Improving Efficiency in Japanese Residency Matching

Haris Aziz[1,2], Serge Gaspers[1], Zhaohong Sun[1,2], and Makoto Yokoo[3]

[1] UNSW Sydney, Australia
[2] Data61 CSIRO, Australia
[3] Kyushu University, Japan

Abstract. We study the two-sided matching problem with regional quotas, motivated by the Japanese hospital-doctor matching market in which hospitals are associated with regions and both hospitals and regions are subject to quotas. In order to achieve a balanced distribution of doctors across regions, hard bounds are imposed by the government to limit the number of doctors that can be placed at each region. However, such hard bounds lead to inefficiency in terms of wasting vacant positions. In this paper, we introduce a novel approach to solve this issue by assigning weights to hospitals. And we propose a novel class of Generalized Deferred Acceptance with Regions (GDA-R) algorithms that deals with regional quotas and weights. We also consider the connection between matching with regional quotas and matching with diversity constraints studied in the context of school choice. We show how to convert one instance of matching with soft diversity constraints into a corresponding instance of matching with hard regional quotas such that feasibility and stability are preserved. This connection implies our GDA-R algorithm also applies to another important field in matching.

1 Introduction

Distributional constraints are pervasive in real-life matching markets. In this growing literature [Goto *et al.*, 2015, 2016, 2017], there are at least two streams of work. The first one is *hospital-doctor matching with regional quotas*, motivated by the Japanese Residency Matching Program [Kamada and Kojima, 2015, 2017]. And the second one is *school choice with diversity constraints*, studied intensely in the controlled school choice problem [Abdulkadiroğlu and Sönmez, 2003; Echenique and Yenmez, 2015].

The Japanese residency matching program (JRMP) was established in 2004 to train newly graduated medical students at hospitals. In 2019, 8986 doctors were matched to 1037 hospitals under the JRMP system.[4] Due to the shortage of doctors distributed to rural areas, the Japanese government introduced regional quotas to limit the number of doctors that can be placed at different regions since 2008. To ensure that the number of doctors matched to a region does not

[4] `https://www.mhlw.go.jp/stf/houdou/0000171153_00003.html`

exceed its regional quota, the Japanese government also imposes a *target capacity* on each hospital which is usually smaller than its real capacity.[5] However, such hard target capacities lead to a waste of vacant positions [Kamada and Kojima, 2015]. In this paper, our first research question is *how to eliminate the waste of vacant positions with minimal modifications to the current system?*

In the problem of school choice with diversity constraints, each student is associated with a set of types that capture traits such as being from a disadvantaged group. To achieve a balanced integration of students from diverse backgrounds, each school typically imposes a maximum quota and minimum quota on each type [Kojima, 2012; Hafalir *et al.*, 2013; Kominers and Sönmez, 2013]. If diversity constraints are viewed as rigid bounds, then there may not exist any outcome that fulfills all type-specific minimum quotas, and it is also impossible to design a mechanism that satisfies desirable properties such as fairness and non-wastefulness [Ehlers *et al.*, 2014]. Recent literature treats diversity constraints as *soft diversity constraints* such that a school may admit more students of some type than its maximum quota or fewer students of some type than its minimum quota [Kurata *et al.*, 2015, 2017; Gonczarowski *et al.*, 2019].

In a recent paper, Aziz *et al.* [2019] showed how to transform an instance with hard diversity quotas into a corresponding instance with hard regional quotas. If we apply their reduction directly, then soft diversity constraints are converted into soft regional quotas. And our second research question is the following one: *what is the connection between soft diversity constraints and hard regional quotas?*

Contributions The contributions of this paper are summarized as follows. First, we come up with a novel approach to solve the inefficiency issue in the JRMP market by assigning weights to hospitals. Different from the previous solution by Kamada and Kojima [2015] that still yields a wasteful outcome and requires a specific ordering over hospitals such that hospitals could pick doctors in a round-robin manner, we make minimal modifications to the current system. Second, we propose a novel class of Generalized Deferred Acceptance with Regions algorithms (GDA-R) that deals with regional quotas and weights over hospitals. The class of GDA-R algorithms not only is a generalization of previous algorithms, but also introduces a new framework for matching with regional quotas based on a novel two-stage process such that contracts proposed by doctors are first shortlisted by hospitals and then further refined by regions. Third, we are the first to show how to convert an instance of matching with soft diversity constraints into a corresponding instance of matching with hard regional quotas such that feasibility and stability are preserved. This connection not only unifies the literature but also implies that our GDA-R algorithms also work for another important field in the two-sided matching.

Related Work Kamada and Kojima [2015] first studied the inefficiency issue[6] in the JRMP market and they proposed a flexible deferred acceptance algorithm

[5] `https://www.mhlw.go.jp/seisaku/2009/08/04.html`

[6] Note that Kamada and Kojima [2015] formally define efficiency as Pareto efficiency, while we consider efficiency with respect to wastefulness.

which does not eliminate the waste of vacant positions. There are other algorithms proposed for matching with regional quotas [Goto *et al.*, 2015, 2016, 2017; Hamada *et al.*, 2017] that work for one particular setting. We propose a novel class of Generalized Deferred Acceptance with Regions algorithms (GDA-R) that provide a new framework for matching with regions.

There are other papers that are more mathematical which consider an abstract and general class of constraints, e.g., constraints that can be represented as a substitute choice function [Hatfield and Milgrom, 2005; Hatfield and Kojima, 2008], an M-convex set [Kojima *et al.*, 2018]. Although representing these two models in an abstract model is possible, but how to encode such constraints and preferences as a choice function in an abstract model is not obvious/trivial and deserves further exploration.

2 Matching with Regional Quotas

In this section, we introduce a new model of matching with regional quotas that incorporates weights over hospitals. We choose the classical hospital-doctor setting for illustration, while our model applies to many matching markets outside the context of hospital-doctor matching. For instance, a university may want to achieve a balance of newly enrolled students across different departments where the university can be considered as a region, and different departments are considered as hospitals within the region.

An instance I^R of matching with regional quotas is composed of a tuple (D, H, q_H, R, $\overline{\delta}$, Y, $\succsim_D$, $\succsim_H$, $\succsim_R$, W). Let D and H denote the set of doctors and the set of hospitals, respectively. A capacity vector $q_H = (q_h)_{h \in H}$ consists of each hospital h's capacity q_h, which is the maximum number of doctors that hospital h can accommodate.

There is a set of regions R where each region $r \in R$ is a subset of hospitals, i.e. $r \subseteq H$. We assume all hospitals are partitioned into disjoint regions as in the Japanese hospital-doctor matching market [Kamada and Kojima, 2015] s.t., for any two $r_i, r_j \in R$, either $r_i = r_j$ or $r_i \cap r_j = \emptyset$. Let $\overline{\delta} = (\overline{\delta}_r)_{r \in R}$ denote a vector consisting of each region's maximum quota $\overline{\delta}_r$ which limits the number of doctors that can be distributed to all hospitals within region r.

Each contract (d, h) is a doctor-hospital pair denoting that doctor d is matched with hospital h. Let $\mathcal{Y} \subseteq D \times H$ be the set of available contracts. An outcome (or a matching) is a set of contracts $Y \subseteq \mathcal{Y}$. For any outcome $Y \subseteq \mathcal{Y}$, let $Y_d = \{(d, h) \in Y | h \in H\}$ denote the set of contracts involving doctor d, $Y_h = \{(d, h) \in Y | d \in D\}$ denote the set of contracts involving hospital h, and $Y_r = \bigcup_{h \in r} Y_h$ denote the set of contracts involving region r.

Let $\succsim_D = \{\succsim_d\}_{d \in D}$ be the preference profile of all doctors. Each doctor d has a preference ordering $\succsim_d$ over $\mathcal{Y}_d \cup \{\emptyset\}$ where $\emptyset$ denotes the null contract indicating that doctor d is unmatched. For any two contracts $x, y \in \mathcal{Y}_d \cup \{\emptyset\}$, $x \succsim_d y$ means that doctor d prefers contract x to contract y or doctor d is indifferent between two contracts, where $\succ_d$ and $\sim_d$ denote the strict and indifferent relation, respectively.

Each hospital h has a priority ordering $\succsim_h$ over $\mathcal{Y}_h \cup \{\emptyset\}$ and each region r has a regional priority ordering $\succsim_r$ over $\mathcal{Y}_r \cup \{\emptyset\}$. Let $\succsim_H = \{\succsim_h\}_{h \in H}$ and $\succsim_R = \{\succsim_r\}_{r \in R}$ denote the priority profile of hospitals H and regions R, respectively. Note that contracts allow us to describe more complicated regional priorities over doctor-hospital pairs instead of doctors.

A contract $(d, h) \in Y$ is *acceptable* to doctor d and hospital h if $(d, h) \succsim_d \emptyset$ and $(d, h) \succsim_h \emptyset$ hold. W.L.O.G, we assume that for any $h \in r$, if a contract $y \in Y_h$ is acceptable to hospital h, then it is also acceptable to region r.

An outcome $Y \subseteq \mathcal{Y}$ is *feasible* for I^R under regional quotas if i) for each doctor d, we have $|Y_d| \leq 1$, ii) for each hospital h, $|Y_h| \leq q_h$ holds, and iii) the outcome Y *respects regional quotas*, i.e., for any region r we have $|Y_r| \leq \overline{\delta}_r$.

Definition 1 (Non-wastefulness). *Given a feasible outcome X, a doctor d claims an empty seat at hospital h if $(d, h) \notin X$ and $X \cup \{(d, h)\} \setminus X_d$ is feasible. A feasible outcome is non-wasteful if no doctor claims an empty seat.*

The main difference from the previous model is that each region r additionally assigns a weight $w(h)$ to each hospital $h \in r$ to specify the importance of hospital h to region r. Let $W = \{w(h)\}_{h \in H}$ denote the set of weights. The intuition of weights over hospitals is that, when there are more doctors applying to hospitals at region r than its regional quota $\overline{\delta}_r$, region r gives higher precedence to the hospitals with a larger weight and lower precedence to the hospitals with a smaller weight. If ties occur, then region r chooses the contract with higher regional priority based on $\succ_r$.

2.1 Inefficiency in Japanese Market

In this subsection, we show how the addition of weights over hospitals provides a suitable way to solve the inefficiency issue in the Japanese hospital-doctor matching market with minimal modification of the current system. To ensure that the number of doctors matched to one region does not exceed the regional quota, the Japanese government also imposes a *target capacity* $\overline{q}_h$ on each hospital, i.e., $\sum_{h \in r} \overline{q}_h \leq \overline{\delta}_r$ [Kamada and Kojima, 2015]. However, such hard target capacities lead to the waste of vacant positions as Example 1 shows.

Example 1. Consider one region r with regional quota $\overline{\delta}_r = 12$ that contains two hospitals h_1 and h_2 with capacity $q_{h_1} = 6$ and $q_{h_2} = 10$, respectively. Suppose the target capacity of each hospital is $\overline{q}_{h_1} = 4$ and $\overline{q}_{h_2} = 8$. There are 12 doctors $D = \{d_1, ..., d_{12}\}$ in which the first 6 doctors prefers hospital h_1 to being unmatched and the latter 6 prefers hospital h_2 to being unmatched, i.e, for $i \in [1, 6], (h_1, d_i) \succ_{d_i} \emptyset$ and for $j \in [7, 12], (h_2, d_j) \succ_{d_j} \emptyset$.

Each doctor could be matched to his favorite hospital without violating the regional maximum quota. However, two doctors from $\{d_1, ..., d_6\}$ cannot be matched to hospital h_1 due to the target capacity $\overline{q}_{h_1} = 4$.

In our setting, we can divide the virtual hospital h_1 into two dummy hospitals h_1^1, h_1^2 where h_1^1 admits doctors up to the target capacity with $q_{h_1^1} = \overline{q}_{h_1} = 4$ and h_1^2 admits doctors after reaching the target capacity with $q_{h_1^2} = q_{h_1} - \overline{q}_{h_1} = 2$.

Similarly, hospital h_2 is divided into h_2^1 and h_2^2 with $q_{h_2^1} = 8$ and $q_{h_2^2} = 2$. Region r gives larger weight to hospital h_1^1 and h_2^1, and smaller weight to hospital h_1^2 and h_2^2. Then all doctors could be matched to their favorite hospitals without exceeding any hospital capacity.

In our approach, the only modification to the system is to treat the target capacity as soft bound so that doctors can still be assigned to some hospital when the target capacity is reached. Each hospital is divided into two dummy hospitals where one dummy hospital has a target capacity and the other one has the remaining capacity. Each region gives higher precedence to all dummy hospitals with target capacity by assigning a larger weight and lower precedence to all dummy hospitals with remaining capacity by assigning a smaller weight.

This slight modification provides the government ability to distribute doctors to underserved regions as the target quotas while eliminating the waste of vacant positions. We can also apply this idea to other matching markets, in which the agents who play the role of regions have the authority to interfere in the process of matching.

3 Generalized Deferred Acceptance with Regions

In this section, we propose a novel class of Generalized Deferred Acceptance with Regions algorithms (GDA-R) that deals with regional priorities and weights. Our algorithm is based on a novel two-stage process such that contracts proposed by doctors are first shortlisted by hospitals and then further refined by regions. We not only come up with a novel and general algorithm but also introduce a new framework for matching with regions.

The intuition is that regions assign weights to hospitals to quantify the importance in terms of achieving a balanced outcome and hospitals first choose doctors based on their own priority ordering. When the number of applicants exceeds the regional quota, region determines which doctors should be selected in a reasonable way that it fills the vacant positions at the hospitals with larger weights whenever possible.

3.1 GDA-R

We illustrate the framework of Generalized Deferred Acceptance with Regions (GDA-R) at a high level. Firstly each doctor selects his favorite contract involving the hospital that has not rejected him yet. Then each hospital selects a set of contracts among all proposals from doctors without exceeding its capacity. Next, each region chooses a set of contracts among the set of contracts selected by each hospital within the region without exceeding its regional quotas. All contracts that are not chosen by regions are rejected. Repeat this procedure until no more contracts are rejected.

Given a set of contracts Y, let $Ch_d(Y)$ denote the choice function of doctor d which selects his most preferred acceptable contract from Y_d. Both the choice

function $Ch_h(Y)$ of hospital h and the choice function $Ch_r(Y)$ of region r select a set of contracts. We extend the choice function of each individual agent to a set of agents by taking the union, i.e., $Ch_D(Y) = \bigcup_{d \in D} Ch_d(Y)$. Armed with these choice functions, we describe the GDA-R algorithm in Algorithm 1.

Input: I^R, Ch_D, Ch_H, Ch_R, a set of contracts Y
Output: An outcome $Z \subseteq Y$

$Re \leftarrow \emptyset, A \leftarrow Y, B \leftarrow \emptyset, Z \leftarrow \emptyset$
while $A \neq Z$ **do**
  $A \leftarrow Ch_D(Y \setminus Re), B \leftarrow Ch_H(A), Z \leftarrow Ch_R(B)$
  $Re \leftarrow Re \cup (Z \setminus A)$
return Z

Algorithm 1: Generalized Deferred Acceptance with Regions

Next, we present one particular way to define $Ch_h(Y)$ and $Ch_r(Y)$ as shown in Algorithm 2 and Algorithm 3, respectively. Note that it is not unique to define the choice functions $Ch_h(Y)$ and $Ch_r(Y)$, and each different method specifies one particular algorithm of GDA-R.

Input: An instance I^R, a set of contracts Y
Output: A set of contracts $Y' \subseteq Y$

$Y' \leftarrow \emptyset$ % remove unacceptable contracts from Y_h
for $y = (d, h) \in Y$ in descending ordering of $\succ_h$ **do** % Ties are broken to derive a strict priority ordering $\succ_h$
  if $|Y'_h| < q_h$ **then**
    $Y' \leftarrow Y' \cup \{y\}$
return Y'

Algorithm 2: Choice function Ch_h of hospital h

The choice function Ch_h in Algorithm 2 works as follows: each hospital selects contracts one by one in accordance with its priority ordering $\succ_h$ until the number of contracts reaches its capacity q_h.

The choice function Ch_r of region r in Algorithm 3 works as follows: First divide all contracts $Y_r = Y_r^1 \cup Y_r^2 ... \cup Y_r^k$ into disjoint groups based on the weights over hospitals s.t. for any two contracts $(d, h) \in Y_r^a, (d', h') \in Y_r^b$, i) if two hospitals have the same weight, then two contracts belong to the same group; ii) if hospital h has higher weight than hospital h', then hospital h belongs to the group with smaller index a. Region r selects contracts from group Y_r^1 first, and then Y_r^2 and so on. For each group Y_r^a, region r selects contracts based on its regional priority $\succ_r$ without exceeding its regional quota.

3.2 Stability

In this subsection, we propose a new stability concept for matching with regional priorities and weights.

Given a feasible outcome Y and a contract $y = (d, h) \notin Y$, we use a function $\alpha(Y_r, y)$ to quantify the importance of the contract y to region r with $h \in r$,

$$\alpha(Y_r, y) = \begin{cases} w(h) & \text{if } |Y_h| < q_h \\ -\infty & \text{otherwise} \end{cases} \quad (1)$$

Input: An instance I^R, a set of contracts Y
Output: A set of contracts $Z \subseteq Y$

1: $Z \leftarrow \emptyset$ % remove unacceptable contracts from Y_r
2: Let $Y_r = Y_r^1 \cup ... \cup Y_r^k$ s.t. $\forall (d,h) \in Y_r^a, \forall (d',h') \in Y_r^b$

- $w(h) = w(h') \Rightarrow a = b$
- $w(h) > w(h') \Rightarrow a < b$

3: **for** each $Y_r^a \in Y_r$, $a \in [1, \cdots, k]$ **do**
4: **for** $y = (d,h) \in Y_r^a$ in descending ordering of $\succ_r$ **do** % Ties are broken to derive a strict ordering $\succ_r$
5: **if** $|Z_r| < \overline{\delta}_r$ **then**
6: $Z \leftarrow Z \cup \{y\}$
7: **return** Z

Algorithm 3: Choice function Ch_r of region r

The function $\alpha(Y_r, y)$ returns the weight of hospital h when the hospital still has a vacant position, and returns negative infinity otherwise. Given a feasible outcome Y and two contracts $y = (d,h) \notin Y, y' = (d',h') \notin Y$ with $h, h' \in r$, we use a function $\beta(Y_r, y, y')$ to compare the weights.

$$\beta(Y_r, y, y') = \alpha(Y_r, y) - \alpha(Y_r, y') \tag{2}$$

Definition 2 (Stability). *Given a feasible outcome Y, a doctor d and a hospital h form a blocking pair if $y = (d,h) \notin Y$, $y \succ_d Y_d$ and one of the two conditions holds: either i) the outcome $Y \cup \{y\} \setminus Y_d$ is feasible; or ii) there exists a contract $y' = (d',h') \in Y$ s.t. $h \in r, h' \in r$ and for the outcome $Y' = Y \setminus \{y'\}$, one of the following cases holds,*

- *ii-a) $h = h'$, $y \succ_h y'$ and $y \succ_r y'$;*
- *ii-b) $h \neq h'$ and $\beta(Y'_r, y, y') > 0$;*
- *ii-c) $h \neq h'$, $\beta(Y'_r, y, y') = 0$ and $y \succ_r y'$.*

A feasible outcome is stable if there is no blocking pair.

Note that condition i) corresponds to non-wastefulness in Definition 1. Definition 2 states that given an outcome Y, a doctor d and a hospital h form a blocking pair if doctor d is not matched to hospital h, while doctor d prefers hospital h to his assignment and either i) moving doctor d to hospital h does not violate the feasibility requirement including hospital h's capacity and region r's regional quota, or ii) there exists another doctor d' who is matched to hospital h' at region r such that ii-a) hospital h and h' are the same, and both hospital h and region r prefer the contract y to y'; ii-b) hospital h has a larger weight than hospital h', or ii-c) hospital h and h' are different but with the same weight, and region r prefers the contract y to y'.

Theorem 1. *The GDA-R algorithm with choice functions ch_h and ch_r defined in Algorithm 2 and 3 yields a stable outcome.*

Proof. We prove Theorem 1 by contradiction. Let Y be the yielded by the GDA-R algorithm and suppose there exists a blocking pair (d, h) where $h \in r$.

Case i) If the outcome $Y \cup (d, h) \setminus Y_d$ is feasible, then both hospital h and region r have not filled its capacity or maximum quota in the outcome Y. Since doctor d prefers (d, h) to Y_d, he must have selected the contract (d, h) before choosing Y_d and the contract (d, h) was rejected by either hospital h or region r. However, whenever a contract is rejected, either the number of contracts reaches the capacity of hospital h or the quota of region r, a contradiction.

Case ii) Suppose there exist another contract $(d', h') \in Y$ with $h' \in r$. Let $Y' = Y \setminus (d', h')$. Note that when (d, h) was rejected, the number of contracts chosen by region r has already reached its maximum quota. If ii-a) $h = h'$ holds, then all contracts matched to hospital h must have higher hospital priority than h, a contradiction. If ii-b) $h \neq h'$ and $\beta(Y'_r, (d, h), (d', h')) > 0$ hold, then hospital h has a larger weight than hospital h'. This leads to a contraction that region r selects one contract involving hospital h' with smaller weight before filling all vacancies at hospital h. If ii-c) $h \neq h'$, $\beta(Y'_r, (d, h), (d', h')) = 0$ and $(d, h) \succ_r (d', h')$ hold, then both hospitals have the same weight to region r, but the contract (d, h) has higher regional priority. By the time contract (d, h) was rejected, for each contract (d'', h'') that is selected by region r, either h'' has a larger weight than h or contract (d'', h'') has higher regional priority. This leads to a contraction that region r selects the contract (d', h') that does not satisfy any of the two conditions.

3.3 Comparison with Previous Algorithms

Kamada and Kojima [2015] proposed flexible deferred acceptance (FDA) for the Japanese residency matching which does not yield a non-wasteful outcome. The FDA algorithm requires a specific ordering over hospitals such that hospitals could pick doctors in a round-robin manner when the number of doctors matched to some hospital has reached the target capacity.

The FDA algorithm is a very special implementation of GDA-R algorithm. We can divide each hospital into multiple dummy hospitals where one dummy has target capacity, and all the others have capacity 1. For all the dummy hospitals with target capacity, they are assigned the largest weight. For the rest of dummy hospitals, each of them has a different weight which is in accordance with the order of round-robin.

There are other algorithms proposed for matching with regional quotas [Goto *et al.*, 2015, 2016, 2017; Hamada *et al.*, 2017] which requires a master list over doctors. The role of the master list is equivalent to imposing a unified regional priority ordering on all regions. And these algorithms are also particular implements of the GDA-R algorithm in which all hospitals have the same weight and all regions have the same regional priority orderings.

4 Transformation from Soft Diversity Constraints to Hard Regional Quotas

In this section, we discuss the connection between matching with soft diversity constraints and matching with hard regional quotas. Different from previous work [Aziz *et al.*, 2019], we show how to transform an instance of matching with *soft* diversity constraints into a corresponding instance of matching with *hard* regional quotas.

4.1 Soft Diversity Constraints

In this subsection, we first describe the model of *matching with soft diversity constraints*. To distinguish from the setting of matching with regional quotas, we choose school choice for illustration.

An instance I^T of matching with soft diversity constraints is composed of a tuple $(S, C, q_C, T, \underline{\eta}, \overline{\eta}, \mathcal{X}, \succsim_S, \succsim_C)$. There is a set of students S and a set of schools C. A capacity vector $q_C = (q_c)_{c \in C}$ consists of each school c's capacity q_c. Let T denote the type space and $T(s)$ represent the set of types to which student s belongs. We followed the setting of [Ehlers *et al.*, 2014] in which each student is associated with one type, i.e., $|T(s)| = 1$.

Each school c imposes a minimum quota $\underline{\eta}_c^t$ and a maximum quota $\overline{\eta}_c^t$ on each type t. Let $\underline{\eta}_c = (\underline{\eta}_c^t)_{t \in T}$ be a vector consisting of all minimum quotas at school c and let $\underline{\eta}$ denote a matrix of all minimum vectors of all schools. Similarly, let $\overline{\eta}$ denote the matrix consisting of all schools' type-specific maximum quotas.

Each contract $x = (s, c)$ is a student-school pair denoting that student s is matched with school c. An *outcome* (or a matching) is a set of contracts. Let $\mathcal{X} \subseteq S \times C$ denote the set of available contracts. Given any $X \subseteq \mathcal{X}$, let X_s denote the set of contracts involving student s, X_c denote the set of contracts involving school c and $X_{c,t}$ denote the set of contracts involving type t and school c.

The preference profile of all students is denoted as $\succsim_S = \{\succsim_{s_1}, ..., \succsim_{s_n}\}$, where each student s has a preference ordering $\succsim_s$ over $\mathcal{X}_s \cup \{\emptyset\}$. Let $\succsim_C = \{\succsim_{c_1}, ..., \succsim_{c_m}\}$ denote the priority profile of all schools, where each school c has a priority ordering $\succsim_c$ over $\mathcal{X}_c \cup \{\emptyset\}$.

An outcome X is *feasible* for I^T if i) each student s is matched to at most one school, i.e., $|X_s| \leq 1$, and ii) each school c admits at most q_c students, i.e., $|X_c| \leq q_c$. Note that we consider soft diversity constraints such that in a feasible outcome, a school may admit more students of type t than its maximum quota $\overline{\eta}_c^t$ or fewer students of type t than its minimum quota $\underline{\eta}_c^t$.

A contract (s, c) is *acceptable* if $(s, c) \succsim_s \emptyset$ and $(s, c) \succsim_c \emptyset$ hold. A feasible outcome X is *individually rational* if each contract $(s, c) \in X$ is acceptable. Without loss of generality, we focus on individually rational outcomes only.

Next, we introduce two important properties that are intensely studied in the context of school choice. The first property is called *non-wastefulness*, that requires that a feasible outcome should make efficient use of vacant school seats.

Definition 3 (Non-wastefulness). *Given a feasible outcome X, student s claims an empty seat of school c if $(s,c) \succ_s X_s$ and $|X_c| < q_c$. A feasible outcome is non-wasteful if no student claims an empty seat.*

Ehlers *et al.* [2014] proposed a fairness concept in Definition 4 for soft diversity constraints that captures a natural idea called *dynamic priorities*: Schools give higher precedence to students whose types have not met the minimum quotas, medium precedence to students whose types have filled the minimum quotas, but not the maximum quotas, and lower precedence to students whose types have reached the maximum quotas[7].

Definition 4 (Fairness). *Given an instance I^T and a feasible outcome X, a student s of type t has* EHYY-justified-envy *towards another student s' of type t' if $(s,c) \succ_s X_s$, $(s',c) \in X$ and either i) $t = t'$ and $(s,c) \succ_c (s',c)$, or ii) $t \neq t'$ and one of the following cases holds,*

- *(a) $|X_{c,t}| < \underline{\eta}_c^t$, $|X_{c,t'}| \leq \underline{\eta}_c^{t'}$ and $(s,c) \succ_c (s,c')$;*
- *(b) $|X_{c,t}| < \underline{\eta}_c^t$ and $|X_{c,t'}| > \underline{\eta}_c^{t'}$;*
- *(c) $\underline{\eta}_c^t \leq |X_{c,t}| < \overline{\eta}_c^t$, $\underline{\eta}_c^{t'} < |X_{c,t'}| \leq \overline{\eta}_c^{t'}$ and $(s,c) \succ_c (s,c')$;*
- *(d) $\underline{\eta}_c^t \leq |X_{c,t}| < \overline{\eta}_c^t$, $|X_{c,t'}| \geq \overline{\eta}_c^{t'}$;*
- *(e) $|X_{c,t}| \geq \overline{\eta}_c^t$, $|X_{c,t'}| > \overline{\eta}_c^{t'}$ and $(s,c) \succ_c (s,c')$.*

An outcome is EHYY-fair if no student has EHYY-justified-envy towards another student.

4.2 Comparison with the Previous Reduction

A recent paper [Aziz *et al.*, 2019] studied the connection between *hard diversity constraints* and *hard regional quotas*. If we apply their reduction directly, then soft diversity constraints are converted into soft regional quotas instead of hard regional quotas. Next, we informally describe our transformation and explain the difference from the previous reduction through Example 2.

Example 2. Consider one school c with capacity $q_c = 10$ and soft maximum quota $\overline{\eta}_t = 6$ of type t. In the reduction of [Aziz *et al.*, 2019], we create one region r corresponding to school c with regional quota $\overline{\delta}_r = 10$ and one hospital h corresponding to type t with capacity $q_h = 10$. Then create one more region r_1 that contains hospital h only with soft regional quota 6. This is because school c could admit more than 6 students of type t without violating feasibility requirement. Thus in the reduction by Aziz *et al.* [2019], *soft diversity constraints* are converted into *soft regional quotas*.

Under our reduction, we create two hospitals h_1, h_2 for type t where hospital h_1 has capacity $q_{h_1} = 6$ corresponding to the soft maximum quota $\overline{\eta}_t$ and hospital h_2 has capacity $q_{h_2} = q_c - \overline{\eta}_t = 4$ corresponding to the remaining capacity after

[7] To distinguish with the priority order of schools over contracts, we refer to the dynamic priorities over student types as precedence.

reaching the soft maximum quota. And region r gives larger weight to hospital h_1 and smaller weight to hospital h_2. Each hospital can be considered as a region that contains itself. Then soft diversity constraints are converted into hard regional quotas.

Next, we proceed to the formal reduction from an instance I^T of matching with soft diversity constraints into to a corresponding instance I^R of matching with regional quotas. And we show how non-wastefulness and fairness in the former setting are preserved as stability in the latter setting.

For each student $s_i \in S$, create a corresponding doctor d_i. Let $D = \bigcup_{s_i \in S} d_i$ denote the set of doctors. For each school $c_j \in C$ and each type $t \in T$, create three hospitals $h^1_{c,t}$, $h^2_{c,t}$ and $h^3_{c,t}$ with capacity $\underline{\eta}^t_{c_j}$, $\overline{\eta}^t_{c_j} - \underline{\eta}^t_{c_j}$ and $q_{c_j} - \underline{\eta}^t_{c_j}$, respectively. Assign weights to each induced hospital as follows: $w(h^1_{c,t}) = 3$, $w(h^2_{c,t}) = 2$ and $w(h^3_{c,t}) = 1$. Let $H_{j,t} = \{h^1_{c,t}, h^2_{c,t}, h^3_{c,t}\}$ be the set of hospitals induced from school c_j for type t.

Intuitively, hospitals $h^1_{c,t}$ corresponds to the case that minimum quota $\underline{\eta}^t_{c_j}$ has not reached yet, hospital $h^2_{c,t}$ corresponds to the case that the minimum quota has reached but not the maximum quota, and hospital $h^3_{c,t}$ corresponds to the remaining capacity after reaching the maximum quota $\overline{\eta}^t_{c_j}$.

For each school $c_j \in C$, create one region r_j that contains all induced hospitals $H_j = \bigcup_{t \in T} H_{j,t}$ from school c_j. The regional maximum quota of region r_j is the capacity q_{c_j} of school c_j. Let $R = \bigcup_{c_j \in C} r_j$ be the set of regions. For each contract $x = (s_i, c_j) \in \mathcal{X}$ with $T(s_i) = t$, create 3 contracts $y^1_{i,j} = (d_i, h^1_{j,t})$, $y^2_{i,j} = (d_i, h^2_{j,t})$, $y^3_{i,j} = (d_i, h^3_{j,t})$ involving each induced hospital $h^k_{j,t}$ for $k \in [1, 3]$.

For each doctor $d_i \in D$, given any two contracts $y^o_{i,j} = (d_i, h^o_{j,t})$ and $y^{o'}_{i,j'} = (d_i, h^{o'}_{j',t})$ involving doctor d_i, i) if $j \neq j'$, then doctor d_i's preference over these two contracts is consistent with student s's preference over corresponding contracts (s_i, c_j) and $(s_i, c_{j'})$; ii) if $j = j'$, doctor d_i is indifferent between two contracts. For each hospital $h^i_{j,t} \in H_j$ induced from school c_j, its priority ordering is consist with $\succ_{c_j}$ of school c_j. For each region r_j induced from school c_j, given any two contracts (d_i, h), $(d_{i'}, h') \in Y_{r_j}$, i) if $d_i \neq d_{i'}$, then region r_j's priority over two contracts is consistent with priority ordering of school c_j over contracts (s_i, c_j) and (s'_i, c_j); ii) if $d_i = d_{i'}$, then region r_j is indifferent between two contracts.

Next, we show how to create a corresponding outcome Y for induced instance I^R from a feasible outcome X for I^T. The general idea is that if student s_i is matched to school c_j, then doctor d_i is matched to region r_j. For the set of doctors who are matched to region r_j and correspond to students of type t, region r_j assigns $\underline{\eta}^t_{c_j}$ doctors with highest regional priority to hospital $h^1_{j,t}$, assigns $\overline{\eta}^t_{c_j} - \underline{\eta}^t_{c_j}$ doctors to hospital $h^2_{j,t}$ with second highest regional priority and assigns $q_{c_j} - \underline{\eta}^t_{c_j}$ doctors with lowest regional priority to hospital $h^3_{j,t}$.

Theorem 2. *A feasible outcome X is fair and non-wasteful for I^T with soft diversity constraints if and only if the induced outcome Y is stable for the corresponding instance I^R with regional quotas.*

Proof. First we prove that if the induced outcome Y is not stable for I^R, then the outcome X either admits justified envy or admits an empty seat. Suppose a doctor d and a hospital h form a blocking pair with $y = (d, h) \notin Y$. i) If the outcome $Y \cup \{(d, h)\} \setminus \{Y_d\}$ is feasible, then region r with $h \in r$ has not reached its regional maximum quota. In other words, school c corresponding to region $r$still has a vacant seat and student s corresponding to doctor d claims an empty seat at school c. Thus the outcome X is wasteful. ii) If there exists a contract $y' = (d', h') \in Y$ with $h, h' \in r$ and for the outcome $Y' = Y \setminus \{y'\}$, one of the three cases holds: ii-a) $h = h'$, $y \succ_h y'$ and $y \succ_r y'$, ii-b) $h \neq h'$ and $\beta(Y'_r, y, y') > 0$; or ii-c) $h \neq h'$, $\beta(Y'_r, y, y') = 0$ and $y \succ_r y'$. Let students s and s' with type t and t' correspond to doctors d and d'. Then either student s's type is more important than s''s type or both types are tied in terms of importance to school c, but the contract involving student s has higher school priority. Thus for all cases ii-a, ii-b, ii-c, student s has justified envy towards student s'.

Next, we prove that if the outcome X is not fair or is wasteful, then the induced outcome Y is not stable. i) If student s has justified envy towards student s', then either the type of student s is more important than the one of s' or both types are tied in terms of importance, but the contract involving student s has higher school priority. Then either hospital h has higher weight than h', or both hospitals have the same weight and contract y has higher regional priority. ii) If the outcome X is wasteful, then doctor d can be placed at region r without exceeding regional maximum quota. For both cases, doctor d and hospital h form a blocking pair.

5 Conclusion

In this paper, we studied the two-sided matching problem with regional quotas. We introduced a novel approach to solve the inefficiency issue in Japanese residency matching by assigning weights to hospitals. We also proposed a novel class of Generalized Deferred Acceptance with Regions (GDA-R) algorithms that deals with regional quotas and weights. We established a connection between matching with soft diversity constraints and matching with hard regional quotas. This connection implies our GDA-R algorithm also applies to another important field in matching.

The GDA-R is a new framework for matching with regional quotas and there are many directions to explore. For instance, what is a sufficient condition that guarantees a stable outcome and guarantees that there is no incentive for doctors to misreport their preferences?

Bibliography

A. Abdulkadiroğlu and T. Sönmez. School choice: A mechanism design approach. *American Economic Review*, 93(3):729–747, 2003.

H. Aziz, S. Gaspers, Z. Sun, and T. Walsh. From matching with diversity constraints to matching with regional quotas. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 377–385. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

F. Echenique and M. B. Yenmez. How to control controlled school choice. *American Economic Review*, 105(8):2679–94, August 2015.

L. Ehlers, I. E. Hafalir, M. B. Yenmez, and M. A. Yildirim. School choice with controlled choice constraints: Hard bounds versus soft bounds. *Journal of Economic Theory*, 153:648–683, 2014.

Y. A. Gonczarowski, N. Nisan, L. Kovalio, and A. Romm. Matching for the Israeli "Mechinot" gap year: Handling rich diversity requirements. In *Proceedings of the 20th ACM Conference on Economics and Computation*, pages 321–321, 2019.

M. Goto, R. Kurata, N. Hamada, A. Iwasaki, and M. Yokoo. Improving fairness in nonwasteful matching with hierarchical regional minimum quotas. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1887–1888, 2015.

M. Goto, A. Iwasaki, Y. Kawasaki, R. Kurata, Y. Yasuda, and M. Yokoo. Strategyproof matching with regional minimum and maximum quotas. *Artificial intelligence*, 235:40–57, 2016.

M. Goto, F. Kojima, R. Kurata, A. Tamura, and M. Yokoo. Designing matching mechanisms under general distributional constraints. *American Economic Journal: Microeconomics*, 9(2):226–62, 2017.

I. E. Hafalir, M. B. Yenmez, and M.A. Yildirim. Effective affirmative action in school choice. *Theoretical Economics*, 8(2):325–363, 2013.

N. Hamada, C. Hsu, R. Kurata, T. Suzuki, S. Ueda, and M. Yokoo. Strategyproof school choice mechanisms with minimum quotas and initial endowments. *Artificial Intelligence*, 249:47–71, 2017.

J. W. Hatfield and F. Kojima. Matching with contracts: Comment. *American Economic Review*, 98(3):1189–94, 2008.

J. W. Hatfield and P. R. Milgrom. Matching with contracts. *American Economic Review*, 95(4):913–935, 2005.

Y. Kamada and F. Kojima. Efficient matching under distributional constraints: Theory and applications. *The American Economic Review*, 105(1):67–99, 2015.

Y. Kamada and F. Kojima. Recent developments in matching with constraints. *The American Economic Review*, 107(5):200–204, 2017.

F. Kojima, A. Tamura, and M. Yokoo. Designing matching mechanisms under constraints: An approach from discrete convex analysis. *Journal of Economic Theory*, 176:803–833, 2018.

F. Kojima. School choice: Impossibilities for affirmative action. *Games and Economic Behavior*, 75(2):685–693, 2012.

S. D. Kominers and T. Sönmez. Designing for diversity in matching. In *Proceedings of the 14th ACM Conference on Economics and Computation (ACM-EC)*, pages 603–604, 2013.

R. Kurata, N. Hamada, A. Iwasaki, and M. Yokoo. Controlled school choice with soft bounds and overlapping types. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 951–957, 2015.

R. Kurata, N. Hamada, A. Iwasaki, and M. Yokoo. Controlled school choice with soft bounds and overlapping types. *Journal of Artificial Intelligence Research*, 58:153–184, 2017.