the classification (assignment) matrix is binary. We denote the binary assignment matrix by $I_{T\times N}$:

$$I_{T\times N}(i,j) = \mathbb{1}\{\text{perfectly observed type-}i\text{ is assigned to priority queue }j\}, \tag{4}$$

and the corresponding cost under perfect type information by $\mathcal{C}(I_{T\times N})$.

With perfectly observed types, the $c\mu$ rule guarantees that it is never optimal to have $N > T$. For any given number of queues $N \leq T$, denote an optimal perfect-information classification matrix that yields minimal cost by $I^*_{T\times N}$:

$$I^*_{T\times N} = \arg\min_{I_{T\times N}} \mathcal{C}(I_{T\times N}) \tag{5}$$

Proposition 1 (Perfect Type Information) *Given T types such that $c_1\mu_1 > c_2\mu_2 > \ldots > c_T\mu_T$, and $N \leq T$ priority queues, then* *(a) There exist $N+1$ indices $0=i_0<i_1<i_2<\ldots\leq i_{N-1}<i_N=T$ such that*

$$I^*_{T\times N}(i,j)=1 \text{ iff } i_{j-1}<i\leq i_j.$$

(b) $\mathcal{C}(I^*_{T\times N})$ is a decreasing function of N, so that it is optimal to have $N=T$ queues with the trivial identity classification matrix, $I^*_{T\times T}(i,i)=i, \forall i\in\mathcal{T}$).

Beja and Sid (1975) study the partition of continuous types. The proof for discrete types is simple and appears in the Appendix. With perfect observation, and in the absence of a practical restriction on the number N of queues, it is optimal then to have as many priority levels as types. The first item implies that it is never optimal to assign a lower priority to a type that ranks higher on $c\mu$.

When job types are not perfectly observed, the jobs are assigned to priority queues using a classifier. A classifier (e.g., a decision tree, a support vector machine) separates jobs into several categories (prediction classes) based on their observed features.

Proposition 2 (Classification errors) *Given a deterministic classifier that maps features $X\in\mathcal{X}$ of T types into K prediction classes that are grouped into $N\leq K$ priority queues, then*

(a) The optimal number of priority queues N^* equals K. Equivalently, it is optimal to consider each prediction class as a priority queue.

(b) Let P denote the classification matrix induced my the classification of features of T types to N queues. It is optimal to prioritize queue m over queue n ($m,n\in\{1,\ldots,N\}$) if and only if

$$\frac{\sum_{i\in\mathcal{T}}P_{im}\lambda_i c_i}{\sum_{i\in\mathcal{T}}P_{im}\rho_i}\geq\frac{\sum_{i\in\mathcal{T}}P_{in}\lambda_i c_i}{\sum_{i\in\mathcal{T}}P_{in}\rho_i}.$$

Part (a) of the above proposition states that post deterministic classification, it cannot be optimal to further group classes into coarser priority queues. It does not comment on the optimal number of classes but only on the optimal number of priority queues ex-post classification.

When classifying T types into N queues, the waiting cost (3) can be written equivalently as the waiting cost of a multi-class system with N-priority queues where queue $m\in\{1,2,\cdots,N\}$ has arrival rate $\lambda_m(P)$, service rate $\mu_m(P)$, and delay cost rate $c_m(P)$:

$$\lambda_m(P)=\sum_{k\in\mathcal{T}}\lambda_k P_{km},\quad \frac{1}{\mu_m(P)}=\sum_{i\in\mathcal{T}}\frac{1}{\mu_i}\frac{\lambda_i P_{im}}{\lambda_m(P)}=\frac{\sum_{i\in\mathcal{T}}\rho_i P_{im}}{\lambda_m(P)}\quad\text{and}\quad c_m(P)=\frac{\sum_{i\in\mathcal{T}}c_i\lambda_i P_{im}}{\lambda_m(P)}.$$

If $N=T$, then the classification matrix P that minimizes the waiting cost is the identity matrix, corresponding to perfect type observation giving an upper bound on performance.

4.1. Feature-Driven Priority Queueing: Performance Bound ($T = N = 2$)

In this section, we construct lower bounds on the waiting costs achievable through feature-driven classification. The lower bounds are tight and lead to interesting revelations. First, they confirm the intuition that feature distributions that are well separable across types result in lower waiting costs. Second, they show that the unbiasedness of a queue classifier is sub-optimal. This second result is counterintuitive since most type classification algorithms are optimized to reduce bias.

For $T = N = 2$, the average waiting cost for a classification matrix P equals

$$\mathcal{C}(P) = \mathcal{C}(P^{FIFO}) - \frac{(c_1\mu_1 - c_2\mu_2)\rho_1\rho_2\mathbb{E}[\mathcal{S}]}{1-\rho} \cdot \frac{P_{11} - P_{21}}{1 - \rho_1 P_{11} - \rho_2 P_{21}}, \text{ where } P = \begin{bmatrix} P_{11}, \, 1 - P_{11} \\ P_{21}, \, 1 - P_{21} \end{bmatrix}. \quad (6)$$

$\mathcal{C}(P^{FIFO})$ is the average waiting cost under a single queue first in first out (FIFO) regime:

$$P^{FIFO} = \begin{bmatrix} 1, \, 0 \\ 1, \, 0 \end{bmatrix}, \; \mathcal{C}(P^{FIFO}) = \frac{(c_1\mu_1\rho_1 + c_2\mu_2\rho_2)\mathbb{E}[\mathcal{S}]}{1-\rho}.$$

The average waiting cost $\mathcal{C}(P)$ depends on the classification probabilities of types 1 and 2 into priority queue 1, which depend on the feature distributions for the two types. For a classifier $Q(\cdot)$, let $A = \{X : Q(X) = 1, \; X \in \mathbb{R}^p\}$ be the set of features that are mapped to priority queue 1. From (1), the classification probabilities induced by the classifier $Q(\cdot)$ satisfy

$$P_{11} = \int_{X \in A} f_1(X)dX, \; P_{21} = \int_{X \in A} f_2(X)dX, \; A \subseteq \mathbb{R}^p. \quad (7)$$

If the feature distributions ($F_1(\cdot)$ and $F_2(\cdot)$) of the types are identical, then for all possible classifiers $Q(\cdot)$, $P_{11} = P_{21}$, and $\mathcal{C}(P) = \mathcal{C}(P^{FIFO})$. In contrast, if the feature distributions are well separated, a greater wait time stratification may be possible. Below, we formalize this intuition. At the minimum, P_{11} and P_{21} satisfy the below constraints:

$$|P_{11} - P_{21}| \leq \sup_A \int_{X \in A} (f_1(X) - f_2(X))dX = \frac{\int_{X \in \mathbb{R}^p} |f_1(X) - f_2(X)| dX}{2} = TV(F_1, F_2) \leq 1 \quad (8)$$

$TV(F_1, F_2)$ is the *total variation distance* (Bhattacharyya et al. (2023)) between $F_1(\cdot)$ and $F_2(\cdot)$.

Therefore, the below minimization problem provides a lower bound on $\mathcal{C}(P)$:

$$\min_{P_{11}, P_{21}} \frac{(c_1\mu_1\rho_1 + c_2\mu_2\rho_2)\mathbb{E}[\mathcal{S}]}{1-\rho} - \frac{(c_1\mu_1 - c_2\mu_2)\mathbb{E}[\mathcal{S}]\rho_1\rho_2}{1-\rho} \cdot \frac{P_{11} - P_{21}}{1 - \rho_1 P_{11} - \rho_2 P_{21}}$$

subject to:

$$|P_{11} - P_{21}| \leq TV(F_1, F_2)$$

$$0 \leq P_{11} \leq 1$$

$$0 \leq P_{21} \leq 1$$

The optimal solution to the above minimization problem is $P^*_{11} = 1,\ P^*_{21} = 1 - TV(F_1, F_2)$. Restricting P to be unbiased imposes an additional constraint $p_1 P_{11} + p_2 P_{21} = p_1$, leading to the optimal solution $P^*_{11} = p_1 + p_2 TV(F_1, F_2),\ P^*_{21} = p_1(1 - TV(F_1, F_2))$.

Proposition 3 (Lower Bound)

1. *For any classification matrix P induced by a classifier that maps features to queues,*

$$\mathcal{C}(P) \geq \frac{(c_1\mu_1\rho_1 + c_2\mu_2\rho_2)\mathbb{E}[S]}{1-\rho} - \frac{(c_1\mu_1 - c_2\mu_2)\mathbb{E}[S]\rho_1\rho_2}{1-\rho}\frac{TV(F_1,F_2)}{1-\rho+\rho_2 TV(F_1,F_2)}$$

2. *For an unbiased classification matrix P that satisfies $p_1 P_{11} + p_2 P_{21} = p_1$,*

$$\mathcal{C}(P) \geq \frac{(c_1\mu_1\rho_1 + c_2\mu_2\rho_2)\mathbb{E}[S]}{1-\rho} - \frac{(c_1\mu_1 - c_2\mu_2)\mathbb{E}[S]\rho_1\rho_2}{1-\rho}\frac{TV(F_1,F_2)}{1-\rho p_1 + (\rho_2 p_1 - \rho_1 p_2) TV(F_1,F_2)}$$

Since $\frac{d}{dx}(\frac{x}{1-\rho+\rho_2 x}) = \frac{1-\rho}{(1-\rho+\rho_2 x)^2} > 0$, the lower bound in Proposition 3 decreases as $TV(F_1, F_2)$ increases i.e., the types are more distinguishable from features. Further, the bound is tight: it holds with equality for $TV(F_1, F_2) = 0$ (the feature distributions are identical and hence no classifier can do better than FIFO) as well as for $TV(F_1, F_2) = 1$ (feature distributions are completely separable and $P^* = I_{2\times 2}$).

The difference between the lower bounds in (1) and (2) is

$$-\frac{(c_1\mu_1 - c_2\mu_2)\mathbb{E}[S]\rho_1\rho_2\rho p_2}{1-\rho}\frac{TV(F_1,F_2)(1-TV(F_1,F_2))}{(1-\rho+\rho_2 TV(F_1,F_2))(1-\rho p_1 + (\rho_2 p_1 - \rho_1 p_2)TV(F_1,F_2))} \leq 0.$$

While the difference is negative for $0 < TV(F_1, F_2) < 1$, it is zero for $TV(F_1, F_2) = 0$ and $TV(F_1, F_2) = 1$. When $TV(F_1, F_2) = 1$, the feature distributions are perfectly separable, and $P^* = I_{2\times 2}$ (which is unbiased). When $TV(F_1, F_2) = 0$, both the lower bounds are equal to the average waiting cost under FIFO.

The reason why the unbiased classifier fares worse (has a higher lower bound on the average waiting cost) is because it does not allow for very high values of P_{11}. For example, because of unbiasedness, the solution $P_{11} = 1$ implies $P_{21} = 0$ which violates the constraint (8) when there is an overlap in feature distributions i.e., $TV(F_1, F_2) < 1$.

4.2. Direct vs Type-First

We compare two approaches to priority queue classification:

1. *Direct Queue Classification* ("direct"), where the classifier predicts the priority queue directly from features and is optimized to minimize the waiting cost.
2. *Type Classification and Queue Optimization* ("type-first") where the classifier is optimized to predict type probabilities and the allocation of predicted type probabilities to queues is optimized to minimize the waiting cost.

We show that both the approaches (direct and type-first) are equivalent if the classifier is "perfect"(meaning it outputs the Bayes' probability of a type given a feature), but if the classifier is imperfect and outputs the Bayes' probability with some error, then the type-first approach is sub-optimal, even when the classifier is unbiased with expected error equal to zero. In Section 4.1, we showed that an unbiased but imperfect type classifier may lead to under-triage where less than optimal number of type-1 and type-2 jobs are classified into queue 1. Here, we demonstrate the same phenomenon in a different setting.

For any feature X, let $\mathbb{P}^T(X)$ be a T-dimensional vector with i^{th} entry $\mathbb{P}(i|X)$, the posterior probability that the true type is i; where,

$$\mathbb{P}(i|X) = \frac{p_i f_i(X)}{\sum_{i\in\mathcal{T}} p_i f_i(X)},\ i \in \mathcal{T} \tag{9}$$

Lemma 1 *For observed feature $X \in \mathcal{X}$, the Bayes type posterior $\mathbb{P}^T(X)$ is statistically sufficient for the set of probability measures on $(\mathcal{X}, S)$ corresponding to the feature densities $\{f_i(X); i \in \mathcal{T}\}$.*

In other words, the feature X does not provide more information about the underling type than the Bayes type posterior $\mathbb{P}^T(X)$.

Lemma 2 *Consider a feature-driven priority mapping such that observed feature X is mapped to queue j with probability $q_j(X)$, $j \in \mathcal{N}$. Then, there exists an equivalent type-posterior based priority mapping such that type posterior $\mathbb{P}^T(X)$ is mapped to queue j with probability $q_j^t(\mathbb{P}^T(X))$:*

$$q_j^t(\mathbb{P}^T(X)) = E[q_j(X)|\mathbb{P}^T(X)]\ \forall j \in \mathcal{N}$$

$$P_{ij} = E[q_j^t(\mathbb{P}^T(X))|X \sim F_i] = E[q_j(X)|X \sim F_i]\ \forall i \in \mathcal{T}, j \in \mathcal{N}$$

where P_{ij} is the probability that type-i is mapped to queue j $\forall i \in \mathcal{T},\ j \in \mathcal{N}$.

Lemma 2 states that any classification matrix, and the corresponding average waiting cost, realized by a mapping of features to priority queues can be equivalently realized by a probabilistic mapping of the Bayes type posterior to priority queues.

Next we investigate the scenario where the the type classifier makes errors in predicting the Bayes type posterior. Consider $T = N = 2$, suppose a classifier predicts the type 1 posterior for feature X as $\mathbb{Z}(X)$. Let the distribution of the classifier signal be such that it has density $b(z)$ at $\mathbb{Z}(X) = z$, and the classifier be unbiased i.e., $\int_0^1 zb(z)dz = p_1$. Argon and Ziya (2009) show that the optimal deterministic prioritization policy that minimizes the expected waiting cost conditioned on $b(\cdot)$ is a threshold policy such that all values of $\mathbb{Z}(X) \in [t, 1]$ are assigned to queue 1 and values in $[0, t)$ are assigned to queue 2. For a threshold t, the classification probabilities are estimated as:

$$\hat{P}_{11}(t) = p_1^{-1} \int_t^1 zb(z)dz;\ \hat{P}_{21}(t) = p_2^{-1} \int_t^1 (1 - z)b(z)dz. \tag{10}$$

From (6), the threshold $\bar{t}$ that minimizes $\mathcal{C}(\hat{P}(t))$ satisfies

$$\bar{t} = \arg\min_t \mathcal{C}(\hat{P}(t)) = argmax_t \frac{\hat{P}_{11}(t) - \hat{P}_{21}(t)}{1-\rho_1\hat{P}_{11}(t) - \rho_2\hat{P}_{21}(t)} = argmax_t \frac{1-\frac{1-\rho\hat{P}_{11}(t)}{1-\rho\hat{P}_{21}(t)}}{\rho_2+\rho_1\frac{1-\rho\hat{P}_{11}(t)}{1-\rho\hat{P}_{21}(t)}} = \arg\min_t \frac{1-\rho\hat{P}_{11}(t)}{1-\rho\hat{P}_{21}(t)} \tag{11}$$

($\because \frac{d}{du}(\frac{1-u}{\rho_2+\rho_1 u}) = -\frac{\rho_1+\rho_2}{(\rho_1+\rho_2 u)^2} < 0$ and $\frac{1-u}{\rho_2+\rho_1 u}$ is decreasing in u). Further, we can show that $\bar{t}$ satisfies

$$\frac{\frac{d\hat{P}_{11}(t)}{dt}}{\frac{d\hat{P}_{21}(t)}{dt}} = \frac{1-\rho\hat{P}_{11}(t)}{1-\rho\hat{P}_{21}(t)} \implies$$

$$\frac{\bar{t}}{1-\bar{t}} = \frac{p_1}{p_2}\frac{1-\rho\hat{P}_{11}(\bar{t})}{1-\rho\hat{P}_{21}(\bar{t})} = \frac{p_1-\rho\int_{\bar{t}}^1 zb(z)dz}{p_2-\rho\int_{\bar{t}}^1 (1-z)b(z)}. \tag{12}$$

Let $\mathbb{P}(1|\mathbb{Z}(X)=z)$ be the probability that the job is of type-1 conditioned on classifier signal $\mathbb{Z}(X)=z$, and $P_{11}(t)$ and $P_{22}(t)$ be the classification probabilities *realized* for threshold t:

$$P_{11}(t) = \mathbb{P}(z\in[t,1]|X\sim F_1) = \int_t^1 \frac{\mathbb{P}(1|\mathbb{Z}(X)=z)b(z)}{\mathbb{P}(X\sim F_1)}dz = p_1^{-1}\int_t^1 \mathbb{P}(1|\mathbb{Z}(X)=z)b(z)dz \tag{13}$$
$$P_{21}(t) = \mathbb{P}(z\in[t,1]|X\sim F_2) = \int_t^1 \frac{\mathbb{P}(X\sim F_2|z)b(z)}{\mathbb{P}(X\sim F_2)}dz = p_2^{-1}\int_t^1 (1-\mathbb{P}(1|\mathbb{Z}(X)=z))b(z)dz$$

The optimal threshold t^* that minimizes the true waiting cost, i.e., $t^* = \arg\min_t \mathcal{C}(P(t^*))$, satisfies:

$$t^* = \arg\min_t \frac{1-\rho P_{11}(t)}{1-\rho P_{21}(t)}.$$

Further, t^* satisfies

$$\frac{\frac{dP_{11}(t)}{dt}}{\frac{dP_{21}(t)}{dt}} = \frac{1-\rho P_{11}(t)}{1-\rho P_{21}(t)} \implies$$

$$\frac{\mathbb{P}(1|\mathbb{Z}(X)=t^*)}{1-\mathbb{P}(1|\mathbb{Z}(X)=t^*)} = \frac{p_1}{p_2}\frac{1-\rho P_{11}(t^*)}{1-\rho P_{21}(t^*)} = \frac{p_1-\rho\int_{t^*}^1 \mathbb{P}(1|\mathbb{Z}(X)=z)b(z)dz}{p_2-\rho\int_{t^*}^1 (1-\mathbb{P}(1|\mathbb{Z}(X)=z))b(z)dz}. \tag{14}$$

1. When $\mathbb{Z}(X)=\mathbb{P}(1|X)$ (the Bayes probability), then $\mathbb{P}(1|\mathbb{Z}(X)=z)=z$, and the estimated classification probabilities (10) are equal to the realized classification probabilities (13), i.e., $P_{11}(t)=\hat{P}_{11}(t),\ P_{21}(t)=\hat{P}_{21}(t)$. The threshold obtained by (12) is equal to t^*, and gives the lowest possible waiting cost:

$$\mathcal{C}(P(t=t^*)) = \frac{(c_1\mu_1\rho_1+c_2\mu_2\rho_2)}{1-\rho} - \frac{(c_1\mu_1-c_2\mu_2)\rho_1\rho_2}{1-\rho}\frac{(p_1-t^*)}{\rho_2 p_1+(\rho p_2-\rho_2)t^*}$$

2. When $\mathbb{Z}(X)\neq\mathbb{P}(1|X)$, then $\mathbb{P}(1|\mathbb{Z}(X)=z)\neq z$, and the estimated and realized classification probabilities are not identical $(\hat{P}_{11}(t)\neq P_{11}(t), \hat{P}_{21}(t)\neq P_{21}(t))$, and (12) does not give the optimal solution t^*. Therefore,

$$\mathcal{C}(P(\bar{t})) \geq \mathcal{C}(P(t^*))$$

Next, we characterize a scenario where a classifier that under-estimates small values of the Bayes posterior and over-estimates large values of the Bayes posterior leads to under-triage such that fewer than the optimal number of type-1 and type-2 jobs are classified into queue 1.

Lemma 3 *Let the classifier signal for feature X be $\mathbb{Z}(X) = \mathbb{P}(1|X) + \zeta(X)$, where $\mathbb{P}(1|X)$ is the Bayes probability of type-1 for feature X, and $\zeta(X)$ is the error in estimating $\mathbb{P}(1|X)$. Let $\zeta(X)$ be such that there exists a mapping $\epsilon(\cdot)$ such that $\zeta(X) = \epsilon(\mathbb{Z}(X))$, i.e.,*

$$\mathbb{P}(1|\mathbb{Z}(X) = z) = z - \epsilon(z)$$

Let $b(z)$ be the probability density of $\mathbb{Z}(X)$ at $\mathbb{Z}(X) = z \in [0,1]$ such that $\epsilon(z)$ is unbiased i.e., $\int_0^1 \epsilon(z)b(z)dz = 0$, and higher values of the classifier signal correspond to higher values of the Bayes posterior, i.e., $z - \epsilon(z)$ is increasing in z. If there exists a threshold $t_u \leq \bar{t}$ (defined in 12) such that $0 \leq \epsilon(z) \leq z$ for all $z \leq t_u$ and $z - 1 \leq \epsilon(z) < 0$ for all $z > t_u$, then prioritization based on z leads to under-triage.

From (14), the optimal threshold (t^*) that minimizes the realized waiting cost satisfies

$$\frac{t^* - \epsilon(t^*)}{1 - (t^* - \epsilon(t^*))} = \frac{(p_1 - \rho\int_{t^*}^1 (z - \epsilon(z))b(z)dz)}{(p_2 - \rho\int_{t^*}^1 (1 - z + \epsilon(z))b(z)dz)} \leq \frac{(p_1 - \rho\int_{\bar{t}}^1 (z - \epsilon(z))b(z)dz)}{(p_2 - \rho\int_{\bar{t}}^1 (1 - z + \epsilon(z))b(z)dz)} \tag{15}$$

($\because t^* = \arg\min_t \frac{1-\rho P_{11}(t)}{1-\rho P_{11}(t)}$). Under the assumptions of Lemma 3, $\int_t^1 \epsilon(z)b(z)dz \leq 0$ for $t > 0$, $\epsilon(\bar{t}) \leq 0$, and

$$\frac{(p_1 - \rho\int_{\bar{t}}^1 zb(z)dz + \rho\int_{\bar{t}}^1 \epsilon(z)b(z)dz)}{(p_2 - \rho\int_{\bar{t}}^1 (1-z)b(z)dz - \rho\int_{\bar{t}}^1 \epsilon(z)b(z)dz)} \leq \frac{(p_1 - \rho\int_{\bar{t}}^1 zb(z)dz)}{(p_2 - \rho\int_{\bar{t}}^1 (1-z)b(z)dz)} = \frac{\bar{t}}{1-\bar{t}} \leq \frac{\bar{t} - \epsilon(\bar{t})}{1 - \bar{t} + \epsilon(\bar{t})} \tag{16}$$

From (15) and (16), we have $t^* - \epsilon(t^*) \leq \bar{t} - \epsilon(\bar{t})$. Since, $z - \epsilon(z)$ is increasing in z, we have $t^* \leq \bar{t}$. Therefore, the probabilities that types 1 and 2 get classified into queue 1, i.e., $\int_{\bar{t}}^1 (z - \epsilon(z))b(z)dz$ and $\int_{\bar{t}}^1 (1 - z + \epsilon(z))b(z)dz$ are smaller than the optimal ($\int_{t^*}^1 (z - \epsilon(z))b(z)dz$ and $\int_{t^*}^1 (1 - z + \epsilon(z))b(z)dz$).

Next, we illustrate a scenario where no threshold-based prioritization of a noisy classifier signal can recover the minimum cost solution. However, even with the same prediction model, the direct approach can asymptotically recover the minimum cost solution by optimizing the feature weights.

Example 2 *Suppose the feature X is one-dimensional. For type-1, X is uniformly distributed over $[0, 2\sigma]$ where $\sigma \in [\frac{1}{2}, 1]$. For type-2, X is uniformly over $[-1, 1]$:*

$$f_1(X) = \frac{1}{2\sigma}\mathbb{1}(X \in [0, 2\sigma]),\ f_2(X) = \frac{1}{2}\mathbb{1}(X \in [-1, 1]);\ \textit{(see Figure 1).}$$

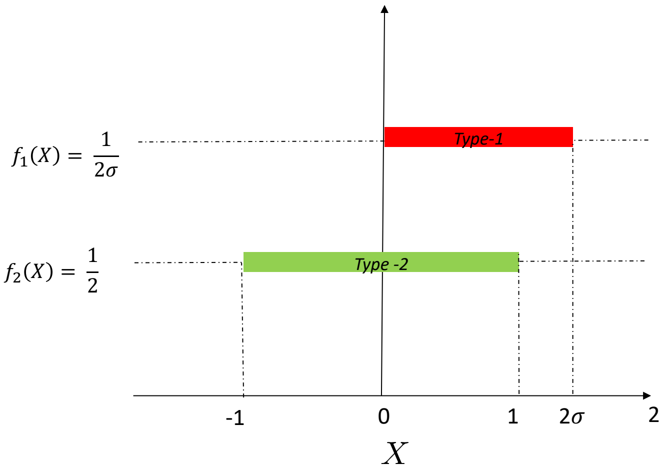

Figure 1 **Feature distributions for the two types in Example 2**

$$\textit{Then: } F_1(x) = \begin{cases} 0 & \textit{if } x < 0; \\ \frac{x}{2\sigma} & \textit{if } 0 \leq x \leq 2\sigma; \\ 1 & \textit{if } x > 2\sigma \end{cases}, \; F_2(x) = \begin{cases} 0 & \textit{if } x < -1; \\ \frac{(x+1)}{2} & \textit{if } -1 \leq x \leq 1; \\ 1 & \textit{if } x > 1 \end{cases}$$

Using an interchange argument, it can be shown that the optimal solution is to assign $X \geq K$ to queue 1 where $K \in [0,1]$. The optimal K^* as per (6)) is:

$$K^* = \arg\min_K \left[\frac{1-\rho P_{11}(K)}{1-\rho P_{21}(K)} = \frac{1-\rho+\rho F_1(K)}{1-\rho+\rho F_2(K)} \right] = 0.$$

The Bayes probability that feature X is of type 1 is

$$\mathbb{P}(1|X) = \begin{cases} 0 & \text{if } X < 0; \\ \frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}} & \text{if } 0 \leq X \leq 1; \\ 1 & \text{if } X > 1 \end{cases}$$

Now, suppose $p_1 \leq p_2$, and a classifier noisily predicts the Bayes probability of type-1:

$$\mathbb{Z}(X) = \begin{cases} 0 & \text{if } X < -1; \\ -\frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}} X & \text{if } -1 \leq X \leq 0; \\ \frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}} - \frac{\frac{p1}{2\sigma}X}{\frac{p1}{2\sigma}+\frac{p2}{2}} & \text{if } 0 \leq X \leq 1; \\ 1 & \text{if } X > 1; \end{cases}$$

Type Classification and Queue Optimization: Suppose we assign $\mathbb{Z}(X) \geq t$ to queue 1, where $t \in (0,1)$. Now,

$$\mathbb{Z}(X) \geq t \iff \begin{cases} \left[-\frac{t}{\frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}}}, 0 \right] \cup \left[0, \frac{\frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}} - t}{\frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}}} \right] & \text{if } t \leq \frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}} \\ X > 1 & \text{if } t > \frac{\frac{p1}{2\sigma}}{\frac{p1}{2\sigma}+\frac{p2}{2}} \end{cases}$$

There does not exist a threshold t such that $\mathbb{Z}(X) \geq t \iff X \geq 0$.

Direct: Maps feature X to queue 1 if $z(\beta_0 + \beta_1 X) \geq 1 - z(\beta_0 + \beta_1 X)$, where the weights β are optimized to minimize the waiting cost. Let $\beta_0 = 1, \beta_1 = M$. We show that the direct approach recovers the optimal policy as $M \to \infty$.

$$z(MX+1) = \begin{cases} 0 & \text{if } MX+1 < -1; \\ -\frac{\frac{p_1}{2\sigma}}{\frac{p_1}{2\sigma}+\frac{p_2}{2}}(MX+1) & \text{if } -1 \leq MX+1 \leq 0; \\ \frac{\frac{p_1}{2\sigma}}{\frac{p_1}{2\sigma}+\frac{p_2}{2}} - \frac{\frac{p_1}{2\sigma}}{\frac{p_1}{2\sigma}+\frac{p_2}{2}}(MX+1) & \text{if } 0 \leq MX+1 \leq 1; \\ 1 & \text{if } MX+1 > 1; \end{cases}$$

$$\lim_{M\to\infty} \mathbb{P}(MX+1 > 1; X > 0) = 1 \textit{ and } \lim_{M\to\infty} \mathbb{P}(MX+1 \geq -1; X < 0) = 0$$

4.3. Optimizing a simple Direct Queue Linear-Classifier

Section 4.2 presents the advantages of direct queue classification. Here, we demonstrate the analytical tractability of optimizing the direct queue classifier. For $T = N = 2$, we show that optimizing a linear classifier for minimizing the average waiting cost can be formulated as a linear programming problem. Let $Q(X) \in \{1, 2\}$ be the priority queue for feature X. We consider a linear classifier that predicts the probability that feature X is of priority queue 1.

$$\mathbb{P}(Q(X) = 1) = \beta' X \tag{17}$$

The model is trained on historic data of jobs' observed features X and corresponding type label t (as classified by a human expert). The training data set S thus consists of s pairs $(X_r, t_r), r \in [s]$ where the true type-t_r is a binary variable (1 for type 1 and 0 for type 2). [4] The empirical estimate of the classification matrix is

$$\hat{P}(\beta) = \begin{bmatrix} \frac{\sum_{r=1}^{s} t_r \beta' X_r}{\sum_{r=1}^{s} t_r} & 1 - \frac{\sum_{r=1}^{s} t_r \beta' X_r}{\sum_{r=1}^{s} t_r} \\ \frac{\sum_{r=1}^{s} (1-t_r) \beta' X_r}{\sum_{r=1}^{s} 1-t_r} & 1 - \frac{\sum_{r=1}^{s} (1-t_r) \beta' X_r}{\sum_{r=1}^{s} 1-t_r} \end{bmatrix} \tag{18}$$

The optimal $\beta = \beta^*$ that minimizes the empirical estimate for the waiting cost satisfies:

$$\beta^* = argmax_\beta \frac{\hat{P}_{11}(\beta) - \hat{P}_{21}(\beta)}{1 - \rho_1 \hat{P}_{11}(\beta) - \rho_2 \hat{P}_{21}(\beta)} = argmax_\beta \frac{1 - \rho \hat{P}_{21}(\beta)}{1 - \rho \hat{P}_{11}(\beta)} \; (see\ (6), (11)) \tag{19}$$

For $\hat{P}_{11}(\beta)$ and $\hat{P}_{21}(\beta)$ to be valid probabilities, we restrict β to satisfy:

$$\begin{aligned} 0 \leq \frac{\sum_{r=1}^{s} t_r \beta' X_r}{\sum_{r=1}^{s} t_r} \leq 1 \\ 0 \leq \frac{\sum_{r=1}^{s} (1-t_r) \beta' X_r}{\sum_{r=1}^{s} 1-t_r} \leq 1 \end{aligned} \tag{20}$$

[4] We normalize all features so that they have zero mean and unit standard deviation by replacing each X_r^j with $\frac{X_r^j - mean(\{X_r^j; 1 \leq r \leq s\})}{sd(\{X_r^j; 1 \leq r \leq s\})}$, $\forall 1 \leq j \leq p$).

Lemma 4 *The optimization problem (*19*) subject to (20) can be reformulated as the below linear programming problem:*

$$
\begin{aligned}
& maximize_{w_0,w}[w_0 - \rho w' u] \\
& subject\ to: \\
& Aw \leq bw_0 \\
& w_0 - \rho w' v = 1 \\
& w_0 \geq 0
\end{aligned}
$$

where $u = \frac{\sum_{r=1}^{s}(1-t_r)X_r}{\sum_{r=1}^{s}1-t_r}, v = \frac{\sum_{r=1}^{s}t_r X_r}{\sum_{r=1}^{s}t_r},\ A = [u, v, -u, -v]', b = [1, 1, 0, 0]'$

5. Analytic Example with 3 Types

For $T = 3$ types, we investigate the structure of an idealized classifier P that can achieve any value satisfying constraints imposed by the statistical distances between the feature distributions. We also compare the waiting costs under three stylized classifiers for 3 and 2 queues to uncover conditions when fewer queues can be better, and when is it better to assign a high (or low) priority to types with medium cost reduction per unit time ($c\mu$).

Regardless of design (direct or type-first), the total waiting cost is determined by the classification matrix P. Each classification matrix P is defined by the number of queues N and the probabilities of types getting classified into the N queues.

The choice of the number of queues N trades-off two competing forces. On one hand, the greater the number of queues, the smaller the waiting cost; recall that greater stratification up to $N = T$ is optimal under perfect information. On the other hand, due to overlap in feature distributions, the classifier learning, and hence the queue assignment accuracy, may drop when the number of queues equals the number of types. Due to the overlap in feature distributions, it is not possible to increase one classification probability without increasing the other, and, at the minimum, the classification probabilities satisfy the below constraints:

$$P_{ij} - P_{mj} \leq \frac{1}{2}\int_{X \in \mathbb{R}^p} |f_i(X) - f_m(X)| dX = TV(F_i, F_m)\ \forall i, m \in \mathcal{T}, j \in \mathcal{N} \tag{21}$$

Further, using triangle inequalities, we can show that $\forall i, k, m \in \mathcal{T}$:

$$TV(F_i, F_k) + TV(F_k, F_m) \leq TV(F_i, F_m)\ and\ |TV(F_i, F_j) - TV(F_j, F_k)| \geq TV(F_i, F_k)$$

We define the parameter vector $\Phi = \{c_1, c_2, c_3, \mu_1, \mu_2, \mu_3, \lambda_1, \lambda_2, \lambda_3\}$, where types are labeled so that $c_1\mu_1 > c_2\mu_2 > c_3\mu_3$. The relative cost reductions per unit time of the types captured by:

$$\text{Summary Parameter } \gamma := \frac{c_1\mu_1 - c_2\mu_2}{c_2\mu_2 - c_3\mu_3} \tag{22}$$

which captures the relative increments in cost reduction per unit time between the types: $\gamma = \infty$ is the case that types 2 and 3 are indistinguishable while $\gamma = 0$ means that types 2 and 1 and are indistinguishable in terms of their cost reduction per unit time. We introduce the total utilization

$$\rho_{tot} = \rho_1 + \rho_2 + \rho_3$$

Even for the idealized case, fully characterizing the optimal P is complex and depends on the relative values of cost reductions per unit time, the utilizations for the three types, and the total variation distances between the feature distributions of the three types. We partially characterize the optimal classification matrix for some specific values of the parameters. For $N = 3$ queues, we focus on high values of γ, when the cost reduction per unit time of type 1 is so high that it is optimal to keep types 2 and 3 separate from the type 1 (as much as possible). Even under sufficiently high γ, we find that the optimal P can take two possible values, depending on the relative cost reductions per unit time of types 2 and 3, as well as total variation distances between feature distributions.

Proposition 4 (3 Queues) *For $T = 3$ and $N = 3$, let $TV(F_1, F_2) < TV(F_1, F_3)$, $TV(F_2, F_3) < TV(F_1, F_3)$. Let the optimal P that minimizes $\mathcal{C}(P)$ under the constraints (21) be P^*. Then, there exists γ_u^* (that depends on $\rho_1, \rho_2, \rho_3, TV(F_1, F_2), TV(F_2, F_3), TV(F_1, F_3)$) such that $\forall\ \gamma > \gamma_u^*$:*

$$P^* = \begin{bmatrix} 1, & 0, & 0 \\ 1 - TV(F_1,F_2), & TV(F_1,F_2), & 0 \\ 1 - TV(F_1,F_3), & TV(F_1,F_3) - TV(F_2,F_3), & TV(F_2,F_3) \end{bmatrix} \text{ if } \Delta \geq 0$$

$$P^* = \begin{bmatrix} 1, & 0, & 0 \\ 1 - TV(F_1,F_2), & TV(F_1,F_2) + TV(F_2,F_3) - TV(F_1,F_3), & TV(F_1,F_3) - TV(F_2,F_3) \\ 1 - TV(F_1,F_3), & 0, & TV(F_1,F_3) \end{bmatrix} \text{ if } \Delta < 0$$

where,

$$\begin{aligned} \Delta = {} & c_2\mu_2\rho_2(TV(F_1,F_3) - TV(F_2,F_3))[\rho_{tot}(1 - \rho_{tot} + \rho_3 TV(F_2,F_3)) - (1-\rho_{tot})TV(F_1,F_2)(\rho_2 + \rho_3)] \\ & + c_3\mu_3\rho_2\rho_3^2 TV(F_1,F_3)TV(F_2,F_3)(TV(F_1,F_2) + TV(F_2,F_3) - TV(F_1,F_3)) \\ & - c_3\mu_3\rho_2\rho_3(1-\rho_{tot})(TV(F_1,F_3) - TV(F_1,F_2))(TV(F_1,F_3) - TV(F_2,F_3)) \end{aligned}$$

The conditions $TV(F_1, F_2) < TV(F_1, F_3)$ *and* $TV(F_2, F_3) < TV(F_1, F_3)$ imply that the total variation distance between the features for types 1 and 3 is the greatest. The quantity Δ is positive when the cost reduction per unit time for type 2 ($c_2\mu_2$) is significantly higher than the cost reduction per unit time for type 3 ($c_3\mu_3$) or when the total utilization (ρ_{tot}) is very high. In this case, it is important to protect type 2 from joining queue 3. When Δ is negative, it is optimal to keep type 3 in queue 3.

Statistical pooling: If it is difficult to distinguish between types from features, the optimal classifier may utilize fewer queues than types. For $\Delta \geq 0$, the utilization of the third queue, $\rho_3 TV(F_2, F_3)$, approaches 0 as $TV(F_2, F_3)$ (the total variation distance between types 2 and 3) approaches zero. For $\Delta < 0$, the utilization of the second queue approaches to zero as $TV(F_1, F_2) + TV(F_2, F_3) - TV(F_1, F_3)$ approaches to zero i.e., when the total variation distance between the feature distributions of types 1 and 2 and between types 2 and 3 is small as compared to the total variation distance between the feature distributions of types 1 and 3. In our case study with chest X-rays, we find that the optimal queue classifier utilizes only 3 out of 4 possible queues (see matrix $\hat{P}^d$ (26) in Section 6).

Next, we characterize the optimal classification matrices for $T = 3$ types and $N = 2$ queues.

Proposition 5 (2 Queues) *For $T = 3$ and $N = 2$, let $TV(F_1, F_2) \leq TV(F_1, F_3)$, $TV(F_2, F_3) \leq TV(F_1, F_3)$. Let the optimal P that minimizes $\mathcal{C}(P)$ under the constraints (21) be P^*:*

1. *If Summary Parameter $\gamma > \frac{\rho_3(1-\rho_{tot}+\rho_{tot}TV(F_1,F_3))}{\rho_1(1-\rho_{tot})}$, then*

$$P^* = \begin{bmatrix} 1, & 0 \\ 1 - TV(F_1, F_2), & TV(F_1, F_2) \\ 1 - TV(F_1, F_3), & TV(F_1, F_3) \end{bmatrix}$$

2. *If Summary Parameter $\gamma < \frac{\rho_3}{\rho_1(1-\rho_{tot}TV(F_1,F_3))}$, then*

$$P^* = \begin{bmatrix} 1, & 0 \\ 1 - TV(F_1, F_3) + TV(F_2, F_3), & TV(F_1, F_3) - TV(F_2, F_3) \\ 1 - TV(F_1, F_3), & TV(F_1, F_3) \end{bmatrix}$$

Over and under prioritization of the medium type: For 2 queues, it is always optimal to keep types 1 and 3 in queues 1 and 2 respectively, to the greatest possible extent under the constraints (21).

1. When the summary parameter γ is higher than a certain threshold (type 1 has a significantly higher cost reduction per unit time than type 2), then it is optimal to *under-prioritize* type 2 and have a greater volume of type 2 in the last queue (as compared to the optimal 3 queue design). The threshold increases with the total utilization ρ_{tot} and the under-prioritization of the medium type gets sub-optimal under heavy traffic. Further, the threshold value of γ increases with $TV(F_1, F_3)$, the total variation distance between the feature distributions of types 1 and 3 i.e., under-prioritization is worse for high classification accuracy and better for low classification accuracy.

2. When the summary parameter γ is smaller than a certain threshold (types 1 and 2 have similar cost reductions per unit time), then it is optimal to *over-prioritize* type 2 and have a greater volume of type 2 in the first queue (as compared to the optimal 3 queue design). The threshold value of γ increases with the total utilization ρ_{tot} and over-prioritization of the medium type gets optimal under heavy traffic. Further, the threshold increases with $TV(F_1, F_3)$ i.e., over-prioritization is better for high classification accuracy and worse for low classification accuracy.

With two queues, Proposition 2 in Argon and Ziya (2009) shows that as the total utilization ρ_{tot} approaches 1, the optimal utilization of the second queue drops to zero. Here, we prove that even for small values of total utilization ρ_{tot}, the utilization of the second queue drops to zero as the statistical distances between the feature distributions ($TV(F_1, F_2)$, $TV(F_1, F_3)$) approach zero.

Sensitivity more important than specificity: The optimal classifier chooses to have high sensitivity (the probability that type 1 is classified as queue 1) even if the specificity (the probability that type 3 is classified as queue 2) is low.

The above results are derived for an idealized classifier that can take all possible values of P that satisfy the constraints (21) imposed by the statistical distances between feature distributions. To test the robustness of these results, we compare the waiting costs under the below choice of classification matrices for 3 and 2 queues to answer two important questions:

1. When does the optimal classifier utilize fewer queues than types?
2. When is it be better to assign a high (or low) priority to the jobs with medium cost reduction per unit time.

$$P(3,\beta_3) = \begin{bmatrix} \beta_3 & \frac{1-\beta_3}{2} & \frac{1-\beta_3}{2} \\ \frac{1-\beta_3}{2} & \beta_3 & \frac{1-\beta_3}{2} \\ \frac{1-\beta_3}{2} & \frac{1-\beta_3}{2} & \beta_3 \end{bmatrix}, P^o(2,\beta_2) = \begin{bmatrix} \beta_2 & 1-\beta_2 \\ \beta_2 & 1-\beta_2 \\ 1-\beta_2 & \beta_2 \end{bmatrix}, P^u(2,\beta_2) = \begin{bmatrix} \beta_2 & 1-\beta_2 \\ 1-\beta_2 & \beta_2 \\ 1-\beta_2 & \beta_2 \end{bmatrix}, \quad (23)$$

where $\frac{1}{3} < \beta_3 \leq 1$ and $\frac{1}{2} < \beta_2 \leq 1$.

The matrix $P(3,\beta_3)$ utilizes $N = T = 3$ queues such that each type gets mapped to its correct priority queue with probability β_3 and to any other queue with probability $(1-\beta_3)/2$. The restriction $\beta_3 > 1/3$ guarantees that the classifier does better than random queue assignment. The matrices $P^o(2,\beta_2)$ and $P^u(2,\beta_2)$ utilize $N = 2$ queues such that types 1 and 3 get mapped to the first and the last queues with probability $\beta_2 > \frac{1}{2}$, respectively. The matrix $P^o(2,\beta_2)$ corresponds to an "over-prioritization" of the medium type where more type 2 join the first queue, whereas $P^u(2,\beta_2)$ corresponds to an "under-prioritization" of the medium type where more type 2 join the last queue.

Under perfect type information, the average waiting cost decreases with the number of priority queues, i.e., $N^* = 3$; see Proposition 1. This is no longer the case with noisy observations because a classifier's accuracy can drop with the number of prediction classes (Cui et al. 2019, Deng et al. 2012). In the extreme case when classification is uninformative any prioritization is the same as FIFO (a single queue) in terms of waiting cost. We prove that for intermediate values of classification accuracy, we can benefit from statistical pooling, meaning that if prediction accuracy increases with fewer queues we must classify into, then sufficient improvement in accuracy from pooling may overcome the loss due to coarse queueing.

Proposition 6 (Statistical pooling) *Suppose $T = 3$ and $P(3,\beta_3), P^u(2,\beta_2), P^o(2,\beta_2)$ are as specified in eq. (23). Then, there exists a threshold $\beta_3^*(\Phi)$ such that:*

1. *For all $\beta_3 \in [0,\beta_3^*(\Phi))$, there exists a threshold value of β_2 equal to $\beta_2^*(\beta_3,\Phi) \in (\beta_3,1)$, such that*
$$\min\{\mathcal{C}(P^u(2,\beta_2)),\mathcal{C}(P^o(2,\beta_2))\} < \mathcal{C}(P(3,\beta_3)) \ \forall \beta_2 \in (\beta_2^*(\beta_3,\Phi),1].$$
The two-queue average waiting cost with classifier accuracy $\beta_2 \in (\beta_2^*(\beta_3,\Phi),1]$ is lower than the three-queue average waiting cost with classifier accuracy β_3.
2. *For all $\beta_3 \in [\beta_3^*(\Phi),1]$,*
$$\mathcal{C}(P(3,\beta_3)) \leq \min\{\mathcal{C}(P^u(2,\beta_2)),\mathcal{C}(P^o(2,\beta_2))\} \forall \beta_2 \in [0,1].$$
The three-queue average waiting cost with classifier accuracy $\beta_3 \in [\beta_3^*(\Phi),1]$ is lower than the two-queue average waiting cost for all classifier accuracies $\beta_2 \in [0,1]$.

As an example, consider the extreme case where classification is purely random, i.e., $\beta_3 = \frac{1}{3}$. In this case the waiting cost does not vary with prioritization and equals that of FIFO. The waiting cost is lower with two priority queues ($N = 2$) as long as $\beta_2 > \frac{1}{2}$. When utilization is substantial, it is increasingly important to protect the customers with high delay costs from joining the lowest priority queue. Even a small gain in accuracy may justify coarse priorities ($N = 2$).

Proposition 7 (Statistical pooling: impact of total utilization) *Suppose $T = 3$ types that have the same load ($\rho_1 = \rho_2 = \rho_3 = \rho$). If $\gamma \leq 1$, the classification threshold $\beta_3^*(\Phi)$ is increasing in ρ (i.e., the range of accuracies β_3 for which 3 queues have lower waiting cost than 2 queues with $\beta_2 = 1$ shrinks as ρ increases). If $\gamma > 1$, the threshold $\beta_3^*(\Phi)$ is non-monotone in ρ: it decreases for total load $3\rho < \frac{\gamma-1}{\gamma}$ and increases thereafter.*

When types 2 and 1 are closer in their cost reductions per unit time as compared to types 2 and 3, then the higher the utilization, the larger the upper limit of β_3 for which three queues with accuracy β_3 are worse than two queues with accuracy 1. When types 2 and 1 differ more in their cost reductions per unit time than types 2 and 3, then for small utilization, the higher the utilization, the smaller the upper limit of β_3 for which three queues with accuracy β_3 are worse than two queues with accuracy 1. However, once the utilization is sufficiently high, the upper limit of β_3 increases with utilization.

In Proposition 5, we show that for the idealized classifier, over-prioritization of the medium type is better than under-prioritization of the medium type for high γ, high classification accuracy (low feature-overlap among types), and low utilization. In Proposition 8, we demonstrate the robustness of that result by proving that it also holds for the choice of classifiers in (23).

Proposition 8 (Over and under prioritization with 2 queues) *Suppose $T = 3$ and $N = 2$ with $P^u(2, \beta_2)$, $P^o(2, \beta_2)$ as specified in eq. (23). Then, there exists an accuracy threshold:*

$$\beta_2^*(\Phi) = \begin{cases} 0 & \textit{if } \dfrac{\gamma\rho_1 - (1-\rho_{tot})\rho_3}{\rho_{tot}(\gamma\rho_1 + \rho_3)} < 0 \\ 1 & \textit{if } \dfrac{\gamma\rho_1 - (1-\rho_{tot})\rho_3}{\rho_{tot}(\gamma\rho_1 + \rho_3)} > 1 \\ \dfrac{\gamma\rho_1 - (1-\rho_{tot})\rho_3}{\rho_{tot}(\gamma\rho_1 + \rho_3)} & \textit{otherwise} \end{cases}, \tag{24}$$

such that it is better to under-prioritize (i.e., $\mathcal{C}(P^u(2, \beta_2)) \leq \mathcal{C}(P^o(2, \beta_2))$) if $\beta_2 \leq \beta_2^*(\Phi)$ and over-prioritize (i.e., $\mathcal{C}(P^u(2, \beta_2)) > \mathcal{C}(P^o(2, \beta_2))$) otherwise. For fixed ρ_1, ρ_2, ρ_3, the threshold $\beta_2^*(\Phi)$ is increasing in γ.

Over-prioritizing type 2 increases the waiting time of type 1 while under-prioritizing type 2 increases the waiting time of type 2. Which of the two is the "lesser evil" depends on the delay costs c_1, c_2 and the relative lengths of the high priority and low priority queues. The low priority queue is $(1-\rho_{tot})^{-1}$ times longer than the high priority queue (from eq. (2)). The range $\gamma > \frac{\rho_3}{\rho_1(1-\rho_{tot})}$ corresponds to $\beta_2^*(\Phi) > 1$. The values of γ are rather large, and are of the same order as the ratio of the low to high priority queue lengths (i.e., $O(1/(1-\rho_{tot}))$. In this range c_1 is significantly larger than c_2 so that the cost of over-prioritization much exceeds the cost of under-prioritization, regardless of the accuracy β_2. In contrast, if $\gamma < \frac{\rho_3}{\rho_1}$, then $\beta_2^*(\Phi) < 0$. In this case, the cost of under-prioritization exceeds the cost of over-prioritization, regardless of accuracy β_2.

Lemma 5 *For fixed γ, and $\rho_1 = \rho_2 = \rho_3 = \rho$, $\beta_2^*(\Phi)$ is decreasing in ρ and approaches $\gamma/(\gamma+1)$ as the total utilization $\rho_{tot} = 3\rho$ approaches 1.*

If $\rho_1 = \rho_2 = \rho_3 = \rho$, then $\beta_2^*(\Phi) = (\gamma - 1 + 3\rho)/(3\rho(\gamma+1)) = ((\gamma-1)/3\rho(\gamma+1)) + (1/(\gamma+1))$, which is decreasing function of ρ for $\gamma > 1$. As the total utilization 3ρ approaches 1, the threshold accuracy $\beta_2^*(\Phi)$ approaches $\gamma/(\gamma+1)$. Since this limiting cut-off is greater than 0.5 for $\gamma > 1$, the question of under or over prioritization persists even in heavy-traffic: as the total utilization approaches 100%—depending on the classifier's accuracy—it might be optimal to assign medium cost jobs as low priority.

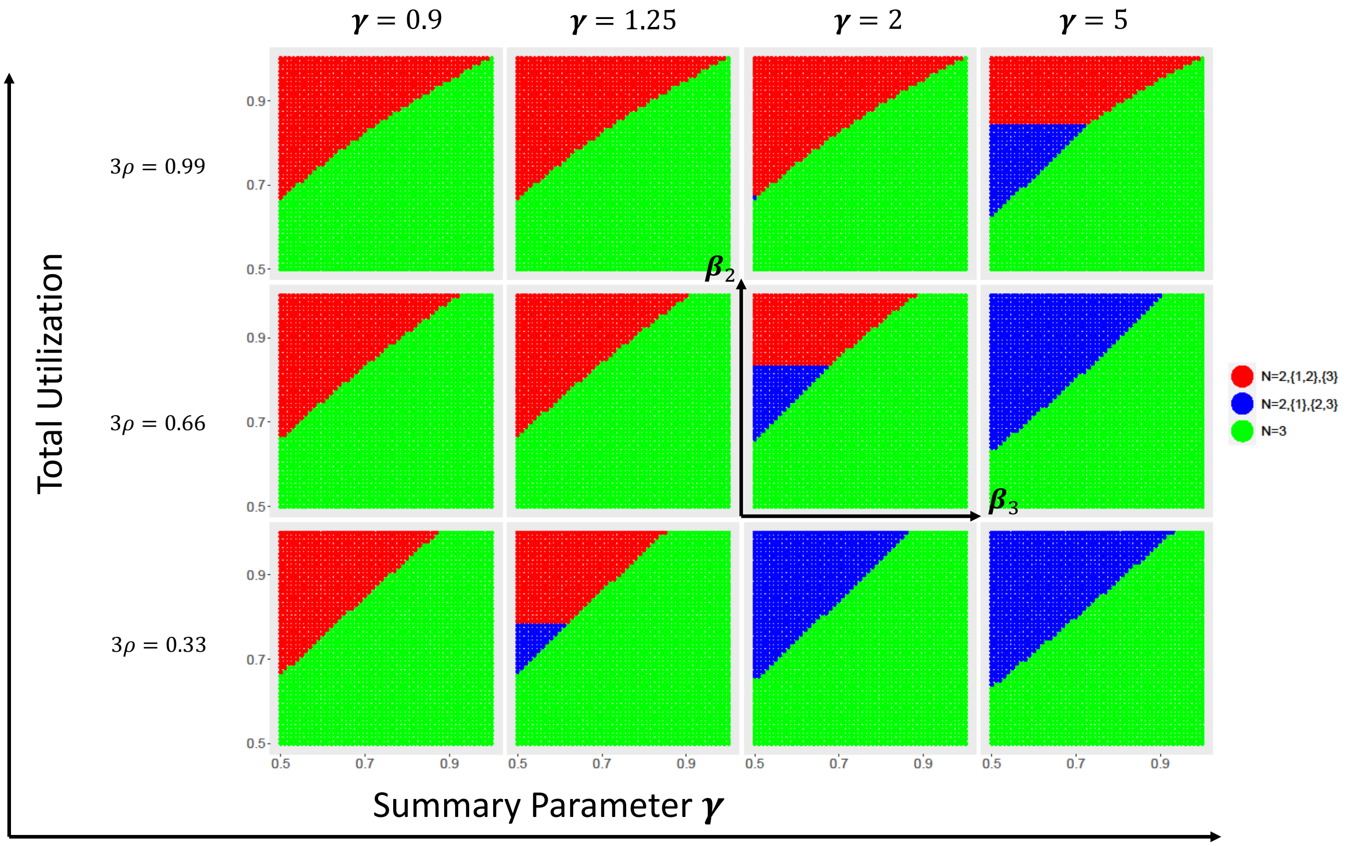

Figure 2 **Each inner plot shows the optimal priority queuing design as a function of the classifier accuracy parameters β_2 (vertical axis) and β_3 (horizontal axis). The design is shown for 4 values of type 2's relative cost γ (increasing from left to right) and three total utilization values given symmetric type-loads $\rho_1 = \rho_2 = \rho_3 = \rho$ (increasing from bottom to top). There are three priority queuing designs: (1) $N = 3$ priority queues is colored green; (2) $N = 2$ with under-prioritization of type 2 is blue; (3) $N = 2$ with over-prioritization of type 2 is red.**

	Summary Parameter γ		Classification Accuracy		Total Utilization	
	High	**Low**	**High**	**Low**	**High**	**Low**
Benefit from statistical pooling	Less	More	Less	More	More	Less
Over-prioritization or Under-prioritization for $N = 2$	Under	Over	Over	Under	Over	Under

Table 1 **Summary of relative benefits of statistical pooling and over/under prioritization of the medium type for 3 types under different regimes—different values of γ (ratio of difference in cost reduction per unit time between types 1 and 2 and difference in cost reduction per unit time between types 2 and 3), classification accuracy/statistical distance between feature distributions, and total utilization.**

Table 1 shows the optimal queue assignment strategy for different parameter regimes. Figure 2 shows the optimal priority queue design for three types, for a special case of $\rho_1 = \rho_2 = \rho_3$, different values of total utilization, summary parameter γ, and classification accuracies under two queues (β_2) and three queues (β_3).

Recall that for $N=2$, the idealized classifier was 100% accurate in assigning type 1 jobs to queue 1, even though it was less accurate in assigning type 3 jobs to queue 2. We show that the result is robust by demonstrating it for two stylized asymmetric classification matrices. Let $N=2$, and

$$P^u(2,\beta_{11},\beta_{22}) = \begin{bmatrix} \beta_{11} & 1-\beta_{11} \\ 1-\beta_{22} & \beta_{22} \\ 1-\beta_{22} & \beta_{22} \end{bmatrix}, \; P^o(2,\beta_{11},\beta_{22}) = \begin{bmatrix} \beta_{11} & 1-\beta_{11} \\ \beta_{11} & 1-\beta_{11} \\ 1-\beta_{22} & \beta_{22} \end{bmatrix} \tag{25}$$

The parameter β_{11} represents the sensitivity of classification (the accuracy of classifying the high priority types as such) and β_{22} represents the specificity (the accuracy of classifying the low priority types as such) of classification. The next proposition shows that sensitivity has a stronger effect on waiting cost compared to specificity.

Proposition 9 (Sensitivity is more important than specificity) *Under the matrices specified in eq.* (25)*, suppose*

$$\mathcal{C}^* = \min\{\mathcal{C}(P^u(2,\beta_{11},\beta_{22})), \mathcal{C}(P^o(2,\beta_{11},\beta_{22}))\}$$

$\mathcal{C}^*$ is decreasing in both β_{11} and β_{22}. Further, $\mathcal{C}^*$ is more sensitive to β_{11} than to β_{22}:

$$\frac{\partial \mathcal{C}^*}{\partial \beta_{11}} < \frac{\partial \mathcal{C}^*}{\partial \beta_{22}} < 0.$$

In our case study with chest X-rays, we find that the optimal queue classifier has much higher sensitivity than specificity (see matrix $\hat{P}^d$ (26) in Section 6).

6. Quantifying Improvements using a Chest X-ray 2D Image Dataset

In this section we explore feature-driven prioritization using state-of-the-art 2D image classifiers. The purpose of our experiments is to estimate the potential gains from queue classification (vs. type classification) for a practical setting with $T=14$ types and $N=4$ queues, and underscore the drivers of these gains.

The publicly available “ChestX-ray14” dataset provided by the National Institutes of Health (NIH) Clinical Center (Wang et al. 2017) consists of 112,120 frontal chest X-ray (CXR) 2D images with text-mined disease labels from the associated radiological reports for 30,805 unique patients. Following standard practice, we use an image resolution of 512×512 pixels with 3 channels for red, blue, and green (RGB) with 8 bits per channel. Each image is stored as a $512 \times 512 \times 3$ array with each entry storing conventional brightness intensities between 0 and 255. The feature space

$\mathcal{X}$ is discrete with the number of possible observed features (numerical representation of images) equal to $512 \times 512 \times 256^3$.

Types: There are 13 diseases with possible co-occurrences or no-occurrence (no finding). Table 2 shows the rank order of delay costs of different diseases, based on discussions with clinicians.[5] In consultation with a radiologist, we define the type of an image with multiple diseases as the rank of the highest delay cost disease present on that image. There are 14 types of images ($T = 14$) with type 14 corresponding to no disease finding. We consider two families of delay costs:

- *Convex delay costs*: The delay cost of type-i is $c_i = \eta^{T-i}$ for $i \in \mathcal{T}$. This family has a single parameter η that we can vary to analyze the impact of cost heterogeneity. We consider $\eta > 1$ (to ensure that the delay cost of type-1 is the highest and that of type-14 is the lowest). For $\eta > 1$, the difference in delay costs of successive types is increasing with the rank order of the types: $c_1 - c_2 = \eta^{12}(\eta - 1) > c_2 - c_3 = \eta^{11}(\eta - 1) > c_3 - c_4 = \eta^{10}(\eta - 1) > \ldots > c_{13} - c_{14} = (\eta - 1)$. We vary $\eta \in \{1.8, 1.5\}$: $\eta = 1.8$ has more type heterogeneity than $\eta = 1.5$.
- *Linear delay costs*: The delay cost of type-i is $c_i = \delta * (T - i) + 1$ for $i \in \mathcal{T}$. The difference in delay costs between any two successive types is δ i.e., $c_1 - c_2 = c_2 - c_3 = c_3 - c_4 = \ldots = c_{13} - c_{14}$. We vary $\delta \in \{1.8, 10\}$: $\delta = 1.8$ implies smaller differences in delay costs than $\delta = 10$.

We do not have arrival times of the CXRs, and for the purpose of this study we assume that the arrivals of CXRs of different types to the radiology center follow Poisson processes. We normalize the total arrival rate to 1 so that the arrival rate of each type equals the fraction of that type in the full dataset of all the images. For a sample data set D_s consisting of s CXRs and type labels, let $D_s = \{(X_r, t_r) : r \in \mathcal{S}\}$ where $t_r \in \mathcal{T}$ is the known true type of the CXR and X_r is an $512 \times 512 \times 3$ data array that contains red, green, and blue color components for each individual pixel of the image. Here $\mathcal{S} = \{1, 2, \cdots, s\}$ is the set of all image indices in the sample. We estimate the arrival rate of type-$i \in \mathcal{T}$ image as $\hat{\lambda}_i = \sum_{r=1}^{s} \mathbb{I}(t_r = i)/s$. Table 3 shows the arrival rates of different types of images.

Focusing on a portion of the day (say peak hours) where arrivals are relatively stationary does not qualitatively change the results. We assume an equal service time of the radiologist for all image types: $\mu = (\sum_{t \in \mathcal{T}} \hat{\lambda}_t)/\rho_{tot} \forall i \in \mathcal{T}$, where ρ_{tot} is the total server utilization.

In large radiology centers and hospitals, it is common to have a separate worklist for CXRs. Each incoming CXR is assigned a priority for the radiologist reading ("triage"). There are typically four priority queues for radiologist review: *Critical* (1) $\succ$ *Urgent* (2) $\succ$ *Important* (3) $\succ$ *Routine* (4).

[5] We acknowledge that the rank ordering of disease findings on delay costs may be subjective and context dependent (for example, (1) some findings may become more urgent in case of an outbreak; (2) a second CXR for a patient showing the same finding from the previous day may no longer be very urgent since it is already diagnosed and being monitored for treatment; (3) a CXR with no finding may be important for diagnostic clarity), and actual implementation of direct in a clinical setting would require collaboration with health experts.

Disease	**Rank Order of Delay Cost**
Pneumothorax	1
Emphysema	2
Pneumonia	3
Edema	4
Consolidation	5
Effusion	6
Infiltration	7
Atelectasis	8
Cardiomegaly	9
Pleural Thickening	10
Fibrosis	11
Mass	12
Nodule	13
No Finding	14

Table 2 Rank order of delay costs for different diseases (from highest to lowest).

Type	**Arrival Rate (Prevalence)**
1	0.05
2	0.02
3	0.01
4	0.02
5	0.04
6	0.09
7	0.12
8	0.04
9	0.01
10	0.01
11	0.01
12	0.02
13	0.02
14	0.54
Total	1.00

Table 3 Type Distribution “ChestX-ray14” dataset

Existing studies (Baltruschat et al. 2019, Wang et al. 2017) and open-source publications (Garyfallos 2019) use deep learning methods or image classification on this same dataset. We use the Mobile Net Architecture in Garyfallos (2019) as a building block for our experiments. Mobile Net is a Convolutional Neural Network (CNN) designed specifically for image classification and mobile vision. It utilizes depth-wise separable convolutions to greatly reduce the number of parameters while retaining sufficient depth levels. The computational savings make Mobile Net a great choice for computer vision on devices with less power, such as cellphones and embedded cameras, and hence a good choice for practical implementation in healthcare settings.

We simulate a single-server priority queuing model of AI/ML (Mobile Net) enabled feature-driven triage of CXRs at a radiology center, and experiment with a range of delay cost families (convex and linear), their parameters η and δ, and total utilization ρ_{tot}. Our python implementation is publicly available.

1 **Classify Type then Optimize Queuing (t)**: We use the Mobile Net classifier[6] to predict the probabilities of different diseases, and hence the probabilities of the 14 types on the images (see Appendix 8.1 for more details). Let $z_r = z(X_r) = [z_{i,r}] = [z_i(X_r)],\ i \in \mathcal{T}$ be the T dimensional probability vector output of the Mobile Net for the image $r \in \mathcal{S}$ with feature vector X_r such that $z_i(X_r)$ is the probability that the image r is of type-i. We map the type probability z_r to queue $Q^t(\beta_1, \beta_2, \beta_3, z_r)$:

$$Q^t(\beta_1,\beta_2,\beta_3,z_r) = \underset{j\in\{1,2,3,4\}}{\arg\max}\, q_j(\beta_1,\beta_2,\beta_3,z_r):$$

$$q_1(\beta_1,\beta_2,\beta_3,z_r) = \frac{exp(\beta_1\cdot z_r)}{exp(\beta_1\cdot z_r)+exp(\beta_2\cdot z_r)+exp(\beta_3\cdot z_r)+1}$$

$$q_2(\beta_1,\beta_2,\beta_3,z_r) = \frac{exp(\beta_2\cdot z_r)}{exp(\beta_1\cdot z_r)+exp(\beta_2\cdot z_r)+exp(\beta_3\cdot z_r)+1}$$

$$q_3(\beta_1,\beta_2,\beta_3,z_r) = \frac{exp(\beta_3\cdot z_r)}{exp(\beta_1\cdot z_r)+exp(\beta_2\cdot z_r)+exp(\beta_3\cdot z_r)+1}$$

$$q_4(\beta_1,\beta_2,\beta_3,z_r) = \frac{1}{exp(\beta_1\cdot z_r)+exp(\beta_2\cdot z_r)+exp(\beta_3\cdot z_r)+1}$$

where β_1, β_2, and β_3 are $T-$dimensional weight vectors.

Now, $0 < q_j(\beta_1,\beta_2,\beta_3,z_r) < 1\ \forall j\{1,2,3,4\},\ r \in \mathcal{S}$, and $\sum_{j\in\{1,2,3,4\}} q_j(\beta_1,\beta_2,\beta_3,z_r) = 1\ \ \forall r \in \mathcal{S}$. Therefore, $q_j(\beta_1,\beta_2,\beta_3,z_r)$ can be interpreted as the probability that image r with predicted type probabilities z_r belongs to queue j, and $Q^t(\beta_1,\beta_2,\beta_3,z_r)$ is the queue with the greatest probability $q_j(\beta_1,\beta_2,\beta_3,z_r)$. We estimate the classification probabilities as: $\forall i \in \mathcal{T},\ j \in \mathcal{N}$

$$\hat{P}_{ij}(\beta_1,\beta_2,\beta_3) = \frac{\sum_{r\in\mathcal{S}} z_{i,r} q_j(\beta_1,\beta_2,\beta_3,z_r)}{\sum_{r\in\mathcal{S}} z_{i,r}}.$$

We find the optimal $\beta_1,\ \beta_2,\ \beta_3$ that minimize the estimated average waiting cost:

$$\beta_1^{t*},\beta_2^{t*},\beta_3^{t*} = \underset{(\beta_1,\beta_2,\beta_3)}{\arg\min}\, \mathcal{C}(\hat{P}(\beta_1,\beta_2,\beta_3)).$$

[6] For training the Mobile Net, we create a re-sampled dataset that is well-balanced across types.

Delay Cost	ρ_{tot}	$\mathcal{C}(\hat{P}^t)$	$\mathcal{C}(\hat{P}^d)$	$\mathcal{C}(P^{full})$	$\mathcal{C}(P^{FIFO})$
$c_i = 1.8(T-i)+1$	0.9	58 (8)	27 (1)	16	59
$c_i = 10(T-i)+1$	0.9	270 (46)	106 (10)	51	290
$c_i = 1.5^{(T-i)}$	0.9	103 (18)	59 (4)	28	166
$c_i = 1.8^{(T-i)}$	0.9	550 (50)	395 (41)	158	1285
$c_i = 1.8^{(T-i)}$	0.6	96 (3)	99 (5)	63	143

Table 4 Estimates of Mean (Standard Deviation) of Average Waiting Costs Using 10 samples of 2000 test images each. Recall that $\hat{P}^t$ corresponds to type-first, and $\hat{P}^d$ corresponds to direct.

2 Direct (d): We use the same Mobile Net architecture as for type-first that predicts the probabilities for each image, but add an activation layer on top that outputs a 4-dimensional probability vector

$$q_1(\beta_1,\beta_2,\beta_3,z(X_r)) = \frac{exp(\beta_1\cdot z(X_r))}{exp(\beta_1\cdot z(X_r))+exp(\beta_2\cdot z(X_r))+exp(\beta_3\cdot z(X_r))+1}$$

$$q_2(\beta_1,\beta_2,\beta_3,z(X_r)) = \frac{exp(\beta_2\cdot z(X_r))}{exp(\beta_1\cdot z(X_r))+exp(\beta_2\cdot z(X_r))+exp(\beta_3\cdot z(X_r))+1}$$

$$q_3(\beta_1,\beta_2,\beta_3,z(X_r)) = \frac{exp(\beta_3\cdot z(X_r))}{exp(\beta_1\cdot z(X_r))+exp(\beta_2\cdot z(X_r))+exp(\beta_3\cdot z(X_r))+1}$$

$$q_4(\beta_1,\beta_2,\beta_3,z(X_r)) = \frac{1}{exp(\beta_1\cdot z(X_r))+exp(\beta_2\cdot z(X_r))+exp(\beta_3\cdot z(X_r))+1}$$

where, β_1, β_2, and β_3 are $T-$dimensional weight vectors. We map feature X_r to queue $Q^d(\beta_1,\beta_2,\beta_3,X_r)$ which corresponds to the queue with the greatest predicted probability

$$Q^d(\beta_1,\beta_2,\beta_3,X_r) = \underset{j\in\{1,2,3,4\}}{\arg\max}\, q_j(\beta_1,\beta_2,\beta_3,z(X_r))$$

We estimate the classification probabilities as: $\forall i\in\mathcal{T},\ j\in\mathcal{N}$

$$\hat{P}_{ij}(\beta_1,\beta_2,\beta_3,z) = \frac{\sum_{r\in\mathcal{S}}\mathbb{1}(t_r=i)q_j(\beta_1,\beta_2,\beta_3,z(X_r))}{\sum_{r\in\mathcal{S}}\mathbb{1}(t_r=i)}.$$

We find the optimal Mobile Net z^* (by optimizing the weights of the Convolutional Neural Network) as well the weights $\beta_1^{d*},\beta_2^{d*},\beta_3^{d*}$ of the queueing activation layer to minimize the average waiting cost:

$$(\beta_1^{d*},\beta_2^{d*},\beta_3^{d*},z^{d*}) = \underset{(\beta_1,\beta_2,\beta_3,z)}{\arg\min}\ \mathcal{C}(\hat{P}(\beta_1,\beta_2,\beta_3,z))$$

To benchmark these, we estimate the average waiting cost under perfect/full type observation with classification matrix P^{full} (full information) and under the single queue first-in-first-out approach with classification matrix P^{FIFO}, where $P^{full} = I^*_{T\times N}$ (see eq. (5) in §3.1), $P^{FIFO}: P_{ij}^{FIFO} = 1\ \forall i\in\mathcal{T}, j=1$.

Table 4 reports the performance for the different schemes implemented by taking multiple samples $D_s = \{(X_r, t_r) : r \in \mathcal{S}\}$. For each sample, we estimate the classification matrices $\hat{P}_1$ and $\hat{P}_2$ generated by the type classification approach (t) and the direct approach (d):

$$\hat{P}^t_{ij} = \frac{\sum_{r\in\mathcal{S}} \mathbb{1}(t_r = i)\mathbb{1}(Q^t(\beta_1^{t*}, \beta_2^{t*}, \beta_3^{t*}, z_r) = j)}{\sum_{r\in\mathcal{S}} \mathbb{1}(t_r = i)}$$

$$\hat{P}^d_{ij} = \frac{\sum_{r\in\mathcal{S}} \mathbb{1}(t_r = i)\mathbb{1}(Q^d(\beta_1^{d*}, \beta_2^{d*}, \beta_3^{d*}, z^{d*}(X_r)) = j)}{\sum_{r\in\mathcal{S}} \mathbb{1}(t_r = i)}$$

and estimate the average waiting costs as $\mathcal{C}(\hat{P}^t)$ and $\mathcal{C}(\hat{P}^d)$, respectively. The results reveal the following:

$$\hat{P}^t = \begin{bmatrix} 0.69 & 0.24 & 0.08 & 0.0 \\ 0.38 & 0.40 & 0.22 & 0.0 \\ 0.27 & 0.46 & 0.27 & 0.0 \\ 0.17 & 0.52 & 0.31 & 0.0 \\ 0.21 & 0.39 & 0.40 & 0.0 \\ 0.14 & 0.40 & 0.47 & 0.0 \\ 0.67 & 0.27 & 0.06 & 0.0 \\ 0.34 & 0.51 & 0.16 & 0.0 \\ 0.48 & 0.48 & 0.03 & 0.0 \\ 0.42 & 0.47 & 0.11 & 0.0 \\ 0.54 & 0.39 & 0.07 & 0.0 \\ 0.25 & 0.47 & 0.28 & 0.0 \\ 0.23 & 0.55 & 0.22 & 0.0 \\ 0.05 & 0.24 & 0.71 & 0.0 \end{bmatrix}, \hat{P}^d = \begin{bmatrix} 0.0 & 0.83 & 0.14 & 0.03 \\ 0.0 & 0.77 & 0.19 & 0.05 \\ 0.0 & 0.78 & 0.17 & 0.05 \\ 0.0 & 0.75 & 0.21 & 0.05 \\ 0.0 & 0.63 & 0.27 & 0.10 \\ 0.0 & 0.61 & 0.28 & 0.10 \\ 0.0 & 0.81 & 0.15 & 0.04 \\ 0.0 & 0.62 & 0.30 & 0.08 \\ 0.0 & 0.63 & 0.27 & 0.10 \\ 0.0 & 0.67 & 0.25 & 0.09 \\ 0.0 & 0.71 & 0.21 & 0.08 \\ 0.0 & 0.63 & 0.26 & 0.11 \\ 0.0 & 0.64 & 0.27 & 0.09 \\ 0.0 & 0.16 & 0.31 & 0.53 \end{bmatrix}, \hat{P}^{full} = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \tag{26}$$

1. **Statistical Pooling at High Utilization**: Queues 4 and 1 in $\hat{P}^t$ and $\hat{P}^d$, respectively, in (26) have a zero utilization, showing that both the approaches — type-first and direct — utilize fewer queues than maximum possible. This validates our theoretical result on statistical pooling (Proposition 6) for a stylized family of classifiers.
2. **Over-triage at High Utilization**: The direct approach gives much lower average waiting cost than type-first for $\rho = 0.9$ (by 30% to 60% in our experiments; see Table 4). It has a significant over-triage of the medium types. The average classification matrices [7] for delay costs of types given by $c_i = 1.8^{(T-i)}$ (26) show that under the direct approach, 61% of the medium type 6 join the first queue and 10% join the last queue. In contrast, under type-first, 14% of the medium type 6 join the first queue and 47% join the last queue.

[7] The average of the classification matrices for the 10 samples of 2000 images

3. **Sensitivity more important than Specificity**: The direct approach correctly classifies 83% of type 1 as queue 1. However, the accuracy of correctly classifying type 14 as the last queue is only 53% (26). In contrast, the type-first approach has approximately the same accuracy of correctly classifying type 1 (69%) as queue 1 and type 14 as last queue (71%).
4. **Type-Driven vs Direct:** The type-first approach gives a significantly lower average waiting cost than a single queue with first in first out regime only under convex cost structures. When the differences in delay costs of successive types are constant and the classifier is unbiased (as is the type classifier), the optimal queue assignment based on type probabilities balances over-triage and under-triage of the medium types. Whereas the optimal solution must have more over-triage of the medium types. In contrast, when the differences in delay costs of successive types are convex, and the classifier is unbiased, the optimal queue assignment based on type probabilities chooses more over-triage — even smaller probabilities for high types are assigned to the higher priority queue.

   The direct approach gives significant savings across both linear as well as convex cost structures. It provides significant cost savings over the type-first approach (30% - 60% in our experiments) under high utilization ($\rho_{tot} = 0.9$), but performs equally or marginally worse than the latter approach for low utilization ($\rho_{tot} = 0.6$). Under the direct design ($\hat{P}^d$), the average waiting times of all types except type 14 (no finding) are in the 2.1–3.9 range (Table 5), whereas type-14 waits for significantly longer (12.4-time units). The optimized classifier is, on its own, leveraging an attribute of the underlying data: more than half (54%) of the total arrival volume is of type 14 (“no finding”) while the prevalence of each of the types 1 to 13 is small. This makes sense as imaging is often prescribed to rule out dangerous conditions. It is a characteristic of priority queues, that the waiting time of high-priority decreases, the larger the low-priority arrival rate as a portion of the total arrival rate is. The queue classifier is smart enough to capture this (whereas the type classifier is not); it effectively creates two wait time categories: separating scans with a finding from those without a finding.[8]

We believe that the performance reported in Table 4 is a conservative estimate of the gains of feature-driven queue prioritization due to the following:

- We did not optimize some design parameters of the Mobile Net neural network—arrangement of dropout layers, dropout rate, learning rate—as they were tuned for disease prediction in Garyfallos (2019), even though we deploy the network for queue prediction. Tuning these parameters specifically for priority queue prediction could possibly lead to better performance.

[8] Even though both the queue and type classifiers are trained on a well-balanced training set with over-sampling of types with lower arrival rates, the queue classifier can capture the imbalance via the actual values of arrival rates used for computing the queuing loss.

	Average Waiting Time		
Type	$\hat{P}^t$	$\hat{P}^d$	$\hat{P}^{full}$
1	2.33	2.11	0.87
2	4.44	2.64	0.87
3	5.21	2.63	0.87
4	5.80	2.64	1.09
5	6.95	3.82	1.09
6	7.93	3.95	1.09
7	2.06	2.33	1.68
8	3.67	3.58	1.68
9	1.89	3.86	1.68
10	2.94	3.50	1.68
11	2.31	3.27	1.68
12	5.34	3.94	1.68
13	4.64	3.65	13.40
14	11.27	12.40	13.40

Table 5 Average waiting times of different types for different classification matrices for delay costs of types given by $c_i = 1.8^{(T-i)}$; $i\{1,2,3,\ldots,14\}$. One time unit corresponds to a normalized total arrival rate of 1. Recall that $\hat{P}^t$ corresponds to type classification and queue optimization, $\hat{P}^d$ corresponds to direct queue classification, and $\hat{P}^{full}$ corresponds to perfect information when the type of each image is known on arrival.

- Due to GPU memory limitations, we used a batch size (the number of images used for one gradient update due to optimization) of 32 in training the classifier. We believe that a larger batch size (for each gradient update) is more beneficial for optimizing waiting cost, as compared to optimizing a standard loss function like the mean squared error. This is because the former requires an estimation of $T \times N$ classification matrix. Increasing the batch-size is practically implementable via multiple-gpu based parallel processing.

This computational study presents evidence that our optimal queue classification design can lead to substantial improvements over the current type classification design in the practical setting of triaging chest x-rays. We do acknowledge that our limited dataset forces us to assume that the priority queue depends only on the features of the image. In reality, both clinical and non-clinical features may determine the priority queue of an image. Also, previous studies (Ibanez et al. 2017) have shown that radiologists often exercise discretion and deviate from assigned queues. In this study, we make the simplistic assumption that delay costs aggregate linearly over time. Exploring non-linear costs could be an interesting and relevant extension, potentially yielding counter-intuitive results, as demonstrated in the experiments in Section 9 of Argon and Ziya (2009).

7. Summary, Discussion and Conclusion

Priority queuing is an unexplored context for the study of direct prediction of decisions. The non-linear queuing externalities—the impact of priority assignment of one job on the waiting times of other jobs——give rise to new questions about the interaction between classification/prediction and priority queues (workflow) design that warrant a detailed study.

In this study, we explicitly link the statistical distances between types' feature distributions with the optimal queue design for two and three types. For two types, we construct tight lower bounds on the waiting costs achievable through feature-based classification, and show that the unbiasedness of a queue classifier is sub-optimal.

Our theoretical analysis of the three-type setting helps us characterize the mechanisms through which direct queueing brings improvements under different levels of system utilization and the relative urgency of the types: 1) choosing between more queues (better waiting time stratification) and fewer queues (more accurate queue classification) depending on the classifier accuracy for more vs. fewer queues, which, in turn, depends on the statistical distances between feature distributions; 2) balancing under and over prioritization; 3) optimizing the classification accuracy for the most important queue. These results are novel because they open the black box and provide interpretability to the mathematical interaction between AI/ML (classification) and priority queuing.

We also present evidence that direct queue classification can yield substantial improvements over type classification and queue optimization in a practical setting of triaging chest x-rays. Our computational performance study may be the first to apply deep learning-based classification of medical images to study the impact of classification errors on optimal priority queue design. We show that feature-driven priority queuing can improve waiting cost significantly under high utilization.

Our research proposes a priority recommendation system and demonstrates its value in a stylized setting assuming the user follows the priority recommendation. Accounting for human deviations from algorithmic recommendations is undoubtedly an exciting problem addressed in some other contexts (Sun et al. 2021) but beyond the scope of this first study on feature-driven design of priority queues.

References

Nilay Tanık Argon and Serhan Ziya. Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management*, 2009.

R.R. Bahadur. Sufficiency and statistical decision functions. *The Annals of Mathematical Statistics*, 1954. URL `https://doi.org/10.1214/aoms/1177728715`.

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific Reports*, 2019.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 2019.

Avraham Beja and Esther Sid. Optimal priority assignment with heterogeneous waiting costs. *Operations Research*, 1975.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 2020.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. On approximating total variation distance. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2023.

Austin Bren and Soroush Saghafian. Data-driven percentile optimization for multiclass queueing systems with model ambiguity: Theory and application. *INFORMS Journal on Optimization*, 2019.

Carri W. Chan, Michael Huang, and Vahid Sarhangian. Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. *Operations Research*, 2021.

A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186, September 1962. doi: 10.1002/nav.3800090303.

Alan Cobham. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 1954.

Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In *NIPS*, 2003.

Sir David Roxbee Cox and Walter L. Smith. *Queues*. John Wiley & Sons, 1961.

Yin Cui, Zeqi Gu, Dhruv Mahajan, Laurens Van Der Maaten, Serge J. Belongie, and Ser-Nam Lim. Measuring dataset granularity. *Computing Research Repository*, 2019.

Jia Deng, Jonathan Krause, Alexander C. Berg, and Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 2021.

Spyros Garyfallos. `https://github.com/paloukari/NIH-Chest-X-rays-Classification`, 2019.

Maria R. Ibanez, Jonathan R. Clark, Robert S. Huckman, and Bradley R. Staats. Discretionary task ordering: Queue management in radiological services. *Management Science*, 2017.

Anton Kleywegt, Alexander Shapiro, and Tito Homem-De-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 2001.

Liwan Liyanage and J. George Shanthikumar. A practical inventory control policy using operational statistics. *Operations Research Letters*, 2005.

Chuen-Teck See and Melvyn Sim. Robust approximation to multiperiod inventory management. *Operations Research*, 2010.

Jiankun Sun, Dennis J. Zhang, Haoyuan Hu, and Jan A. Van Mieghem. Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, 2021.

Zhankun Sun, Nilay Tanik Argon, and Serhan Ziya. When to triage in service systems with hidden customer class identities? *Production and Operations Management*, 2019.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Computing Research Repository*, 2017.

S.P. Van Der Zee and Henri Theil. Priority assignment in waiting-line problems under conditions of misclassification. *Operations Research*, 1961.

8. Appendix

8.1. Type-Classification for Triage of Chest X-rays

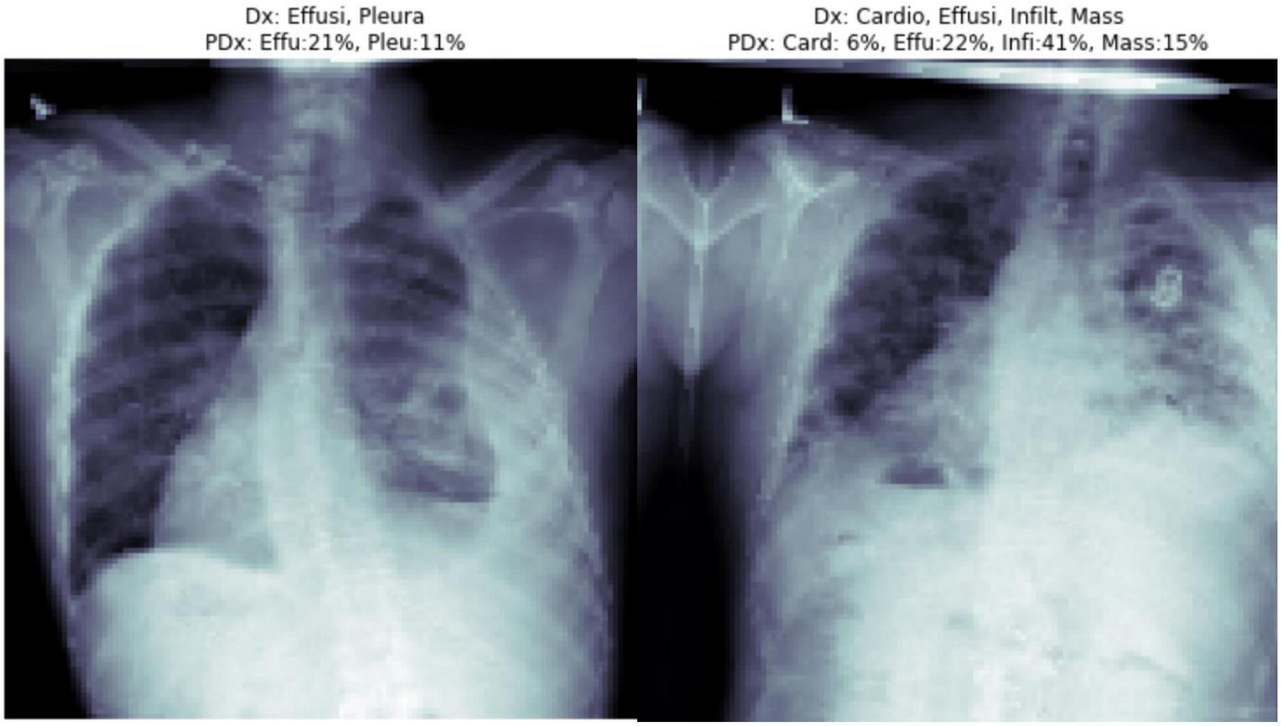
Dx: Effusi, Pleura
PDx: Effu:21%, Pleu:11%
Dx: Cardio, Effusi, Infilt, Mass
PDx: Card: 6%, Effu:22%, Infi:41%, Mass:15%

Figure 3 **2D chest X-rays labeled with diseases and Mobile Net predictions. Dx: disease findings extracted from their radiologist reports. PDx: predicted probabilities by Mobile Net*. Image Source: Garyfallos (2019).**

*Mobile Net outputs probabilities for all 14 findings but the figure shows only those for the actual findings.

We use a Mobile Net (multiple binarizers in Garyfallos (2019)) [9] to predict the probability of each disease (Figure 3 shows two CXRs labeled with the disease findings extracted from their radiologist reports and the Mobile Net output for the images). The classification is optimized to minimize the binary cross entropy loss. Figure 4 reports the performance of Mobile Net. For all diseases, the results (in terms of the area under the curve (AUC) for the receiver operating characteristic curve (ROC)) are equal or better than those reported in Wang et al. (2017). The probability of a type is equal to the probability that

[9] For training the Mobile Net, we create a re-sampled dataset that is well-balanced across the types.

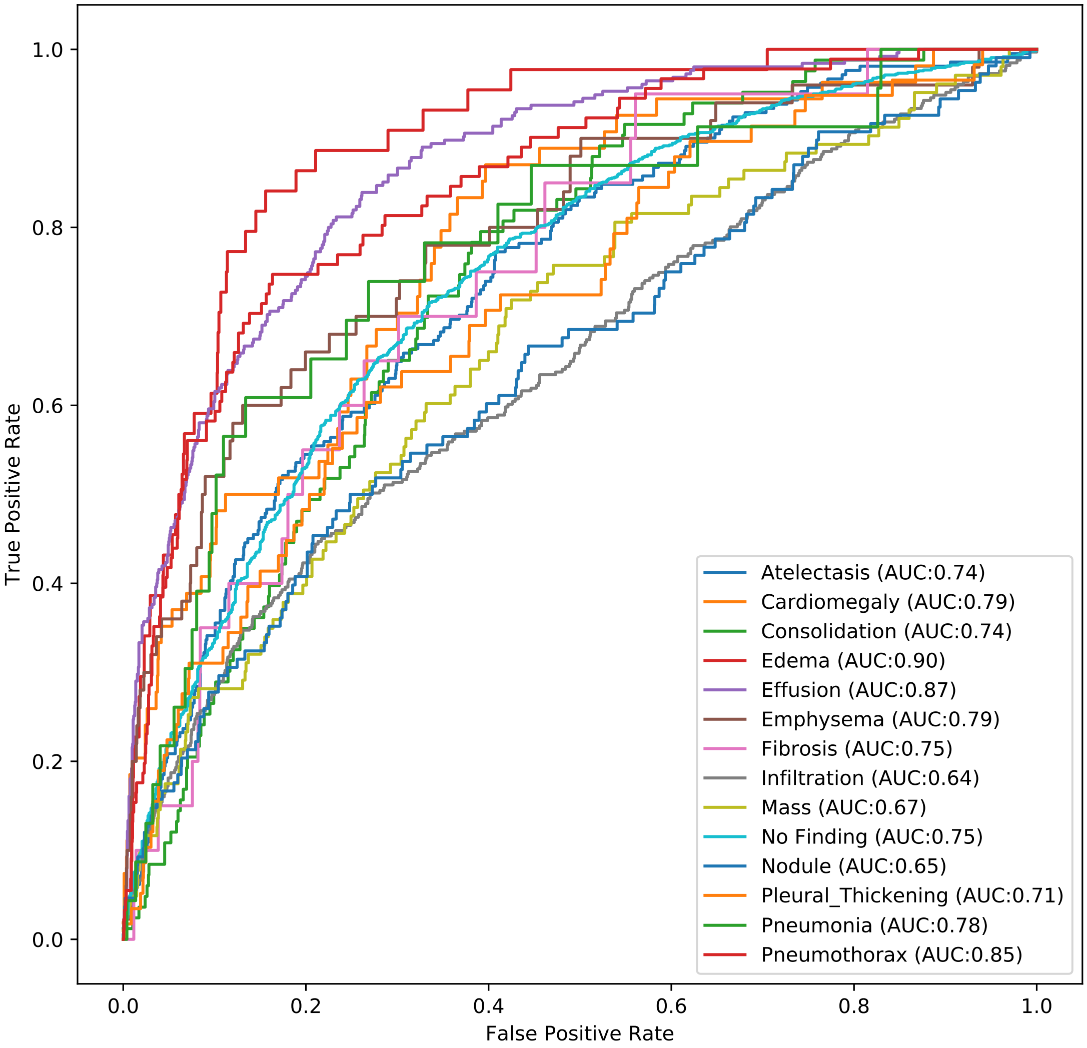

Figure 4 **ROC curves for classification of different diseases for a sample of 2000 images from the ChestX-ray14 dataset.**

8.2. Proofs

Proof of Proposition 1. We use a standard interchange argument. Towards contradiction, suppose that the optimal partition $I^*_{T\times N}$ has $j, j+1 \in \mathcal{N},\ i,k \in \mathcal{T}$ such that $c_i\mu_i < c_k\mu_k$ but i has higher priority, that is $I^*_{T\times N}(i,j) = I^*_{T\times N}(k,j+1) = 1$ (in the optimal policy type-i goes to queue j and type-k goes to queue $j+1$).

We will prove that the expected stationary waiting cost is strictly lower under at least one of the two modifications to $I^*_{T\times N}$: a) shifting type-i to queue $j+1$ i.e., $I^1_{T\times N}(i,j+1)=1$; b) shifting type-k to queue j i.e., $I^1_{T\times N}(k,j)=1$. This would then contradict the optimality of $I^*_{T\times N}$.

Define

$$A := \{l : I^*_{T\times N}(l,j)=1, l\neq i\},\ c_A\mu_A := \frac{\sum_{l\in A}\rho_l c_l \mu_l}{\sum_{l\in A}\rho_l},\ \rho_A := \sum_{l\in A}\rho_l,$$

and similarly

$$B := \{l : I^*_{T\times N}(l,j+1)=1, l\neq k\},\ c_B\mu_B := \frac{\sum_{l\in B}\rho_l c_l \mu_l}{\sum_{l\in B}\rho_l},\ \rho_B := \sum_{l\in B}\rho_l,$$

where $c_A\mu_A=\rho_A=0$ and $c_B\mu_B=\rho_B=0$ if $A=\phi$ or $B=\phi$, respectively. and, finally,

$$\bar{\rho} := \sum_{l:\exists q\leq j-1 \text{ and } I^*_{T\times N}(l,q)=1} \rho_l,$$

with $\bar{\rho}=0$ if $j\leq 1$.

Then,

$$\begin{aligned}
&\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^1_{T\times N})\\
&\qquad=\frac{c_A\mu_A\rho_A+c_i\mu_i\rho_i}{(1-\bar{\rho})(1-\bar{\rho}-\rho_A-\rho_i)}+\frac{c_k\mu_k\rho_k+c_B\mu_B\rho_B}{(1-\bar{\rho}-\rho_A-\rho_i)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)}\\
&\qquad\quad-\frac{c_A\mu_A\rho_A}{(1-\bar{\rho})(1-\bar{\rho}-\rho_A)}-\frac{c_i\mu_i\rho_i+c_k\mu_k\rho_k+c_B\mu_B\rho_B}{(1-\bar{\rho}-\rho_A)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)}\\
&\qquad=\rho_i\left[\frac{c_A\mu_A\rho_A+c_i\mu_i(1-\bar{\rho}-\rho_A)}{(1-\bar{\rho})(1-\bar{\rho}-\rho_A-\rho_i)(1-\bar{\rho}-\rho_A)}+\frac{c_k\mu_k\rho_k+c_B\mu_B\rho_B-c_i\mu_i(1-\bar{\rho}-\rho_A-\rho_i)}{(1-\bar{\rho}-\rho_A-\rho_i)(1-\bar{\rho}-\rho_A)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)}\right]\\
&\qquad=\rho_i\left[\frac{\Delta_1}{(1-\bar{\rho})(1-\bar{\rho}-\rho_A-\rho_i)(1-\bar{\rho}-\rho_A)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)}\right],
\end{aligned}$$

where

$$\begin{aligned}
\Delta_1 :=&c_A\mu_A\rho_A(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)+(c_k\mu_k\rho_k+c_B\mu_B\rho_B)(1-\bar{\rho})\\
&-c_i\mu_i(\rho_A(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)+(\rho_k+\rho_B)(1-\bar{\rho})).
\end{aligned}$$

Similarly,

$$\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^2_{T\times N})=\rho_k\left[\frac{\Delta_2}{(1-\bar{\rho})(1-\bar{\rho}-\rho_A-\rho_i)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)}\right],$$

where

$$\begin{aligned}
\Delta_2 :=&-(c_A\mu_A\rho_A+c_i\mu_i\rho_i)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)\\
&-c_B\mu_B\rho_B(1-\bar{\rho})+c_k\mu_k(\rho_B(1-\bar{\rho})+(\rho_A+\rho_i)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)).
\end{aligned}$$

The utilization for each type-$l\in\mathcal{T}$ satisfies $0\leq\rho_l<1$ and the total utilization (across arrivals for all types) is less than 1, therefore $0<1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B\leq 1-\bar{\rho}-\rho_A-\rho_i-\rho_k<1-\bar{\rho}-\rho_A-\rho_i<$

$1-\bar{\rho}-\rho_A \leq 1-\bar{\rho}$. The signs (positive or negative) of $\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^1_{T\times N})$ and $\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^2_{T\times N})$ depend only on the signs of Δ_1 and Δ_2,respectively. Further, recalling that $c_k\mu_k > c_i\mu_i$:

$$\Delta_1+\Delta_2 = (c_k\mu_k - c_i\mu_i)\Big[(\rho_k+\rho_B)(1-\bar{\rho})+(\rho_A+\rho_i)(1-\bar{\rho}-\rho_A-\rho_i-\rho_k-\rho_B)\Big] > 0.$$

so we conclude that at either $\Delta_1 > 0$ or $\Delta_2 > 0$ or both. In turn,

$$(\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^1_{T\times N}))+(\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^1_{T\times N}))>0,$$

implying that either $\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^1_{T\times N})>0$, or $\mathcal{C}(I^*_{T\times N})-\mathcal{C}(I^1_{T\times N})>0$ contradicting the optimality of $I^*_{T\times N}$. Q.E.D.

Proof of Proposition 2(a). Let us assume that there are $N^* < K$ queues with prioritization order and let G be a priority group with cardinality $|G|>1$ (i.e., G has more than one class).

Consider an alternative with N^*+1 queues where one of the prediction classes in G, let us call it j, is given priority over all classes in $G\backslash\{j\}$ i.e. $1>2>3,..>G-1>\{j\}>G/j>G+1>..N$. The difference between the costs of the optimal design and this alternative design:

$$\begin{aligned}
&=\sum_{i=1}^{T}c_i\lambda_i\mathbb{E}[\mathcal{S}]\Big(\frac{P_{iG}}{(1-\sum_{m=1}^{G-1}\rho_m(P))(1-\sum_{m=1}^{G}\rho_m(P))}\Big)\\
&-\sum_{i=1}^{T}c_i\lambda_i\mathbb{E}[\mathcal{S}]\Big(\frac{P_{ij}}{(1-\sum_{m=1}^{G-1}\rho_m(P))(1-\sum_{m=1}^{G-1}\rho_m(P)-\rho_j(P))}+\frac{P_{iG}-P_{ij}}{(1-\sum_{m=1}^{G-1}\rho_m(P)-\rho_j)(1-\sum_{m=1}^{G}\rho_m(P))}\Big)\\
&=\sum_{i=1}^{T}c_i\lambda_i\mathbb{E}[\mathcal{S}]\Big(\frac{P_{ij}\rho_G(P)-P_{iG}\rho_j}{(1-\sum_{m=1}^{G-1}\rho_m(P))(1-\sum_{m=1}^{G}\rho_m(P))(1-\sum_{m=1}^{G-1}\rho_m(P)-\rho_j)}\Big)
\end{aligned}$$

We claim that $\exists j$ such that the above expression is positive. This is because $\exists j\in G$ for which $P_{ij}\rho_G(P)-P_{iG}\rho_j$ is positive, which is because $\sum_{j\in G}P_{ij}\rho_G(P)-P_{iG}\rho_j = P_{iG}\rho_G(P)-P_{iG}\rho_G(P)=0$ Q.E.D.

Proof of Proposition 2(b). The difference in average waiting cost per unit time on interchanging priority queues m and n is equal to

$$\begin{aligned}
&\frac{\sum_{i\in\mathcal{T}}P_{im}\lambda_i c_i\mathbb{E}[\mathcal{S}]}{(1-\sum_{q\in[m-1]}\rho_q(P))(1-\sum_{q\in[m-1]}\rho_q(P)-\rho(m,P))}+\frac{\sum_{i\in\mathcal{T}}P_{in}\lambda_i c_i\mathbb{E}[\mathcal{S}]}{(1-\sum_{q\in[n-1]}\rho_q(P))(1-\sum_{q\in[n-1]}\rho_q(P)-\rho(n,P))}\\
&-\frac{\sum_{i\in\mathcal{T}}P_{in}\lambda_i c_i\mathbb{E}[\mathcal{S}]}{(1-\sum_{q\in[m-1]}\rho_q(P))(1-\sum_{q\in[m-1]}\rho_q(P)-\rho(n,P))}-\frac{\sum_{i\in\mathcal{T}}P_{im}\lambda_i c_i\mathbb{E}[\mathcal{S}]}{(1-\sum_{q\in[n-1]}\rho_q(P))(1-\sum_{q\in[n-1]}\rho_q(P)-\rho(n,P))}.
\end{aligned}$$

This difference is negative iff

$$\frac{\sum_{i\in\mathcal{T}}P_{im}\lambda_i c_i}{\sum_{i\in\mathcal{T}}P_{im}\lambda_i\dfrac{1}{\mu_i}} > \frac{\sum_{i\in\mathcal{T}}P_{in}\lambda_i c_i}{\sum_{i\in\mathcal{T}}P_{in}\lambda_i\dfrac{1}{\mu_i}}$$

Q.E.D.

Proof of Proposition 3 (a) The minimization problem can be reformulated as follows (see 11):

$$\max \frac{\rho P_{11}-1}{1-\rho P_{21}}$$

subject to:

$$TV(F_1,F_2)-P_{11}+P_{21}\geq 0$$

$$1-P_{11}\geq 0$$

$$1-P_{21}\geq 0$$

$$P_{11}\geq 0$$

$$P_{21}\geq 0$$

(We restrict to cases where $P_{11}\geq P_{21}$). The Lagrangian is equal to

$$\mathcal{L}(P_{11},P_{21},u_1,u_2,u_3,u_4)=\frac{\rho P_{11}-1}{1-\rho P_{21}}+u_1(TV(F_1,F_2)-P_{11}+P_{21})+u_2(1-P_{11})+u_3(1-P_{21})+u_4P_{11}+u_5P_{21}$$

$$\frac{\partial\mathcal{L}(P_{11},P_{21},u_1,u_2,u_3,u_4)}{P_{11}}=0\implies\frac{\rho}{1-\rho P_{21}}-u_1-u_2+u_4=0$$
$$\frac{\partial\mathcal{L}(P_{11},P_{21},u_1,u_2,u_3,u_4)}{P_{21}}=0\implies\frac{\rho(\rho P_{11}-1)}{(1-\rho P_{21})^2}+u_1-u_3+u_5=0$$

The conditions imply that $u_1+u_2>0$, and $u_1+u_5>0$. If $u_1=0$, then $u_2,u_5>0$, and $P_{11}=1$ and $P_{21}=0$, violating the constraint $TV(F_1,F_2)-P_{11}+P_{21}\geq 0$. Therefore, $u_1>0$, and $TV(F_1,F_2)-P_{11}+P_{21}=0$. The objective simplifies to

$$\frac{\rho P_{11}-1}{1-\rho P_{11}+\rho TV(F_1,F_2)}$$

Further,

$$\frac{d}{dP_{11}}\left(\frac{\rho P_{11}-1}{1-\rho P_{11}+\rho TV(F_1,F_2)}\right)=\frac{\rho(1-\rho P_{11}+\rho TV(F_1,F_2))+\rho(\rho P_{11}-1)}{(1-\rho P_{11}+\rho TV(F_1,F_2))^2}>0.$$

The maximum occurs at the greatest possible value of P_{11} which is 1. Q.E.D.

Proof of Proposition 3 (b) Adding the constraint $p_1P_{11}+p_2P_{21}=p_1$, the problem reduces to

$$\max\frac{\rho P_{11}-1}{p_2-\rho p_1+\rho p_1P_{11}}$$

subject to:

$$p_2TV(F_1,F_2)-P_{11}+p_1\geq 0$$

$$1-P_{11}\geq 0$$

$$P_{11}\geq 0$$

Now,

$$\frac{d}{dP_{11}}\frac{\rho P_{11}-1}{p_2-\rho p_1+\rho p_1 P_{11}}=\frac{\rho(p_2-\rho p_1+\rho p_1 P_{11})-\rho p_1(\rho P_{11}-1)}{(p_2-\rho p_1+\rho p_1 P_{11})^2}>0$$

Hence the maximum occurs at $P_{11}=p_1+p_2 TV(F_1,F_2)$ Q.E.D.

Proof of Lemma 1 From Corollary 6.1 in Bahadur (1954), the necessary and sufficient condition that the statistic $\mathbb{P}^T(X)$ is sufficient for the probability measures corresponding to densities $\{f_i(X); i\in\mathcal{T}\}$ is that there is a non-negative S-measurable function h on X and a set $\{g_i : i\in\mathcal{T}\}$ of S-measurable non-negative functions on the range of $\mathbb{P}^T(\cdot)$ such that:

$$f_i(x)=h(x)g_i(\mathbb{P}^T(X))\ \forall i\in\mathcal{T}$$

We know that $\forall i\in\mathcal{T}$

$$f_i(x)=(\sum_{i=1}^{T}p_if_i(x))(\frac{1}{p_i}\frac{p_if_i(x)}{\sum_{i=1}^{T}p_if_i(x)})=(\sum_{i=1}^{T}p_if_i(x))(\frac{1}{p_i}e_i'\mathbb{P}^T(X))$$

The corresponding values of $h(x)$ and $g_i(\mathbb{P}^T(X))$ are $\sum_{i=1}^{T}p_if_i(x)$ and $\frac{1}{p_i}e_i'\mathbb{P}^T(X)$, respectively. Both are non negative measurable functions. Q.E.D.

Proof of Lemma 2 $E[q_j^t(\mathbb{P}^T(X)|X\sim F_i]=E[E[q_j(X)|\mathbb{P}^T(X)]|X\sim F_i]=E[E[q_j(X)|\mathbb{P}^T(X),X\sim F_i]|X\sim F_i]$=$E[q_j(X)|X\sim F_i]$ (because $\mathbb{P}^T(X)$ is a sufficient statistic, $E[q_j(X)|\mathbb{P}^T(X)]$ is independent of the underlying type i, hence $E[q_j(X)|\mathbb{P}^T(X),X\sim F_i]=E[q_j(X)|\mathbb{P}^T(X)\ \forall i\in\mathcal{T})$. Q.E.D.

Proof of Lemma 4. The optimization problem can be reformulated as the linear fractional program:

$$maximize_{\beta}[\tfrac{1-\rho\beta' u}{1-\rho\beta' v}] \tag{27}$$

$$subject\ to:$$

$$A\beta\leq b$$

where $u=\frac{\sum_{r=1}^{s}(1-t_r)X_r}{\sum_{r=1}^{s}1-t_r}, v=\frac{\sum_{r=1}^{s}t_rX_r}{\sum_{r=1}^{s}t_r}$, $A=[u,v,-u,-v]', b=[1,1,0,0]'$. By using Charnes-Cooper linearization (Charnes and Cooper (1962)), and substituting $w_0=\frac{1}{1-\rho\beta' v}$, $w=\frac{\beta}{1-\rho\beta' v}$, the above optimization problem can be reformulated as:

$$maximize_{w_0,w}[w_0-\rho w' u]$$

$$subject\ to:$$

$$Aw\leq bw_0$$

$$w_0-\rho w' v=1$$

$$w_0\geq 0$$

Proof of Proposition 4: Consider $T = N = 3$. The difference in average waiting waiting costs under classification matrices P^{FIFO} (a single queue with first in first out regime such that $P_{ij}^{FIFO} = 1 \forall i \in \{1,2,3\}, j = 1$) and P is

$$\Delta\mathcal{C}(P) = \mathcal{C}(\mathcal{P}^{\mathcal{FIFO}}) - \mathcal{C}(\mathcal{P}) = (c_1\mu_1 - c_3\mu_3)\Delta W_1(P) + (c_2\mu_2 - c_3\mu_3)\Delta W_2(P) \tag{28}$$

where $\Delta W_1(P)$ and $\Delta W_2(P)$ are the reduction in average waiting times of types 1 and 2 under P as compared to a random classifier P^{FIFO}:

$$\begin{aligned}\Delta W_1(P) = [&\tfrac{\rho_1}{1-\rho} - \tfrac{\rho_1 P_{11}}{1-\rho_1 P_{11}-\rho_2 P_{21}-\rho_3 P_{31}} - \tfrac{\rho_1 P_{12}}{(1-\rho_1 P_{11}-\rho_2 P_{21}-\rho_3 P_{31})(1-\rho_1(P_{11}+P_{12})-\rho_2(P_{21}+P_{22})-\rho_3(P_{31}+P_{32})} \\ &- \tfrac{\rho_1 P_{13}}{(1-\rho_1(P_{11}+P_{12})-\rho_2(P_{21}+P_{22})-\rho_3(P_{31}+P_{32})))(1-\rho)}] \\ \Delta W_2(P) = [&\tfrac{\rho_2}{1-\rho} - \tfrac{\rho_2 P_{21}}{1-\rho_1 P_{11}-\rho_2 P_{21}-\rho_3 P_{31}} - \tfrac{\rho_2 P_{22}}{(1-\rho_1 P_{11}-\rho_2 P_{21}-\rho_3 P_{31})(1-\rho_1(P_{11}+P_{12})-\rho_2(P_{21}+P_{22})-\rho_3(P_{31}+P_{32})} \\ &- \tfrac{\rho_2 P_{23}}{(1-\rho_1(P_{11}+P_{12})-\rho_2(P_{21}+P_{22})-\rho_3(P_{31}+P_{32})))(1-\rho)}]\end{aligned} \tag{29}$$

The optimal P^* solves the following optimization problem

$$\begin{aligned}
&maximize (c_1\mu_1 - c_3\mu_3)\Delta W_1(P) + (c_2\mu_2 - c_3\mu_3)\Delta W_2(P) \iff \\
&subject\ to: \\
&TV(F_1,F_2) - P_{11} + P_{21} \geq 0,\ TV(F_1,F_2) - P_{21} + P_{11} \geq 0 \\
&TV(F_1,F_2) - P_{12} + P_{22} \geq 0,\ TV(F_1,F_2) - P_{22} + P_{12} \geq 0 \\
&TV(F_1,F_3) - P_{11} + P_{31} \geq 0,\ TV(F_1,F_3) - P_{31} + P_{11} \geq 0 \\
&TV(F_1,F_3) - P_{32} + P_{12} \geq 0,\ TV(F_1,F_3) - P_{12} + P_{32} \geq 0 \\
&TV(F_2,F_3) - P_{21} + P_{31} \geq 0,\ TV(F_2,F_3) - P_{31} + P_{21} \geq 0 \\
&TV(F_2,F_3) - P_{22} + P_{32} \geq 0,\ TV(F_2,F_3) - P_{32} + P_{22} \geq 0 \\
&1 - P_{11} \geq 0,\ 1 - P_{12} \geq 0 \\
&1 - P_{21} \geq 0,\ 1 - P_{22} \geq 0 \\
&1 - P_{31} \geq 0,\ 1 - P_{32} \geq 0 \\
&1 - P_{11} - P_{12} \geq 0,\ 1 - P_{21} - P_{22} \geq 0,\ 1 - P_{31} - P_{32} \geq 0 \\
&P_{11} \geq 0,\ P_{12} \geq 0 \\
&P_{21} \geq 0,\ P_{22} \geq 0 \\
&P_{31} \geq 0,\ P_{32} \geq 0 \\
&P_{11} + P_{12} - P_{21} - P_{22} + TV(F_1,F_2) \geq 0 \\
&P_{21} + P_{22} - P_{11} - P_{12} + TV(F_1,F_2) \geq 0
\end{aligned}$$

$$P_{11} + P_{12} - P_{31} - P_{32} + TV(F_1, F_3) \geq 0$$

$$P_{32} + P_{31} - P_{11} - P_{12} + TV(F_1, F_3) \geq 0$$

$$P_{21} + P_{22} - P_{31} - P_{32} + TV(F_2, F_3) \geq 0$$

$$P_{31} + P_{32} - P_{21} - P_{22} + TV(F_2, F_3) \geq 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{11}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{11}} - \lambda_1 + \lambda_2 - \lambda_5 + \lambda_6 - \lambda_{13} - \lambda_{19} + \lambda_{22} + \lambda_{28} - \lambda_{29} + \lambda_{30} - \lambda_{31} = 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{12}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{12}} - \lambda_3 + \lambda_4 + \lambda_7 - \lambda_8 - \lambda_{14} - \lambda_{19} + \lambda_{23} + \lambda_{28} - \lambda_{29} + \lambda_{30} - \lambda_{31} = 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{21}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{21}} + \lambda_1 - \lambda_2 - \lambda_9 + \lambda_{10} - \lambda_{15} - \lambda_{20} + \lambda_{24} - \lambda_{28} + \lambda_{29} + \lambda_{32} - \lambda_{33} = 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{22}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{22}} + \lambda_3 - \lambda_4 - \lambda_{11} + \lambda_{12} - \lambda_{16} - \lambda_{20} + \lambda_{25} - \lambda_{28} + \lambda_{29} + \lambda_{32} - \lambda_{33} = 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{31}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{31}} + \lambda_5 - \lambda_6 + \lambda_9 - \lambda_{10} - \lambda_{17} - \lambda_{21} + \lambda_{26} - \lambda_{30} + \lambda_{31} - \lambda_{32} + \lambda_{33} = 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{32}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{32}} - \lambda_7 + \lambda_8 + \lambda_{11} - \lambda_{12} - \lambda_{18} - \lambda_{21} + \lambda_{27} - \lambda_{30} + \lambda_{31} - \lambda_{32} + \lambda_{33} = 0$$

$$\frac{\partial \Delta W_1(P)}{\partial P_{11}} = \frac{\rho_1}{(1-\rho(1,P)-\rho(2,P))(1-\rho)} - \frac{\rho_1}{1-\rho(1,P)} - \frac{\rho_1^2 P_{11}}{(1-\rho(1,P))^2} - \frac{\rho_1^2 P_{12}}{(1-\rho(1,P))^2(1-\rho(1,P)-\rho(2,P))}$$

$$- \frac{\rho_1^2 P_{12}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2} - \frac{\rho_1^2 P_{13}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_1(P)}{\partial P_{12}} = \frac{\rho_1}{(1-\rho(1,P)-\rho(2,P))(1-\rho)} - \frac{\rho_1}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))}$$

$$- \frac{\rho_1^2 P_{12}}{(1-\rho(1,P)(1-\rho(1,P)-\rho(2,P))^2} - \frac{\rho_1^2 P_{13}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_1(P)}{\partial P_{21}} = -\frac{\rho_1\rho_2 P_{11}}{(1-\rho(1,P))^2} - \frac{\rho_1\rho_2 P_{12}}{(1-\rho(1,P))^2(1-\rho(1,P)-\rho(2,P))} - \frac{\rho_1\rho_2 P_{12}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2}$$

$$- \frac{\rho_1\rho_2 P_{13}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_1(P)}{\partial P_{22}} = -\frac{\rho_1\rho_2 P_{12}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2} - \frac{\rho_1\rho_2 P_{13}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_2(P)}{\partial P_{21}} = \frac{\rho_2}{(1-\rho(1,P)-\rho(2,P))(1-\rho)} - \frac{\rho_2}{1-\rho(1,P)} - \frac{\rho_2^2 P_{21}}{(1-\rho(1,P))^2} - \frac{\rho_2^2 P_{22}}{(1-\rho(1,P))^2(1-\rho(1,P)-\rho(2,P))}$$

$$- \frac{\rho_2^2 P_{22}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2} - \frac{\rho_2^2 P_{23}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_2(P)}{\partial P_{22}} = \frac{\rho_2}{(1-\rho(1,P)-\rho(2,P))(1-\rho)} - \frac{\rho_2}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))} -$$

$$\frac{\rho_2^2 P_{22}}{(1-\rho(1,P)(1-\rho(1,P)-\rho(2,P))^2} - \frac{\rho_2^2 P_{23}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_2(P)}{\partial P_{11}} = -\frac{\rho_1\rho_2 P_{21}}{(1-\rho(1,P))^2} - \frac{\rho_1\rho_2 P_{22}}{(1-\rho(1,P))^2(1-\rho(1,P)-\rho(2,P))} - \frac{\rho_1\rho_2 P_{22}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2}$$

$$- \frac{\rho_1\rho_2 P_{23}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_2(P)}{\partial P_{12}} = -\frac{\rho_1\rho_2 P_{22}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2} - \frac{\rho_1\rho_2 P_{23}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_2(P)}{\partial P_{31}} = -\frac{\rho_3\rho_2 P_{21}}{(1-\rho(1,P))^2} - \frac{\rho_3\rho_2 P_{22}}{(1-\rho(1,P))^2(1-\rho(1,P)-\rho(2,P))} - \frac{\rho_3\rho_2 P_{22}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2}$$

$$- \frac{\rho_3\rho_2 P_{23}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

$$\frac{\partial \Delta W_2(P)}{\partial P_{32}} = -\frac{\rho_3\rho_2 P_{22}}{(1-\rho(1,P))(1-\rho(1,P)-\rho(2,P))^2} - \frac{\rho_3\rho_2 P_{23}}{(1-\rho(1,P)-\rho(2,P))^2(1-\rho)}$$

Proof (i)

First, we focus on solutions where $P_{11}=1, P_{12}=0, P_{13}=0$

1. $\frac{\partial \Delta W_1(P)}{\partial P_{11}} > 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{12}} > 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{21}} < 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{22}} = 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{31}} < 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{32}} < 0$
2. $\frac{\partial \Delta W_2(P)}{\partial P_{11}} < 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{12}} < 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{21}} > 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{22}} > 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{31}} < 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{32}} < 0$
3. If $\frac{c_1\mu_1 - c_2\mu_2}{c_2\mu_2 - c_3\mu_3}$ is sufficiently large, then

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{11}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{11}} > 0 \implies \lambda_1 + \lambda_5 + \lambda_{13} + \lambda_{19} + \lambda_{29} + \lambda_{31} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{12}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{12}} > 0 \implies \lambda_3 + \lambda_8 + \lambda_{14} + \lambda_{19} + \lambda_{29} + \lambda_{31} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{21}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{21}} < 0 \implies \lambda_1 + \lambda_{10} + \lambda_{24} + \lambda_{39} + \lambda_{32} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{22}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{22}} > 0 \implies \lambda_4 + \lambda_{11} + \lambda_{16} + \lambda_{20} + \lambda_{28} + \lambda_{33} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{31}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{31}} < 0 \implies \lambda_5 + \lambda_9 + \lambda_{26} + \lambda_{31} + \lambda_{33} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{32}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{32}} < 0 \implies \lambda_8 + \lambda_{11} + \lambda_{27} + \lambda_{31} + \lambda_{33} > 0$$

Table 6 shows all possible values of P_{11}, P_{12}, P_{21}, P_{22}, P_{31}, and P_{32} that satisfy the above constraints. First, we consider the solutions where $P_{21}+P_{22}=1-TV(F_1,F_2)$ or $P_{31}+P_{32}=1-TV(F_1,F_3)$

(i) Since $P_{21} \geq 1-TV(F_1,F_2)$, so one solution is $P_{21}=1-TV(F_1,F_2)$, $P_{22}=0$, $P_{23}=TV(F_1,F_2)$. Since $P_{22}=0$, the only possibility (check under P_{22}) is $P_{31}+P_{32}=1-TV(F_1,F_2)-TV(F_2,F_3)$ but then it contradicts $|P_{13}-P_{33}| \leq TV(F_1,F_3)$ and due to triangle inequality $TV(F_1,F_3) \leq TV(F_1,F_2)+TV(F_2,F_3)$

(ii) Now the only possibility is $P_{31}=1-TV(F_1,F_3),\ P_{32}=0,\ P_{33}=TV(F_1,F_3)$

(iii) Since $|P_{11}-P_{21}| \leq TV(F_1,F_2)$, $P_{21}=1-TV(F_1,F_2)$

(iv) $P_{22}=TV(F_1,F_2)$ (not possible because $P_{23}-P_{33} \leq TV(F_2,F_3)$)

(v) $P_{22}=TV(F_2,F_3)$ (possible only when $TV(F_2,F_3) < TV(F_1,F_2)$), then $P_{23}=TV(F_1,F_2)-TV(F_2,F_3)$ and $|P_{33}-P_{23}|=|TV(F_1,F_3)-TV(F_1,F_2)+TV(F_2,F_3)|$, which is not possible

P_{11}	P_{12}	P_{21}
$P_{11}=P_{21}+TV(F_1,F_2)$	$P_{12}=P_{22}+TV(F_1,F_2)$	$P_{21}=P_{11}-TV(F_1,F_2)$
$P_{11}=P_{31}+TV(F_1,F_3)$	$P_{12}=P_{32}+TV(F_1,F_3)$	$P_{21}=P_{31}-TV(F_2,F_3)$
$P_{11}=1$	$P_{12}=1$	$P_{21}=0$
$P_{11}=1-P_{12}$	$P_{12}=1-P_{11}$	$P_{21}=0$
$P_{11}+P_{12}=P_{21}+P_{22}+TV(F_1,F_2)$	$P_{11}+P_{12}=P_{21}+P_{22}+TV(F_1,F_2)$	$P_{21}+P_{22}=P_{11}+P_{12}-TV(F_1,F_2)$
$P_{11}+P_{12}=P_{31}+P_{32}+TV(F_1,F_3)$	$P_{11}+P_{12}=P_{31}+P_{32}+TV(F_1,F_3)$	$P_{21}+P_{22}=P_{31}+P_{32}-TV(F_2,F_3)$
P_{22}	P_{31}	P_{32}
$P_{22}=P_{12}+TV(F_1,F_2)$	$P_{31}=P_{11}-TV(F_1,F_3)$	$P_{32}=P_{12}-TV(F_1,F_3)$
$P_{22}=P_{32}+TV(F_2,F_3)$	$P_{31}=P_{21}-TV(F_2,F_3)$	$P_{32}=P_{22}-TV(F_2,F_3)$
$P_{22}=1$	$P_{31}=0$	$P_{32}=0$
$P_{22}=1-P_{21}$		
$P_{21}+P_{22}=P_{11}+P_{12}+TV(F_1,F_2)$	$P_{31}+P_{32}=P_{11}+P_{12}-TV(F_1,F_3)$	$P_{31}+P_{32}=P_{11}+P_{12}-TV(F_1,F_3)$
$P_{21}+P_{22}=P_{31}+P_{32}+TV(F_2,F_3)$	$P_{31}+P_{32}=P_{21}+P_{22}-TV(F_2,F_3)$	$P_{31}+P_{32}=P_{21}+P_{22}-TV(F_2,F_3)$

Table 6

(vi) Hence $P_{22}=TV(F_1,F_2)+TV(F_2,F_3)-TV(F_1,F_3)$, $P_{23}=TV(F_1,F_3)-TV(F_2,F_3)$

$$P_1=\begin{bmatrix} 1, & 0, & 0 \\ 1-TV(F_1,F_2), & TV(F_1,F_2)+TV(F_2,F_3)-TV(F_1,F_3), & TV(F_1,F_3)-TV(F_2,F_3) \\ 1-TV(F_1,F_3), & 0, & TV(F_1,F_3) \end{bmatrix}$$

Next, we consider the solutions where $P_{21}=P_{11}-TV(F_1,F_2)$ or $P_{21}=P_{31}-TV(F_2,F_3)$

(i) Let $P_{21}=1-TV(F_1,F_2)$

(a) $P_{31}=P_{21}-TV(F_2,F_3)$, 0 are not possible

(b) Let $P_{31}=1-TV(F_1,F_3)$ and $P_{22}=TV(F_1,F_2)$. For $P_{32}=TV(F_1,F_2)-TV(F_2,F_3)$ to be feasible, $TV(F_1,F_2)-TV(F_2,F_3)\geq 0$. But then $P_{33}=$ $TV(F_1,F_3)-TV(F_1,F_2)+TV(F_2,F_3)>TV(F_2,F_3)$ (not possible because $TV(F_1,F_3)\geq TV(F_2,F_3)$). If $P_{32}=0$, then $P_{33}=TV(F_1,F_3)>TV(F_2,F_3)$ (not possible). Next, if $P_{32}=TV(F_1,F_3)-TV(F_2,F_3)$, and $P_{33}=P_{23}$. So a feasible solution is

$$P_2=\begin{bmatrix} 1, & 0, & 0 \\ 1-TV(F_1,F_2), & TV(F_1,F_2), & 0 \\ 1-TV(F_1,F_3), & TV(F_1,F_3)-TV(F_2,F_3), & TV(F_2,F_3) \end{bmatrix}$$

(c) Let $P_{31}=1-TV(F_1,F_3)$ and $P_{22}=P_{32}+TV(F_2,F_3) \implies |P_{23}-P_{33}|=TV(F_2,F_3)+TV(F_1,F_3)-TV(F_1,F_2)>TV(F_2,F_3)$

(d) Let $P_{31}=1-TV(F_1,F_3)$ and $P_{21}+P_{22}=P_{31}+P_{32}+TV(F_2,F_3) \implies P_{22}=TV(F_1,F_2)-TV(F_1,F_3)+TV(F_2,F_3)+P_{32}$. The solution is of the form

$$P=\begin{bmatrix} 1, & 0, & 0 \\ 1-TV(F_1,F_2), & TV(F_1,F_2)+TV(F_2,F_3)-TV(F_1,F_3)+P_{32}, & TV(F_1,F_3)-TV(F_2,F_3)-P_{32} \\ 1-TV(F_1,F_3), & P_{32}, & TV(F_1,F_3)-P_{32} \end{bmatrix}$$

where $0 \leq P_{32} \leq TV(F_1,F_3) - TV(F_2,F_3)$. We can show that $\mathcal{C}(P)$ is a linear fractional in P_{32}, and is monotonic in P_{32}. Hence the optimal solution occurs at $P_{32} = 0$ (which leads to $P = P_1$) or at $P_{32} = TV(F_1,F_3) - TV(F_2,F_3)$ (which leads to $P = P_2$).

(ii) Let $P_{21} = P_{31} - TV(F_2,F_3)$. There is no feasible value of P_{31}.

Next, we focus on the solutions where $P_{11} + P_{12} = 1, P_{12} > 0$

1. $\frac{\partial \Delta W_1(P)}{\partial P_{11}} > 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{12}} > 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{21}} < 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{22}} < 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{31}} < 0$, $\frac{\partial \Delta W_1(P)}{\partial P_{32}} < 0$
2. $\frac{\partial \Delta W_2(P)}{\partial P_{11}} < 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{12}} < 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{21}} > 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{22}} > 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{31}} < 0$, $\frac{\partial \Delta W_2(P)}{\partial P_{32}} < 0$
3. We consider solutions where:

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{11}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{11}} > 0 \implies \lambda_1 + \lambda_5 + \lambda_{13} + \lambda_{19} + \lambda_{29} + \lambda_{31} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{12}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{12}} > 0 \implies \lambda_3 + \lambda_8 + \lambda_{14} + \lambda_{19} + \lambda_{29} + \lambda_{31} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{21}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{21}} < 0 \implies \lambda_1 + \lambda_{10} + \lambda_{24} + \lambda_{39} + \lambda_{32} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{22}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{22}} < 0 \implies \lambda_3 + \lambda_{12} + \lambda_{25} + \lambda_{29} + \lambda_{32} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{31}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{31}} < 0 \implies \lambda_5 + \lambda_9 + \lambda_{26} + \lambda_{31} + \lambda_{33} > 0$$

$$(c_1\mu_1 - c_3\mu_3)\frac{\partial \Delta W_1(P)}{\partial P_{32}} + (c_2\mu_2 - c_3\mu_3)\frac{\partial \Delta W_2(P)}{\partial P_{32}} < 0 \implies \lambda_8 + \lambda_{11} + \lambda_{27} + \lambda_{31} + \lambda_{33} > 0$$

P_{11}	P_{12}	P_{21}
$P_{11} = P_{21} + TV(F_1,F_2)$	$P_{12} = P_{22} + TV(F_1,F_2)$	$P_{21} = P_{11} - TV(F_1,F_2)$
$P_{11} = P_{31} + TV(F_1,F_3)$	$P_{12} = P_{32} + TV(F_1,F_3)$	$P_{21} = P_{31} - TV(F_2,F_3)$
$P_{11} = 1$	$P_{12} = 1$	$P_{21} = 0$
$P_{11} = 1 - P_{12}$	$P_{12} = 1 - P_{11}$	
$P_{11} + P_{12} = P_{21} + P_{22} + TV(F_1,F_2)$	$P_{11} + P_{12} = P_{21} + P_{22} + TV(F_1,F_2)$	$P_{21} + P_{22} = P_{11} + P_{12} - TV(F_1,F_2)$
$P_{11} + P_{12} = P_{31} + P_{32} + TV(F_1,F_3)$	$P_{11} + P_{12} = P_{31} + P_{32} + TV(F_1,F_3)$	$P_{21} + P_{22} = P_{31} + P_{32} - TV(F_2,F_3)$
P_{22}	P_{31}	P_{32}
$P_{22} = P_{12} - TV(F_1,F_2)$	$P_{31} = P_{11} - TV(F_1,F_3)$	$P_{32} = P_{12} - TV(F_1,F_3)$
$P_{22} = P_{32} - TV(F_2,F_3)$	$P_{31} = P_{21} - TV(F_2,F_3)$	$P_{32} = P_{22} - TV(F_2,F_3)$
$P_{22} = 0$	$P_{31} = 0$	$P_{32} = 0$
$P_{21} + P_{22} = P_{11} + P_{12} - TV(F_1,F_2)$	$P_{31} + P_{32} = P_{11} + P_{12} - TV(F_1,F_3)$	$P_{31} + P_{32} = P_{11} + P_{12} - TV(F_1,F_3)$
$P_{21} + P_{22} = P_{31} + P_{32} - TV(F_2,F_3)$	$P_{31} + P_{32} = P_{21} + P_{22} - TV(F_2,F_3)$	$P_{31} + P_{32} = P_{21} + P_{22} - TV(F_2,F_3)$

Table 7

Table 7 shows all possible values of P_{11}, P_{12}, P_{21}, P_{22}, P_{31}, and P_{32} that satisfy the above constraints. Next, we explore the different possible solutions:

1. Let $P_{21} = P_{11} - TV(F_1,F_2)$
   (a) $P_{22} = P_{12} - TV(F_1,F_2)$ is not possible

(b) Let $P_{22} = P_{32} - TV(F_2, F_3)$. Now $P_{32} = P_{12} - TV(F_1, F_3)$, $P_{32} = P_{22} - TV(F_2, F_3)$, $P_{32} = 0$, $P_{31} + P_{32} = P_{21} + P_{22} - TV(F_2, F_3)$ are not possible. Let $P_{31} + P_{32} = P_{11} + P_{12} - TV(F_1, F_3)$. Therefore, $P_{12} - P_{32} - TV(F_1, F_3) = P_{31} - P_{11} \geq -TV(F_1, F_3) \implies P_{12} \geq P_{32} \geq P_{22} + TV(F_2, F_3)$. But $P_{11} \geq P_{21} + TV(F_1, F_2)$. Not possible.

(c) Let $P_{22} = 0$. Then $P_{23} = P_{12} + TV(F_1, F_2)$ but $P_{13} = 0$, hence $P_{23} - P_{13} > 0$ (not possible)

(d) Let $P_{21} + P_{22} = P_{11} + P_{12} - TV(F_1, F_2)$. This implies $P_{22} = P_{12}$, $P_{23} = TV(F_1, F_2)$. Now $P_{32} = P_{12} - TV(F_1, F_3) = P_{22} - TV(F_1, F_3)$ is not possible. Let $P_{32} = P_{22} - TV(F_2, F_3) = P_{12} - TV(F_2, F_3)$. Now $P_{31} = P_{11} - TV(F_1, F_3)$, $P_{31} = P_{21} - TV(F_2, F_3)$ are not possible. If $P_{31} = 0$, then $P_{33} = 1 - P_{12} + TV(F_1, F_3) = P_{11} + TV(F_1, F_3)$ (not possible because $P_{13} = TV(F_1, F_3)$

(e) Let $P_{21} + P_{22} = P_{31} + P_{32} - TV(F_2, F_3)$. Hence $P_{31} + P_{32} = P_{11} + P_{12} - TV(F_1, F_3)$ is not possible. Now, $P_{31} = P_{11} - TV(F_1, F_3)$ and $P_{32} = 0$ are not possible because then $P_{33} = 1 - P_{11} + TV(F_1, F_3)$ (not possible). $P_{31} = P_{11} - TV(F_1, F_3)$ and $P_{32} = P_{22} - TV(F_2, F_3)$ implies $P_{21} = P_{31} - 2TV(F_2, F_3)$ (not possible). $P_{31} = P_{21} - TV(F_2, F_3)$ (not possible). If $P_{31} = 0$, then $P_{32} = P_{12} - TV(F_1, F_3)$ implies $P_{33} = P_{11} + TV(F_1, F_3)$, not possible. If $P_{31} = 0$, then $P_{32} = P_{22} - TV(F_2, F_3)$ implies $P_{21} = P_{31} - 2TV(F_2, F_3)$. If $P_{31} = 0$, then $P_{32} = 0$ is not possible.

2. Let $P_{21} = P_{31} - TV(F_2, F_3)$.

(a) The only feasible P_{31} satisfies $P_{31} + P_{32} = P_{11} + P_{12} - TV(F_1, F_3)$ or $P_{31} + P_{32} = P_{21} + P_{22} - TV(F_2, F_3)$.

(b) First, let $P_{31} + P_{32} = P_{11} + P_{12} - TV(F_1, F_3)$ i.e., $P_{33} = TV(F_1, F_3)$. Now, $P_{22} = P_{32} - TV(F_2, F_3)$ is not possible. If $P_{22} = P_{12} - TV(F_1, F_2)$, then $P_{23} = P_{11} - P_{31} + TV(F_2, F_3) + TV(F_1, F_2)$ (not possible because for the optimal solution $P_{11} - P_{31} \geq 0$). If $P_{22} = 0$, then $P_{23} = P_{32} + P_{33} + TV(F_2, F_3) = P_{32} + TV(F_1, F_3) + TV(F_2, F_3)$ (not possible). Now, if P_{22} satisfies $P_{21} + P_{22} = P_{11} + P_{12} - TV(F_1, F_2)$, then $P_{21} + P_{22} = P_{31} + P_{32} + TV(F_1, F_3) - TV(F_1, F_2)$ i.e., $P_{23} = P_{33} - TV(F_1, F_3) + TV(F_1, F_2) = TV(F_1, F_2)$. Then, $P_{22} = 1 - P_{31} + TV(F_2, F_3) - TV(F_1, F_2)$, $P_{32} = 1 - P_{31} - TV(F_1, F_3)$, and $P_{22} - P_{32} = TV(F_1, F_3) - TV(F_1, F_2) + TV(F_2, F_3) > TV(F_2, F_3)$ (not possible). Now, if P_{22} satisfies $P_{21} + P_{22} = P_{31} + P_{32} - TV(F_2, F_3) = P_{11} + P_{22} - TV(F_1, F_3) - TV(F_2, F_3)$ (not possible).

(c) Next, let $P_{31} + P_{32} = P_{21} + P_{22} - TV(F_2, F_3) \implies P_{32} = P_{22} - 2TV(F_2, F_3)$. (not possible)

3. Let $P_{21} = 0$

(a) $P_{22} = P_{12} - TV(F_1, F_2)$, $P_{23} = 1 - P_{12} + TV(F_1, F_2) = P_{11} + TV(F_1, F_2)$ (not possible because $P_{13} = 0$).

(b) $P_{22} = P_{32} - TV(F_2, F_3)$ is not possible

(c) $P_{22} = 0$ is not possible, because then $P_{23} = 1$

(d) $P_{21}+P_{22}=P_{11}+P_{12}-TV(F_1,F_2)$, $P_{22}=1-TV(F_1,F_2)$, $P_{23}=TV(F_1,F_2)$. Now, let $P_{31}=P_{11}-TV(F_1,F_3)$. For this to be possible, $P_{11}>TV(F_1,F_3)$. But then $P_{21}=0$ (so not possible)

(e) $P_{21}+P_{22}=P_{31}+P_{32}-TV(F_2,F_3)$. Now $P_{31}=P_{11}-TV(F_1,F_3)$, $P_{32}=0$ is not possible, because then $P_{33}=P_{12}+TV(F_1,F_3)$. The only possible combination is $P_{31}=P_{11}-TV(F_1,F_3)$ and $P_{32}=P_{22}-TV(F_2,F_3)=P_{31}+P_{32}-2TV(F_2,F_3)$, hence $P_{31}=2TV(F_2,F_3)$ (not possible because $P_{21}=0$)

4. Let $P_{21}+P_{22}=P_{11}+P_{12}-TV(F_1,F_2)$

(a) If $P_{31}+P_{32}=P_{11}+P_{12}-TV(F_1,F_3)$, then the objective is to maximize

$$\frac{c_1\mu_1\rho_1(\bar{\rho}P_{11}-1)+c_2\mu_2\rho_2(\bar{\rho}P_{21}-(1-TV(F_1,F_2)))+c_3\mu_3\rho_3(\bar{\rho}P_{31}-(1-TV(F_1,F_3)))}{1-\rho_1P_{11}-\rho_2P_{21}-\rho_3P_{31}}$$

where $\bar{\rho}=\rho-\rho_2TV(F_1,F_2)-\rho_3TV(F_1,F_3)$. Since it is optimal to keep P_{31} as small as possible, we have $P_{31}=P_{11}-TV(F_1,F_3)$. When $\frac{c_1\mu_1-c_2\mu_2}{c_2\mu_2-c_3\mu_3}$ is sufficiently high, we can show that that it is optimal to have $P_{21}=P_{11}-TV(F_1,F_2)$. In this case, we can prove that the objective is increasing in P_{11} i.e., the optimal value is at $P_{11}=1$,$P_{12}=0$.

(b) $P_{31}+P_{32}=P_{21}+P_{22}-TV(F_2,F_3)=P_{11}+P_{12}-TV(F_1,F_2)-TV(F_2,F_3)$ is not possible

5. Let $P_{21}+P_{22}=P_{31}+P_{32}-TV(F_2,F_3)$.

(a) Let $P_{31}=P_{11}-TV(F_1,F_3)$, then $P_{32}=P_{12}-TV(F_1,F_3)$ or $P_{32}=0$ are not possible. Let $P_{32}=P_{22}-TV(F_2,F_3)$, then $P_{21}+P_{22}=P_{31}+P_{22}-2TV(F_2,F_3)$, or $P_{21}=P_{31}-2TV(F_2,F_3)$ (not possible)

(b) Let $P_{31}=P_{21}-TV(F_2,F_3)$, then $P_{22}=P_{32}-2TV(F_2,F_3)$ (not possible)

(c) Let $P_{31}=0$, then $P_{32}=P_{12}-TV(F_1,F_3)$ is not possible, because $P_{33}=P_{11}+TV(F_1,F_3)$ (not possible). If $P_{32}=P_{22}-TV(F_2,F_3)$, then $P_{21}=-2TV(F_2,F_3)$ (not possible). $P_{32}=0$ is not possible.

The candidates for the optimal solution are P_1 and P_2. Through numerical experiments, we find that both the matrices are possible. Q.E.D.

Proof of Proposition 5:

$$minimize\mathcal{C}(P)=maximize\frac{c_1\mu_1\rho_1(\rho P_{11}-1)+c_2\mu_2\rho_2(\rho P_{21}-1)+c_3\mu_3\rho_3(\rho P_{31}-1)}{(1-\rho_1P_{11}-\rho_2P_{21}-\rho_3P_{31})(1-\rho)}$$

$$subject\ to:$$

$$TV(F_1,F_2)-P_{11}+P_{21}\geq 0$$

$$TV(F_1,F_3)-P_{11}+P_{31}\geq 0$$

$$TV(F_2,F_3)-P_{21}+P_{31}\geq 0$$

$$1-P_{11}\geq 0$$

$$1 - P_{21} \geq 0$$

$$1 - P_{31} \geq 0$$

The lagrangian is:

$$\mathcal{L}(P_{11}, P_{21}, P_{31}, \Lambda) = \frac{c_1\mu_1\rho_1(\rho P_{11} - 1) + c_2\mu_2\rho_2(\rho P_{21} - 1) + c_3\mu_3\rho_3(\rho P_{31} - 1)}{(1 - \rho_1 P_{11} - \rho_2 P_{21} - \rho_3 P_{31})(1 - \rho)}$$
$$+ \lambda_1(TV(F_1, F_2) - P_{11} + P_{21}) + \lambda_2(TV(F_1, F_3) - P_{11} + P_{31}) + \lambda_3(TV(F_2, F_3) - P_{21} + P_{31})$$
$$+ \lambda_4(1 - P_{11}) + \lambda_5(1 - P_{21}) + \lambda_6(1 - P_{31})$$
$$\frac{\partial\mathcal{L}(P_{11}, P_{21}, P_{31}, \Lambda)}{\partial P_{11}} = \frac{\rho_1\rho_2(c_1\mu_1 - c_2\mu_2)(1 - \rho P_{21}) + \rho_1\rho_3(c_1\mu_1 - c_3\mu_3)(1 - \rho P_{31})}{(1 - \rho_1 P_{11} - \rho_2 P_{21} - \rho_3 P_{31})^2(1 - \rho)^2} - \lambda_1 - \lambda_2 - \lambda_4 = 0$$
$$\frac{\partial\mathcal{L}(P_{11}, P_{21}, P_{31}, \Lambda)}{\partial P_{21}} = \frac{\rho_1\rho_2(c_2\mu_2 - c_1\mu_1)(1 - \rho P_{11}) + \rho_2\rho_3(c_2\mu_2 - c_3\mu_3)(1 - \rho P_{31})}{(1 - \rho_1 P_{11} - \rho_2 P_{21} - \rho_3 P_{31})^2(1 - \rho)^2} + \lambda_1 - \lambda_3 - \lambda_5 = 0$$
$$\frac{\partial\mathcal{L}(P_{11}, P_{21}, P_{31}, \Lambda)}{\partial P_{31}} = \frac{\rho_1\rho_3(c_3\mu_3 - c_1\mu_1)(1 - \rho P_{11}) + \rho_2\rho_3(c_3\mu_3 - c_2\mu_2)(1 - \rho P_{21})}{(1 - \rho_1 P_{11} - \rho_2 P_{21} - \rho_3 P_{31})^2(1 - \rho)^2} + \lambda_2 - \lambda_3 - \lambda_6 = 0$$

1. $$\frac{\rho_1\rho_3(c_3\mu_3 - c_1\mu_1)(1 - \rho P_{11}) + \rho_2\rho_3(c_3\mu_3 - c_2\mu_2)(1 - \rho P_{21})}{(1 - \rho_1 P_{11} - \rho_2 P_{21} - \rho_3 P_{31})^2(1 - \rho)^2} < 0 \implies \lambda_2 > 0 \implies P_{31} = P_{11} - TV(F_1, F_3)$$

2. Consider $\rho_1\rho_2(c_2\mu_2 - c_1\mu_1)(1 - \rho P_{11}) + \rho_2\rho_3(c_2\mu_2 - c_3\mu_3)(1 - \rho P_{31}) =$
$$\rho_2\rho_1(c_2\mu_2 - c_3\mu_3)\left(\frac{\rho_3(1 - \rho P_{11} + \rho TV(F_1, F_3))}{\rho_1(1 - \rho P_{11})} - \gamma\right)$$
For $TV(F_1, F_3) \leq P_{11} \leq 1$,
$$\frac{\rho_3}{\rho_1(1 - \rho TV(F_1, F_3))} - \gamma \leq \frac{\rho_3(1 - \rho P_{11} + \rho TV(F_1, F_3))}{\rho_1(1 - \rho P_{11})} - \gamma \leq \frac{\rho_3(1 - \rho + \rho TV(F_1, F_3))}{\rho_1(1 - \rho)} - \gamma$$

3. If $\gamma > \frac{\rho_3(1-\rho+\rho TV(F_1,F_3))}{\rho_1(1-\rho)}$, then
$$\rho_1\rho_2(c_2\mu_2 - c_1\mu_1)(1 - \rho P_{11}) + \rho_2\rho_3(c_2\mu_2 - c_3\mu_3)(1 - \rho P_{31}) < 0 \implies \lambda_1 > 0 \implies P_{21} = P_{11} - TV(F_1, F_2)$$

4. If $\gamma > \frac{\rho_3(1-\rho+\rho TV(F_1,F_3))}{\rho_1(1-\rho)}$, then $P_{31} = P_{11} - TV(F_1, F_3),\ P_{21} = P_{11} - TV(F_1, F_2) \implies \lambda_3 = \lambda_5 = \lambda_6 = 0$

5. $$\frac{\partial\mathcal{L}(P_{11}, P_{21}, P_{31}, \Lambda)}{\partial P_{11}} + \frac{\partial\mathcal{L}(P_{11}, P_{21}, P_{31}, \Lambda)}{\partial P_{21}} + \frac{\partial\mathcal{L}(P_{11}, P_{21}, P_{31}, \Lambda)}{\partial P_{31}} = 0 \implies$$
$$\frac{\rho_1\rho_2(c_1\mu_1 - c_2\mu_2)(\rho P_{11} - \rho P_{21}) + \rho_1\rho_3(c_1\mu_1 - c_3\mu_3)(\rho P_{11} - \rho P_{31}) + \rho_2\rho_3(c_2\mu_2 - c_3\mu_3)(\rho P_{21} - \rho P_{31})}{(1 - \rho_1 P_{11} - \rho_2 P_{21} - \rho_3 P_{31})^2(1 - \rho)^2} - \lambda_4 = 0$$

6. The optimal solution has $P_{11} > P_{21}$ and $P_{11} > P_{31}$. If in addition $P_{21} > P_{31}$, then $\lambda_4 > 0$ i.e., $P_{11} = 1$ (which holds since $TV(F_1, F_2) \leq TV(F_1, F_3)$).

7. If $\gamma < \frac{\rho_3}{\rho_1(1-\rho TV(F_1,F_3))}$, then $\lambda_1 - \lambda_3 - \lambda_5 < 0$ i.e. $\lambda_3 > 0$ or $\lambda_5 > 0$

8. If $\gamma < \frac{\rho_3}{\rho_1(1-\rho TV(F_1,F_3))}$, then the optimal P_{21} is 1 or $P_{31}+TV(F_2,F_3) = P_{11} - TV(F_2,F_3) + TV(F_1,F_3)$
9. Since $|P_{21}-P_{31}| \leq TV(F_2,F_3)$, $P_{21} = P_{11} + TV(F_2,F_3) - TV(F_1,F_3)$
10. The optimal matrix is $\begin{bmatrix} P_{11}, & 1-P_{11} \\ P_{11}+TV(F_2,F_3)-TV(F_1,F_3), & 1-P_{11}+TV(F_1,F_3)-TV(F_2,F_3) \\ P_{11}-TV(F_1,F_3), & 1-P_{11}+TV(F_1,F_3) \end{bmatrix}$
11. $P_{11}-P_{21} = TV(F_1,F_3)-TV(F_2,F_3) \leq TV(F_1,F_2)$ (triangle inequality)
12. Hence the problem simplifies to maximizing the below over $P_{11} \in [TV(F_1,F_3),1]$

$$\frac{c_1\mu_1\rho_1(\rho P_{11}-1)+c_2\mu_2\rho_2(\rho(P_{11}+TV(F_2,F_3)-TV(F_1,F_3))-1)+c_3\mu_3\rho_3(\rho(P_{11}-TV(F_1,F_3))-1)}{(1-\rho_1 P_{11}-\rho_2(P_{11}+TV(F_2,F_3)-TV(F_1,F_3))-\rho_3(P_{11}-TV(F_1,F_3)))(1-\rho)}$$

Note that $\frac{d}{dx}(\frac{ax+b}{cx+d}) = \frac{ad-bc}{(cx+d)^2}$, in this case $x = P_{11}$ and $ad-bc =$

$$(c_1\mu_1\rho_1+c_2\mu_2\rho_2+c_3\mu_3\rho_3)\rho(1+\rho_2(TV(F_1,F_3)-TV(F_2,F_3))+\rho_3 TV(F_1,F_3))-$$
$$\rho(c_1\mu_1\rho_1+c_2\mu_2\rho_2(1+\rho(TV(F_1,F_3)-TV(F_2,F_3)))+c_3\mu_3\rho_3(1+\rho TV(F_1,F_3))) > 0$$

Therefore, the maximum occurs at $P_{11} = 1$ Q.E.D.

Proof of Proposition 6: First, we prove the following result: For $N = T = 3$, and $P(3,\beta_3)$ specified in eq. (23):

1 $\mathcal{C}(P(3,\beta_3)) \geq \mathcal{C}(I_{3\times3})$ $\forall \beta_3 \in [0,1]$.

2 $\exists \beta^*(\Phi) \in [0,1)$, such that $\beta_3^1 \geq \beta_3^2 \implies \mathcal{C}(P(3,\beta_3^1)) \leq \mathcal{C}(P(3,\beta_3^2))$ $\forall \beta_3^2 \geq \beta^*(\Phi)$.

3 If $\frac{\lambda_1}{\mu_1} = \frac{\lambda_2}{\mu_2} = \frac{\lambda_3}{\mu_3}$, then $\beta_3^1 \geq \beta_3^2 \implies \mathcal{C}(P(3,\beta_3^1)) \leq \mathcal{C}(P(3,\beta_3^2))$ $\forall \beta_3^2 \geq 0$.

Let $\delta = (1-\beta_3)/2$ (equivalently, $\beta_3 = 1-2\delta$). $0 \leq \beta_3 \leq 1 \iff 0 \leq \delta \leq 1/2$. Let $\Delta(\delta) = \mathcal{C}(P(3,\beta_3 = 1)) - \mathcal{C}(P(3,\beta_3))$:

$$\begin{aligned}\Delta(\delta) = & \frac{c_1\mu_1\rho_1}{1-\rho_1} + \frac{c_2\mu_2\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{c_3\mu_3\rho_3}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)} \\ & - \frac{c_1\mu_1\rho_1(1-2\delta)+c_2\mu_2\rho_2\delta+c_3\mu_3\rho_3\delta}{1-\rho_1(1-2\delta)-\rho_2\delta-\rho_3\delta} - \frac{c_1\mu_1\rho_1\delta+c_2\mu_2\rho_2(1-2\delta)+c_3\mu_3\rho_3\delta}{(1-\rho_1(1-2\delta)-\rho_2\delta-\rho_3\delta)(1-\rho_1(1-\delta)-\rho_2(1-\delta)-\rho_3(2\delta))} \\ & - \frac{c_1\mu_1\rho_1\delta+c_2\mu_2\rho_2\delta+c_3\mu_3\rho_3(1-2\delta)}{(1-\rho_1(1-\delta)-\rho_2(1-\delta)-\rho_3(2\delta))(1-\rho_1-\rho_2-\rho_3)} \\ & = \frac{(c_1\mu_1-c_2\mu_2)(\rho_1\rho_2(3\delta^2-4\delta)+\rho_1\rho_3(-2\delta-3\delta^2))+(c_2\mu_2-c_3\mu_3)(\rho_2\rho_3(3\delta-4\delta^2)+\rho_1\rho_3(-2\delta-3\delta^2))}{(1-\rho_1)(1-\rho_1(1-2\delta)-\rho_2\delta-\rho_3\delta)(1-\rho_1(1-\delta)-\rho_2(1-\delta)-\rho_3(2\delta))(1-\rho_1-\rho_2-\rho_3)},\end{aligned}$$

where $\rho_1 = \frac{\lambda_1}{\mu_1}$, $\rho_2 = \frac{\lambda_2}{\mu_2}$, $\rho_3 = \frac{\lambda_3}{\mu_3}$.

Since $(c_1\mu_1-c_2\mu_2) \geq 0$, $(c_2\mu_2-c_3\mu_3) \geq 0$, and $3\delta^2-4\delta \leq 0$, $-2\delta-3\delta^2 \leq 0$, $(1-\rho_1(1-2\delta)-\rho_2\delta-\rho_3\delta) \geq 0$, $(1-\rho_1(1-\delta)-\rho_2(1-\delta)-\rho_3(2\delta)) \geq 0$, $\forall \delta \in [0,1/2]$, we have that $\Delta(\delta) \leq 0$ $\forall \delta \in [0,1/2]$ i.e., $\mathcal{C}(P(3,\beta_3)) \geq \mathcal{C}(P(3,\beta_3 = 1))$ $\forall \beta_3 \in [0,1]$. This proves part 1 of the proposition.

Next, $\mathcal{C}(P(3,\beta_3))$ is decreasing in β_3 if and only if $\Delta(\delta)$ is decreasing in δ, which holds if and only if $f(\delta)$ is decreasing in δ, where

$$f(\delta) = \frac{\gamma\rho_1(\rho_2(3\delta^2-4\delta)+\rho_3(-2\delta-3\delta^2))+\rho_3(\rho_2(3\delta^2-4\delta)+\rho_1(-2\delta-3\delta^2))}{(1-\rho_1(1-2\delta)-\rho_2\delta-\rho_3\delta)(1-\rho_1(1-\delta)-\rho_2(1-\delta)-\rho_3(2\delta))},$$

The derivative $f^{'}(\delta)$ of $f(\delta)$ w.r.t δ satisfies

$$f^{'}(\delta) = \frac{g(\delta)}{(1-\rho_1(1-2\delta)-\rho_2\delta-\rho_3\delta)^2(1-\rho_1(1-\delta)-\rho_2(1-\delta)-\rho_3(2\delta))^2},$$

where $g(\delta) = (A_1 + A_2)\delta^2 + (B_1 + B_2)\delta + C_1 + C_2$:

$$A_1 = \gamma\rho_1 \left[3(\rho_2-\rho_3)[(1-\rho_1)(\rho_1+\rho_2-2\rho_3)+(1-\rho_1-\rho_2)(2\rho_1-\rho_2-\rho_3)]+2(2\rho_2+\rho_3)(\rho_1+\rho_2-2\rho_3)(2\rho_1-\rho_2-\rho_3)\right.$$

$$A_2 = \rho_3 \left[3(\rho_2-\rho_1)[(1-\rho_1)(\rho_1+\rho_2-2\rho_3)+(1-\rho_1-\rho_2)(2\rho_1-\rho_2-\rho_3)]+2(2\rho_2+\rho_1)(\rho_1+\rho_2-2\rho_3)(2\rho_1-\rho_2-\rho_3)\right]$$

$$B_1 = 6\rho_1(\rho_2-\rho_3)(1-\rho_1)(1-\rho_1-\rho_2)$$

$$B_2 = 6\rho_3(\rho_2-\rho_1)(1-\rho_1)(1-\rho_1-\rho_2)$$

$$C_1 = -2\rho_1(2\rho_2+\rho_3)(1-\rho_1)(1-\rho_1-\rho_2)$$

$$C_2 = -2\rho_3(2\rho_2+\rho_1)(1-\rho_1)(1-\rho_1-\rho_2)$$

Now, $f(\delta)$ is decreasing in δ wherever $g(\delta) \leq 0$. Further, $B_1\delta + C_1 \leq 0$ and $B_2\delta + C_2 \leq 0$ $\forall\delta \in [0, 1/2]$. If $A_1 + A_2 \leq 0$, then $g(\delta) \leq 0$ $\forall\delta \in [0, 1/2]$, and $f(\delta)$ and $\Delta(\delta)$ are decreasing in δ $\forall\delta \in [0, 1/2]$, and $\mathcal{C}(P(3, \beta_3))$ is decreasing in β_3 $\forall\beta_3 \in [0, 1]$. If $A_1 + A_2 \geq 0$, then $g(\delta)$ is convex in δ. Since $g(0) = C_1 + C_2 < 0$, if $g(1/2) \leq 0$, then $g(\delta) \leq 0$ $\forall\delta \in [0, 1/2]$, and $\mathcal{C}(P(3, \beta_3))$ is decreasing in β_3 $\forall\beta_3 \in [0, 1]$. Otherwise, if $g(1/2) > 0$ (as is for $\rho_1 = 0.46, \rho_2 = 0.46, \rho_3 = 0.01, \gamma = 100$), then $\exists\delta^*(\Phi) \in (0, 1/2]$ such that $g(\delta) \leq 0$ $\forall\delta \in [0, \delta^*(\Phi)]$, and equivalently, $\mathcal{C}(P(3, \beta_3))$ is decreasing in β_3 $\forall\beta_3 \in [\beta^*(\Phi) = 1 - 2\delta^*(\Phi), 1]$. This proves part 2 of the proposition.

Finally, if $\rho_1 = \rho_2 = \rho_3$, then $g(\delta) = C_1 + C_2 < 0$ $\forall\delta \in [0, 1/2]$, and equivalently, $\mathcal{C}(P(3, \beta_3))$ is decreasing in β_3 $\forall\beta_3 \in [0, 1]$. This proves part 3 of the proposition. Q.E.D.

Next, we prove the statement of Proposition 6. For $T = 3, N = 3$ and accuracy $\beta_3 = q$, let $P = P(3, q)$. If we cluster the predicted classes 1 and 2 into priority queue 1 and predicted class 3 into priority queue 2, it would lead to a 3×2 classification matrix $P^o(2, \frac{1+q}{2}, q)$ (as defined in eq. (25)):

$$P^o_{ij}(2, \frac{1+q}{2}, q) = \begin{cases} \frac{1+q}{2} & i = 1, 2, j = 1 \\ \frac{1-q}{2} & i = 1, 2, j = 2 \\ q & i = 3, j = 2 \\ 1-q & i = 3, j = 1 \end{cases}$$

Similarly, if we cluster the predicted classes 2 and 3 into priority queue 2 and predicted class 1 into priority queue 1, it would lead to a 3×2 classification matrix $P^u(2, q, \frac{1+q}{2})$ (as defined in eq. (25)):

$$P^u_{ij}(2, q, \frac{1+q}{2}) = \begin{cases} \frac{1+q}{2} & i = 2, 3, j = 2 \\ \frac{1-q}{2} & i = 2, 3, j = 2 \\ q & i = 1, j = 1 \\ 1-q & i = 1, j = 2 \end{cases}$$

From Proposition 2, we know that both these classification matrices have a higher average waiting cost than the original 3×3 classifier $P(3,q)$ i.e.,

$$\mathcal{C}(P(3,q)) \leq \min\{\mathcal{C}(P^o(2,\frac{1+q}{2},q)),\mathcal{C}(P^u(2,q,\frac{1+q}{2}))\}.$$

From proposition 9, we know that,

$$\mathcal{C}(P^o(2,\frac{1+q}{2},q)) \leq \mathcal{C}(P^o(2,q,q)) = \mathcal{C}(P^o(2,q))$$

, and

$$\mathcal{C}(P^u(2,q,\frac{1+q}{2}))\} \leq \mathcal{C}(P^u(2,q,q)) = \mathcal{C}(P^u(2,q)).$$

Hence,

$$\mathcal{C}(P(3,q)) \leq \min\{\mathcal{C}(P^o(2,q)),\mathcal{C}(P^u(2,q))\}.$$

The average waiting cost for $N=3$ and $\beta_3 = \frac{1}{3}$ is equal to that under $N=2$ and $\beta_2 = \frac{1}{2}$, as both are equivalent to a single queue with first in first out (FIFO) system with no prioritization. Therefore, the average waiting cost at $N=3$ and $\beta_3 = \frac{1}{3}$ is more than that under $N=2$ and $\beta_2 = 1$ ($\min\{\mathcal{C}(P^o(2,1)),\mathcal{C}(P^u(2,1))\}$); see Proposition 1. The average waiting cost under $N=3$ and $\beta_3 = 1$ is less that under $N=2$ and $\beta_2 = 1$ (from Proposition 1). Now, if $\beta^*(\Phi)$ is less than 1/3, then $\mathcal{C}(P(3,\beta_3))$ is decreasing $\forall \beta_3 \in [1/3,1]$. Therefore, there exists a value $\beta_3^*(\Phi)$ between 1/3 and 1 such that cost under $N=3$ and β_3 is less than the cost under $N=2$ and $\beta_2 = 1$ $\forall \beta_3 \in [\beta_3^*(\Phi),1]$. If, $\beta^*(\Phi)$ is more than 1/3, then $\mathcal{C}(P(3,\beta^*(\Phi))) \geq \mathcal{C}(P(3,1/3))$. Since $\mathcal{C}(P(3,\beta_3))$ is decreasing $\forall \beta_3 \in [\beta^*(\Phi),1]$. Therefore, there exists a value $\beta_3^*(\Phi)$ between $\beta^*(\Phi)$ and 1 such that cost under $N=3$ and β_3 is less than the cost under $N=2$ and $\beta_2 = 1$ $\forall \beta_3 \in [\beta_3^*(\Phi),1]$. Q.E.D.

Proof of Proposition 7: When under-prioritization of type 2 is better than over-prioritization for $\beta_2 = 1$, i.e., $\mathcal{C}(P^u(2,1)) \leq \mathcal{C}(P^o(2,1))$, then the value of β_3 for which $N=3$ has same cost as $N=2$, $\beta_2 = 1$ ($\mathcal{C}(P(3,\beta_3)) = \mathcal{C}(P^u(2,1))$ satisfies:

$$\beta_3^*(\Phi) = \frac{(3c_1\lambda_1 - 2c_2\lambda_2 - c_3\lambda_3) - \rho(6c_1\lambda_1 - 3c_2\lambda_2 - 3c_3\lambda_3)}{(3c_1\lambda_1 - 3c_3\lambda_3) - \rho(6c_1\lambda_1 - 3c_2\lambda_2 - 3c_3\lambda_3)}$$

$\frac{1}{3} \leq \beta_3^*(\Phi) \leq 1$ for $0 \leq \rho \leq \frac{1}{3}$, and decreases with ρ.

When over-prioritization of type 2 is better than under-prioritization, i.e., $\mathcal{C}(P^o(2,1)) \leq \mathcal{C}(P^u(2,1))$, then the value of β_3 for which $N=3$ has same cost as $N=2$, $\beta_2 = 1$ ($\mathcal{C}(P(3,\beta_3)) = \mathcal{C}(P^o(2,1))$ satisfies:

$$\beta_3^*(\Phi) = \frac{(2c_1\lambda_1 + c_2\lambda_2 - 3c_3\lambda_3) - \rho(3c_1\lambda_1 - 3c_3\lambda_3)}{(3c_1\lambda_1 - 3c_3\lambda_3) - \rho(6c_1\lambda_1 - 3c_2\lambda_2 - 3c_3\lambda_3)}$$

$\frac{1}{3} \le \beta_3^*(\Phi) \le 1$ for $0 \le \rho \le \frac{1}{3}$, and increases with ρ. In fact,

$$\beta_3^*(\Phi) = \begin{cases} \frac{(3c_1\lambda_1 - 2c_2\lambda_2 - c_3\lambda_3) - \rho(6c_1\lambda_1 - 3c_2\lambda_2 - 3c_3\lambda_3)}{(3c_1\lambda_1 - 3c_3\lambda_3) - \rho(6c_1\lambda_1 - 3c_2\lambda_2 - 3c_3\lambda_3)} & \text{if } \rho < \frac{\gamma - 1}{3\gamma} \\ \frac{(2c_1\lambda_1 + c_2\lambda_2 - 3c_3\lambda_3) - \rho(3c_1\lambda_1 - 3c_3\lambda_3)}{(3c_1\lambda_1 - 3c_3\lambda_3) - \rho(6c_1\lambda_1 - 3c_2\lambda_2 - 3c_3\lambda_3)} & \text{otherwise} \end{cases}$$

Q.E.D.

Proof of Proposition 8: Let $\Delta = \mathcal{C}(P^u(2,\beta_2)) - \mathcal{C}(P^o(2,\beta_2))$:

$$\begin{aligned}\Delta =& \left[\frac{c_1\lambda_1\beta_2 + c_2\lambda_2(1-\beta_2) + c_3\lambda_3(1-\beta_2)}{1-\rho_1\beta_2 - \rho_2(1-\beta_2) - \rho_3(1-\beta_2)} + \frac{c_1\lambda_1(1-\beta_2) + c_2\lambda_2\beta_2 + c_3\lambda_3\beta_2}{(1-\rho_1\beta_2 - \rho_2(1-\beta_2) - \rho_3(1-\beta_2))(1-\rho_{tot})}\right] \\ & - \left[\frac{c_1\lambda_1\beta_2 + c_2\lambda_2\beta_2 + c_3\lambda_3(1-\beta_2)}{1-\rho_1\beta_2 - \rho_2\beta_2 - \rho_3(1-\beta_2)} + \frac{c_1\lambda_1(1-\beta_2) + c_2\lambda_2(1-\beta_2) + c_3\lambda_3\beta_2}{(1-\rho_1\beta_2 - \rho_2\beta_2 - \rho_3(1-\beta_2))(1-\rho_{tot})}\right]\end{aligned}$$

$$\begin{aligned}\Delta =& (1-2\beta_2)\Bigg[\frac{(c_1\lambda_1\rho_2(1-\beta_2\rho_{tot}) + c_2\lambda_2(\rho_{tot}\beta_2(\rho_1 - \rho_3) + \rho_{tot}(\rho_3 - 1) + \rho_2)}{(1-\rho_1\beta_2 - \rho_2\beta_2 - \rho_3(1-\beta_2))(1-\rho_1\beta_2 - \rho_2(1-\beta_2) - \rho_3(1-\beta_2))(1-\rho_{tot})} \\ & + \frac{c_3\lambda_3\rho_2(1-(1-\beta_2)\rho_{tot})}{(1-\rho_1\beta_2 - \rho_2\beta_2 - \rho_3(1-\beta_2))(1-\rho_1\beta_2 - \rho_2(1-\beta_2) - \rho_3(1-\beta_2))(1-\rho_{tot})}\Bigg]\end{aligned}$$

Where $\rho_{tot} = \rho_1 + \rho_2 + \rho_3$. Substituting $\lambda_1 = \rho_1\mu_1, \lambda_2 = \rho_2\mu_2, \lambda_3 = \rho_3\mu_3$, $\forall \beta_2 \in [0.5, 1]$:

$$\Delta \le 0 \iff \beta_2\rho_{tot}[(c_1\mu_1 - c_2\mu_2)\rho_1 + (c_2\mu_2 - c_3\mu_3)\rho_3] \le (c_1\mu_1 - c_2\mu_2)\rho_1 + (c_3\mu_3 - c_2\mu_2)\rho_3(1-\rho_{tot})$$

Q.E.D.

Proof of Proposition 9. Let the value of $\mathcal{C}(P^u(2,\beta_{11},\beta_{22})) = \mathcal{C}(P)$:

$$\mathcal{C}(P) = \mathbb{E}[\mathcal{S}]\left(\frac{c_1\lambda_1\beta_{11} + (c_2\lambda_2 + c_3\lambda_3)(1-\beta_{22})}{1-\rho_1\beta_{11} - (\rho_2+\rho_3)(1-\beta_{22})} + \frac{c_1\lambda_1(1-\beta_{11}) + (c_2\lambda_2 + c_3\lambda_3)\beta_{22}}{(1-\rho_1\beta_{11} - (\rho_2+\rho_3)(1-\beta_{22}))(1-\rho_1-\rho_2-\rho_3)}\right)$$

$$\frac{\partial \mathcal{C}(P)}{\partial \beta_{11}} = \mathcal{C}(P)\left[\frac{-\rho_1((c_1\mu_1 - c_2\mu_2)\rho_2 + (c_1\mu_1 - c_3\mu_3)\rho_3)(1-(1-\beta_{22})(\rho_1+\rho_2+\rho_3))}{((c_1\lambda_1\beta_{11} + (c_2\lambda_2 + c_3\lambda_3)(1-\beta_{22}))(1-\rho_1-\rho_2-\rho_3) + c_1\lambda_1(1-\beta_{11}) + (c_2\lambda_2 + c_3\lambda_3)\beta_{22})}\right] < 0$$

$$\frac{\partial \mathcal{C}(P)}{\partial \beta_{22}} = \mathcal{C}(P)\left[\frac{-\rho_1((c_1\mu_1 - c_2\mu_2)\rho_2 + (c_1\mu_1 - c_3\mu_3)\rho_3)(1-\beta_{11}(\rho_1+\rho_2+\rho_3))}{((c_1\lambda_1\beta_{11} + (c_2\lambda_2 + c_3\lambda_3)(1-\beta_{22}))(1-\rho_1-\rho_2-\rho_3) + c_1\lambda_1(1-\beta_{11}) + (c_2\lambda_2 + c_3\lambda_3)\beta_{22})}\right] < 0$$

$$\frac{\partial \mathcal{C}(P)}{\partial \beta_{11}} < \frac{\partial \mathcal{C}(P)}{\partial \beta_{22}} \iff \beta_{11} + \beta_{22} - 1 > 0$$

$$\frac{\partial \mathcal{C}(P)}{\partial \beta_{11}} < \frac{\partial \mathcal{C}(P)}{\partial \beta_{22}} \implies \left|\frac{\partial \mathcal{C}(P)}{\partial \beta_{11}}\right| > \left|\frac{\partial \mathcal{C}(P)}{\partial \beta_{22}}\right|$$

The proof for $\mathcal{C}(P^o(2,\beta_{11},\beta_{22}))$ is similar and hence omitted. Q.E.D.