Fairness in Operations: From Theory to Practice

by

Nikolaos K. Trichakis

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author . . . .
Sloan School of Management
May 11, 2011

Certified by . . . .
Dimitris J. Bertsimas
Boeing Professor of Operations Research
Co-director, Operations Research Center
Thesis Supervisor

Certified by . . . .
Vivek F. Farias
Robert N. Noyce Career Development Assistant Professor of Management
Thesis Supervisor

Accepted by . . . .
Patrick Jaillet
Dugald C. Jackson Professor, Department of Electrical Engineering and Computer Science
Co-director, Operations Research Center

Fairness in Operations: From Theory to Practice

by

Nikolaos K. Trichakis

Submitted to the Sloan School of Management
on May 11, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

This thesis deals with two basic issues in resource allocation problems. The first issue pertains to how one approaches the problem of designing the "right" objective for a given resource allocation problem. The notion of what is "right" can be fairly nebulous; we consider two issues that we see as key: efficiency and fairness. We approach the problem of designing objectives that account for the natural tension between efficiency and fairness in the context of a framework that captures a number of problems of interest to operations managers. We state a precise version of the design problem, provide a quantitative understanding of the tradeoff between efficiency and fairness inherent to this design problem and demonstrate the approach in a case study that considers air traffic management.

Secondly, we deal with the issue of designing implementable policies that serve such objectives, balancing efficiency and fairness in practice. We do so specifically in the context of organ allocation for transplantation. In particular, we propose a scalable, data-driven method for designing national policies for the allocation of deceased donor kidneys to patients on a waiting list, in a fair and efficient way. We focus on policies that have the same form as the one currently used in the U.S., that are policies based on a point system, which ranks patients according to some priority criteria, *e.g.*, waiting time, medical urgency, etc., or a combination thereof. Rather than making specific assumptions about fairness principles or priority criteria, our method offers the designer the flexibility to select his desired criteria and fairness constraints from a broad class of allowable constraints. The method then designs a point system that is based on the selected priority criteria, and approximately maximizes medical efficiency, *i.e.*, life year gains from transplant, while simultaneously enforcing selected fairness constraints.

Using our method, we design a point system that has the same form, uses the same criteria and satisfies the same fairness constraints as the point system that was recently proposed by U.S. policymakers. In addition, the point system we design delivers an 8% increase in extra life year gains. We evaluate the performance of all policies under consideration using the same statistical and simulation tools and data as the U.S. policymakers use. We perform a sensitivity analysis which demonstrates that the increase in extra life year gains by relaxing certain fairness constraints can be as high as 30%.

Thesis Supervisor: Dimitris J. Bertsimas
Title: Boeing Professor of Operations Research
Co-director, Operations Research Center

Thesis Supervisor: Vivek F. Farias
Title: Robert N. Noyce Career Development Assistant Professor of Management

Acknowledgments

I would like to thank my academic advisors, Dimitri Bertsima and Vivek Farias, for their phenomenal support and guidance over the course of my doctoral work at MIT. I find the amount of effort that both Dimitris and Vivek invested in my academic and personal development truly remarkable, and I would like to thank them for that. Interacting and working with such great teachers, researchers and mentors as Dimitris and Vivek are, has been an absolute joy and privilege.

I would like to thank the other two members of my thesis committee, Georgia Perakis and Itai Ashlagi, who have helped me greatly in improving this work. Georgia has been extremely supportive and helpful, along academic and non-academic dimensions, and I would like to particularly thank her for that.

Every ORC faculty member I interacted with has been very supportive, but I would like to specifically thank Retsef Levi and Stephen Graves for all their help and our fruitful interaction those years.

I am convinced that the ORC is a truly unique environment and I consider myself very fortunate to be a part of it. I would like to thank Xuan Vinh Doan, Chai Bandi, Shashi Mittal, Andre Calmon, Matthew Fontana, Apostoli Ferti, Shubham and Vishal Gupta, Nick Howard, Phil Keller, Jon Kluberg, Wei Sun, Joline Uichanco, Gareth Williams, Eric Zarybnisky and Juliane Dunkel. I would also like to thank my close friends and future colleagues, Kosta Bimpiki, Dan Iancu, Doug Fearing, Ilan and Ruben Lobel, Dave Goldberg, Sasha Rikun and Dimitri Bisia.

I always thought that my adventure in the U.S. would be a short-lived one, but living in Cambridge the last 4 years has admittedly changed that, primarily because of my interaction with unique people and friends like Anastasia and Tom Trikalino, Sofia and Marko Tricha, Amalia and Vag Koutsolelo, Alexandro Micheli, Yanni Bertsato, Yanni Simaiaki, Yuri Gagarin[1], Bettina and Pol Ypodimatopoulo, Antoni Vytinioti, Gerry Tsoukala, Dimitri Bisia, Kosta Bimpiki, Ruben Lobel, Dan Iancu, Sasha Rikun, Dave Goldberg, Gareth Williams[2], Nikola Pyrgioti and Vasili Papapostolou.

[1] aka Giorgos Papachristoudis.

[2] Best remembered for his timely DT drops.

Finally, I would like to thank my wife Eleni for her continuous love and support throughout those years. I always thought from day 1 that I would dedicate this thesis to her; reflecting back on all those years however, I now believe that all this has really been a team effort between the two of us and I can no longer dedicate to her something that she has ownership of. I want to thank her again for her patience, support and love.

Together with my wife Eleni, we would like to thank our parents, Kosta, Theologia, Achillea and Anastasia, and our brothers, Pavlo, Saki, Maroudia and Niko for their unconditional love.

Contents

1 Introduction 13

1.1 Literature Review . . . 18

2 A General Framework for Resource Allocation 23

2.1 Utilitarian and Fair Allocations . . . 25

2.1.1 Using the Framework . . . 29

3 Fairness Schemes 33

3.1 Proportional Fairness . . . 36

3.2 Max-min Fairness . . . 36

3.3 α-fairness . . . 38

4 The Efficiency-Fairness Tradeoff 41

4.1 The Price of Fairness . . . 42

4.1.1 Worst-case Examples . . . 58

4.2 The Price of Efficiency . . . 62

4.2.1 Worst-case Examples . . . 64

4.3 Discussion . . . 65

5 Fairness in Air Traffic Management 67

5.1 The Model . . . 69

5.2 Numerical Experiments . . . 70

6 Fairness in Organ Allocation for Kidney Transplantation 75

6.1 Overview 75

6.1.1 Literature Review 79

6.2 Distribution and Allocation Policies 80

6.2.1 National Allocation Policies 81

6.3 Designing Allocation Policies 86

6.3.1 Methodology 87

6.3.2 Discussion 91

6.4 Numerical Evidence and the Design of a New Allocation Policy . . . 94

6.4.1 Data, Statistical Models and Tools 94

6.4.2 Methodology 95

6.4.3 Results 97

6.4.4 Sensitivity Analysis 100

7 Concluding Remarks 105

A Technical Notes 109

A.1 More on Near Worst-case Examples for the Price of Fairness 109

A.2 Auxiliary Results 112

B A Model for Air Traffic Flow Management 127

List of Figures

4-1 An example of a two-dimensional utility set, with the points of interest and the associated supporting hyperplanes used in the proof of Theorem 1, in Section 4.1. 44

4-2 Bounds on the price of α-fairness for different values of α implied by Theorem 2, in Section 4.1. The bounds are plotted as functions of the number of players n. 59

4-3 The network flow topology in case of $n = 5$ and $y = 3$, for the bandwidth allocation example in Section 4.1.1. 60

4-4 The price of α-fairness for constructed examples (markers) in Section 4.1.1, for different values of α. The corresponding bounds are also plotted (lines). The values/bounds are plotted as functions of the number of players. 61

4-5 Bounds on the price of efficiency of α-fair allocations for different values of the number of players n implied by Theorem 3, in Section 4.2. The bounds are plotted as functions of inequality aversion parameter α. . 64

5-1 (a) The price of α-fairness and (b) the coefficient of variation of the individual utilities for the numerical experiments in Section 5.2, for different values of the parameter α. 72

6-1 An illustration of the functionality of the proposed method for designing scoring rule based policies for the allocation of deceased donor kidneys to patients for transplantation in Section 6.3. 86

6-2 Simulated percentage distribution of transplant recipients across the 11 distinct geographical regions that UNOS has divided the U.S. into, under the KTC policy (white) and the policy designed in Section 6.4 (black), for an out-of-sample period of 6 months in 2008. 98

6-3 Simulated life years from transplant gains for policies (designed by our method) with relaxed constraints on all patient groups of different dialysis time, for various values of the slack parameter s; for more details see Section 6.4.4. The results are for an out-of-sample period of 6 months in 2008. The marker corresponds to the operational point of the policy proposed by the UNOS policymakers. 102

6-4 Simulated life years from transplant gains for policies (designed by our method) with relaxed constraints on all patient groups of different sensitization levels, for various values of the slack parameter s; for more details see Section 6.4.4. The results are for an out-of-sample period of 6 months in 2008. The marker corresponds to the operational point of the policy proposed by the UNOS policymakers. 103

List of Tables

2.1 Preview of results for Chapter 5. For each airline, we report the actual delay (in minutes) across its flights on that day (under current practice), and the delay reductions that different α-fair allocations would achieve. 31

5.1 Numerical results for the case study in Section 5.2, for 2 days and 4 airlines. For each airline, we report the actual delay (in minutes) across its flights on that day (under the RBS policy), and the delay reductions that different allocations would achieve. 71

6.1 Simulated allocation results of the KTC policy and the policy designed in Section 6.4, for an out-of-sample period of 6 months in 2008. . . . 99

Chapter 1

Introduction

Operations managers are frequently concerned with problems of resource allocation. They must build quantitative decision models for such problems, calibrate these models, and then use suitable decision support/optimization tools to make implementable decisions or allocations. There is a vast amount of academic research in operations management and associated fields available to complement each of the steps above. At the risk of belaboring the obvious, the following examples serve to specify this connection with resource allocation:

- Call center design: Pools of specialized agents must be utilized to provide service to various classes of customers. Decisions include staffing levels across agent pools and routing protocols to assign customers to agents. If delays experienced by customers are associated with dollar values, a natural objective is minimizing the expected delay costs incurred across customer classes.

- Healthcare scheduling: Beds (and the associated resources of doctors, nurses and equipment) must be allocated over time to patients in need of care. In the case of an operating room, a natural objective might be (and, at least nominally, frequently is) the maximization of throughput. In an urgent care setting, one may care about delay related objectives. For instance, in the case of scheduling a specialized ICU, a natural objective is minimizing the expected waiting time for a bed. In more sophisticated settings, the objectives may be directly related to physiological outcomes – for instance, minimizing mortality.

- Management of large scale traffic and communication networks: Available network capacity (or bandwidth) must be allocated to traffic flows. A natural objective is the maximization of throughput; *i.e.*, the total flow routed through the network.

- Air traffic control: In case of inclement weather, the U.S. Federal Aviation Administration (FAA) needs to re-allocate landing and takeoff slots among the airlines. Delays on the ground and in the air are associated with dollar values and a natural objective to consider is then re-allocating slots in a manner that minimizes the total dollar impact of the resulting delays.

- Allocation of cadaveric organs: The United Network for Organ Sharing (UNOS) oversees the allocation of cadaveric organs (*e.g.*, kidneys, livers etc.) to patients in need of them. Medical researchers and statisticians have built sophisticated models that predict the physiological outcome of allocating a specific organ (as measured by a number of attributes) to specific patients. These outcomes are frequently measured in terms of the number of quality adjusted life years (QUALYs) the transplant will add to the patient's life. A natural objective is to assign organs in a manner that maximizes the expected QUALYs added across the population over time.

The list above is somewhat idiosyncratic – there are a number of other examples that one could list. What the examples above do share in common, however, is their undoubted relevance from the perspective of the social utility at stake in their solution. Academic work on these problems frequently tends to focus on decision support related issues. For example, how does one design a routing scheme that minimizes delays in a particular queueing model? Or how does one make organ allocation decisions given the uncertainties in supply, demand and the acceptance behavior of patients? In other words, the question is, given a particular objective, how does one devise an implementable policy that serves that objective in practice?

Another basic issue however, pertains to the selection of the actual objective. That is, how does one come up with the right operational objective in each of the scenarios above? Is the "obvious" objective the right one? To return to the examples above, it is hard to argue that minimizing the dollar impact of delays is not a noble objective for the FAA – in fact, a vast number of proposals attempt to do just that. Of course, this noble objective fails to account for the outcome an individual airline might have to endure as part of such an allocation. Similarly, in the case of organ allocation, it is difficult to argue against the value of an allocation scheme that maximizes the number of life years generated via transplantation activities. Unfortunately, this objective fails to account for inequities such a scheme might imply for a particular group of patients (based, for instance, on their age, or particular physiological characteristics). Designing the right objective is a first-order issue, and the tensions inherent in designing the "right" objective are frequently complex as the examples we have just noted make clear. This crucial design task is nonetheless frequently executed in an ad-hoc fashion.

The contributions of the present work are along the two aforementioned central questions: for a resource allocation problem (a) how does one select/ design the right operational objective, and, given such a selection, (b) how does one find an implementable policy that serves this objective in practice?

More specifically, for the first question this work attempts to provide some structure to guide the underlying desing task as follows:

- An abstract framework: We view resource allocation problems through the lens of welfare economics. In particular, we imagine that any resource allocation problem may be viewed as one where the system designer (or operations manager, in this case) must decide on an allocation of *utilities* to several parties from some set of feasible utilities. How might we select an allocation from among the many efficient allocations possible? A little reflection shows that the criterion implicitly employed in the examples above is a *utilitarian* criterion – one simply seeks to maximize the sum of utilities. We will return to this notion later, but for now simply note that this criterion can in many situations be unambiguously interpreted as *the* criterion by which to measure efficiency. Put mathematically, the manager's job is selecting an allocation of utilities to n parties, $u \in \mathbf{R}^n$ from some set of feasible utilities U. The utilitarian criterion seeks to find an allocation u to maximize $\sum_j u_j$.

- Inequity: The utilitarian criterion is neutral towards inequity. Coupled with the fact that in many of the examples we have encountered above, an auxiliary mechanism for monetary compensation is not implementable, this inequity is the root cause of the tensions in designing an appropriate objective. Fortunately, we have available to us an axiomatic treatment of attitudes towards inequity. This axiomatic treatment has deep roots in early philosophy, and has quantitatively culminated over the last fifty years in a family of *social welfare functions* parametrized by a single parameter that measures the attitude of the system designer towards inequities. This family is given by [1]

$$\sum_{j=1}^{n} \frac{u_j^{1-\alpha}}{1-\alpha}.$$

  The parameter α measures an aversion to inequity. This family of "α-fair" welfare functions subsumes the well known Nash ($\alpha \to 1$) and Kalai-Smorodinsky ($\alpha \to \infty$) solutions, also referred to as proportional and max-min fairness.

- The design problem: The above setting allows us to reduce the problem of designing an appropriate objective to the selection of a single parameter, or, equivalently, of a fairness scheme. A natural tradeoff implicit in selecting this parameter or scheme (at least, as seen from the operational perspective), is the loss in total system utility, or loosely, efficiency, incurred in the pursuit of equity. We seek to quantify this tradeoff. In particular, we show that this loss (measured in relative terms) scales like $1 - \Theta\left(n^{-\frac{\alpha}{\alpha+1}}\right)$, where n is the number of parties and α a design choice that measures the importance of equity. Conversely, another tradeoff that arises from the selection of the parameter is the loss in

[1] It is tempting to confuse this welfare function with the well-known Arrow-Pratt utility function; it is important to not conflate the notions of a utility function and welfare function.

fairness incurred in the pursuit of efficiency. To this end, we show that a natural measure of fairness, namely the minimum utility that every party is guaranteed to derive, degrades (measured in relative terms) like $1 - \Theta\left(n^{-\frac{1}{\alpha}}\right)$. The above quantifications are among the principal theoretical contributions in this work, and to the best of our knowledge are the first general characterizations of the very natural underlying tradeoff curves.

Furthermore, we provide a concrete illustration of the value of the framework above by implementing it in the context of the air traffic management problem mentioned earlier. In particular, we present a concrete, quantitative statement of the design problem a system manager seeking the "right" operational objective might solve, and then explore the consequences of various solutions in a study using detailed historical air traffic data.

For the second question pertaining to the implementation of policies that account for fairness and efficieny, we focus on a specific problem: the problem faced by the UNOS in allocating deceased-donor kidneys to patients on a waiting list. In this setup, the identification of the ideal fairness objective is very challenging and, to some extent, subjective. Additionally, due to the dynamic and stochastic nature of the problem, the design of an allocation policy that (a) complies with a fairness notion and (b) is simultaneously as efficient[2] as possible, is potentially even more challenging. Our contribution is a mechanism that takes fairness constraints as input and designs in a systematic way an allocation policy that approximately maximizes anticipated net life year gains of the patients, satisfying the fairness constraints. Moreover, the designed allocation policies have the same form as the current allocation rule in use nationwide, namely the form of a point system that is used to rank patients. Such a point system is easy to communicate to physicians and patients and is eminently implementable.

From a practical perspective, the contribution is particularly important, as the UNOS is also currently revising the national allocation policy. Using the mechanism,

[2] Efficiency of allocation policies in the context of organ allocation for transplantation is typically measured by the number of life year gains garnered by transplantation acitivities.

we design an allocation policy that matches the fairness properties of the so far proposed policy by the UNOS, relies on the same criteria (point system), and achieves a relative increase of 8% in anticipated life year gains. The performance gain is established by using the exact same data and simulation tools as the UNOS, obtained from the Scientific Registry of Transplant Recipients. Moerover, we use our method to perform a sensitivity analysis that explores the consequences from relaxing or introducing fairness constraints. In the case of some constraints, relaxations of fairness constraints can result in life year gains on the order of 30%.

The structure of this thesis is as follows. In the next section, we review relevant applications in the literature where the need for the design of objectives that balance equity and "efficiency" is apparent. We will also review important developments in the welfare economics and bargaining literature that yield the foundations of our framework. Chapters 2-5 deal with the problem of designing an operational objective: in Chapter 2, we introduce our framework rigorously, placing it in the context of welfare economics, and simultaneously relating it to a couple of concrete operational problems. We review relevant fairness schemes in Chapter 3. Chapter 4 establishes the tradeoff curves that, as we have discussed, can guide the design of an equitable objective. Chapter 5 considers a concrete design problem in this vain in the context of air traffic management. This case-study uses actual air traffic data and illustrates the value of our framework. Finally, Chapter 6 deals with the implementation question and discusses the kidney allocation problem. Concluding remarks are included in Chapter 7.

1.1 Literature Review

Economic Theory: A typical setting in welfare economics concerns the scenario where a central planner must make an allocation of goods in an economy to a number of distinct entities. The planner is aware of the preferences of the entities, and one typically assumes these are described via cardinal utilities. The central problem in welfare economics is then concerned with how the central planner should go about making these allocations. Samuelson [56] provided the first formulation in which the relevant constraint set for the planner was the set of achievable utility allocations, or the *utility possibility set*; an idea which became central in this area. In fact, our framework is based on exactly that notion. The welfare economics problem can then be stated as the problem of picking a point in the utility set (for more details see Chapter 2).

One prominent way of addressing the allocation problem above has been the identification of a real-valued social welfare function of the allocation of utilities, which is used by the central decision maker to rank allocations. The approach in which the welfare function reflects the distributional value judgement of the central planner was first taken by Bergson [8] and Samuelson [56]. Some of the most important instances of social welfare functions are the utilitarian, maximin and constant elasticity functions. For the merits of the utilitarian function, see [27]. The maximin function is based on the *Rawlsian justice*, introduced by Rawls [50]. For details regarding the constant elasticity function, see Chapter 3.3. We refer the reader to [77] and [58] for a thorough overview of the above work. Mas-Colell et al. [36] provides a nice introduction.

Another approach to dealing with the allocation problem is provided by *bargaining* theory. Here one formulates axioms that any allocation must satisfy and then seeks allocation rules that satisfy these axioms. The standard form of the bargaining problem was first posed by Nash [38]. Nash [38] provided a set of axioms that an allocation must satisfy, and demonstrated the unique allocation rule satisfying these axioms, all in a two-player setting. An alternative solution (and axiomatic system) for the two-player problem was introduced by Kalai and Smorodinsky [29]. The work by Lensberg [34] extended these solutions to a setting with multiple players. For other axiomatic formulations see [54]. Finally, see [77] and [36] for surveys of the literature.

Applications: As is evident from our introductory remarks, the need to design resource allocation objectives that in addition to being “efficient” in an appropriate sense, are also equitable is ubiquitous. Below we discuss a biased sample of related applications:

Healthcare: The fundamental question in this area is how to balance equity in health provision and medical utility, which typically corresponds to the aggregate health of the population; see [72]. This natural dichotomy between equity and efficiency is apparent across a wide spectrum of healthcare operations. For instance, in managing operations in a hospital's intensive care unit, one cannot simply maximize throughput without accounting for fairness and medical urgency; see [65], [19]. Furthermore, in their book, "Medicine and the market: equity v. choice", Callahan and Wasunna [17] discuss the use of markets and government funding to balance efficiency and equity (respectively), for the purposes of insurance policies and a healthcare reform. See also [46] for a related discussion. The efficiency-fairness tradeoff is also particularly important in the allocation of research funds by the National Institutes of Health (NIH) of the United States over various biomedical research projects. Each of the projects deals with improving the care provided to patients of particular diseases (e.g., cancer, HIV, etc.). A primary goal of the allocation is then to maximize clinical efficiency, that is, to allocate the funds such that the resulting research gains lead to the highest possible anticipated increase in quality adjusted life years of the population. Such practice however, may potentially be unethical and result in age or race discrimination. To ensure an equitable health treatment, the NIH needs to diversify its allocation, trading off clinical efficiency and fairness (see [51] and [14]). Finally, similar considerations arise in the allocation of deceased-donor kidneys to patients on a waiting list; see [60] and [62] for a detailed discussion.

Service Operations: Other settings where the equity-efficiency tradeoff is of importance include call center design and other associated queuing problems, supply chain and service applications. As discussed previously, the maximization of the throughput or the minimization of average waiting time are the typical objectives for a service manager in designing a queuing system. Several studies have acknowledged the importance of accounting for inequity in these settings by employing alternative objectives such as the variability in service times or queue lengths, etc. (see [59], [2], [20]). Within the supply chain literature, Cui et al. [22] incorporate the concept of fairness into the conventional dyadic channel to investigate how fairness may affect the interactions between the manufacturer and the retailer. Finally, Wu et al. [76] study the impact of fair processes on the motivation of employees and their performance in execution. They examine the tradeoffs involved and study under which circumstances management should use fair processes or not.

Yet another application, revisited in Chapter 5 for a case study, is the air traffic control problem, alluded to in the discussion above. There is an extant body of research devoted to formulating and solving the problem of minimizing the total system delay cost (see [41], [12]). While this objective is natural, a somewhat surprising fact is that existing practice (at least within the United States) does not take into account delays in making such re-allocation decisions. The emphasis rather, is on an allocation that may be viewed as equitable or fair to the airlines concerned. Recent research work deals with combining those two objectives (see [71], [6], [10]).

Networks: The tradeoff between efficiency and fairness is hardly specific to just operations management problems. In particular, it is well recognized and studied in many engineering applications as well, ranging from networking and bandwidth allocation, job scheduling to load balancing. For instance, the network utility maximization problem has been heavily studied in the literature. In that problem, a network administrator needs to assign transmission rates to clients sharing bandwidth over a network, accounting for efficiency (*e.g.*, net throughput of the network) and fairness (*e.g.*, "equal" bandwidth assignment). For more details, see [9], [31] and [37].

Worst Case Analysis: Recent work has focused on studying the worst-case degradation of the utilitarian objective, *i.e.*, the sum of the utilities, under a fair allocation compared to the allocation that maximizes the utilitarian objective. Butler and Williams [16] show that the degradation is zero under a max-min fair allocation for a specialized facility location problem. Correa et al. [21] also analyze the degradation under a max-min fair allocation for network flow problems with congestion. Chakrabarty et al. [18] show that when the set of achievable "utilities" is a polymatroid, the worst-case degradation is zero under all Pareto resource allocations. This is a somewhat restrictive condition and a general class of resource allocation problems that satisfy this condition is not known. Relative to the above literature, the present work provides the first analysis that is simultaneously applicable to *general* resource allocation problems for a *general* family of allocation rules.

Chapter 2

A General Framework for Resource Allocation

We describe a general framework that captures the majority of the applications that are relevant to this work and are discussed in the Introduction. We then review allocation mechanisms that account for the objectives of equity and "efficiency" alluded to in the Introduction, introduce the notions of the price of fairness and the price of efficiency and conclude by highlighting the usefulness of our framework.

Consider a resource allocation problem, in which a *central decision maker* (CDM) needs to decide on the allocation of scarce resources among n players. Each player derives a nonnegative utility, depending on the allocation decided by the CDM (*e.g.*, via means of a utility function). For a given allocation of resources, there is thus a corresponding *utility allocation* $u \in \mathbf{R}_+^n$, with u_j equal to the utility derived by the jth player, $j = 1, \ldots, n$.

A utility allocation $u \in \mathbf{R}_+^n$ is *feasible* if and only if there exists an allocation of resources for which the utilities derived by the players are $u_1, u_2, \ldots, u_n$ accordingly. We define the *utility set* $U \subset \mathbf{R}_+^n$ as the set of all feasible utility allocations. Encapsulated in the notion of the utility set are the preferences of the players and the way they derive utility, as well as individual constraints of the players or the CDM, constraints on the resources, etc. Thus, the utility set provides a condensed way of describing a general resource allocation problem. Given the utility set, the CDM then needs to decide which utility allocation among the players to select, or, equivalently which point from the utility set to select. The notion of the utility set was introduced by Samuelson [56].

The above setup has been studied within the research areas of fair bargaining and welfare economics (see Chapter 1.1). Note that these utilities may not be quasi-linear; that is to say, there is no reason to assume that an allocation to a specific party might be substituted by a cash payment to that party. To illustrate the applicability of the setup, we discuss below two concrete applications.

Example 1. *As a concrete application of the model above, consider the call center design problem alluded to in the Introduction. An operations manager (the central decision maker) needs to decide on staffing levels across agent pools (the scarce resources) and routing protocols in order to serve n different customer classes (the players). Suppose that a specific set of decisions results in the jth customer class experiencing an expected waiting time of w_j, $j = 1, \ldots, n$, during steady-state operation of the center. The vector of steady-state expected waiting times of the customer classes is commonly referred to as the* performance vector. *Suppose also that the utility derived by the jth customer class is $v_j - c_j w_j$, where v_j is the constant nominal utility derived by that particular class for the service and c_j is effectively the value of time to the jth class. Let W be the set of achievable performance vectors, known as the* achievable performance set *or* space. *Note that the description of W might be very complex. The utility set in that case is*

$$U = \{v_1 - c_1 w_1, \ldots, v_n - c_n w_n \,|\, w \in W\}.$$

Note that a lot of work has been devoted to providing tractable descriptions of the underlying achievable performance space, W, or approximations of it, under different settings. Results of that kind are very powerful, as they allow one to maximize concave functions of the waiting times (e.g., utilities) very efficiently. We refer the reader to [26], [68] and [11] for early results in that field.

Example 2. *Consider the fund allocation problem faced by the NIH, discussed in*

Chapter 1.1. The NIH plays the role of the central decision maker, by allocating B monetary units (the scarce resources) to n different biomedical research projects. Each research project aims to improve the treatment of a particular disease, and thus serve the group of patients (the players) affected by that disease. The improvement is measured by the extra life years that patients of the associated disease gain, and is equal to the utility of that "player", i.e., the group of patients. For illustration purposes, suppose that for every monetary unit invested on the jth project, the anticipated gain of the patients having the associated disease is q_j extra life years. Suppose also that due to some regulations, the funding of the first k projects should be at least L monetary units. The utility set in this example is then

$$U = \left\{ (q_1 x_1, \ldots, q_n x_n) \,\middle|\, \sum_{j=1}^{n} x_j \leq B, \quad \sum_{j=1}^{k} x_j \geq L, \quad x \geq 0 \right\}.$$

The problem for the NIH is then to decide which allocation $u \in U$ to pick.

In the next section, we review social welfare functions and allocation mechanisms that give rise to "efficient" and "fair" allocations.

2.1 Utilitarian and Fair Allocations

A natural objective for the central decision maker is to maximize an efficiency metric of the system (defined appropriately). On the other hand, in an environment where self-interested parties are involved, such a practice might result in inequalities among the utilities derived by different players that in a typical economic setup would be compensated for via monetary transfers. Absent the ability to make such transfers, accounting for equity will most likely have a negative impact on the efficiency of the system. Indeed, the efficiency-fairness tradeoff is a central issue of resource allocation problems, see [30].

In this work, we adopt the sum of utilities (derived by the players) as our metric of system efficiency. This is referred to as the "utilitarian" criterion. Our rationale in doing so is two-fold:

- The utilitarian criterion emerges as the natural efficiency metric employed in practice. The examples alluded to thus far are cases in point, and by themselves are sufficient to justify this benchmark.

- In a general setting where monetary transfers are allowed as a mechanism to compensate for inequity, the sum of utilities is the *only* admissible criterion of efficiency. It stands to reason then, that the allocation induced by such a criterion may be viewed as efficient, whether or not monetary transfers are possible.

We next discuss utilitarian and fair allocations.

Utilitarian allocation: Given a utility set U, a utilitarian allocation corresponds to an optimal solution of the problem

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T u \\ \text{subject to} & u \in U, \end{array}$$

with variable $u \in \mathbf{R}_+^n$ and $\mathbf{1}$ is the vector of all ones. We denote the optimal value of this problem with SYSTEM (U), *i.e.*,

$$\text{SYSTEM}\,(U) = \sup\left\{\mathbf{1}^T u \,\middle|\, u \in U\right\}.$$

As discussed above, we will regard this value as corresponding to the highest possible level of system efficiency (or social utility) achievable.

The sum of utilities is among the most well studied social welfare functions, and is known as the Bentham utilitarian function given the philosophical justification of this criterion provided by Jeremy Bentham. The utilitarian principle of maximizing the sum of utilities is neutral towards inequalities among the utilities derived by the players. It is therefore possible that the utilitarian solution is achieved at the expense of some players. As a result, it is considered to lack fairness considerations, see [77].

Fair allocation: Alternatively to classical utilitarianism, the central decision maker might decide on the utility allocation incorporating fairness considerations. Depending on the nature of the problem and her own perception about fairness, the CDM picks a fairness scheme of her preference, that is, a set of rules or properties (*e.g.*, total equity, under which every player derives exactly the same utility). The selected allocation then needs to be compatible with the fairness scheme.

To make this more precise, we model a fairness scheme as a set of rules and a corresponding set function $\mathcal{S} : 2^{\mathbf{R}_+^n} \to \mathbf{R}_+^n$, that takes a utility set as an input, and maps it into an element of the utility set. Given a utility set U, $\mathcal{S}(U) \in U$ is then an allocation that abides to the set of rules of the fairness scheme in consideration.

Due to the subjective nature of fairness and different possible interpretations of equity, there is no scheme that is universally accepted as "the most fair". In particular, there has been a plethora of proposals in the literature under axiomatic bargaining, welfare economics, as well as in applications ranging from networks, air traffic management, healthcare and finance. We review many of those proposed fairness schemes in Chapter 3 and refer the reader to [77] for a more detailed exposition.

The aforementioned work has focused on proposing fairness principles and analyzing the fairness properties of various scheme (*e.g.*, via an axiomatic characterization). What is lacking however, is a precise understanding of the efficiency-fairness tradeoff implicit in a selection of a fairness scheme; an understanding of this tradeoff would provide the CDM with a useful design tool. In particular, in order to make decisions, the CDM needs to understand

(a) what the *efficiency loss* might be, and

(b) what the *fairness loss* might be

for a specific fairness scheme. This work sheds light on exactly those matters, by quantifying what the maximum efficiency and fairness loss can be for various widely-used fairness schemes. We next formally define the notions of the efficiency and fairness loss.

Efficiency loss and the price of fairness: As the central decision maker incorporates fairness considerations, the efficiency of the system (measured as the sum of utilities), is likely to decrease, compared to the efficiency under the utilitarian solution.

Suppose the CDM adopts a particular fairness scheme $\mathcal{S}$. Recall that in that case, $\mathcal{S}(U)$ will denote the associated fair allocation. Then, the efficiency of the system under fairness scheme $\mathcal{S}$ will be the sum of the components of the associated utility allocation, denoted by

$$\mathrm{FAIR}(U;\mathcal{S}) = \mathbf{1}^T \mathcal{S}(U).$$

The *efficiency loss* is the difference between the maximum system efficiency, that is SYSTEM (U), and the efficiency under the fair scheme, FAIR$(U;\mathcal{S})$. The efficiency loss relative to the maximum system efficiency is the so called *price of fairness*, defined as

$$\mathrm{POF}(U;\mathcal{S}) = \frac{\mathrm{SYSTEM}(U) - \mathrm{FAIR}(U;\mathcal{S})}{\mathrm{SYSTEM}(U)}.$$

This price is a number between 0 and 1, and corresponds to the percentage efficiency loss compared to the maximum system efficiency. It is a key quantity to understanding the efficiency-fairness tradeoff.

Fairness loss and the price of efficiency: In order to quantify and compare the fairness properties of different schemes, we need to select a fairness metric to adopt. Due to the subjective nature of fairness, note that such a selection can be nebulous, in a similar way to the selection of a fairness scheme (see discussion above).

The fairness metric we adopt in this work is the minimum utility. That is, given a utility allocaion u, we measure its fairness properties by $\min_j u_j$. This fairness metric was advocated by Rawls [50] and can be interpreted as a minimum guarantee of utility to all the players. The rationale behind this selection is presented in Chapter 3.3.

For a particular utility set U, our fairness metric attains its highest value for an allocation that has the maximum minimum utility guarantee for all players and is equal to

$$\max_{u \in U} \min_{j=1,\ldots,n} u_j.$$

As the CDM selects a different fairness scheme and fair allocation for that matter, the minimum utility guarantee is likely to dercrease.

Suppose the CDM adopts scheme $\mathcal{S}$; under the associated allocation $\mathcal{S}(U)$, the fairness metric evaluates to

$$\min_{j=1,\dots,n} \mathcal{S}(U)_j.$$

The *fairness loss* is the difference between the maximum value of the fairness metric and the metric evaluated at $\mathcal{S}(U)$. We then call the fairness loss relative to the maximum value of the fairness metric as the *price of efficiency*, defined as

$$\mathrm{POE}(U;\mathcal{S}) = \frac{\max\limits_{u\in U} \min\limits_{j=1,\dots,n} u_j - \min\limits_{j=1,\dots,n} \mathcal{S}(U)_j}{\max\limits_{u\in U} \min\limits_{j=1,\dots,n} u_j}.$$

The price of efficiency can be interpreted as the percentage loss in the minimum utility guarantee compared to the maximum minimum utility guarantee.

2.1.1 Using the Framework

In Chapter 3 we discuss various fairness schemes and in Chapter 4 we present worst-case analyses of the efficiency and fairness loss those schemes result in for a very broad class of problems. Note that in order to balance social utility and fairness in practice, a manager needs to understand the dependence of the efficiency and fairness loss on the choice of the fairness scheme (as discussed above). As it turns out however, either loss can be substantially different from instance to instance even within the same class of problems (a fact that is illustrated in the case study in Chapter 5). The worst-case analysis of the efficiency and fairness loss then provides a theoretical tool that can assist the understanding of the behavior of those quantities. Prior to presenting our contributions in detail, we review the framework described in this chapter and its usefulness in management practice.

The framework and the analysis we provide can be utilized by a central decision maker in order to design the appropriate operational objective in a resource allocation problem that strikes the right balance between social utility and fairness. Our approach delivers a systematic way of dealing with this problem and is summarized below.

(a) The central decision maker identifies the feasible ways of allocating the resources among the players and evaluates the associated profits (or utilities) of the players. Note that this step is routinely carried out in any quantitative analysis of a resource allocation problem.

(b) The central decision maker allocates resources (or, equivalently, utilities) according to a selected fairness scheme. The choice of the scheme reflects the balance between social utility and fairness and can be guided as follows.

  - The literature (*e.g.*, existing axiomatic treatment in the economics literature) provides an understanding of what fairness properties each scheme possesses. Additionally, our worst-case analysis of the price of efficiency further enhances this understanding and provides a characterization of the fairness loss.
  - Our worst-case analysis of the price of fairness characterizes the efficiency loss associated with a particular scheme.

Note that among other merits, our approach provides a unifying way of incorporating fairness in resource allocation problems. Moreover, our analysis of the efficiency and fairness loss complements the existing analyses of fairness properties of various schemes, and thus provides the central decision maker with powerful tools that assist her task of balancing efficiency and fairness.

In Chapter 5 we show how one can utilize the above ideas in practice, specifically in the context of air traffic management. In particular, we analyze the problem of allocating landing and takeoff slots to airlines in a way that reduces delays compared to current practice. The underlying model we use to formulate the allocation borrows from the work of Bertsimas and Patterson [12]. Then, we demonstrate how our framework can be adopted as a natural extension. The usage of our framework highlights how one can think about and incorporate axiomatically justified notions of fairness under a complicated setting as is the air traffic management problem.

In the above setup of air traffic management, the efficiency metric of an allocation corresponds to the achieved delay reductions compared to current practice, whereas fairness corresponds to an equitable split of those reductions among the airlines. In the case study we discuss how our worst-case analysis can guide the balance of efficiency and fairness. Finally, we use historical data to study the performance of a particular parameterized fairness scheme, called α-fairness (see Chapter 3), under realistic problem instances. Particularly, we are interested in how the delay reductions are split among the airlines for different values of the parameter α (which is used to balance efficiency and fairness), and what the associated efficiency loss, or price of fairness is. Table 2.1 serves as a preview of our results. In particular, Table 2.1 includes the actual delays experienced by 4 airlines on a particular day in the past, together with the possible delay reductions in case one implemented the α-fairness scheme for $\alpha = 0$ and $\alpha = 1$. One can see that the delay reductions are more evenly split under the higher value of α, but that comes at a price of lower aggregate reductions among all airlines (efficiency loss). More details and a discussion are included in Chapter 5.

Table 2.1: Preview of results for Chapter 5. For each airline, we report the actual delay (in minutes) across its flights on that day (under current practice), and the delay reductions that different α-fair allocations would achieve.

		Actual delay	**Delay reduction**	
		(under current practice)	Utilitarian	Prop. fair
			($\alpha = 0$)	($\alpha = 1$)
05/13/05	Airline 1	12,722	0	420
	Airline 2	6,252	0	420
	Airline 3	13,613	990	540
	Airline 4	9,470	1,155	630
	Total	42,057	2,145	2,010

Chapter 3

Fairness Schemes

Fairness in allocation problems has been extensively studied through the years in many areas, notably in social sciences, welfare economics and engineering. A plethora of fairness criteria have been proposed. Due to multiple (subjective) interpretations of the concepts of fairness, and the different characteristics of allocation problems, there is no single principle that is universally accepted. Nevertheless, there are general theories of justice and equity that figure prominently in the literature, on which most fairness schemes are based. Moreover, there has been a body of literature that deals with axiomatic foundations of the concepts of fairness. In this chapter, we briefly review the most important theories and axioms, and then focus on proportional and max-min fairness, the two criteria that emerge from the axiomatic foundations, and α-fairness, a unifying parameterized family of schemes. For more details, see [77] and [58].

Among the most prominent, the oldest theory of justice is Aristotle's equity principle, according to which, resources should be allocated in proportion to some pre-existing claims, or rights to the resources that each player has. Another theory, widely considered in economics in the 19th century, is classical utilitarianism, which dictates an allocation of resources that maximizes the sum of utilities (see Chapter 2.1). A third approach is due to Rawls [50]. The key idea of Rawlsian justice is to give priority to the players that are the least well off, so as to guarantee the highest minimum utility level that every player derives. Finally, Nash introduced the Nash standard of comparison, which is the percentage change in a player's utility when he receives a small additional amount of the resources. A transfer of resources between two players is then justified, if the gainer's utility increases by a larger percentage than the loser's utility decreases.

Aristotle's equity principle is used in the majority of cases where players have specific pre-existing claims or rights to the resources (for example, split of profits among shareholders). In this work, we do not deal with such cases, hence the Aristotelian principle does not apply. The utilitarian principle has been criticized (see [77]) since it is not clear that it is ethically sound: in maximizing the sum of utilities, the utility of some players might be greatly reduced in order to confer a benefit to the system. Finally, the max-min and proportional fairness schemes that we will discuss are based on the Rawlsian justice and the Nash standard respectively, which are in line with the common perception of equity and fairness.

In addition to using theories of justice and common perception, researchers have also established sets of axioms that a fairness scheme should ideally satisfy. The main work in this area is within the literature of fair bargains in economics (see [77] and references therein). We now briefly present the most well studied set of axioms in the case of a two-player problem ($n = 2$). In the axioms that follow, we denote the utility set U and define the *maximum achievable utility* of the jth player, $u_j^\star$, according to

$$u_j^\star = \sup \{ u_j \,|\, u \in U \} .$$

Axiom 1. *(Pareto Optimality) The fair solution $\mathcal{S}(U)$ is Pareto optimal, that is, there does not exist an allocation $u \in U$, such that $u \geq$ [1]$\mathcal{S}(U)$ and $u \neq \mathcal{S}(U)$.*

Axiom 2. *(Symmetry) If $\mathcal{I} : \mathbf{R}^2 \to \mathbf{R}^2$ is the permutation operator defined by $\mathcal{I}\left((u_1, u_2)\right) = (u_2, u_1)$, then the fair allocation under the permuted system, $\mathcal{S}\left(\mathcal{I}(U)\right)$[2], is equal to the permutation of the fair allocation under the original system, $\mathcal{I}\left(\mathcal{S}(U)\right)$. That is, $\mathcal{S}\left(\mathcal{I}(U)\right) = \mathcal{I}\left(\mathcal{S}(U)\right)$.*

[1] The inequality sign notation for vectors is used for componentwise inequality.

[2] If $g : \mathbf{R}^n \to \mathbf{R}^n$ is an operator and $A \subset \mathbf{R}^n$ is a set, then $g(A) = \{g(x) \,|\, x \in A\} \subset \mathbf{R}^n$.

Axiom 3. *(Affine invariance) If $A : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ is an affine operator defined by $A(u_1, u_2) = (A_1(u_1), A_2(u_2))$, with $A_i(u) = c_i u + d_i$ and $c_i > 0$, then the fair allocation under the affinely transformed system, $\mathcal{S}\left(A(U)\right)$, is equal to the affine transformation of the fair allocation under the original system, $A\left(\mathcal{S}(U)\right)$. That is, $\mathcal{S}\left(A(U)\right) = A\left(\mathcal{S}(U)\right)$.*

Axiom 4. *(Independence of irrelevant alternatives) If U and W are two utility sets such that $U \subset W$, and $\mathcal{S}(W) \in U$, then $\mathcal{S}(U) = \mathcal{S}(W)$.*

Axiom 5. *(Monotonicity) Let U and W be two utility sets, under which the maximum achievable utility of player 1 is equal, i.e., $u_1^\star = w_1^\star$. If for every utility level that player 1 may demand, the maximum achievable utility that player 2 can derive simultaneously, is bigger or equal under W, then the utility level of player 2 under the fair allocation should also be bigger or equal under W, i.e., $\mathcal{S}(U)_2 \leq \mathcal{S}(W)_2$.*

Pareto optimality ensures that there is no wastage. By symmetry, the central decision maker does not differentiate the players by their names. The affine invariance requirement means that the scheme is invariant to a choice of numeraire. According to the independence of irrelevant alternatives, preferring option A over option B is independent of other available options. Finally, by monotonicity, if for every utility level that player 1 may demand, the maximum utility level that player 2 can simultaneously derive is increased, then the utility level assigned to player 2 under the fair scheme should also be increased. For a more detailed discussion about monotonicity, see [29].

The main result in this area is that, under mild assumptions on the utility set, there does not exist a scheme that satisfies all axioms; see [38] and [29] for more details. Moreover, the unique scheme that satisfies Axioms 1-4 is the *Nash solution* [38]; the unique scheme that satisfies Axioms 1-3, and 5 is the *Kalai-Smorodinsky solution* [29]. Proportional and max-min fairness are direct generalizations of those schemes, and are studied next.

3.1 Proportional Fairness

Proportional fairness (PF) is the generalization of the Nash solution for a two-player problem. The Nash solution is the unique scheme that satisfies Axioms 1-4, and is based on the Nash standard of comparison. Under the Nash standard, a transfer of resources between two players is favorable and fair if the percentage increase in the utility of one player is larger than the percentage decrease in utility of the other player. Proportional fairness is the generalized Nash solution for multiple players. In that setting, the fair allocation should be such that, if compared to any other feasible allocation of utilities, the aggregate proportional change is less than or equal to zero. In mathematical terms,

$$\sum_{j=1}^{n} \frac{u_j - \mathcal{S}^{\mathrm{PF}}(U)_j}{\mathcal{S}^{\mathrm{PF}}(U)_j} \leq 0, \quad \forall u \in U.$$

In case U is convex, the fair allocation under proportional fairness $\mathcal{S}^{\mathrm{PF}}(U)$ can be obtained as the (unique) optimal solution of the problem

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{j=1}^{n} \log u_j \\ \text{subject to} & u \in U, \end{array}$$

since the necessary and sufficient first order optimality condition for this problem is exactly the Nash standard of comparison principle for n players.

Proportional fairness has been extensively studied and used in the areas of telecommunications and networks, especially after the paper of Kelly et al. [31].

3.2 Max-min Fairness

Max-min fairness (MMF) is a generalization of the Rawlsian justice and the Kalai-Smorodinsky (KS) solution in the two-player problem. The KS solution is the unique solution that satisfies Axioms 1-3, and 5. In settings where the maximum achievable utility levels of the two players are equal, the KS solution corresponds to maximizing the minimum utility the players derive simultaneously. Otherwise, the central decision maker decides on the allocation in the same way, but by considering a scaled, normalized system, under which the players have equal maximum achievable utility levels. In other words, under the KS solution the players simultaneously derive the largest possible equal fraction of their respective maximum achievable utilities. For simplicity, for the rest of this section, we deal with normalized problems where the players have equal maximum achievable utilities.

In a setting that involves more than two players, such an allocation may not be Pareto optimal, thus indicating a waste of resources. That can happen for instance in case there exist players that can derive higher utility levels without affecting the others, and their allocated resources are not optimized. Max-min fairness generalizes the above criteria to account for this potential loss of efficiency, and always yields Pareto optimal allocations.

Under max-min fairness, the central decision maker tries at the first step to maximize the lowest utility level among all the players. After ensuring that all players derive (at least) that level, the second lowest utility level among the players is maximized, and so on. The resulting allocation yields a distribution of utility levels among the players that has the following property: the distribution of the utility levels of any other allocation that achieves a strictly higher utility for a specific level, is such that there exists a lower level of utility that has been strictly decreased. In other words, any other allocation can only benefit the rich at the expense of the poor (in terms of utility).

Intuitively, max-min fairness maximizes the minimum utility that all players derive. In situations where an efficient allocation exists that results in equal utility for all players, MMF converges to this equitable allocation. In cases where some players can achieve higher utility levels, without depriving others of the minimum utility performance, MMF equitably and efficiently allows them to increase their utility, in a similar fashion, by maximizing a new minimum utility level that all improving players derive.

In mathematical terms, let $T : \mathbf{R}^n \to \mathbf{R}^n$ be the sorting operator, that is

$$T(y) = \left(y_{(1)}, \ldots, y_{(n)}\right), \quad y_{(1)} \leq \ldots \leq y_{(n)},$$

where $y_{(i)}$ is the ith smallest element of y. The max-min fairness scheme corresponds to lexicographically[3] maximizing $T(u)$ over U, that is, finding an allocation $\mathcal{S}^{\text{MMF}}(U) \in U$ such that its resulting sorted utility distribution is lexicographically largest among all sorted feasible utility distributions. We then have

$$T(\mathcal{S}^{\text{MMF}}(U)) \succeq_{\text{lex}} T(u), \quad \forall u \in U.$$

The existence of a max-min fair allocation is guaranteed under mild conditions (*e.g.*, if U is compact), and the Pareto optimality of the allocation follows by its construction, see [49] for more details. Efficient algorithms for computing an MMF allocation have also been developed and studied in the literature. The computations involve a sequential optimization procedure, that identifies the corresponding utility levels at each step. For more details, see [42].

Max-min fairness was first implemented in networking and telecommunications applications and has also initiated a lot of research in this area (see [9], [15], [35]). It has many applications in bandwidth allocation, routing and load balancing problems, as well as in general resource allocation or multiobjective optimization problems.

3.3 α-fairness

We now present a unifying, parameterized family of fairness schemes, which subsumes proportional, max-min fairness and classical utilitarianism as special cases.

The *α-fairness* scheme was studied early on by Atkinson [5], building on notions of individual risk-aversion introduced by Pratt [48] and Arrow [3], and using these instead as notion of aversion to inequity (see also [36] and [7] for more details). Ac-

[3] $a = (a_1, \ldots, a_n) \succ_{\text{lex}} (b_1, \ldots, b_n) = b$ if $\exists\, k$: $a_i = b_i$, $\forall\, i < k$, and $a_k > b_k$. $a \succeq_{\text{lex}} b$ if $a \succ_{\text{lex}} b$ or $a = b$.

cording to α-fairness, the CDM decides on the allocation by maximizing the *constant elasticity social welfare function* W_α, parameterized by $\alpha \geq 0$, and defined for $u \in \mathbf{R}_+^n$ as

$$W_\alpha(u) = \begin{cases} \sum_{j=1}^{n} \frac{u_j^{1-\alpha}}{1-\alpha}, & \text{for } \alpha \geq 0,\ \alpha \neq 1, \\ \sum_{j=1}^{n} \log(u_j), & \text{for } \alpha = 1. \end{cases}$$

A resulting utility allocation, denoted by $z(\alpha)$, is such that

$$z(\alpha) \in \operatorname*{argmax}_{u \in U} W_\alpha(u), \tag{3.1}$$

and is referred to as an *α-fair allocation.*

Under the constant elasticity welfare function, the proportional increase in welfare attributed to a given player for a given proportional increase of her utility, is the same at all utility levels. Moreover, since the constant elasticity function is concave and component-wise increasing, it exhibits diminishing marginal welfare increase as utilities increase. The rate at which marginal increases diminish is controlled by the parameter α, which is called the *inequality aversion parameter* for that reason.

The α-fairness scheme can be useful in practice for a CDM, as it facilitates an understanding of the efficiency-fairness tradeoff. In particular, a higher value of the inequality aversion parameter is thought to correspond to a "fairer" scheme (see [67], [7], [33]). Note that for the smallest value of $\alpha = 0$, we recover the utilitarian principle, which is neutral towards inequalities. Thus, the CDM can adjust attitudes towards inequalities by means of a single parameter.

Furthermore, the α-fairness scheme captures as special cases the two important fair bargaining solutions we discussed above; for $\alpha = 1$, the scheme corresponds to proportional fairness, whereas for $\alpha \to \infty$, the α-fair allocation converges to the utility allocation suggested by max-min fairness (see [36]).

Although the α-fairness scheme has been studied both from a theoretical and a practical perspective, most prominently in networks ([37], [15]) and healthcare ([72]), the underlying efficiency-fairness tradeoff is still not well understood. Recent work by Lan et al. [33] has been devoted to theoretically characterizing what it actually means for a higher value of α to be more fair.

We conclude this chapter by remarking on the choice of the minimum utility as our fairness metric in this work (see Chapter 2.1). For the family of α-fair allocations, a natural measure of fairness could be the associated inequity aversion parameter α, or the constant elasticity welfare function (see [5] and [33]). Those measures however, are not easy to interpret. Other fairness metrics that have been well studied in the literature and are perharps easier to interpret include the minimum utility, the difference between the maximum and the minimum utility, the standard deviation or the coefficient of variation of the utilities, the Jain index, the Theil index, the mean log deviation of the utilities, the Gini coefficient, etc.

To guide the selection of a fairness metric, we further require that under it the max-min fair allocation (*i.e.*, the α-fair allocation for $\alpha \to \infty$) is optimal among all Pareto allocations for any utility set. Recall that under the premises of α-fairness, the max-min fair allocation is deemed as the “most fair” allocation (see discussion above). Thus, we require that the max-min fair allocation preserves this property under the selected fairness metric as well. To this end, the fairness metric we adopt in this work is the minimum utility as the only fairness metric from the ones discussed above that is easy to interpret and also satisfies the max-min fair allocation optimality requirement[4].

[4] Consider the utility set consisting of two points, $U = \{[0.5\,0.5\,0.8]^T, [0.45\,0.55\,0.72]^T\}$. The first point is the max-min fair allocation of U, however, under the fairness metrics of the difference between the maximum and the minimum utility, the standard deviation, the coefficient of variation of the utilities, the Jain index, the Theil index, the mean log deviation of the utilities and the Gini coefficient the second point is preferred.

Chapter 4

The Efficiency-Fairness Tradeoff

In this chapter, we present the main theoretical results of this work: upper bounds on the price of fairness and the price of efficiency for the fairness schemes we discussed in Chapter 3, namely proportional, max-min and α-fairness. As it will turn out, these bounds are applicable under mild assumptions and are essentially tight so that they will provide the tradeoff curves we seek.

Although there is some understanding on how the fairness properties of the proportional, max-min and α-fairness schemes behave (*e.g.*, with respect to varying α), there is no theoretical work focusing on the potential efficiency degradation (see Chapter 3). As such, the selection of a scheme in a practical setting can be very challenging. The results below pertaining to the price of fairness shed light towards this direction. Furthermore, the results pertaining to the price of efficiency enhance the understanding of the fairness properties of the shemes we study.

Consider a resource allocation problem, as described in Chapter 2. In the results below we assume that the utility set is compact and convex. This assumption is standard in the literature of fair bargains and also very frequently satisfied in practice. In particular, compactness of the utility set follows from limited resources and bounded and continuous functions that map resource allocations to utility for each player. Also, in case of nonconvex utility sets, randomization over possible utility allocations results in a convex set (of expected utilities). For more details, we refer the reader to [77].

4.1 The Price of Fairness

We first analyze the worst-case efficiency degradation for the proportional and max-min fairness schemes. Recall that the maximum achievable utility of the jth player is defined as

$$u_j^\star = \sup\left\{u_j \,|\, u \in U\right\}, \text{ for all } j = 1, \ldots, n.$$

Theorem 1. *Consider a resource allocation problem with n players, $n \geq 2$. Let the utility set, denoted by $U \subset \mathbf{R}_+^n$, be compact, convex and such that the players have equal maximum achievable utilities (greater than zero). Then,*

(a) the price of proportional fairness is bounded by

$$\mathrm{POF}(U; \mathcal{S}^{\mathrm{PF}}) \leq 1 - \frac{2\sqrt{n} - 1}{n},$$

(b) the price of max-min fairness is bounded by

$$\mathrm{POF}(U; \mathcal{S}^{\mathrm{MMF}}) \leq 1 - \frac{4n}{(n+1)^2}.$$

The bound under proportional fairness is tight if $\sqrt{n} \in \mathbf{N}$, and the bound under max-min fairness is tight for all n.

Proof. By assumption, the players have equal maximum achievable utilities. We assume further that they are equal to 1, *i.e.*,

$$u_j^\star = \max\left\{u_j \,|\, u \in U\right\} = 1, \quad \forall j = 1, \ldots, n. \tag{4.1}$$

This is without loss of generality, and can be achieved simply by scaling. As a result,

$$0 \leq u \leq \mathbf{1}, \quad \forall u \in U. \tag{4.2}$$

Without loss of generality, we assume that U is monotone[1]. This is because all schemes we consider, namely utilitarian, proportional and max-min fairness yield

[1] A set $A \subset \mathbf{R}_+^n$ is called monotone if $\{b \in \mathbf{R}^n | 0 \leq b \leq a\} \subset A$, $\forall a \in A$.

Pareto optimal allocations. In particular, suppose there exist allocations $a \in U$ and $b \notin U$, with allocation a dominating allocation b, *i.e.*, $0 \leq b \leq a$. Note that allocation b can thus not be Pareto optimal. Then, we can equivalently assume that $b \in U$, since b cannot be selected by any of the schemes.

Note that the monotonicity assumption and (4.1) also imply that $0 \in U$ and $e_j \in U$ for all $j = 1, \dots, n$. By the convexity assumption, we also have $\frac{1}{n}\mathbf{1} \in U$.

(a) *Proportional fairness.* Let $u^{\mathrm{PF}} \in U$ be the utility distribution under the proportionally fair solution. By definition, we have

$$\mathrm{FAIR}(U; \mathcal{S}^{\mathrm{PF}}) = \mathbf{1}^T \mathcal{S}^{\mathrm{PF}}(U) = \mathbf{1}^T u^{\mathrm{PF}}. \tag{4.3}$$

By the first order optimality condition (see Chapter 3.1), we have

$$\sum_{j=1}^{n} \frac{u_j - u_j^{\mathrm{PF}}}{u_j^{\mathrm{PF}}} \leq 0, \quad \forall u \in U.$$

Equivalently,

$$\left(\gamma^{\mathrm{PF}}\right)^T u \leq 1, \quad \forall u \in U, \tag{4.4}$$

where

$$\gamma_j^{\mathrm{PF}} = \frac{1}{n u_j^{\mathrm{PF}}}. \tag{4.5}$$

This defines a hyperplane that supports U at u^{PF}. Figure 4-1 illustrates u^{PF} and the hyperplane in the case of a two-dimensional example.

Since $u^{\mathrm{PF}} \in U$, using (4.2) we have that $u_j^{\mathrm{PF}} \leq 1 \Rightarrow \gamma_j^{\mathrm{PF}} \geq \frac{1}{n}$, for all j. Moreover, since $e_j \in U$ for all j, using (4.4) we have $\left(\gamma^{\mathrm{PF}}\right)^T e_j \leq 1 \Rightarrow \gamma_j^{\mathrm{PF}} \leq 1$. Without loss of generality, we also assume that the elements of γ^{PF} are ordered. To summarize, we have

$$\frac{1}{n} \leq \gamma_1^{\mathrm{PF}} \leq \ldots \leq \gamma_n^{\mathrm{PF}} \leq 1. \tag{4.6}$$

The supporting hyperplane we identified can now be used to bound the sum of

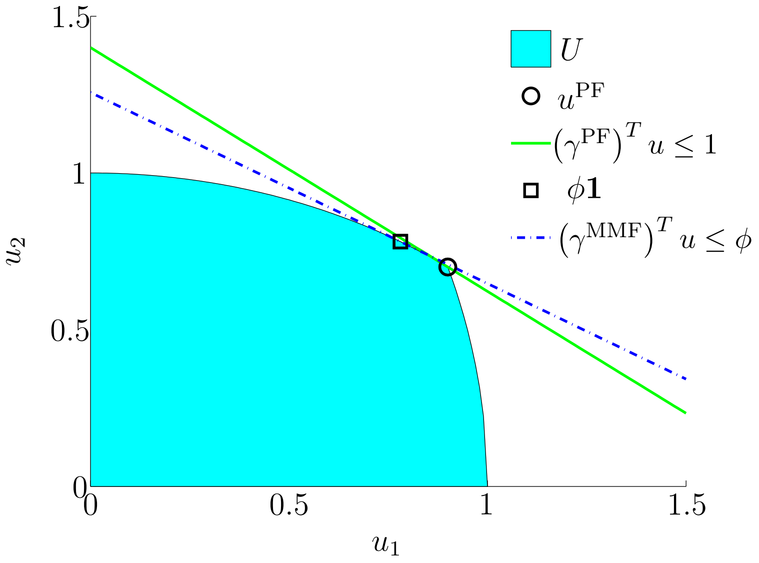


Figure 4-1: An example of a two-dimensional utility set, with the points of interest and the associated supporting hyperplanes used in the proof of Theorem 1, in Section 4.1.

utilities under the utilitarian solution. In particular, using (4.2) and (4.4) we get that

$$\begin{aligned} \mathrm{SYSTEM}(U) &= \max\left\{\mathbf{1}^T u \,\middle|\, u \in U\right\} \\ &\leq \max\left\{\mathbf{1}^T u \,\middle|\, 0 \leq u \leq \mathbf{1}, \left(\gamma^{\mathrm{PF}}\right)^T u \leq 1\right\}, \end{aligned} \tag{4.7}$$

where the right hand side is the optimal value of the linear relaxation of the well-studied knapsack problem, a version of which we review next.

Let $w \in \mathbf{R}^n$ and $B \in \mathbf{R}$ be such that $0 \leq w_1 \leq \ldots \leq w_n \leq B$, $\mathbf{1}^T w \geq 1$, $\frac{1}{n} \leq B \leq 1$. Then, one can show (see [13]) that the linear program

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T y \\ \text{subject to} & w^T y \leq B \\ & 0 \leq y \leq \mathbf{1}, \end{array} \tag{4.8}$$

has an optimal value equal to $\ell(w,B)+\delta(w,B)$, where

$$\ell(w,B) = \max\left\{ i \,\middle|\, \sum_{j=1}^{i} w_j \leq B,\; i \leq n-1 \right\} \in \{1,\ldots,n-1\} \tag{4.9}$$

$$\delta(w,B) = \frac{B - \sum_{j=1}^{\ell(w,B)} w_j}{w_{\ell(w,B)+1}} \in [0,1]. \tag{4.10}$$

Using this observation, we can rewrite (4.7) as

$$\text{SYSTEM}(U) \leq \ell(\gamma^{\text{PF}},1) + \delta(\gamma^{\text{PF}},1). \tag{4.11}$$

We can now provide an upper bound to the price of fairness:

$$\begin{aligned}
\text{POF}(U;\mathcal{S}^{\text{PF}}) &= \frac{\text{SYSTEM}(U) - \text{FAIR}(U;\mathcal{S}^{\text{PF}})}{\text{SYSTEM}(U)} \\
&= 1 - \frac{\text{FAIR}(U;\mathcal{S}^{\text{PF}})}{\text{SYSTEM}(U)} \\
&= 1 - \frac{\sum_{j=1}^{n} z_j^{\text{PF}}}{\text{SYSTEM}(U)} && \text{(from (4.3))} \\
&= 1 - \frac{\sum_{j=1}^{n} \frac{1}{n\gamma_j^{\text{PF}}}}{\text{SYSTEM}(U)} && \text{(from (4.5))} \\
&\leq 1 - \frac{\sum_{j=1}^{n} \frac{1}{n\gamma_j^{\text{PF}}}}{\ell(\gamma^{\text{PF}},1) + \delta(\gamma^{\text{PF}},1)}. && \text{(from (4.11))}
\end{aligned}$$

Let $g : \mathbf{R}^n \to \mathbf{R}$ be defined as

$$g(\gamma) = \frac{\sum_{j=1}^{n} \frac{1}{n\gamma_j}}{\ell(\gamma,1) + \delta(\gamma,1)}.$$

Using this definition and (4.6), the bound can now be rewritten as

$$\text{POF}(U;\mathcal{S}^{\text{PF}}) \leq 1 - g\left(\gamma^{\text{PF}}\right) \leq 1 - \inf_{\frac{1}{n} \leq \gamma_1 \leq \cdots \leq \gamma_n \leq 1} g(\gamma),$$

and it suffices to show that

$$F_1 = \inf_{\frac{1}{n} \leq \gamma_1 \leq \ldots \leq \gamma_n \leq 1} g(\gamma) \geq \frac{2\sqrt{n} - 1}{n}.$$

Let $p : \mathbf{R}^2 \to \mathbf{R}$ be defined as

$$p(y) = \frac{\frac{y_1}{y_2} + n - y_1}{n y_1},$$

and

$$F_2 = \inf_{\substack{y_1 y_2 \leq 1 \\ 1 \leq y_1 \leq n \\ \frac{1}{n} \leq y_2 \leq 1}} p(y).$$

We will first show that $F_1 \geq F_2$. To do that, it is sufficient to show that for any γ such that $\frac{1}{n} \leq \gamma_1 \leq \ldots \leq \gamma_n \leq 1$, there exists a $y \in \mathbf{R}^2$, such that $y_1 y_2 \leq 1$, $1 \leq y_1 \leq n$, $\frac{1}{n} \leq y_2 \leq 1$, and $g(\gamma) \geq p(y)$. Let $y_1 = \ell(\gamma, 1) + \delta(\gamma, 1)$. By the ranges of $\ell(\gamma, 1)$ and $\delta(\gamma, 1)$, it follows that $1 \leq y_1 \leq n$. Moreover, let

$$y_2 = \frac{y_1}{\frac{1}{\gamma_1} + \ldots + \frac{1}{\gamma_{\ell(\gamma,1)}} + \frac{\delta(\gamma,1)}{\gamma_{\ell(\gamma,1)+1}}}.$$

Since $\gamma_j \geq \frac{1}{n}$, we get

$$y_2 = \frac{y_1}{\frac{1}{\gamma_1} + \ldots + \frac{1}{\gamma_{\ell(\gamma,1)}} + \frac{\delta(\gamma,1)}{\gamma_{\ell(\gamma,1)+1}}} \geq \frac{y_1}{n(\ell(\gamma, 1) + \delta(\gamma, 1))} = \frac{1}{n}.$$

A similar argument utilizing that $\gamma_j \leq 1$ shows that $y_2 \leq 1$. To show that $y_1 y_2 \leq 1$, consider the following convex optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{v_1} + \ldots + \frac{1}{v_{\ell(\gamma,1)}} + \frac{\delta(\gamma,1)}{v_{\ell(\gamma,1)+1}} \\ \text{subject to} & v_1 + \ldots + v_{\ell(\gamma,1)} + \delta(\gamma, 1) v_{\ell(\gamma,1)+1} = 1 \\ & v \geq 0, \end{array}$$

with variable $v \in \mathbf{R}^{\ell(\gamma,1)+1}$. Note that γ is feasible for this problem, since by

(4.10) we have

$$\gamma_1 + \ldots + \gamma_{\ell(\gamma,1)} + \delta(\gamma,1)\gamma_{\ell(\gamma,1)+1} = 1.$$

We will show that

$$\bar{v} = \frac{1}{\ell(\gamma,1) + \delta(\gamma,1)}\mathbf{1}$$

is an optimal solution. Feasibility is immediate, and the necessary and sufficient first order optimaltiy conditions are also satisfied: Noting that $\bar{v}_1 = \bar{v}_j$ for all $j = 1, \ldots, \ell(\gamma,1)+1$, we have that for any $v \geq 0$, with $v_1 + \ldots + v_{\ell(\gamma,1)} + \delta(\gamma,1)v_{\ell(\gamma,1)+1} = 1$,

$$\begin{aligned}
&\sum_{j=1}^{\ell(\gamma,1)} \frac{(\bar{v}_j - v_j)}{\bar{v}_j^2} + \frac{\delta(\gamma,1)\left(\bar{v}_{\ell(\gamma,1)+1} - v_{\ell(\gamma,1)+1}\right)}{\bar{v}_{\ell(\gamma,1)+1}^2} = \\
&\frac{1}{\bar{v}_1^2}\left(\left(\bar{v}_1 + \ldots + \bar{v}_{\ell(\gamma,1)} + \delta(\gamma,1)\bar{v}_{\ell(\gamma,1)+1}\right) - \left(v_1 + \ldots + v_{\ell(\gamma,1)} + \delta(\gamma,1)v_{\ell(\gamma,1)+1}\right)\right) \\
&= 0.
\end{aligned}$$

Since γ is feasible and $\bar{v}$ optimal, it follows that

$$\begin{aligned}
\frac{y_1}{y_2} &= \frac{1}{\gamma_1} + \ldots + \frac{1}{\gamma_{\ell(\gamma,1)}} + \frac{\delta(\gamma,1)}{\gamma_{\ell(\gamma,1)+1}} \\
&\geq \frac{1}{\bar{v}_1} + \ldots + \frac{1}{\bar{v}_{\ell(\gamma,1)}} + \frac{\delta(\gamma,1)}{\bar{v}_{\ell(\gamma,1)+1}} \\
&= \frac{\ell(\gamma,1) + \delta(\gamma,1)}{\bar{v}_1} = \left(\ell(\gamma,1) + \delta(\gamma,1)\right)^2 = y_1^2.
\end{aligned}$$

Finally,

$$
\begin{aligned}
g(\gamma) &= \frac{\sum_{j=1}^{n} \frac{1}{n\gamma_j}}{\ell(\gamma,1)+\delta(\gamma,1)} \\
&= \frac{\frac{1}{\gamma_1}+\ldots+\frac{1}{\gamma_{\ell(\gamma,1)}}+\frac{\delta(\gamma,1)}{\gamma_{\ell(\gamma,1)+1}}+\frac{1-\delta(\gamma,1)}{\gamma_{\ell(\gamma,1)+1}}+\frac{1}{\gamma_{\ell(\gamma,1)+2}}+\ldots+\frac{1}{\gamma_n}}{n\left(\ell(\gamma,1)+\delta(\gamma,1)\right)} \\
&= \frac{\frac{y_1}{y_2}+\frac{1-\delta(\gamma,1)}{\gamma_{\ell(\gamma,1)+1}}+\frac{1}{\gamma_{\ell(\gamma,1)+2}}+\ldots+\frac{1}{\gamma_n}}{ny_1} \\
&\geq \frac{\frac{y_1}{y_2}+n-\ell(\gamma,1)-\delta(\gamma,1)}{ny_1} && \text{(from (4.6))} \\
&\geq \frac{\frac{y_1}{y_2}+n-y_1}{ny_1} = p(y).
\end{aligned}
$$

We now evaluate F_2:

$$
F_2 = \inf_{\substack{y_1y_2\leq 1\\ 1\leq y_1\leq n\\ \frac{1}{n}\leq y_2\leq 1}} \frac{\frac{y_1}{y_2}+n-y_1}{ny_1} = \inf_{\substack{y_1y_2\leq 1\\ 1\leq y_1\leq n\\ \frac{1}{n}\leq y_2\leq 1}} \left(\frac{1}{ny_2}+\frac{1}{y_1}-\frac{1}{n}\right).
$$

Clearly, the infimum is attained, and at the optimum $y_1y_2 = 1$, *i.e.*, $\frac{1}{y_2} = y_1$, and

$$
F_2 = \inf_{1\leq y_1\leq n} \left(\frac{y_1}{n}+\frac{1}{y_1}-\frac{1}{n}\right) = \frac{2\sqrt{n}-1}{n}.
$$

The proof is complete by noting that $F_1 \geq F_2$. Section 4.1.1 includes examples that show that the bound is tight in case $\sqrt{n} \in \mathbf{N}$.

(b) *Max-min fairness.* Consider the ray $r\mathbf{1}$, $r \geq 0$. Since $0 \in U$ and $\frac{1}{n}\mathbf{1} \in U$, by convexity of U we have that $r\mathbf{1} \in U$, for $0 \leq r \leq \frac{1}{n}$. Since $U \subset [0,1]^n$ is compact, there exists a $\phi \in \left[\frac{1}{n}, 1\right]$ such that $\phi\mathbf{1} \in \mathbf{bd}(U)^2$. Note that ϕ corresponds to the maximum minimum achievable utility level that all players can derive simultaneously. Under max-min fairness, the utility derived by all players is at least ϕ, as discussed in Chapter 3.2, that is,

$$
\mathcal{S}^{\text{MMF}}(U) \geq \phi\mathbf{1}. \tag{4.12}
$$

[2] The boundary of a set A is denoted by $\mathbf{bd}(A)$.

We can thus use ϕ to bound the sum of utilities under the max-min fair allocation,

$$\mathrm{FAIR}(U;\mathcal{S}^{\mathrm{MMF}}) = \mathbf{1}^T\mathcal{S}^{\mathrm{MMF}}(U) \geq \mathbf{1}^T\left(\phi\mathbf{1}\right) = n\phi. \tag{4.13}$$

Similarly to the derivation for proportional fairness, we will identify a hyperplane that supports U at $\phi\mathbf{1}$. In particular, since U is convex and $\phi\mathbf{1} \in \mathbf{bd}(U)$, by the supporting hyperplane theorem, $\exists\, \gamma^{\mathrm{MMF}} \in \mathbf{R}^n \setminus \{0\}$ such that

$$\left(\gamma^{\mathrm{MMF}}\right)^T u \leq \left(\gamma^{\mathrm{MMF}}\right)^T(\phi\mathbf{1}), \quad \forall u \in U. \tag{4.14}$$

Applying the above equation to $0 \in U$,

$$0 \in U \Rightarrow \left(\gamma^{\mathrm{MMF}}\right)^T 0 \leq \left(\gamma^{\mathrm{MMF}}\right)^T(\phi\mathbf{1}) \Rightarrow \mathbf{1}^T\gamma^{\mathrm{MMF}} \geq 0.$$

Suppose that $\mathbf{1}^T\gamma^{\mathrm{MMF}} = 0$. Combining this fact with (4.14) for every $e_j \in U$, we get

$$e_j \in U \Rightarrow \left(\gamma^{\mathrm{MMF}}\right)^T e_j \leq \left(\gamma^{\mathrm{MMF}}\right)^T(\phi\mathbf{1}) \Rightarrow \gamma_j^{\mathrm{MMF}} \leq 0.$$

Together with the assumption $\mathbf{1}^T\gamma^{\mathrm{MMF}} = 0$, that leads to $\gamma^{\mathrm{MMF}} = 0$, a contradiction. Hence, $\mathbf{1}^T\gamma^{\mathrm{MMF}} > 0$, and we can assume without loss that

$$\mathbf{1}^T\gamma^{\mathrm{MMF}} = 1.$$

The equation that defines the supporting hyperplane to U, (4.14), can now be rewritten as

$$\left(\gamma^{\mathrm{MMF}}\right)^T u \leq \phi, \quad \forall u \in U. \tag{4.15}$$

Figure 4-1 again illustrates the point $\phi\mathbf{1}$ and the supporting hyperplane in the case of a two-dimensional example.

We will now show that $\gamma^{\mathrm{MMF}} \geq 0$. Suppose that $\gamma_j^{\mathrm{MMF}} < 0$, and let $y = \phi\mathbf{1} - \frac{\phi}{2}e_j$.

Since $0 \leq y \leq \phi\mathbf{1}$, we have $y \in U$, by monotonicity of U. But,

$$\left(\gamma^{\text{MMF}}\right)^T y = \left(\gamma^{\text{MMF}}\right)^T \left(\phi\mathbf{1} - \frac{\phi}{2} e_j\right) = \phi - \frac{\phi}{2}\gamma_j^{\text{MMF}} > \phi,$$

a contradiction to (4.15), since $y \in U$. Hence, $\gamma^{\text{MMF}} \geq 0$.

Furthermore, since $e_j \in U$ for all j, using (4.15) we have

$$\left(\gamma^{\text{MMF}}\right)^T e_j \leq \phi \Rightarrow \gamma_j^{\text{MMF}} \leq \phi.$$

Without loss, we can assume similarly to the proportional fairness case, that the elements of γ^{MMF} are ordered. To summarize, if we let

$$C = \left\{ (y, B) \in \mathbf{R}^n \times \mathbf{R} \,\middle|\, 0 \leq y_1 \leq \ldots \leq y_n \leq B,\, \mathbf{1}^T y = 1,\, \frac{1}{n} \leq B \leq 1 \right\},$$

then $\left(\gamma^{\text{MMF}}, \phi\right) \in C$.

Similar to the analysis for the case of proportional fairness, using (4.2), (4.15) and the analysis of (4.8) we get

$$\begin{aligned} \text{SYSTEM}(U) &\leq \max\left\{ \mathbf{1}^T u \,\middle|\, 0 \leq u \leq \mathbf{1}, \left(\gamma^{\text{MMF}}\right)^T u \leq \phi \right\} \\ &= \ell(\gamma^{\text{MMF}}, \phi) + \delta(\gamma^{\text{MMF}}, \phi). \end{aligned} \tag{4.16}$$

It follows that

$$\begin{aligned} \text{POF}(U; \mathcal{S}^{\text{MMF}}) &= 1 - \frac{\text{FAIR}(U; \mathcal{S}^{\text{MMF}})}{\text{SYSTEM}(U)} \\ &\leq 1 - \frac{n\phi}{\text{SYSTEM}(U)} && \text{(from (4.13))} \\ &\leq 1 - \frac{n\phi}{\ell(\gamma^{\text{MMF}}, \phi) + \delta(\gamma^{\text{MMF}}, \phi)} && \text{(from (4.16))} \\ &\leq 1 - \inf_{(\gamma,\phi)\in C} \frac{n\phi}{\ell(\gamma, \phi) + \delta(\gamma, \phi)}. \end{aligned}$$

We will show that

$$\ell(\gamma, \phi) + \delta(\gamma, \phi) \leq n + 1 - \frac{1}{\phi}, \quad \forall\, (\gamma, \phi) \in C.$$

That will imply that for any such γ and ϕ,

$$\frac{n\phi}{\ell(\gamma, \phi) + \delta(\gamma, \phi)} \geq \frac{n\phi}{n + 1 - \frac{1}{\phi}} \geq \frac{4n}{(n+1)^2},$$

and the proof will be complete. Note that the last inequality follows by simply minimizing over $\phi \in \left[\frac{1}{n}, 1\right]$.

Fix any $(\gamma, \phi) \in C$. If $\ell(\gamma, \phi) + \delta(\gamma, \phi) < n$, let

$$y = \frac{(1 - \delta(\gamma, \phi))\gamma_{\ell(\gamma,\phi)+1} + \gamma_{\ell(\gamma,\phi)+2} + \ldots + \gamma_n}{n - \ell(\gamma, \phi) - \delta(\gamma, \phi)}.$$

Note that since $\gamma_j \leq \phi$, we get $y \leq \phi$. Then,

$$\begin{aligned}
1 &= \mathbf{1}^T\gamma \\
&= \gamma_1 + \ldots + \gamma_{\ell(\gamma,\phi)} + \delta(\gamma, \phi)\gamma_{\ell(\gamma,\phi)+1} + (1 - \delta(\gamma, \phi))\gamma_{\ell(\gamma,\phi)+1} + \ldots + \gamma_n \\
&= \phi + (1 - \delta(\gamma, \phi))\gamma_{\ell(\gamma,\phi)+1} + \gamma_{\ell(\gamma,\phi)+2} + \ldots + \gamma_n \\
&= \phi + (n - \ell(\gamma, \phi) - \delta(\gamma, \phi))y \\
&\leq \phi + (n - \ell(\gamma, \phi) - \delta(\gamma, \phi))\phi,
\end{aligned}$$

which demonstrates that $\ell(\gamma, \phi) + \delta(\gamma, \phi) \leq n + 1 - \frac{1}{\phi}$. If $\ell(\gamma, \phi) + \delta(\gamma, \phi) = n$, we get $1 = \mathbf{1}^T\gamma = \phi$, and hence $\ell(\gamma, \phi) + \delta(\gamma, \phi) = n = n + 1 - \frac{1}{\phi}$, and the proof is complete.

Section 4.1.1 includes examples that show that the bound is tight for all $n \geq 2$. □

We now discuss the case of α-fairness. Let POF $(U; \alpha)$ denote the associated price of fairness for $\alpha \geq 0$.

For $\alpha = 0$, the α-fairness scheme corresponds to the utilitarian principle, since W_0

is the sum of utilities, $W_0(u) = \mathbf{1}^T u$. Hence, for any compact utility set U, the sum of utilities is the same under both schemes, *i.e.*, $\text{SYSTEM}\,(U) = \text{FAIR}\,(U; 0)$, and

$$\text{POF}\,(U; 0) = 0.$$

For $\alpha > 0$, we have the following result.

Theorem 2. *Consider a resource allocation problem with n players, $n \geq 2$. Let the utility set, denoted by $U \subset \mathbf{R}^n_+$, be compact, convex and such that the players have equal maximum achievable utilities (greater than zero). For the α-fairness scheme, $\alpha > 0$, the price of fairness is bounded by*

$$\text{POF}\,(U; \alpha) \leq 1 - \min_{x \in [1,n]} \frac{x^{1+\frac{1}{\alpha}} + n - x}{x^{1+\frac{1}{\alpha}} + (n-x)x} = 1 - \Theta\left(n^{-\frac{\alpha}{\alpha+1}}\right).$$

Proof. Without loss of generality, we assume that U is monotone. This is because both schemes we consider, namely utilitarian and α-fairness yield Pareto optimal allocations. In particular, suppose there exist allocations $a \in U$ and $b \notin U$, with allocation a dominating allocation b, *i.e.*, $0 \leq b \leq a$. Note that allocation b can thus not be Pareto optimal. Then, we can equivalently assume that $b \in U$, since b cannot be selected by any of the schemes.

We also assume that the maximum achievable utilities of the players are equal to 1; the proof can be trivially modified otherwise.

By combining the above two assumptions, we get

$$e_j \in U, \quad \forall\, j = 1, \ldots, n, \tag{4.17}$$

where e_j is the unit vector in $\mathbf{R}^n$, with the jth component equal to 1.

Fix $\alpha > 0$ and let $z = z(\alpha) \in U$ be the unique allocation under the α-fairness criterion (since W_α is strictly concave for $\alpha > 0$), and assume, without loss of generality, that

$$z_1 \geq z_2 \geq \ldots \geq z_n. \tag{4.18}$$

The necessary first order condition for the optimality of z can be expressed as

$$\nabla W_\alpha(z)^T(u-z) \le 0 \Rightarrow \sum_{j=1}^{n} z_j^{-\alpha}(u_j - z_j) \le 0, \quad \forall u \in U,$$

or equivalently

$$\gamma^T u \le 1, \quad \forall u \in U, \tag{4.19}$$

where

$$\gamma_j = \frac{z_j^{-\alpha}}{\sum_i z_i^{1-\alpha}}, \quad j = 1, \ldots, n. \tag{4.20}$$

Note that (4.18) implies

$$\gamma_1 \le \gamma_2 \le \ldots \le \gamma_n. \tag{4.21}$$

Using (4.17) and (4.19) we also get

$$\gamma_j = \gamma^T e_j \le 1, \quad j = 1, \ldots, n. \tag{4.22}$$

We now use (4.19), and the fact that each player has a maximum achievable utility of 1 to bound the sum of utilities under the utilitarian principle as follows:

$$\begin{aligned} \text{SYSTEM}\,(U) &= \max\left\{\mathbf{1}^T u \,\middle|\, u \in U\right\} \\ &\le \max\left\{\mathbf{1}^T u \,\middle|\, 0 \le u \le \mathbf{1}, \gamma^T u \le 1\right\}. \end{aligned} \tag{4.23}$$

Using the above inequality,

$$\begin{aligned} \text{POF}\,(U;\alpha) &= \frac{\text{SYSTEM}\,(U) - \text{FAIR}\,(U;\alpha)}{\text{SYSTEM}\,(U)} \\ &= 1 - \frac{\text{FAIR}\,(U;\alpha)}{\text{SYSTEM}\,(U)} \\ &= 1 - \frac{\sum_{j=1}^{n} z_j}{\text{SYSTEM}\,(U)} \\ &\le 1 - \frac{\sum_{j=1}^{n} z_j}{\max\left\{\mathbf{1}^T u \,|\, 0 \le u \le \mathbf{1}, \gamma^T u \le 1\right\}}. \end{aligned} \tag{4.24}$$

The optimization problem in (4.24) is the linear relaxation of the well-studied knapsack problem, a version of which we review next. Let $w \in \mathbf{R}_+^n$ be such that $0 < w_1 \leq \ldots \leq w_n \leq 1$ (in particular, γ satisfies those conditions). Then, one can show (see [13]) that the linear optimization problem

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T y \\ \text{subject to} & w^T y \leq 1 \\ & 0 \leq y \leq \mathbf{1}, \end{array} \tag{4.25}$$

has an optimal value equal to $\ell(w) + \delta(w)$, where

$$\ell(w) = \max \left\{ i \,\middle|\, \sum_{j=1}^{i} w_j \leq 1,\, i \leq n-1 \right\} \in \{1, \ldots, n-1\} \tag{4.26}$$

$$\delta(w) = \frac{1 - \sum_{j=1}^{\ell(w)} w_j}{w_{\ell(w)+1}} \in [0, 1]. \tag{4.27}$$

We can apply the above result to compute the optimal value of the problem in (4.24),

$$\max \left\{ \mathbf{1}^T u \,\middle|\, 0 \leq u \leq \mathbf{1}, \gamma^T u \leq 1 \right\} = \ell(\gamma) + \delta(\gamma). \tag{4.28}$$

The bound from (4.24) can now be rewritten,

$$\mathrm{POF}\,(U; \alpha) \leq 1 - \frac{\sum_{j=1}^{n} z_j}{\ell(\gamma) + \delta(\gamma)}. \tag{4.29}$$

Consider the set S in the $(n+3)$-dimensional space, defined by the following constraints with variables $d \in \mathbf{R}$, $\lambda \in \mathbf{N}$ and $x_1, \ldots, x_\lambda, \overline{x}_{\lambda+1}, \underline{x}_{\lambda+1}, x_{\lambda+2}, \ldots, x_n \in \mathbf{R}$:

$$0 \leq d \leq 1 \tag{4.30a}$$

$$1 \leq \lambda \leq n-1 \tag{4.30b}$$

$$0 \leq x_n \leq \ldots \leq x_{\lambda+2} \leq \underline{x}_{\lambda+1} \leq \overline{x}_{\lambda+1} \leq x_\lambda \leq \ldots \leq x_1 \leq 1 \tag{4.30c}$$

$$x_n^{-\alpha} \leq x_1^{1-\alpha} + \ldots + x_\lambda^{1-\alpha} + d\,\overline{x}_{\lambda+1}^{1-\alpha} + (1-d)\,\underline{x}_{\lambda+1}^{1-\alpha} + x_{\lambda+2}^{1-\alpha} + \ldots + x_n^{1-\alpha} \tag{4.30d}$$

$$\begin{aligned} & x_1^{-\alpha} + \ldots + x_\lambda^{-\alpha} + d\,\overline{x}_{\lambda+1}^{-\alpha} \leq \\ & \qquad x_1^{1-\alpha} + \ldots + x_\lambda^{1-\alpha} + d\,\overline{x}_{\lambda+1}^{1-\alpha} + (1-d)\,\underline{x}_{\lambda+1}^{1-\alpha} + x_{\lambda+2}^{1-\alpha} + \ldots + x_n^{1-\alpha}. \end{aligned} \tag{4.30e}$$

We show that

$$\frac{\sum_{j=1}^{n} z_j}{\ell(\gamma)+\delta(\gamma)} \geq \min_{(d,\lambda,x)\in S} \frac{x_1+\ldots+x_\lambda + d\,\overline{x}_{\lambda+1} + (1-d)\,\underline{x}_{\lambda+1} + x_{\lambda+2}+\ldots+x_n}{\lambda+d}. \tag{4.31}$$

We pick values for d, λ and x that are such that (a) they are feasible for S, and (b) the function argument of the minimum, if evaluated at (d,λ,x), is equal to the left-hand side of (4.31). In particular, let

$$\begin{aligned} d &= \delta(\gamma), & \lambda &= \ell(\gamma), \\ x_j &= z_j, \quad j \neq \lambda+1, & \overline{x}_{\lambda+1} &= \underline{x}_{\lambda+1} = z_{\lambda+1}. \end{aligned}$$

Then, (4.30a), (4.30b) and (4.30c) are satisfied because of (4.27), (4.26) and (4.18) respectively. By the definition of γ and the selected value of x, (4.30d) can be equivalently expressed as

$$\gamma_n \leq 1,$$

which is implied by (4.22). Similarly, (4.30e) is equivalent to

$$\gamma_1+\ldots+\gamma_{\ell(\gamma)}+\delta(\gamma)\gamma_{\ell(\gamma)+1} \leq 1,$$

which again holds true (by (4.27)). The function argument of the minimum, evaluated at the selected point, is clearly equal to the left-hand side of (4.31). Finally, the minimum is attained by the Weierstrass Theorem, since the function argument is continuous, and S is compact. Note that (4.30d) in conjunction with (4.30c) bound x_n away from 0. In particular, if $\alpha \geq 1$, we get

$$x_n^{-\alpha} \leq x_1^{1-\alpha}+\ldots+x_n^{1-\alpha} \leq n x_n^{1-\alpha} \Rightarrow x_n \geq \frac{1}{n}.$$

Similarly, for $\alpha < 1$ we get

$$x_n \geq \left(\frac{1}{n}\right)^{\frac{1}{\alpha}}.$$

To evaluate the minimum in (4.31), one can assume without loss of generality that for a point $(d', \lambda', x') \in S$ that attains the minimum, we have

$$x'_1 = \ldots = x'_\lambda = \overline{x}'_{\lambda+1}, \quad \underline{x}'_{\lambda+1} = x'_{\lambda+2} = \ldots = x'_n. \tag{4.32}$$

Technical details are included in the Appendix, Section A.2. Using this observation, we can further simplify (4.31). In particular, consider the set $T \subset \mathbf{R}^3$, defined by the following constraints (with variables x_1, x_2 and y):

$$0 \le x_2 \le x_1 \le 1 \tag{4.33a}$$

$$1 \le y \le n \tag{4.33b}$$

$$x_2^{-\alpha} \le y x_1^{1-\alpha} + (n-y) x_2^{1-\alpha} \tag{4.33c}$$

$$y x_1^{-\alpha} \le y x_1^{1-\alpha} + (n-y) x_2^{1-\alpha}. \tag{4.33d}$$

We show that

$$\begin{aligned} &\min_{(d,\lambda,x)\in S} \frac{x_1 + \ldots + x_\lambda + d\,\overline{x}_{\lambda+1} + (1-d)\,\underline{x}_{\lambda+1} + x_{\lambda+2} + \ldots + x_n}{\lambda + d} \ge \\ &\ge \min_{(x_1,x_2,y)\in T} \frac{y x_1 + (n-y) x_2}{y}. \end{aligned} \tag{4.34}$$

Let $(d', \lambda', x') \in S$ be a point that attains the minimum of the left hand side above, satisfying (4.32). We construct a point $(x_1, x_2, y) \in T$, for which the objective of the minimum on the right hand side of (4.34) is equal to the minimum of the left hand side. Let $x_1 = x'_1$, $x_2 = x'_n$ and $y = \lambda' + d'$. Using (4.32) and the selected values for x_1, x_2 and y, we have that (4.30c) implies (4.33a) and that (4.30a - 4.30b) imply (4.33b). Similarly, (4.30d - 4.30e) imply (4.33c - 4.33d) respectively. To show that the minimum of the right hand side of (4.34) is attained, one can use a similar argument as in showing (4.31).

If we combine (4.29), (4.31), (4.34) we get

$$\mathrm{POF}\,(U; \alpha) \le 1 - \min_{(x_1,x_2,y)\in T} \frac{y x_1 + (n-y) x_2}{y}. \tag{4.35}$$

The final step is the evaluation of the minimum above. Let $(x_1^\star, x_2^\star, y^\star) \in T$ be a point that attains the minimum. Then, we have

$$y^\star < n, \quad x_2^\star < x_1^\star. \tag{4.36}$$

To see this, suppose that $x_2^\star = x_1^\star$. Then, the minimum is equal to $\frac{nx_1^\star}{y^\star}$. But, constraint (4.33d) yields that $nx_1^\star \geq y^\star$, in which case the minimum is greater than or equal to 1. Then, (4.35) yields that the price of fairness is always 0, a contradiction. If $y^\star = n$, (4.33d) suggests that $x_1^\star = 1$. Also, the minimum is equal to $x_1^\star = 1$, a contradiction.

We now show that (4.33c-4.33d) are active at $(x_1^\star, x_2^\star, y^\star)$. We argue for $\alpha \geq 1$ and $\alpha < 1$ separately.

$\alpha \geq 1$: Suppose that (4.33c) is inactive. Then, a small enough reduction in the value of $x_2^\star$ preserves feasibility (with respect to T), and also yields a strictly lower value for the minimum (since $y^\star < n$, by (4.36)), thus contradicting that the point attains the minimum. Similarly, if (4.33d) is inactive, a small enough reduction in the value of $x_1^\star$ leads to a contradiction.

$\alpha < 1$: Suppose that (4.33d) is inactive at $(x_1^\star, x_2^\star, y^\star)$. Then, we increase $y^\star$ by a small positive value, such that (4.33d) and (4.33b) are still satisfied. Constraint (4.33c) is then relaxed, since $(x_1^\star)^{1-\alpha} > (x_2^\star)^{1-\alpha}$. The minimum then has a strictly lower value, a contradiction. Hence, (4.33d) is active at any point that attains the minimum. If we solve for y and substitute back, the objective of the minimum becomes

$$x_1 + x_2^\alpha(x_1^{-\alpha} - x_1^{1-\alpha}), \tag{4.37}$$

and the constraints defining the set T simplify to

$$0 \leq x_2 \leq x_1 \leq 1 \tag{4.38a}$$

$$x_1^{-\alpha} - x_1^{1-\alpha} + x_2^{1-\alpha} \leq n x_1^{-\alpha} x_2. \tag{4.38b}$$

In particular, constraint (4.38b) correspond to constraint (4.33c). In case (4.33c) is not active at a minimum, so is (4.38b). But then, a small enough reduction in the value of $x_2^\star$ leads to a contradiction.

Since for any point that attains the minimum constraints (4.33c-4.33d) are active, we can use the corresponding equations to solve for x_1 and x_2. We get

$$x_1 = \frac{y^{\frac{1}{\alpha}}}{n - y + y^{\frac{1}{\alpha}}}, \tag{4.39}$$

$$x_2 = \frac{1}{n - y + y^{\frac{1}{\alpha}}}. \tag{4.40}$$

If we substitute back to (4.35), we get

$$\mathrm{POF}\,(U;\alpha) \leq 1 - \min_{x \in [1,n]} \frac{x^{1+\frac{1}{\alpha}} + n - x}{x^{1+\frac{1}{\alpha}} + (n-x)x}.$$

The asymptotic analysis is included in the Appendix, Section A.2. □

One can also show that the negative of the function that needs to be minimized in order to compute the exact bound in the Theorem above, is unimodal (see Appendix, Section A.2). As such, one can efficiently compute the unique minimizer and the associated minimum function value. Figure 4-2 depicts bounds on the price of fairness implied by Theorem 2, for different values of the inequality aversion parameter α, as functions of the number of players n. The graph illustrates the dependence of the bound on the number of players, for different values of α; in particular, the worst-case price is increasing with the number of players and the value of α.

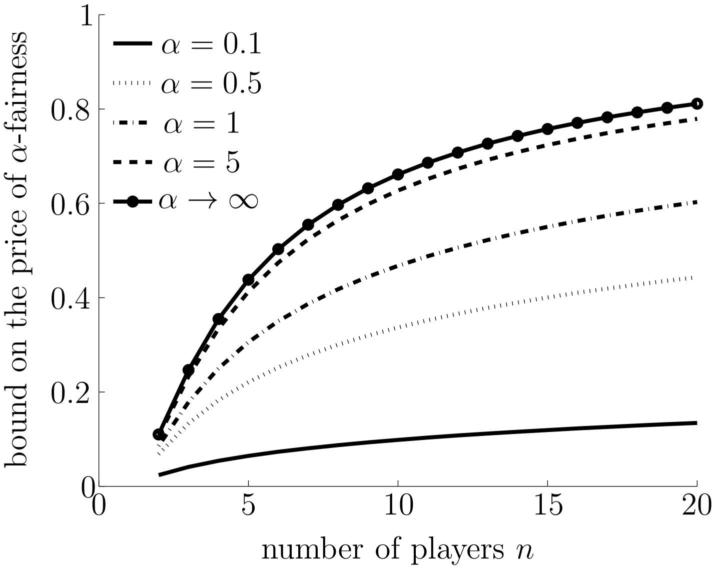
A natural question arising with regard to the results of Theorem 2 is whether the bounds are tight. The surprising fact is that the bounds are very strong, near-tight.

We next discuss (near) worst-case examples for the proportional, max-min and α-fairness schemes.

4.1.1 Worst-case Examples

We discuss the construction of (near) worst-case examples under which the price of fairness is equal or very close to the bounds implied by Theorems 1 and 2, for any values of the problem parameters; the number of players n and the value of the

Figure 4-2: Bounds on the price of α-fairness for different values of α implied by Theorem 2, in Section 4.1. The bounds are plotted as functions of the number of players n.

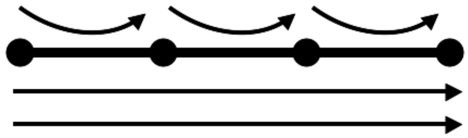


inequality aversion parameter α. To illustrate the fact that the examples are not pathological by any means, but rather have practical significance, we present them in a realistic setup under the context of network management. The setup is relevant to many other applications including traffic management and routing. After discussing the structure of the near worst-case examples, we compare their price of fairness with the established bounds and demonstrate that the bounds are essentially tight. Technical details of the construction of the examples are included in the Appendix, Section A.1.

Near worst-case bandwidth allocation

Consider a network consisting of hubs (nodes) that are connected via capacitated links (edges). Clients, or flows, wish to establish transmission from one hub to another over the network, via a pre-specified and fixed route. The network administrator needs to decide on the transmission rate assigned to each flow, subject to capacity constraints. The resources to be allocated in this case are the available bandwidth of the links, the players are the flows, and the central decision maker is the network administrator. The utility derived by each player is equal to his assigned transmission

Figure 4-3: The network flow topology in case of $n = 5$ and $y = 3$, for the bandwidth allocation example in Section 4.1.1.



rate.

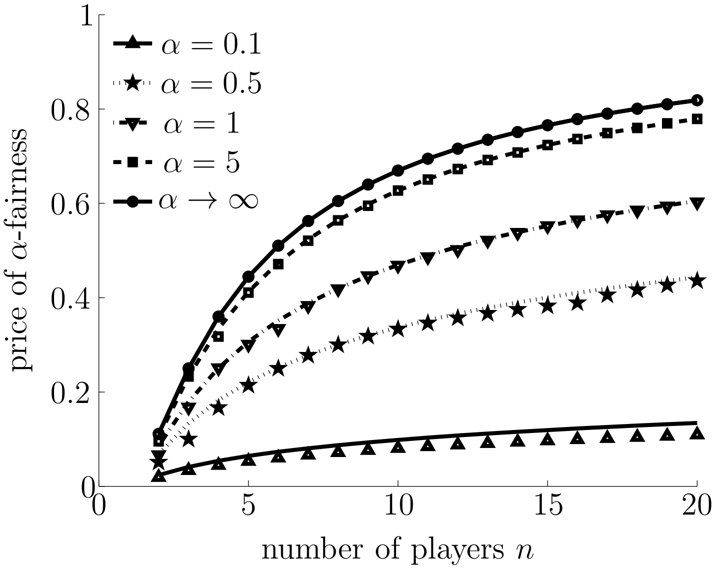
For the purposes of constructing near worst-case examples, we study a line-graph network, which is a specific network topology that has received a lot of attention in the literature and in practice (see [15], [67]). Specifically, suppose we have n players or flows. The network consists of y links of unit capacity, where the routes of the first y flows are disjoint and they all occupy a single (distinct) link. The remaining $n - y$ flows have routes that utilize all y links. The described network topology is shown in Figure 4-3, for $y = 3$ and $n = 5$. Each flow derives a utility equal to its assigned nonnegative rate. Note that in this setup, each player has a maximum achievable utility of 1, which is trivially achieved if all other flows are assigned zero rates. Thus, Theorems 1 and 2 apply.

Suppose we further fix a desired inequality aversion parameter $\alpha > 0$. In that case, one can select y (under some technical conditions) so that the price of α-fairness is exactly equal to bound implied by Theorem 2. Technical details about the selection of y are included in the Appendix, Section A.1. For $\alpha = 1$, the required condition is $y = \sqrt{n} \in \mathbf{N}$. For max-min fairness, and n odd, we select $y = \frac{n+1}{2}$. For n even, similar tight bounds can be obtained, by studying the utility set

$$W = \left\{ u \in \mathbf{R}_+^n \,\middle|\, \frac{1}{n}u_1 + \ldots + \frac{1}{n}u_{n/2} + u_{n/2+1} + \ldots + u_n \leq 1,\, u \leq \mathbf{1} \right\}.$$

Note that the described worst-case network topology pertains to a case of resources shared by n players, who can be of two types; players of the first type (short flows) consume resources at a lower rate, for a unit of utility, compared to players of the second type (long flows). This can be generalized as follows. Consider a knapsack-

Figure 4-4: The price of α-fairness for constructed examples (markers) in Section 4.1.1, for different values of α. The corresponding bounds are also plotted (lines). The values/bounds are plotted as functions of the number of players.



style problem where a unit of a single resource is shared by n players. Players $1, \ldots, \ell$ consume the resource at a rate of γ_1 for a unit of utility, whereas players $\ell + 1, \ldots, n$ consume the resource at a rate of γ_2. The described utility set is then

$$U = \left\{ u \in \mathbf{R}_+^n \,\middle|\, \gamma_1 u_1 + \ldots + \gamma_1 u_\ell + \gamma_2 u_{\ell+1} + \ldots + \gamma_2 u_n \le 1, \quad u \le \mathbf{1}\ \forall j \right\}.$$

In the Appendix, Section A.1, we present a simple algorithmic procedure of selecting parameters ℓ, γ_1 and γ_2 for a fixed number of players n and α, such that the price of α-fairness $\mathrm{POF}(U; \alpha)$ for the set U is very close to the price implied by Theorem 2. Figure 4-4 illustrates the prices achieved by following that procedure for various values of α and n. The average discrepancy between the bound and the values is 0.005, and the largest discrepancy is 0.023.

4.2 The Price of Efficiency

We now analyze the worst-case degradation of the minimum utility guarantee among all players. Note that for the max-min fairness scheme the degradation is always equal to zero, as the max-min fair allocation maximizes the minimum utility metric.

For the general α-fairness scheme we have the following result.

Theorem 3. *Consider a resource allocation problem with n players, $n \geq 2$. Let the utility set, denoted by $U \subset \mathbf{R}_{+}^{n}$, be compact, convex and such that the players have equal maximum achievable utilities (greater than zero). For the α-fairness scheme, $\alpha > 0$, the price of efficiency is bounded by*

$$\mathrm{POE}\,(U;\alpha) \leq 1 - \min_{x \in [\rho,1]} \frac{(n-1)x + x^{1-\alpha}}{n - 1 + x^{1-\alpha}} = 1 - \Theta\left(n^{-\frac{1}{\alpha}}\right),$$

where ρ is the unique root of $n - 1 + x^{-\alpha}(x-1) = 0$ in $(0,1)$.

Proof. We follow similar steps to the ones in the proof of Theorem 2. Thus, assume that U is monotone, the maximum achievable utilities of the players are equal to 1 and that $z_1 \geq z_2 \geq \ldots \geq z_n$ (where $z = z(\alpha) \in U$ is the unique α-fair allocation). Then, for the variable γ (defined as in (4.20)), we similarly have

$$\gamma^T u \leq 1, \quad \forall u \in U,$$

and

$$\gamma_1 \leq \gamma_2 \leq \ldots \gamma_n \leq 1.$$

We use the above to bound the maximum value of the fairness metric

$$\max\left\{ \min_{j=1,\ldots,n} u_j \,\middle|\, u \in U \right\} \leq \max\left\{ \min_{j=1,\ldots,n} u_j \,\middle|\, 0 \leq u \leq \mathbf{1},\, \gamma^T u \leq 1 \right\} = \frac{1}{\mathbf{1}^T \gamma},$$

where the equality follows from $z \leq \mathbf{1}$ and $\mathbf{1}^T \gamma \geq 1$.

We bound the price of efficiency using $z_1 \geq \ldots \geq z_n$, $\gamma_n \leq 1$ and the inequality above as follows:

$$\begin{aligned}
\operatorname{POE}(U;\alpha) &= \frac{\max\limits_{u\in U}\min\limits_{j=1,\ldots,n} u_j - \min\limits_{j=1,\ldots,n} z_j(\alpha)}{\max\limits_{u\in U}\min\limits_{j=1,\ldots,n} u_j} \\
&= 1 - \frac{z_n}{\max\limits_{u\in U}\min\limits_{j=1,\ldots,n} u_j} \\
&\leq 1 - z_n\mathbf{1}^T\gamma \\
&= 1 - \frac{z_n\left(z_1^{-\alpha} + z_2^{-\alpha} + \ldots + z_n^{-\alpha}\right)}{z_1^{1-\alpha} + z_2^{1-\alpha} + \ldots + z_n^{1-\alpha}} \\
&= 1 - f^\star,
\end{aligned}$$

where $f^\star$ is the optimal value of the problem

$$\begin{array}{ll}
\text{minimize} & \dfrac{z_n\left(z_1^{-\alpha} + z_2^{-\alpha} + \ldots + z_n^{-\alpha}\right)}{z_1^{1-\alpha} + z_2^{1-\alpha} + \ldots + z_n^{1-\alpha}} \\
\text{subject to} & 0 \leq z_n \leq z_{n-1} \leq \ldots \leq z_1 \leq 1 \\
& z_n^{-\alpha} \leq z_1^{1-\alpha} + z_2^{1-\alpha} + \ldots + z_n^{1-\alpha}.
\end{array} \tag{4.41}$$

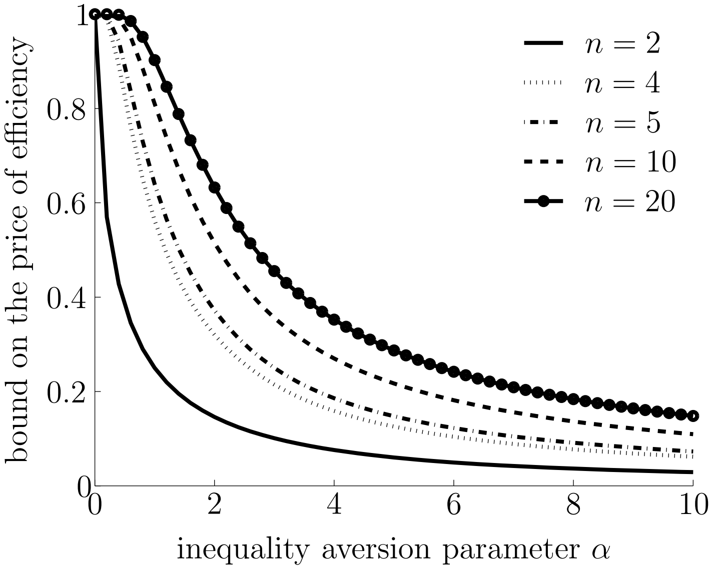
Let $z^\star$ be an optimal solution of (4.41) (guaranteed to exist by the Weierstrass Theorem). Then, without loss of generality we can assume that (a) $z_1^\star = z_2^\star = \ldots = z_{n-1}^\star$ and (b) $z_1^\star = 1$. Technical details are included in the Appendix, Section A.2. Using those two assumptions, $f^\star$ is then equal to

$$\begin{array}{ll}
\text{minimize} & \dfrac{(n-1)x + x^{1-\alpha}}{n-1+x^{1-\alpha}} \\
\text{subject to} & 0 \leq x \leq 1 \\
& x^{-\alpha} \leq n - 1 + x^{1-\alpha}.
\end{array} \tag{4.42}$$

Finally, note that for $x \in [0,1]$ the function $x^{-\alpha} - x^{1-\alpha} - n - 1$ is strictly decreasing, is positive for x small and negative for $x = 1$. Hence, for $x \in [0,1]$ the constraint $x^{-\alpha} \leq n - 1 + x^{1-\alpha}$ is equivalent to $x \geq \rho$. As a result,

$$f^\star = \min_{\rho \leq x \leq 1} \frac{(n-1)x + x^{1-\alpha}}{n-1+x^{1-\alpha}}.$$

Figure 4-5: Bounds on the price of efficiency of α-fair allocations for different values of the number of players n implied by Theorem 3, in Section 4.2. The bounds are plotted as functions of inequality aversion parameter α.



The asymptotic analysis is similar to the analysis in Theorem 2 and is omitted. □

Similarly to Theorem 2, one can show that the negative of the function that needs to be minimized in order to compute the exact bound in Theorem 3, is unimodal, and thus, the minimum function value can be efficiently computed. Figure 4-5 depicts bounds on the price of efficiency implied by Theorem 3, for different values of the number of players n, as functions of the inequality aversion parameter α. The graph illustrates the dependence of the bound on the inequality aversion parameter, for different values of n; in particular, the worst-case price is increasing with the number of players and decreasing with the value of α.

Finally, the bounds on the price of efficiency presented in Theorem 3 are tight.

4.2.1 Worst-case Examples

For any values of the problem parameters, *i.e.*, the number of players n and the value of the inequality aversion parameter α, one can construct worst-case examples under which the price of efficiency is equal to the bounds implied by Theorem 3.

The setup of the worst-case examples for the price of efficiency is identical with the setup discussed in Section 4.1.1 for the price of fairness. In particular, the worst-case topology pertains to a case of a single resource shared by n players, with $n-1$ of them consuming the resource at a rate of γ_1 for a unit of utility, whereas the nth player consumes the resource at a rate of γ_2, for

$$\gamma_1 = \frac{1}{n-1+\xi^{-\alpha}}, \quad \gamma_2 = \frac{\xi^{1-\alpha}}{n-1+\xi^{-\alpha}},$$

where ξ is the (unique) minimizer from Theorem 3. The proof is similar to the proof of Proposition 1 in the Appendix, Section A.1, and is omitted.

4.3 Discussion

We conclude by discussing other important facets of our results.

The results of Theorem 2 subsume the results of Theorem 1, since proportional and max-min fairness are both captured within the α-fairness framework, for $\alpha = 1$ and $\alpha \to \infty$ respectively. Technical details are included in the Appendix, Section A.2.

The results of Theorem 3 constitute, to the best of our knowledge, the first theoretical analysis of a fairness metric for a rich family of fairness schemes and a broad range of related resource allocation problems.

Before we summarize and discuss the importance of our results, note that in case players have unequal maximum achievable utilities, one can generalize our framework to deal with this case, albeit at the expense of additional technical effort. For instance, if under the same setup of Theorem 2, we also assume that the maximum achievable utilities of the players satisfy

$$L \leq \min_{j=1,\dots,n} u_j^\star \leq \max_{j=1,\dots,n} u_j^\star = B,$$

for some $0 < L \leq B$, we have that for $\alpha \geq 1$,

$$\mathrm{POF}\left(U; \alpha\right) \leq 1 - \min_{x \in [1,n]} \frac{\left(\frac{B}{L}\right)^{\frac{1}{\alpha}} x^{1+\frac{1}{\alpha}} + n - x}{\left(\frac{B}{L}\right)^{\frac{1}{\alpha}} x^{1+\frac{1}{\alpha}} + (n-x)\left(\frac{B}{L}\right) x}.$$

The case we focus on, however, (equal maximum achievable utilities) is particularly important, since utility levels of different players are commonly normalized, so as the intercomparison of utilities between them becomes meaningful, see [36], [27].

The theoretical development in this chapter, in conjunction with our formalization of the problem faced by a system manager designing an appropriate operational objective, provide valuable design tools. In particular:

- A system designer can prescribe a level of equity she wishes to inject into the system while accounting for the impact this will have on efficiency.

- She can measure the impact of her equity decision using only a small number of characteristic features of her specific problem; primarily, convexity and the number of distinct interests/ parties. That is to say, the designer's estimates of the price she pays for equity will be robust across a family of problems as opposed to being dependent on a specific instance.

- The designer can quickly recognize instances of her problem that are likely to be particularly costly in terms of the inefficiency introduced due to the requirement of equity (see Section 4.1.1), or conversely, costly in terms of the unfairness introduced due the requirement of efficiency (see Section 4.2.1).

Chapter 5

Fairness in Air Traffic Management

The tools we have introduced thus far provide a principled (as opposed to ad-hoc) approach to the design of appropriate operational objectives. This chapter is devoted to illustrating this value concretely in the context of the air traffic flow management problem. This problem presents the opportunity to save many billion dollars of unnecessary delay costs every year and is viewed as a key priority for the U.S. Federal Aviation Administration (FAA).

The Problem: Consider the problem faced by the FAA in allocating landing and takeoff slots to airlines, as well as routing them across U.S. airspace, in case of reduced capacity due to unpredictable inclement weather. By this allocation, the FAA is effectively allocating unavoidable delays across airlines, as allocations of unfavorable slots result in delayed flights. Currently, the FAA is allocating slots using a *Ration by Schedule* (RBS) principle, which prioritizes flights based on the original schedule, and is considered as fair. Proposals in the literature however, promise to reduce total delay by a significant amount (close to 10%), by using mathematical programming models to minimize total delay (see [12], [41]). Despite the rising delay costs (see [4]), none of these proposals have been implemented. One of the principal reasons for this is that those models do not address the question of whether the gains from optimization will be equitably split among the stakeholders. To this end, recent work deals with minimizing system delay in a fair way to all airlines (see [71], [6], [10]). The notion of what it means to be fair in these pieces of work is ad-hoc.

We now consider a principled approach to solving the above problem, based on the model and analysis presented in Chapters 2-4. The relative merits of such an approach, compared with the proposals in the literature, are the following:

- The notions of fairness we consider are eminently defensible.

- It will be possible to present a clear analysis of the tradeoff inherent in injecting "equity"; presumably this will provide a meaningful basis for the design of a suitable allocation mechanism.

Our framework will apply to this setting in the following way: The airlines correspond to the players, and the FAA to the central decision maker. Since the current policy debate centers around departures from the RBS policy, a natural choice for the utility of each airline is its delay reduction, compared to the RBS policy, which attempts to follow the original schedule in a first-come first-serve fashion. If for an airline a new allocation results in a delay reduction by x minutes, compared to the RBS policy, then that airline derives x units of utility[1]. With this definition in place, proposals that minimize total system delay, correspond to the utilitarian principle that maximizes the sum of utilities of the players. Accordingly, the FAA can incorporate fairness considerations by utilizing the α-fairness scheme; that is, the FAA carries out the allocation by maximizing the constant elasticity welfare function of the airlines' utilities. By the choice of α, one can then trade off efficiency for fairness.

Furthermore, in case the maximum achievable utilities (*i.e.*, delay reductions) of the airlines are equal, the bounds on the maximum relative efficiency loss established in Theorem 2 can be applied. Numerical studies indicated that indeed the maximum achievable utilities of similar-sized airlines are for all practical purposes equal (see Numerical Experiments below).

[1] As it turns out, there is also an agreed upon dollar figure associated with this delay. Also, note that we can easily extend the framework to account for differential costs for airborne and ground delay.

5.1 The Model

To characterize the utility set, we use a well accepted model introduced by Bertsimas and Patterson [12]. The model is highly detailed and specifies a schedule for each flight. In particular, the model specifies for each flight its scheduled location across the national airspace sectors or airports, for every time step. The model accounts for the forecasted capacity of each sector and airport, the maximum and nominal speed of the aircraft used for each flight, as well as potential connectivity of flights (through common usage of aircrafts or crew). A self-contained mathematical description of the model is included in the Appendix, Section B. We refer the reader to the original paper by Bertsimas and Patterson [12] for more details.

We model the utilities as follows. We have a set of flights, $\mathscr{F} = \{1, \ldots, F\}$, operated by a set of airlines, $\mathscr{A} = \{1, \ldots, A\}$ over a discrete time period. Let $\mathscr{F}_a \subset \mathscr{F}$ be the set of flights operated by airline $a \in \mathscr{A}$. The flights utilize a capacitated airspace that is divided into sectors, indexed by j. The decision variables used in the model are defined as

$$w_{ft}^{j} = \begin{cases} 1, & \text{if flight } f \text{ arrives at sector } j \text{ by time step } t, \\ 0, & \text{otherwise.} \end{cases}$$

We denote the scheduled departure and arrival time of flight f with d_f and r_f, and the origin and destination airports with o_f and k_f, respectively. Then, the associated ground and airborne delays experienced by flight f are

$$\begin{aligned} g_f &= \sum_t t(w_{ft}^{o_f} - w_{f,t-1}^{o_f}) - d_f, \\ b_f &= \sum_t t(w_{ft}^{k_f} - w_{f,t-1}^{k_f}) - r_f - g_f. \end{aligned}$$

The net delay experienced by flight f is $g_f + b_f$. The utility of the ath airline, that is the reduction of the cumulative delay of its flights compared to the RBS scheme, is then equal to

$$u_a = \sum_{f \in \mathscr{F}_a} \mathrm{RBS}_f - \sum_{f \in \mathscr{F}_a} (g_f + b_f) \,,$$

where RBS_f is the delay of flight f under the RBS scheme.

Accordingly, the utilitarian objective of minimizing total delay in the system, corresponds to minimize the objective

$$\sum_{a \in \mathscr{A}} \sum_{f \in \mathscr{F}_a} (g_f + b_f) \,.$$

Our framework provides a means to account for fairness in this fairly complicated setup. In particular, the framework focuses only on the utilities of the airlines, that is the important outcomes of the allocation.

5.2 Numerical Experiments

We focus on scheduling flights over a course of a day for 4 airlines (as many as the large airlines currently in the U.S.), which operate at 54 airports, administering in total around $4,000$ flights. We use historical data of scheduled and actual flight departure/arrival times, on different days to study the performance of α-fairness. In particular, we use the model described above to implement the solution that minimizes total delay, or equivalently in our setting, maximizes the total delay reduction, or sum of utilities (utilitarianism). We then implement the α-fairness scheme for different values of the parameter α.

We record the maximum possible system delay reduction (for $\alpha = 0$), and the system delay reduction under the α-fairness scheme, for various positive values of α, particularly 0.5, 1 (proportional fairness), 2. We also implement the max-min fairness scheme ($\alpha \to \infty$) and record the system delay reduction. In order to evaluate the fairness properties of the different schemes, beyond the interpretation based on the value of α, we also record the individual delay reductions of the airlines and their coefficient of variation. Intuitively, a high value for the coefficient indicates that delay reductions are unevenly split, whereas a lower coefficient suggests otherwise.

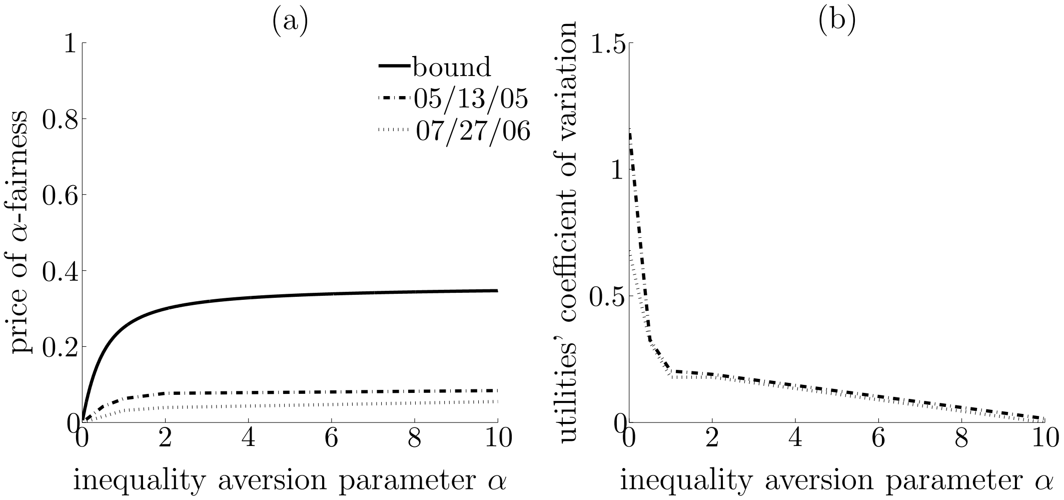
Table 5.1 summarizes the numerical results for 2 representative (actual) days, on which inclement weather severely affected operations across the United States. For each day and airline, the actual cumulative delay (in minutes) across its flights on that day is reported. We calculate the delay reductions that the utilitarian and the α-fairness schemes would achieve, for different values of α, including the special cases of proportional ($\alpha = 1$) and max-min fairness ($\alpha \to \infty$). The utilitarian scheme achieves roughly a 5% reduction compared to the current RBS policy, which is the largest possible for the schemes we consider. Note that this number is highly pessimistic, as we take the worst case scenarios in all calculations of the available capacities. The α-fair allocations yield lower delay reductions, but still the price is relatively small and increasing with α. Note also that the distribution of delay reductions changes rapidly as we are varying α. In particular, note that the utilitarian scheme does not equitably split the gains from optimization, since some airlines incur the same delay as in RBS, and others achieve large reductions. On the contrary, under max-min fairness, all airlines are granted almost the same delay reduction.

Table 5.1: Numerical results for the case study in Section 5.2, for 2 days and 4 airlines. For each airline, we report the actual delay (in minutes) across its flights on that day (under the RBS policy), and the delay reductions that different allocations would achieve.

		RBS delay	**Delay reduction**				
		(under RBS)	Util.	α-fair	PF	α-fair	MMF
			$(\alpha = 0)$	$(\alpha = 0.5)$	$(\alpha = 1)$	$(\alpha = 2)$	$(\alpha \to \infty)$
05/13/05	Airline 1	12,722	0	420	420	420	495
	Airline 2	6,252	0	435	420	420	495
	Airline 3	13,613	990	435	540	525	495
	Airline 4	9,470	1,155	765	630	615	480
	Total	42,057	2,145	2,055	2,010	1,980	1,965
07/27/06	Airline 1	11,099	195	390	390	390	450
	Airline 2	7,761	255	375	480	390	450
	Airline 3	8,511	555	420	405	495	450
	Airline 4	7,961	900	690	570	615	450
	Total	35,332	1,905	1,875	1,845	1,830	1,800

Figure 5-1(a) illustrates the price of α-fairness for the numerical experiments

Figure 5-1: (a) The price of α-fairness and (b) the coefficient of variation of the individual utilities for the numerical experiments in Section 5.2, for different values of the parameter α.



above, as a function of α. We also plot the worst-case bound implied by Theorem 2. Figure 5-1(b) depicts the coefficient of variation of the individual delay reductions for the numerical experiments, for different values of α. As expected, increase of the inequality aversion parameter α yields an increase in the efficiency loss and a decrease in the variation of the individual utilities.

Finally, to support our claim that the maximum achievable utilities of similar-sized airlines were for all practical purposes equal for our experiments, note that their coefficients of variation were 0.015 and 0.007 for day 05/13/05 and 07/27/06 respectively.

We conclude with a few takeaways from our case-study:

1. In retrospect, the framework of Chapter 2 provided a simple way of approaching fairness in what is a fairly complex setup.

2. The price of fairness, even in this model, apparently varies considerably from instance to instance (in this case, from day to day). This is apparent from Figure 5-1. The price of fairness tradeoff curve from Chapter 4 provides a convenient worst-case understanding. Even this worst-case price can be quite modest (for instance, consider the case of proportional fairness for $\alpha = 1$).

3. If one is aware of additional invariants in the decision problem, this information could be used to further constrain the description of the utility set considered in Chapter 2 and one could then hope to computationally compute a tradeoff curve as we did analytically for the case where U is simply required to be convex and compact.

Chapter 6

Fairness in Organ Allocation for Kidney Transplantation

In this chapter we deal with the question of how one designs implementable policies that account for efficiency and fairness in practice. We do so in the context of organ allocation for kidney transplanation. Section 6.1 provides an overview of kidney transplantation, our contributions and relevant work in the literature. Section 6.2 provides background information on the distribution of kidneys, the current allocation policy in the United States, as well as updates on the recent development of a new proposed policy. In Section 6.3, we discuss our method for designing allocation policies in detail. Section 6.4 includes numerical evidence of the usefulness of our work through the design of a new policy, the evaluation of its performance via simulation and a sensitivity analysis. A list of acronyms used appears at the end of this chapter.

6.1 Overview

Renal or kidney transplantation and maintenance dialysis are the only treatments for *end-stage renal disease* (ESRD), a terminal disease affecting over $500,000$ people currently in the United States, see [70]. Despite being a major surgical procedure, transplantation is the treatment of choice for ESRD patients, as a successful transplantation improves their quality of life. In particular, dialysis treatment requires that the patient visits a dialysis center for at least 12 hours each week, whereas transplantation typically allows the patient to resume regular life activities. Furthermore, a multitude of research and clinical studies have statistically demonstrated that transplantation also reduces the mortality risk for patients, see [64], [57], [47], [43]. Thus, a kidney transplant is considered by many as a potentially life-saving gift.

The two sources of kidneys for transplantation are living donors (*e.g.*, family members or friends of the patient) and deceased or cadaveric donors. The majority of patients are unsuccessful in finding living donors, and thus join a pool of patients waiting for a deceased donor organ. Of course, while in the living donor case the donation is typically made to a specific patient, in the deceased donor case an important allocation problem arises. In particular, once an organ is procured from a deceased donor, there can be thousands of medically compatible and available recipients the organ can be allocated to. The problem becomes even more significant, if one accounts for the organ shortage and the size of the pool of waiting patients in the United States: On October 20th 2010, $86,391$ patients were waiting for a kidney transplant. In 2009, there were $33,671$ new additions, but only $16,829$ transplantations were performed, from which $10,442$ transplants were from deceased donors. For more information and statistical details we refer the reader to [69].

In recognition of the aforementioned allocation problem and the growing difficulty of matching supply and demand, the U.S. Congress passed the *National Organ Transplant Act* (NOTA) in 1984. According to this legislation, deceased donor organs are viewed as national resources in the U.S., and as such, their allocation has to be based on fair and equitable policies. Moreover, the sale of organs as well as money transfers of any nature in the acquisition of organs are strictly prohibited. Instead, the policy for allocating the organs should utilize waiting lists and have the form of a *priority method*. That means that patients in need of a transplant register on waiting lists. Then, once an organ is procured, all medically compatible patients are ranked according to some priority rules and the organ is successively offered to them according to their ranking, until it is accepted by a patient. Subsequent to the NOTA, the U.S. Congress established in 1984 the *Organ Procurement and Transplantation Network*

(OPTN) in order for it to maintain a national registry for organ matching and develop allocation policies.

Naturally, the aforementioned allocation policies are of central importance and have to accomplish major objectives in alleviating human suffering, prolonging life and providing nondiscriminatory, fair and equal access to organs for all patients, independent of their race, age, blood group or other peculiar physiological characteristics. Some of the main challenges in designing a kidney allocation policy are the following:

- *Fairness constraints*: What does fair and equal access to organs mean? Due to the subjective nature of fairness, there is no single fairness criterion that is universally accepted by policymakers and academics alike. As such, a great challenge lies in identifying the appropriate fairness constraints that the allocation outcomes of a policy should ideally satisfy. An example of such a constraint could be a lower bound on the percentage of organs allocated to a particular group of patients – say, requiring that at least 47% of all transplants are received by recipients of blood type O. In the absence of such a constraint these groups would otherwise be handicapped and not have access to organs because of their physiological characteristics. A number of such criteria have been studied by OPTN policymakers (see [45], [52]).

- *Efficiency*: As a successful transplantation typically prolongs the life of a patient, while also improving his quality of life, the policy needs to ensure that the number of quality adjusted life year gains garnered by transplantation activities is as high as possible. This is also in line with the view of organs as national resources. Again, this objective is of paramount importance to the current policy design [45].

- *Prioritization criteria*: The policy needs to be based on medically justified criteria and physiological characteristics of patients and organs in order to rank patients. However, ethical rules disallow the use of criteria that can be deemed as discriminatory (*e.g.*, race, gender, etc.).

- *Simplicity*: Patients need to make important decisions about their treatment options, together with their physicians. To this end, they need to be able to estimate the probability of receiving an organ, or at least understand the allocation mechanism. For that reason, the priority method that is used needs to be simple and easy to communicate.

- *Implementation*: Suppose that one has selected his desired fairness constraints, prioritization criteria and a simple priority method. How does he then balance the emphasis put on the different prioritization criteria, so as to design a policy whose allocation outcomes would maximize efficiency, while satisfying the fairness constraints?

All the above challenges were faced by the OPTN policymakers in 2004, when they initiated the development of a new national allocation policy that will eventually replace the current one. In 2008, the OPTN released a concrete proposal in a *Request for Information* publication [52] that is currently under consideration by the U.S. Department of Health and Human Services.

In this work, we deal with the implementation challenge in designing a national allocation policy, while accounting for all the other challenges above. In particular, we focus on perhaps the simplest, most common and currently in use priority method, namely a point system. We make the following contributions:

1. We present a novel method for designing allocation policies based on point systems in a systematic, data-driven way. Our method offers the flexibility to the policymaker to select the fairness constraints he desires, as well as the prioritization criteria on which the point system will be based on. The method then outputs a conforming point system policy that approximately maximizes medical efficiency, while satisfying the fairness constraints.

2. We use our method to design a policy that (a) matches the fairness constraints of the recently proposed policy by U.S. policymakers, and (b) is based on the same criteria and simple scoring rule format. Critically though, it achieves

an 8% increase in anticipated extra life year gains, as demonstrated by our numerical simulations, which are based on the statistical and simulation tools currently in use by U.S. policymakers (see below).

3. We use our method to perform a sensitivity analysis that explores the consequences from relaxing or introducing fairness constraints – for instance, what is the impact of reducing the percentage of transplants to patients on dialysis for greater than 15 years by 1%? In the case of some constraints, relaxations of fairness constraints can result in life year gains on the order of 30%. As such, we believe this is a tool of great value in the policy design process.

Performance in all our numerical studies is evaluated using the same statistical and simulation tools, as well as data, as the U.S. policymakers use. Those tools and datasets were obtained directly from their developers, namely the *United Network for Organ Sharing* (UNOS), which is the non-profit organization that operates the OPTN, and the *Scientific Registry of Transplant Recipients* (SRTR).

6.1.1 Literature Review

The model-based analysis of the organ allocation process has attracted significant interest in the academic literature. One of the first papers in this vain is by Ruth et al. [55], in which the authors develop a simulation model to study the problem. Righter [53] and David and Yechiali [24] formulate the problem as a stochastic assignment problem and analyze stylized models that fit into that framework. Zenios et al. [79] introduce a fluid model approximation of the organ allocation process that allows them to explicitly account for fairness and medical efficiency in the allocation.

Another stream of research focuses on the decision-making behavior of patients, by dealing with organ acceptance policies. David and Yechiali [23] model the candidate's problem as an optimal stopping problem. Similar acceptance policies are developed by Ahn and Hornberger [1] and Howard [28]. The present work will test policies on a simulator developed by SRTR for OPTN; this simulator assumes a specific, exogenous acceptance model for patients built from historical data. While the acceptance model ignores endogeneity it allows us to simulate outcomes in precisely the manner policy makers currently do.

Recent work by Su and Zenios [63], [61] attempts to combine the above streams of research by explicitly accounting for the acceptance behavior of patients in the development of an allocation policy. In a similar vein, Su and Zenios [62] propose an allocation mechanism that elicits the utilities of the patients. For more details, we refer the reader to the thorough review by Zenios [78].

In all the above referenced work dealing with organ allocation policies, the authors design general near optimal dynamic policies. These papers take the important perspective of designing a *fundamentally new* allocation system from the ground up. In our work, we restrict our attention to policies that comply with the precise constraints imposed by current practice. That is, we focus our attention on policies based on simple point systems of the precise format as the ones currently in use and proposed by U.S. policymakers. Moreover, instead of designing a particular policy, we develop a framework that admits various fairness constraints and prioritization criteria. In other words, we design a mechanism that can fit directly in the *current* decision-making process of the U.S. policymakers.

6.2 Distribution and Allocation Policies

In this section, we briefly review the distribution process and the operation of the UNOS/OPTN as coordinators and developers of national policies for the allocation of deceased donor kidneys to patients. We then discuss the requirements such policies need to meet, and focus on policies that are based on point systems or scoring rules. Finally, we review the current policy in use in the U.S. (which itself is based on a scoring rule), as well as updates on the development of a new scoring rule based national policy.

In the U.S., the non-profit *Organ Procurement Organizations* (OPOs) are directly responsible for evaluating, procuring and allocating donated organs within their respective designated service area. Once consent is obtained and an organ is procured by an OPO, the OPTN computerized national registry automatically generates a list of patients who are medically compatible with the procured organ. Medical compatibility of patients is determined by the physiological characteristics they are listed with and those of the procured organ (*e.g.*, accounting for ABO incompatibility[1], weight and size, unacceptable antigens, etc.). Subsequently, the priority method used by the OPO determines the order in which the organ will be offered to patients. Once a kidney is procured, it can be typically preserved for up to 36-48 hours, after which the organ can no longer be used for transplantation. For that reason, priority is given to local patients, although there are rules that determine when priority should be given to non local patients. After an offer is being made to a patient, he has to decide with his surgeon whether to accept it or not within a limited amount of time. In case of rejection, the organ is offered to the next patient according to the specified order and so on. In case no patient accepts the organ within 36-48 hours, the organ is discarded.

In addition to using the OPTN national registry, the activities of the OPOs, and their allocation policies in particular, are coordinated and regulated by the OPTN. That is, the OPTN provides general guidelines and lays out a national allocation policy that is suggested to all OPOs. The allocation policy of every OPO then needs to be consistent with the national policy, although minor alterations are possible subject to approval by the OPTN.

6.2.1 National Allocation Policies

National policies for the allocation of the deceased donor kidneys are developed by the OPTN *Kidney Transplantation Committee* (KTC), and are approved by the U.S. Department of Health & Human Services. Policies need to account for numerous legal, economic, institutional, ethical, and other societal factors; the requirements for an allocation policy are included in the OPTN Final Rule [25]. Below we summarize

[1] ABO incompatibility is a reaction of the immune system that occurs if two different and not compatible blood types are mixed together, see `http://www.nlm.nih.gov/medlineplus/ency/article/001306.htm`.

the most important guidelines that policies have to conform to as per the OPTN Final Rule. In particular, the allocation

1. Shall seek to achieve the best use of donated organs, and avoid organ wastage;
2. Shall set priority rankings based on sound medical judgment;
3. Shall balance medical efficiency (extra life years) and equity (waiting time), without discriminating patients based on their race, age and blood type;
4. Shall be reviewed periodically and revised as appropriate.

Additionally, the priority method in place needs to be simple and easy to communicate, as discussed in the Introduction. As such, the ranking of patients is typically achieved by means of a *point system* or *scoring rule*: all national allocation policies that have been used in practice have been based on scoring rules. We formally define next the notion of a scoring rule based policy and then discuss the current national policy and suggested revisions.

Point system or Scoring rule based policies. Under a policy based on a scoring rule, patients are ranked according to a calculated score, commonly referred to in this context as the *Kidney Allocation Score* (KAS). Specifically, a scoring rule consists of *score components* and scalar constant *score weights.* A score component can be any function of the characteristics of a patient and/or an organ. Then, once an organ is procured and needs to be allocated, one calculates the individual score components for each patient and the particular procured organ. The KAS for each patient is evaluated as the weighted sum of his score components (using the score weights). To introduce some notation, given a patient p and an organ o, we denote the jth score component with $f_{j,(p,o)}$, and the jth score weight with w_j. The KAS of patient p for receiving organ o, $\text{KAS}(p,o)$, is then calculated as

$$\text{KAS}(p,o) = \sum_j w_j f_{j,(p,o)}.$$

For instance, examples of score components can be the number of years the patient has been registered on the waiting list for, the life expectancy of the patient in case he remained on dialysis, or the life expectancy in case he received the procured organ, etc.

One can think of a scoring rule based policy as a priority method that awards points to patients based on different criteria (the score components); patients are also potentially awarded different amounts of points per criterion, based on the score weights. The ranking is then achieved based on the number of points collected by each patient. The current policy in use and the one recently proposed by U.S. policymakers are both examples of scoring rule based policies and are discussed next.

Current allocation policy. The current policy has been in existence for more than 20 years. It is based on a scoring rule that utilizes waiting time, a measure of the patient's sensitization[2] and tissue matching[3] of the organ and the patient as score components. The rationale behind this rule is as follows. Points are given for waiting time and sensitization in order to serve the fairness objective of the allocation and to provide equal access to organs to all patients (note that highly sensitized patients have reduced medical compatibility with donors). On the other hand, since tissue matching is an indication for a successful transplantation, the points given to matched patients serve the medical efficiency objective of the allocation. For more details we refer the reader to [40].

Recent advances in medicine and changes in patients' needs have rendered the current policy inappropriate. More specifically, these changes have rendered the current policy inconsistent with the OPTN Final Rule, see [39] and [52]. For instance, the long waiting times experienced by the patients, coupled with advances in medicine that have prolonged the survivability of patients on dialysis, have resulted in the accumulation of points for waiting time by the patients. This accumulation of points has then created an imbalance between the efficiency and fairness objectives of the

[2] Potential recipients are "sensitized" if their immune system makes antibodies against potential donors. Sensitization usually occurs as a consequence of pregnancy, blood transfusions, or previous transplantation. Highly sensitized patients are more likely to reject an organ transplant than are unsensitized patients. For more information, see `http://www.ustransplant.org/`

[3] When two people share the same human leukocyte antigens (abbreviated as HLA), they are said to be a "match", that is, their tissues are immunologically compatible with each other. HLA are proteins that are located on the surface of the white blood cells and other tissues in the body. For more information, see `http://www.stanford.edu/dept/HPS/transplant/html/hla.html`

allocation, see [44]. In response to that, and in line with the requirement of the OPTN Final Rule for periodic review of the policy, the KTC has been reviewing the policy for the past few years and is currently in the process of developing a new policy, see [44].

Development of a new policy. The OPTN has set the primary objective of the new policy to be the design of a scoring rule that strikes the right balance between fairness and medical efficiency. As mentioned above, the design of a scoring rule involves the identification of the appropriate score components, and the corresponding score weights that form the Kidney Allocation Score. The KTC has selected to base the score components on the following criteria. For a patient p and an organ o, the criteria are

1. *Life years from transplant* $\mathrm{LYFT}(p, o)$, which is equal to the expected incremental quality-adjusted life years gain of patient p from receiving organ o, compared to remaining on dialysis (for a precise definition, we refer the reader to [75]);

2. *Dialysis time* $\mathrm{DT}(p)$, which is equal to the years the patient has already spent on dialysis;

3. *Donor profile index* $\mathrm{DPI}(o)$, which is a number between 0 and 1, indicating the quality of the donated organ (0 corresponds to an organ of highest quality);

4. *Calculated panel reactive antibody* $\mathrm{CPRA}(p)$, which is a number between 0 and 100, measuring the sensitization of the patient (0 corresponds to the lowest sensitization level).

This selection is currently being reviewed by the Office for Civil Rights of the U.S. Department of Health & Human Services for approval, see [69].

Meanwhile, the KTC considered more than 28 different scoring rules based on the above criteria, and utilized simulation to evaluate their performance and identify the appropriate weights (see [45]). The dominant proposal up to this point, published in 2008 in a Request For Information document ([52]), entails the following formula for the Kidney Allocation Score:

$$\begin{aligned}\mathrm{KAS}(p,o) = {} & 0.8\,\mathrm{LYFT}(p,o) \times (1-\mathrm{DPI}(o)) \\ & + 0.8\,\mathrm{DT}(p) \times \mathrm{DPI}(o) \\ & + 0.2\,\mathrm{DT}(p) \\ & + 0.04\,\mathrm{CPRA}(p).\end{aligned}$$

The rule is comprised of four components. The first two components are the life years from transplant and dialysis time, scaled by the donor profile index. The scaling ensures that in case of a high quality organ (DPI close to 0), emphasis is given on life years from transplant, whereas in case of a low quality organ (DPI close to 1), emphasis is given on dialysis time. The last two components are the dialysis time and calculated panel reactive antibody score of the patient. More information and motivating aspects can be found within the Request For Information document [52]. As an example, consider an organ o of medium quality, with $\mathrm{DPI}(o) = 0.55$. Then, patients are awarded $0.8 \times (1 - 0.55) = 0.36$ points for every quality adjusted incremental life year they would gain in expectation, $0.8 \times 0.55 + 0.2 = 0.64$ points for every year they have spent on dialysis, and 0.04 points for every point of their CPRA score.

While medical expertise and the OPTN Final Rule can guide the identification of the appropriate score components, the task of finding the right weights is more involved, as the experimentation of the OPTN KTC with more than 28 different rules suggests. A natural question in response to the proposed scoring rule is whether this is the best we can do. In particular, does there exist another scoring rule of the same simple format that dominates the proposed one, *i.e.*, is equally or more fair and more efficient? In all fairness, this is an involved question to answer; to illustrate that, consider only changing the weights in the proposed scoring rule above. The outcomes by such a change can perhaps be evaluated only via simulation; simulating a single specific scoring rule takes hours. Our proposed methodology makes a step towards answering those questions and is discussed next.

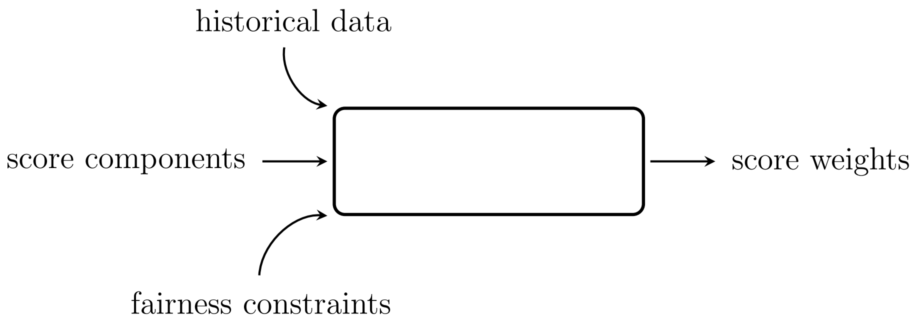


Figure 6-1: An illustration of the functionality of the proposed method for designing scoring rule based policies for the allocation of deceased donor kidneys to patients for transplantation in Section 6.3.

6.3 Designing Allocation Policies

We propose a method for designing scoring rule based policies for the allocation of deceased donor kidneys to patients. Specifically, we propose a data-driven method that computes in a systematic way score weights associated to pre-specified score components, so that the resulting policy achieves a near-optimal medical utility (measured by life years from transplant gains). In other words, after one has decided upon the components he wishes to include in a scoring rule, our method utilizes historical data to efficiently compute associated weights, so as to maximize the efficiency of the policy. In addition, our method can also take as input fairness constraints on the allocation outcomes; while we defer the precise definition of the class of admissible constraints for Section 6.3.1, we point here that our method captures a multitude of important and commonly studied constraints of interest to policymakers. Then, the method computes the score weights, so that the resulting policy is as efficient as possible, and the fairness constraints are approximately satisfied.

Figure 6-1 illustrates the functionality of the proposed method. Typically, policymakers select their desired score components that would feature in the scoring rule and constraints that the allocation outcomes need to satisfy. Our method provides an efficient, scalable and systematic way of striking the right balance between the selected score components by designing a policy that approximately maximizes medical efficiency, subject to the selected constraints.

As an application of our method, we use historical data from 2008, to construct a scoring rule based policy that utilizes the same criteria for components as the current proposal by the OPTN Kidney Transplantation Committee. We also ensure that the resulting policy has similar fairness characteristics with the KTC proposal. Numerical studies then suggest that the policy constructed by our method achieves an 8% improvement in life years from transplant, using the same statistical and simulation tools and data as U.S. policymakers use. Furthermore, we perform a tradefoff analysis by considering deviations from the fairness constraints of the proposed policy. In particular, we study the effect in life year gains of the policy, in case of emphasizing or deemphasizing the priority given to patients who have been waiting for a long time or are sensitized. Our method efficiently redesigns the policy accordingly. The results indicate that the performance gain in life years from transplant can be as high as 30% in that case. Details on the application of our method and simulation studies are included in Section 6.4.

We next present our proposal in full detail.

6.3.1 Methodology

Given a list of n score components, related historical data of patients' and donated organs' characteristics, and constraints on the allocation outcomes (precisely defined below), we calculate score weights $w_1, \ldots, w_n$, such that the resulting scoring rule policy satisfies the constraints approximately, while maximizing life years from transplant.

Consider a fixed time period over which we have complete (ex facto) information about all patients registered in the waitlist (pre-existing and arriving) in that time period. In particular, we know their physiological characteristics, the time of their initial registration, as well as the evolution of their medical status and availability for a transplant during that time period. Suppose we also have complete information about the organs that are procured during the period, that is the time at which they are procured and their physiological characteristics. We index the patients by $p = 1, \ldots, P$ and the organs by $o = 1, \ldots, O$. We say that patient p is *eligible* to receive organ o, or equivalently that the patient-organ pair (p, o) is eligible for transplantation, if at the time of the organ procurement all conditions below are met:

1. The patient is registered at the waitlist for a transplant;
2. The patient is actively waiting for a transplant and his medical status is appropriate for transplantation;
3. The patient is medically compatible with the organ.

Let $\mathcal{C}$ be the set of patient-organ pairs eligible for transplantation, *i.e.*,

$$\mathcal{C} = \{(p, o) : \text{ patient } p \text{ is eligible to receive } o\} .$$

Note that one can construct $\mathcal{C}$ simply by using the arrival information and characteristics of the organs and the patients, and the evolution of the availability and medical status of the patients.

Additionally, one can also compute the score components for each eligible patient-organ pair, as well as the life years from transplant. Let $f_{j,(p,o)}$ be the value of the jth component score, $j = 1, \ldots, n$, and $\text{LYFT}(p, o)$ the life years from transplant for pair $(p, o) \in \mathcal{C}$.

We now define the class of admissible constraints on the allocation outcomes, alluded to thus far. First, let $x_{(p,o)}$ be defined for every eligible patient-organ pair (p, o) as

$$x_{(p,o)} = \begin{cases} 1, & \text{if organ } o \text{ is assigned to patient } p, \\ 0, & \text{otherwise.} \end{cases}$$

A constraint is admissible for our method if it is *linear*, that is if it can be modeled as a linear constraint with respect to variable x. The class of constraints that can be modeled in this way is very broad, and captures the majority of constraints a social planner might wish to incorporate; for instance, one can impose lower bounds for a specific group of patients on

- the probability of receiving a transplant,

- the average life years from transplant gained among the actual transplant recipients,
- the average time spent on dialysis among the actual transplant recipients.

As an example, a lower bound L on the number of organs allocated to a specific group of patients $\mathcal{G} \subset \{1, \ldots, P\}$, can be expressed as

$$\sum_{p \in \mathcal{G}} \sum_{o:(p,o) \in \mathcal{C}} x_{(p,o)} \geq L;$$

for instance setting $\mathcal{G}$ to be the set of all patients of blood type O could enforce a lower bound on transplants for patients of this blood type.

We denote the input fairness constraints with $Ax \leq b$, for some matrix A and vector b.

We now present our method. Consider a social planner with foresight who has knowledge of the set of all eligible pairs $\mathcal{C}$ and the life years from transplant score for every pair in the set. Suppose also that patients accept all organs offered to them. In this setup, the problem of allocating organs to patients so as to maximize medical efficiency, *i.e.*, life years from transplant, subject to fairness constraints $Ax \leq b$, can be formulated as a linear optimization problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{(p,o) \in \mathcal{C}} \text{LYFT}(p,o) x_{(p,o)} \\ \text{subject to} \quad & \sum_{o:(p,o) \in \mathcal{C}} x_{(p,o)} \leq 1, \quad \forall p \\ & \sum_{p:(p,o) \in \mathcal{C}} x_{(p,o)} \leq 1, \quad \forall o \\ & Ax \leq b \\ & x \geq 0. \end{aligned} \tag{6.1}$$

Note that a fractional value for $x_{(p,o)}$ can be interpreted as the probability of assigning organ o to patient p in a randomized policy.

By linear optimization duality, if y is the vector of optimal dual multipliers associated with the constraints $Ax \leq b$ for problem (6.1), then problem (6.1) is equivalent with the one below:

$$
\begin{aligned}
\text{maximize} \quad & \sum_{(p,o)\in\mathcal{C}} \text{LYFT}(p,o)x_{(p,o)} - y^T A x + y^T b \\
\text{subject to} \quad & \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} \leq 1, \quad \forall p \\
& \sum_{p:(p,o)\in\mathcal{C}} x_{(p,o)} \leq 1, \quad \forall o \\
& x \geq 0.
\end{aligned}
\tag{6.2}
$$

Note that problem (6.2) is a matching problem. We equivalently rewrite the objective of (6.2) as $c^T x + y^T b$, utilizing the cost vector c defined as

$$c_{(p,o)} = \text{LYFT}(p,o) - \left(y^T A\right)_{(p,o)}, \quad \forall (p,o) \in \mathcal{C}.$$

Note that our goal is to design a policy that approximately solves the above matching problem online, *i.e.*, a policy that sequentially matches organs at their time of procurement to available patients without utilizing any future information. One possible way of achieving that is by greedily matching procured organs to patients based on the coefficients c. However, those coefficients are calculated above utilizing all information available. Moreover, our goal is to rank patients not by any artificial score coefficients, but rather based on the selected score components. To this end, one can calculate the appropriate score weights, such that the linear combination of the score components based on them is as close as possible to the coefficients c. Specifically, the score weights $w_1, \ldots, w_n$ are found by running a standard linear regression, with the values of the score components for each eligible patient-organ pair as independent variables, and the coefficients of c as dependent variables. That is, we compute the weights such that for every eligible patient-organ pair,

$$c_{(p,o)} \approx w_0 + w_1 f_{1,(p,o)} + \ldots + w_n f_{n,(p,o)}.$$

The method is summarized as Procedure 1.

Procedure 1 Computation of score weights

```
Input: list of $n$ score components, data for linear constraints $(A, b)$, historical data:
    set of eligible patient-organ pairs $\mathcal{C}$, life years for transplant $\mathrm{LYFT}(p,o)$ and values
    of score components, $f_{j,(p,o)}$, $j = 1, \ldots, n$, for every eligible pair $(p,o)$.
Output: weights for scoring rule, $w_1, \ldots, w_n$.
1: solve problem (6.1)
2: $y \leftarrow$ vector of optimal dual multipliers associated with constraints $Ax \leq b$
3: $c_{(p,o)} \leftarrow c_{(p,o)} = \mathrm{LYFT}(p,o) - \left(y^T A\right)_{(p,o)}, \quad \forall (p,o) \in \mathcal{C}$
4: use linear regression to find $w_0, w_1, \ldots, w_n$, such that for all $(p,o) \in \mathcal{C}$
       $$c_{(p,o)} \approx w_0 + w_1 f_{1,(p,o)} + \ldots + w_n f_{n,(p,o)}.$$
```

6.3.2 Discussion

In this section, we discuss (a) why one should expect the proposed method to perform well in practice, and (b) the relative merits of our contribution.

Consider the airline network revenue management setting analyzed in [66]. In that setting, an airline is operating flights and is selling different itinerary tickets to incoming customers, so as to maximize net expected profits from sales subject to capacity constraints (which correspond to the numbers of seats on the different aircrafts operating the flights). The authors analyze a simple control policy that decides whether to sell an itinerary ticket to a passenger or not, and demonstrate that the policy is asymptotically optimal under some conditions. For the organ allocation problem, a simplified version of the policy that we described in the previous section can be cast in the same framework as in [66]; one can then derive a similar result of asymptotic optimality, following the same procedure. In particular, in [66], the authors analyze the performance of the following simple bid-price control policy: one first solves a capacity allocation problem assuming that demand is deterministic and equal to the mean demand. Based on the optimal dual multipliers associated with the resource capacity constraints in that problem, one then calculates a "bid price" for every unit of a particular resource. An itinerary ticket is then sold to a customer if the money offered by the customer exceed the sum of prices of the resources he would consume. In our procedure, if we ignore the regression step, we also assume deterministic demand and solve a similar allocation problem[4]. We then calculate "bid prices" y associated to the fairness constrains and assign the organ to the patient who achieves the highest profit (LYFT), adjusted for the "bid prices". For more details, we refer the reader to [66].

Apart from the above discussion regarding the performance of our method in practice, we next provide numerical evidence. Before that, we summarize the relative merits of our contribution.

1. The proposed method uses detailed historical medical data to extract near optimal score weights in an efficient manner. In particular, the method is highly scalable and can learn the parameters from potentially highly detailed and complicated historical datasets, with no need for simplifications, clustering or grouping of patients' and/or organs' characteristics.

2. The method offers the flexibility and allows policymakers to focus only on identifying score components and desired fairness properties of allocation outcomes in the design of a new policy. The method undertakes the more involved part of finding the appropriate weights and balancing the score components. Although medical intuition can help in making educated guesses for the weights, there is little guarantee that a policy designed in such way would yield the desired results. Furthermore, even if a set of weights yields a policy with the desired outcomes, there can be another policy delivering a superior performance. Due to the computational intensity of simulations, one simply cannot explore all possible combinations of weights. Our contribution is towards this direction, by using mathematical tools to automatically extract near-optimal weights from historical data.

3. In this work, we develop our method in the context of kidney allocation. However, the same procedure can be generalized in a straightforward manner for

[4] Specifically, consider the deterministic linear optimization model analyzed in [66], where the different customer classes correspond to patient classes, the profits correspond to life years from transplant and the network capacity constraints correspond to the fairness constraints. If we instead use historical samples rather than averages, we recover formulation (6.1).

other organs as well. Thus, our methodology is particularly useful for any organ allocation policy one wishes to design based on scoring rules.

4. The failure of the current kidney allocation policy in place to keep up with advances in medicine and the changes in patients' needs throughout the years, has demonstrated that in such a dynamic and complex environment, revisions to policies are likely to be required in the future as well, a fact that is also recognized by the OPTN Final Rule. Furthermore, even in the current process of developing a new policy, there is no guarantee that the Office of Civil Rights will approve the criteria of life years from transplant, dialysis time, etc., suggested by the OPTN policymakers. In both cases, our method will expedite the development of a new policy, as it would require only an updated list of score components and fairness properties to be specified.

5. Our method allows for sensitivity analysis; specifically, one can efficiently evaluate the outcomes of relaxing some or introducing new fairness constraints. In the next section, we provide such an analysis that reveals the dependence of medical efficiency on fairness concepts, and illustrate how it can be used in practice by policymakers. In particular, note that one of the main goals that the OPTN policymakers have set for a new national policy is to deemphasize the role of waiting time and increase medical efficiency (see Section 6.2.1). Our analysis provides a characterization of the tradeoffs involved.

In the next section, we provide numerical evidence of the usefulness of the described method. In particular, we use historical data to create a new scoring policy that performs better than the one proposed by the Kidney Transplantation Committee and also explore other options by means of a sensitivity analysis.

6.4 Numerical Evidence and the Design of a New Allocation Policy

We utilize the method described in the previous section to design a new scoring-rule based allocation policy. We set as benchmark the dominant proposal of the OPTN Kidney Transplantation Committee (referred to also as the KTC policy in this section), presented in the Request for Information document in 2008 (see Section 6.2). We design a policy that uses the same criteria as score components and achieves an 8% increase in life years from transplant, while exhibiting similar fairness properties. Finally, we perform a tradeoff analysis by considering deviations from the fairness properties of the proposed policy. To ensure a fair comparison, we evaluate the performance of the policies by using the same statistical models and tools, as well as datasets with the OPTN KTC policymakers. We first provide details about the data and models, and then present our methodology and results.

6.4.1 Data, Statistical Models and Tools

This work uses highly detailed historical data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donor, wait-listed candidates, and transplant recipients in the U.S., submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The datasets include all the various physiological and demographic characteristics of wait-listed patients and donors that are needed for our study, as well as the evolution of the medical status of the patients, and the arrival process of the donated organs.

In addition, the SRTR has developed sophisticated survivability models for ESRD patients using historical survival rates. The models provide an estimate for the anticipated lifespan of a patient in case he remained on dialysis, or in case he received a particular kidney, based on a plethora of physiological attributes (*e.g.*, the patient's age, body mass index, diagnosis, as well as tissue matching, the donor's age, cause of death, etc.). For more information and a detailed study of the statistical performance of the models, we refer the reader to [75] and [74]. The SRTR has also developed an acceptance model that predicts the probability of a particular patient accepting a particular organ offered to him, based on the physiological characteristics of the patient and the donor, the distance, etc.

The above datasets and statistical models have also been utilized by the SRTR in the development of the *Kidney-Pancreas Simulated Allocation Model* (KPSAM). The KPSAM is an event-driven simulator that simulates the entire allocation process using historical data, for different allocation policies. It was developed in order to support studies of alternative policies. The KPSAM is the platform that the OPTN KTC is utilizing to evaluate the performance of their proposed policies, see [44]. For more details on the data and the simulator, we refer the reader to [73] and [32].

For the purposes of this study, we obtained the KPSAM and utilized its simulation engine in order to obtain realistic allocation outcomes of the policies we consider. The life years from transplant gains are estimated using the aforementioned survivability models, embedded in the KPSAM.

6.4.2 Methodology

Using the KPSAM we simulate the KTC policy for the 2008 dataset. We record the number of transplantations occurring and the net life years from transplant. To explore the fairness properties of the policy, we record the percentage distribution of transplant recipients across different races, age groups, blood types, sensitization groups, as well as diagnosis types, years spent on dialysis and geographical regions. Note that this practice is in line with the comparison criteria studied by the OPTN policymakers (see [45], [52]).

To design a new policy based on our method described in Section 6.3, we use the following as input:

- Historical data: We use the first 6 months of data of the 2008 dataset as input to our method (training data). The data pertaining to the remaining 6 months is used to evaluate performance.

- Score components: We use the life years from transplant (LYFT), dialysis time (DT) [5] and calculated panel reactive antibody (CPRA) as the score components. Note that the components are based on exactly the same criteria as in the KTC policy [6]. In summary, the scoring rule rewards the patient with w_1 points per life year from transplant gained, w_2 per year on dialysis for the first 5 years, w_3 per year on dialysis for years 5-10 and w_4 per year beyond the 10th, and w_5 points per percentage point of CPRA.

- Fairness constraints: To ensure that the fairness properties are similar to the KTC policy, we use the recorded percentage distributions (see above) for the KTC policy as input constraints. We use lower bound constraints on the percentage of organs allocated to the following groups: Caucasian, African-American, Hispanic or patients of another race; patients aged between 18-34, 34-49, 49-64 and above 64 years; patients who have spent less than 5, 5-10, 10-15 or more than 15 years on dialysis; blood type O, A, B, AB patients; patients diagnosed with nephritis, hypertension, polycystic kidney disease, diabetes or other disease; patients with a sensitization level (CPRA) of 0-10, 10-80 or 80-100; patients registered at each of the 11 distinct geographical regions[7] in the U.S. in which UNOS operates. For instance, consider the fairness constraints pertaining to dialysis time. The recorded percentage distribution of recipients for the KTC policy is as follows: 55.4% of the recipients have spent less than 5 years on dialysis, 28.8% between 5-10 years, 10.2% between 10-15 years and 5.6% more than 15 years. The constraints we add then as input to our method

[5] Our scoring rule is piece-wise linear in this component.

[6] The Donor Profile Index (DPI) component proves superfluous.

[7] For its own operational purposes, UNOS has divided the U.S. into 11 distinct geographical regions. Region 1 for instance includes all the states in New England. For more information, see `http://www.unos.org/docs/Article_IX.pdf`.

are:

$$
\begin{aligned}
\sum_{p:\, 0 \le \mathrm{DT}(p) \le 5} \ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\ge \frac{55.4}{100} \sum_{(p,o)\in\mathcal{C}} x_{(p,o)},\\
\sum_{p:\, 5 \le \mathrm{DT}(p) \le 10} \ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\ge \frac{28.8}{100} \sum_{(p,o)\in\mathcal{C}} x_{(p,o)},\\
\sum_{p:\, 10 \le \mathrm{DT}(p) \le 15} \ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\ge \frac{10.2}{100} \sum_{(p,o)\in\mathcal{C}} x_{(p,o)},\\
\sum_{p:\, \mathrm{DT}(p) \ge 15} \ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\ge \frac{5.6}{100} \sum_{(p,o)\in\mathcal{C}} x_{(p,o)}.
\end{aligned}
\tag{6.3}
$$

To evaluate the performance of the method, we use the KPSAM to simulate the output policy for the 6 months of 2008 that were not used as input. We record the number of transplantations occurring and the net life years from transplant. To compare the fairness properties of the policy, we also record the same percentage distributions of transplant recipients as for the KTC policy (see above).

6.4.3 Results

The output of the method is the scoring rule assigning the Kidney Allocation Score to a patient-organ pair (p, o) of

$$\mathrm{KAS}(p,o) = \mathrm{LYFT}(p,o) + g\left(\mathrm{DT}(p)\right) + 0.12\,\mathrm{CPRA}(p),$$

where

$$g\left(\mathrm{DT}\right) = \begin{cases} 0.55\,\mathrm{DT}, & 0 \le \mathrm{DT} \le 5,\\ 2.75 + \mathrm{DT}, & 5 \le \mathrm{DT} \le 10,\\ 7.75 + 0.25\,\mathrm{DT}, & 10 \le \mathrm{DT}. \end{cases}$$

According to the above scoring rule, patients are awarded 1 point for every life year from transplant gain, 0.12 points per point of their sensitization score and points based on their dialysis time as follows: 0.55 points for the first 5 years, 1 point for every additional year up to 10 years and 0.25 points for every additional year beyond that.

We use simulation to compare the performance of the above designed policy and the KTC proposed policy for 6 months in 2008 (see Methodology above). The simu-

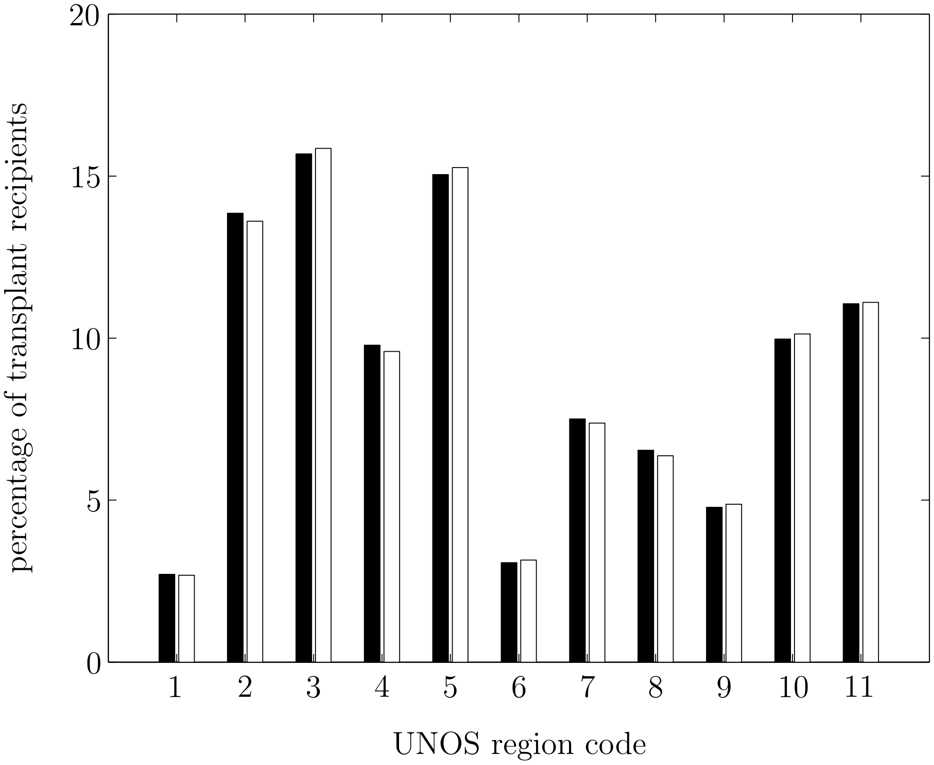


Figure 6-2: Simulated percentage distribution of transplant recipients across the 11 distinct geographical regions that UNOS has divided the U.S. into, under the KTC policy (white) and the policy designed in Section 6.4 (black), for an out-of-sample period of 6 months in 2008.

lation results are presented in Table 6.1.

Compared to the KTC proposed policy, the one designed by our method delivers an almost identical performance in terms of percentage distributions of transplant recipients, but results in an important 8.2% increase in life years from transplant. Both policies appear to have the same performance in number of transplantations. Figure 6-2 illustrates the distribution of recipients under the two policies across the 11 distinct geographical regions that UNOS has divided the U.S. into. Again, in comparison with the KTC policy, the designed policy has an almost identical performance in terms of percentage distributions of transplant recipients across all regions.

In comparing the policies further, consider the organ recipients who spent less than 5 years on dialysis prior to receiving their transplant and the organ recipients who spent more than 5 years. Through the scaling of the LYFT and DT components

	KTC policy	**Designed policy**
transplantations and efficiency		
number of transplantations	5, 746	5, 796
net life years from transplant	34, 290	37, 092
avg life years from transplant	5.95	6.4
racial distribution		
caucasian	41.8%	43.5%
african-american	35.7%	33.3%
hispanic	13.9%	14.5%
other	8.6%	8.7%
age distribution		
18-34 yrs	5%	4.2%
34-49 yrs	26.4%	24.8%
49-64 yrs	50.6%	52.8%
64+ yrs	18%	18.2%
dialysis time distribution		
0-5 yrs	55.4%	56%
5-10 yrs	28.8%	28%
10-15 yrs	10.2%	10.1%
15+ yrs	5.6%	5.9%
blood type distribution		
O	47.9%	47.7%
A	37.9%	37.6%
B	11.7%	12%
AB	2.5%	2.7%
diagnosis type distribution		
nephritis	19.5%	18.9%
hypertension	21.6%	19.2%
polycystic	9.9%	11.8%
other	23%	25%
diabetes	26%	25.1%
sensitization level distribution		
CPRA 0-10	55.4%	54.9%
CPRA 10-80	24.7%	24.9%
CPRA 80+	19.9%	20.2%

Table 6.1: Simulated allocation results of the KTC policy and the policy designed in Section 6.4, for an out-of-sample period of 6 months in 2008.

with the donor profile index (DPI), the KTC policy directs better quality organs to patients with a higher LYFT score, whereas organs of marginal quality are offered to patients who spent many years on dialysis, as discussed in Section 6.2.1; for more details see [45], [52]. As a result, one might expect that under the KTC policy the organ recipients who spent less than 5 years on dialysis to be systematically allocated organs of better quality compared to those who spent more than 5 years. This is indeed reflected by our simulation results: the average life years from transplant gain of recipients who spent less than 5 years on dialysis is 14.2% higher than the average gain of all recipients; in contrast, the average life years from transplant gain of recipients who spent more than 5 years is 19.4% smaller than the average gain of all recipients. Under our policy however, the gain differences across those two groups are smaller: the differences are 9.7% higher than the average and 13.4% smaller than the average respectively for the two groups. That demonstrates that our policy provides a more equitable distribution of the organs, at least in that sense.

6.4.4 Sensitivity Analysis

We conclude our numerical experiments by demonstrating how our method can be used to perform a sensitivity analysis with respect to imposed fairness constraints. Similarly, one can perform an analysis with respect to changes in the score components used.

Specifically, we explore the dependence of life years from transplant gains on the priority given for dialysis time and sensitization. To this end, we redesign the policy by using the same procedure and input as above, but by considering slightly modified fairness constraints. In particular, we firstly use all the constraints used above, but relax the constraints pertaining to patient groups of different dialysis time, *i.e.*, constraints (6.3). The relaxation is performed by introducing a slack parameter s in the percentage requirements of recipients of different groups, that is, the relaxed constraints take the form

$$\begin{aligned}
\sum_{p\colon 0\leq \mathrm{DT}(p)\leq 5}\ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\geq \frac{55.4-s}{100}\sum_{(p,o)\in\mathcal{C}} x_{(p,o)},\\
\sum_{p\colon 5\leq \mathrm{DT}(p)\leq 10}\ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\geq \frac{28.8-s}{100}\sum_{(p,o)\in\mathcal{C}} x_{(p,o)},\\
\sum_{p\colon 10\leq \mathrm{DT}(p)\leq 15}\ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\geq \frac{10.2-s}{100}\sum_{(p,o)\in\mathcal{C}} x_{(p,o)},\\
\sum_{p\colon \mathrm{DT}(p)\geq 15}\ \sum_{o:(p,o)\in\mathcal{C}} x_{(p,o)} &\geq \frac{5.6-s}{100}\sum_{(p,o)\in\mathcal{C}} x_{(p,o)}.
\end{aligned} \tag{6.4}$$

Clearly, for $s = 0$ one would recover the policy that was designed previously. For $s > 0$, the requirement on matching the percentage distribution (with regard to patient groups of different dialysis time) achieved by the KTC policy is relaxed. Thus one should expect that policies designed with such relaxed requirements would achieve higher life years from transplant gains. Using our method, we design policies for various values of the slack parameter s and quantify how the gains in medical efficiency depend on deviations from the selected fairness constraints. Secondly, we follow the same procedure to examine the dependence of medical efficiency on the priority given to sensitized patients. We again use all the constraints as in the previous subsection, but this time relax the constraints pertaining to patient groups of different sensitization levels. The relaxation is again performed using a slack parameter s. Note that one can potentially perform a sensitivity analysis though many other different ways of relaxing the constraints; for illustration purposes we focus here only on the described method of uniformly relaxing the constraints by a slack parameter.

The results we obtain in the aforementioned scenarios are depicted in Figures 6-3 and 6-4. Figure 6-3 shows the life years from transplant gains (for the 6 months period we consider) of policies designed with relaxed constraints on patient groups of different dialysis time, for various values of the slack parameter s. Similarly, Figure 6-4 shows the life years from transplant gains of policies designed with relaxed constraints on patient groups of different sensitization, for various values of the slack parameter s. The two figures also depict the operational points of the KTC proposed policy. Comparing the two, one can observe that the dependence of medical efficiency is

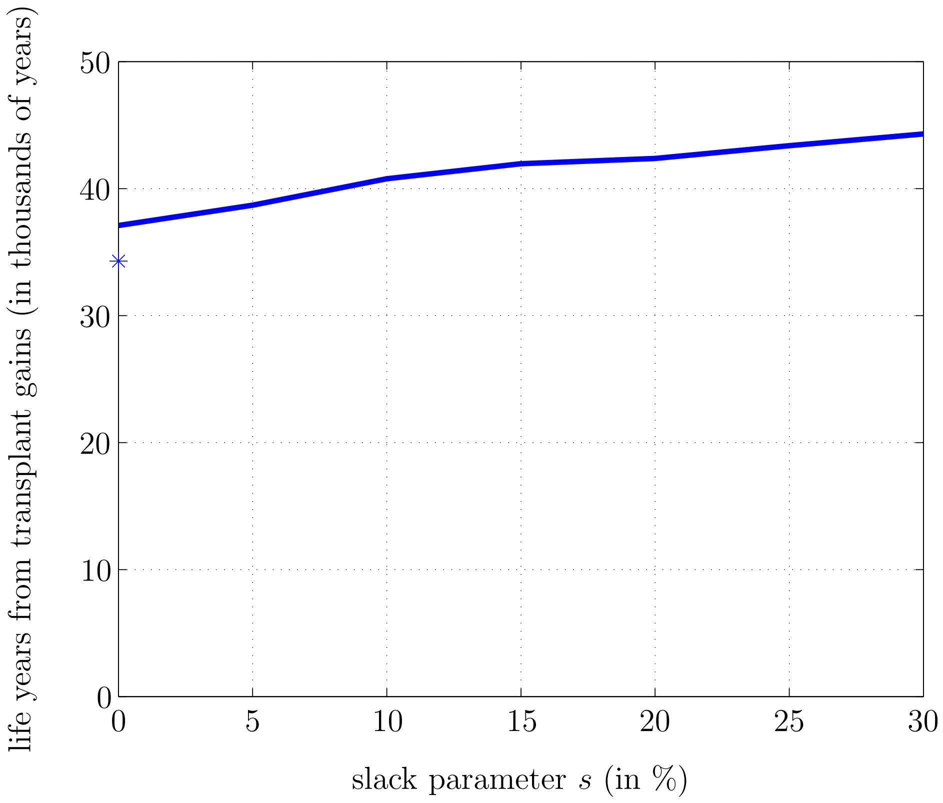


Figure 6-3: Simulated life years from transplant gains for policies (designed by our method) with relaxed constraints on all patient groups of different dialysis time, for various values of the slack parameter s; for more details see Section 6.4.4. The results are for an out-of-sample period of 6 months in 2008. The marker corresponds to the operational point of the policy proposed by the UNOS policymakers.

stronger on dialysis time. Also, the life years from transplant gains can be as high as $44,300$ years, which are 30% larger than the gains of the KTC policy. Note that although such a policy might not be implementable, the analysis can provide insights to policymakers and facilitate their decision process.

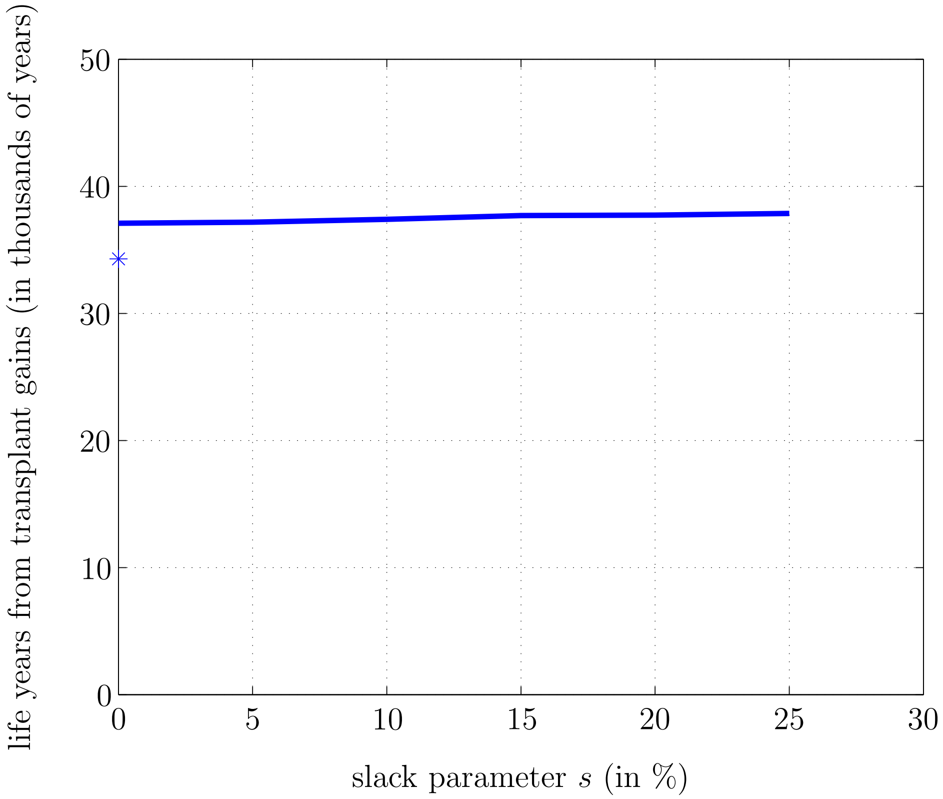


Figure 6-4: Simulated life years from transplant gains for policies (designed by our method) with relaxed constraints on all patient groups of different sensitization levels, for various values of the slack parameter s; for more details see Section 6.4.4. The results are for an out-of-sample period of 6 months in 2008. The marker corresponds to the operational point of the policy proposed by the UNOS policymakers.

List of Acronyms

CPRA	Calculated Panel Reactive Antibody
DPI	Donor Profile Index
DT	Dialysis Time
ESRD	End-Stage Renal Disease
KAS	Kidney Allocation Score
KPSAM	Kidney-Pancreas Simulated Allocation Model
KTC	Kidney Transplantation Committee
LYFT	Life Years From Transplant
NOTA	National Organ Transplant Act
OPO	Organ Procurement Organization
OPTN	Organ Procurement and Transplantation Network
RFI	Request For Information
SRTR	Scientific Registry of Transplant Recipients
UNOS	United Network for Organ Sharing

Disclaimer

The data reported here have been supplied by the Arbor Research Collaborative for Health (Arbor Research) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

Chapter 7

Concluding Remarks

We dealt with the problem of balancing efficiency and fairness in the context of resource allocation. Specifically, we addressed the following two central questions: for a resource allocation problem (a) how does one select/ design the right operational objective, and, given such a selection, (b) how does one find an implementable policy that serves this objective in practice?

We reviewed a plethora of problems in the broad area of operations management, for which the dichotomy between efficiency and fairness constitutes a central issue.

Despite the fact that fairness is of a subjective nature, we identify notions of fairness that are well-documented in the welfare economics literature and are of practical interest: the notions of proportional, max-min and α-fairness. The notion of α-fairness in particular provides a family of welfare functions that is canonical in that it captures the utilitarian allocation, the proportionally and max-min fair allocations. It also permits the decision maker to tradeoff efficiency for fairness by means of a single parameter.

For the above notions, we provide near-tight upper bounds on the relative efficiency loss compared to the efficiency-maximizing solution, where we measure efficiency as the sum of player utilities. In a similar fashion, we provide tight upper bounds on the relative fairness loss, where we measure fairness by the minimum utility guarantee. The bounds are applicable to a broad family of problems; they also suggest when the loss is likely to be small, and illustrate its dependence on the numbers of parties involved and the chosen balance between efficiency and fairness. Such a contribution has been elusive in the literature, to the best of our knowledge, and now provides the means for central decision makers to select their attitudes towards fairness and efficiency using quantitative arguments.

To highlight the above theoretical framework, we studied the problem of air traffic scheduling under limited capacity.

Furthermore, to deal with the question of designing implementable policies, we focused on the important problem of allocating deceased donor kidneys to waitlisted patients, in a fair and efficient way. We discussed the national allocation policy in the United States and the recent effort to revise the current policy in place.

Particularly, we studied allocation policies that are based on point systems; under those policies patients are awarded points according to some priority criteria, and patients are then prioritized by the number of points awarded. We identified the challenges in designing a point system, specifically the relative emphasis put on each criterion such that the resulting policy strikes the right balance between efficiency and fairness.

Our main contribution was a scalable, data-driven method of designing point system based allocation policies in an efficient and systematic way. The method does not presume any particular fairness scheme, or priority criterion. Instead, it offers the flexibility to the designer to select his desired fairness constraints and criteria under which patients are awarded points. Our method then balances the criteria and extracts a near-optimal point system policy, in the sense that the policy outcomes yield approximately the maximum number of life years gains (medical efficiency), while satisfying the fairness constraints.

Using our method, we designed a new policy that matches in fairness properties and priority criteria the policy that was recently proposed by the U.S. policymakers. Critically, our policy delivers an 8% relative increase in life years gains. The performance gain was established via simulation, utilizing the same statistical tools and data as the U.S. policymakers.

Finally, we presented a tradeoff analysis that revealed the dependence of medical efficiency on the important fairness concepts of prioritizing patients who have either spent a lot of time waiting, or are medically incompatible with the majority of donors.

Appendix A

Technical Notes

A.1 More on Near Worst-case Examples for the Price of Fairness

We demonstrate how one can construct near worst-case examples, for which the price of fairness is very close to the bounds implied by Theorem 2, for any values of the problem parameters; the number of players n and the value of the inequality aversion parameter α. We then provide details about the bandwidth allocation problem in Chapter 4.1.1.

For any $n \in \mathbf{N} \setminus \{0, 1\}$, $\alpha > 0$, we create a utility set using Procedure 2.

Procedure 2 Creation of near worst-case utility set

```
Input: $n \in \mathbf{N} \setminus \{0, 1\}$, $\alpha > 0$
Output: utility set $U$
compute $y := \operatorname*{argmin}_{x \in [1,n]} \frac{x^{1+\frac{1}{\alpha}} + n - x}{x^{1+\frac{1}{\alpha}} + (n-x)x}$
$x_1 \leftarrow \frac{y^{\frac{1}{\alpha}}}{n-y+y^{\frac{1}{\alpha}}}$ (as in (4.39))
$x_2 \leftarrow \frac{1}{n-y+y^{\frac{1}{\alpha}}}$ (as in (4.40))
$\ell \leftarrow \min\{\text{round}(y), n-1\}$
$\gamma_i \leftarrow \frac{x_i^{-\alpha}}{yx_1^{1-\alpha}+(n-y)x_2^{1-\alpha}}$ for $i = 1, 2$
$U \leftarrow \left\{u \in \mathbf{R}_+^n \,\middle|\, \gamma_1 u_1 + \ldots + \gamma_1 u_\ell + \gamma_2 u_{\ell+1} + \ldots + \gamma_2 u_n \leq 1, \quad u \leq \mathbf{1} \; \forall j\right\}$
```

The following proposition demonstrates why Procedure 2 creates utility sets that achieve a price of fairness very close to the bounds implied by Theorem 2.

Proposition 1. *For any $n \in \mathbf{N} \setminus \{0,1\}$, $\alpha > 0$, the output utility set U of Procedure 2 satisfies the conditions of Theorem 2. If $y \in \mathbf{N}$, the output utility set U satisfies the bound of Theorem 2 with equality.*

Proof. The output utility set U is a bounded polyhehron, hence convex and compact. Boundedness follows from positivity of γ_1 and γ_2.

Note that the selection of x_1, x_2 and y in Procedure 2 corresponds to a point that attains the minimum of (4.35), hence all properties quoted in the proof of Theorem 2 apply. In particular, by (4.30d) we have $\gamma_2 \leq 1$ and (4.33d) is tight, $y\gamma_1 = 1$. Moreover, the bound from Theorem 2 can be expressed as

$$\mathrm{POF}\,(U;\alpha) \leq 1 - \frac{yx_1 + (n-y)x_2}{y}.$$

The maximum achievable utility of the jth player is equal to 1. To see this, note that the definition of U includes the constraint $u_j \leq 1$, so it suffices to show that $e_j \in U$. For $j \leq \ell$, we have $\gamma_1 \leq \gamma_1 y = 1$. For $j > \ell$, we have $\gamma_2 \leq 1$. It follows that U satisfies the conditions of Theorem 2.

Suppose that $y \in \mathbf{N}$. By (4.36) and the choice of ℓ in Procedure 2, we get $\ell = y$. Consider the vector $z \in \mathbf{R}^n$ with $z_1 = \ldots = z_\ell = x_1$ and $z_{\ell+1} = \ldots = z_n = x_2$. Then, the sufficient first order optimality condition for z to be the α-fair allocation of U is satisfied, as for any $u \in U$

$$\sum_{j=1}^{n} z_j^{-\alpha}(u_j - z_j) = x_1^{-\alpha}(u_1 + \ldots + u_\ell) + x_2^{-\alpha}(u_{\ell+1} + \ldots + u_n) - yx_1^{1-\alpha} - (n-y)x_2^{1-\alpha} \leq 0,$$

since $\gamma_1(u_1 + \ldots + u_\ell) + \gamma_2(u_{\ell+1} + \ldots + u_n) \leq 1$. Hence,

$$\mathrm{FAIR}\,(U;\alpha) = \mathbf{1}^T z = yx_1 + (n-y)x_2.$$

For the efficiency-maximizing solution, since $y\gamma_1 = 1$, we get

$$\text{SYSTEM}\,(U) = y.$$

Then,

$$\text{POF}\,(U; \alpha) = 1 - \frac{y x_1 + (n-y) x_2}{y},$$

which is exactly the bound from Theorem 2. □

The above result demonstrates why one should expect Procedure 2 to generate examples that have a price of fairness very close to the established bounds. In particular, Proposition 1 shows that the source of error between the price of fairness for the utility sets generated by Procedure 2 and the bound is the (potential) non-integrality of y. In case that error is "large", one can search in the neighborhood of parameters γ_1 and γ_2 for an example that achieves a price closer to the bound, for instance by using finite-differencing derivatives and a gradient descent method (respecting feasibility).

Near worst-case bandwidth allocation

We utilize Proposition 1 and Procedure 2 to construct near worst-case network topologies. In particular, one can show that the line-graph discussed in Chapter 4.1.1, actually corresponds to a worst-case topology in this setup.

Suppose that we fix the number of players $n \geq 2$, the desired inequality aversion parameter $\alpha > 0$, and follow Procedure 2. Further suppose that $y \in \mathbf{N}$, as in Proposition 1. Consider then a network with y links of unit capacity, in a line-graph topology: the routes of the first y flows are disjoint and they all occupy a single (distinct) link. The remaining $n - y$ flows have routes that utilize all y links. Each flow derives a utility equal to its assigned nonnegative rate, which we denote $u_1, \ldots, u_n$. We next show that the price of fairness for this network is equal to the bound of Theorem 2.

The output utility set of Procedure 2 achieves the bound, by Proposition 1, since $y \in \mathbf{N}$. Moreover, we also get that $y\gamma_1 = 1$ and $\gamma_2 = 1$. Hence, the output utility set that achieves the bound can be formulated as

$$U = \{u \geq 0 \,|\, u_1 + \ldots + u_y + y\,(u_{y+1} + \ldots + u_n) \leq y,\, u \leq \mathbf{1}\}\,.$$

The utility set corresponding to the line-graph example above can be expressed using the nonnegativity constraints of the flow rates, and the capacity constraints on each of the y links as follows,

$$\overline{U} = \{u \geq 0 \,|\, u_j + u_{y+1} + \ldots + u_n \leq 1,\, j = 1, \ldots, y\}\,.$$

Clearly, the maximum sum of utilities under both sets is equal to y, simply by setting the first y components of u to 1. It suffices then to show that the two sets also share the same α-fair allocation. In particular, by symmetry of U and strict concavity of W_α, if u^F is its α fair allocation, then $u_1^F = \ldots = u_y^F$, and $u_{y+1}^F = \ldots = u_n^F$. As a result, it follows that $u^F \in \overline{U}$. Finally, noting that all inequalities in the definition of U are also valid for $\overline{U}$, it follows that $\overline{U} \subset U$ and that u^F is also the α-fair allocation of $\overline{U}$.

A.2 Auxiliary Results

Proposition 2. *For a point $(d, \lambda, x) \in S$ that attains the minimum of (4.31),*

(a) if $\lambda + 1 < n$, then without loss of generality

$$\underline{x}_{\lambda+1} = x_{\lambda+2} = \ldots = x_n, \text{ and,}$$

(b) without loss of generality

$$x_1 = \ldots = x_\lambda = \overline{x}_{\lambda+1}.$$

Proof. (a) We drop the underline notation for $\underline{x}_{\lambda+1}$ to simplify notation. Suppose that $x_j > x_{j+1}$, for some index $j \in \{\lambda + 1, \ldots, n - 1\}$. We will show that there always exists a new point, $(d, \lambda, x') \in S$, for which $x'_i = x_i$, for all $i \in \{1, \ldots, n\} \setminus \{j, j+1\}$, and which either achieves the same objective with $x'_j = x'_{j+1}$, or it achieves a strictly lower objective.

If $j = \lambda + 1$ and $d = 1$, we set $x'_j = x'_{j+1} = x_{j+1}$. The new point is feasible, and the objective attains the same value.

Otherwise, let $x'_j = x_j - \epsilon$, for some $\epsilon > 0$. We have two cases.

$\alpha \geq 1$: Let $x'_{j+1} = x_{j+1}$ and pick ϵ small enough, such that $x'_j \geq x'_{j+1}$. Moreover, for the new point (compared to the feasible starting point) the left-hand sides of (4.30d) and (4.30e) are unaltered, whereas the right-hand sides are either unaltered (for $\alpha = 1$) or greater, since $x_j^{1-\alpha} < (x_j - \epsilon)^{1-\alpha}$ for $\alpha > 1$. Hence, the new point is feasible. It also achieves a strictly lower objective value.

$\alpha < 1$: Let $x'_{j+1} = x_{j+1} + \rho b \epsilon$, where

$$b = \begin{cases} 1 - d, & \text{if } j = \lambda + 1, \\ 1, & \text{otherwise,} \end{cases}$$

$$\rho \in \left(\frac{x_j^{-\alpha}}{x_{j+1}^{-\alpha}}, 1 \right).$$

For ϵ small enough, we have $x'_j \geq x'_{j+1}$. For the new point, the left-hand side of (4.30d) either decreases (if $j + 1 = n$), or remains unaltered. The left-hand side of (4.30e) remains also unaltered. For the right-hand sides, since the only terms that change are those involving x_j and x_{j+1}, we use a first order Taylor series expansion to get

$$\begin{aligned} b\left(x'_j\right)^{1-\alpha} + \left(x'_{j+1}\right)^{1-\alpha} &= b\left(x_j - \epsilon\right)^{1-\alpha} + \left(x_{j+1} + \rho b \epsilon\right)^{1-\alpha} \\ &= b x_j^{1-\alpha} - b\epsilon(1-\alpha) x_j^{-\alpha} + x_{j+1}^{1-\alpha} + \rho b \epsilon (1-\alpha) x_{j+1}^{-\alpha} + O(\epsilon^2) \\ &= \left(b x_j^{1-\alpha} + x_{j+1}^{1-\alpha}\right) + b(1-\alpha)\left(\rho x_{j+1}^{-\alpha} - x_j^{-\alpha}\right)\epsilon + O(\epsilon^2). \end{aligned}$$

By the selection of ρ, the coefficient of the first order term (with respect to ϵ)

above is positive, and hence, for small enough ϵ we get

$$b\left(x'_j\right)^{1-\alpha}+\left(x'_{j+1}\right)^{1-\alpha}>bx_j^{1-\alpha}+x_{j+1}^{1-\alpha}.$$

That shows that the right hand side increases, and the new point is feasible. Finally, the difference in the objective value is $-b\epsilon+\rho b\epsilon$, and thus negative.

(b) We drop the overline notation for $\overline{x}_{\lambda+1}$ to simplify notation. Suppose that $x_j > x_{j+1}$, for some index $j \in \{1,\ldots,\lambda\}$.

We will show that there always exists a new point, $(d,\lambda,x') \in S$, for which $x'_i = x_i$, for all $i \in \{1,\ldots,n\} \setminus \{j,j+1\}$, and which either achieves the same objective with $x'_j = x'_{j+1}$, or it achieves a strictly lower objective.

If $j+1=\lambda+1$ and $d=0$, we set $x'_j = x'_{j+1} = x_j$. The new point is feasible, and the objective attains the same value.

Otherwise, let

$$\begin{aligned} x'_j &= x_j - \epsilon \\ x'_{j+1} &= x_{j+1} + \rho c\epsilon, \end{aligned}$$

for some $\epsilon > 0$, where

$$\begin{aligned} \rho &\in \left(\frac{x_{j+1}}{x_j}, \frac{x_{j+1}^{-\alpha}}{x_j^{-\alpha}}\right) \\ c &= \frac{x_j^{-\alpha}}{bx_{j+1}^{-\alpha}} \\ b &= \begin{cases} d, & \text{if } j+1=\lambda+1, \\ 1, & \text{otherwise.} \end{cases} \end{aligned}$$

For ϵ small enough, we have $x'_j \geq x'_{j+1}$. For the new point, the left-hand side of (4.30d) remains unaltered. For the left-hand side of (4.30e) we use a first order Taylor series expansion (similarly as above) to get

$$\begin{aligned}\left(x_j'\right)^{-\alpha} + b\left(x_{j+1}'\right)^{-\alpha} &= \left(x_j - \epsilon\right)^{-\alpha} + b\left(x_{j+1} + \rho c\epsilon\right)^{-\alpha} \\ &= x_j^{-\alpha} + \epsilon\alpha x_j^{-\alpha-1} + bx_{j+1}^{-\alpha} - b\rho c\epsilon\alpha x_{j+1}^{-\alpha-1} + O(\epsilon^2) \\ &= \left(x_j^{-\alpha} + bx_{j+1}^{-\alpha}\right) + \epsilon\alpha x_j^{-\alpha-1} - \rho\epsilon\alpha x_j^{-\alpha} x_{j+1}^{-1} + O(\epsilon^2) \\ &= \left(x_j^{-\alpha} + bx_{j+1}^{-\alpha}\right) + \alpha x_j^{-\alpha-1}\left(1 - \rho\frac{x_j}{x_{j+1}}\right)\epsilon + O(\epsilon^2).\end{aligned}$$

By the selection of ρ, the coefficient of the first order term (with respect to ϵ) above is negative, and hence, for small enough ϵ we get that the left-hand side decreases.

For the right-hand side of (4.30d) and (4.30e), we similarly get that

$$\begin{aligned}\left(x_j'\right)^{1-\alpha} + b\left(x_{j+1}'\right)^{1-\alpha} &= \left(x_j - \epsilon\right)^{1-\alpha} + b\left(x_{j+1} + \rho c\epsilon\right)^{1-\alpha} \\ &= x_j^{1-\alpha} - \epsilon(1-\alpha)x_j^{-\alpha} + bx_{j+1}^{1-\alpha} + b\rho c\epsilon(1-\alpha)x_{j+1}^{1-\alpha} + O(\epsilon^2) \\ &= \left(x_j^{1-\alpha} + bx_{j+1}^{1-\alpha}\right) + (1-\alpha)x_j^{-\alpha}\left(\rho - 1\right)\epsilon + O(\epsilon^2).\end{aligned}$$

If for $\alpha > 1$ we pick $\rho < 1$, and for $\alpha < 1$ we pick $\rho > 1$, the first order term (with respect to ϵ) above is positive, and hence, for small enough ϵ we get that the right-hand side increases for $\alpha \neq 1$. For $\alpha = 1$, the right-hand side remains unaltered.

In all cases, the new point is feasible, and the difference in the objective value is

$$-\epsilon + \rho cb\epsilon = \left(\rho cb - 1\right)\epsilon = \left(\rho\frac{x_j^{-\alpha}}{x_{j+1}^{-\alpha}} - 1\right)\epsilon,$$

and thus negative (by the selection of ρ). □

Proposition 3. *Let $n \in \mathbf{N} \setminus \{0,1\}$ and $f : [1,n] \to \mathbf{R}$ be defined as*

$$f(x;\alpha,n) = \frac{x^{1+\frac{1}{\alpha}} + n - x}{x^{1+\frac{1}{\alpha}} + (n-x)x}.$$

For any $\alpha > 0$,

(a) $-f$ is unimodal over $[1,n]$, and thus has a unique minimizer $\xi^\star \in [1,n]$.

(b) $\min_{x\in[1,n]} f(x;\alpha,n) = f(\xi^\star;\alpha,n) = \Theta\left(n^{-\frac{\alpha}{\alpha+1}}\right)$.

Proof. (a) The derivative of f is

$$f'(x;\alpha,n) = \frac{g(x)}{\left(x^{1+\frac{1}{\alpha}} + (n-x)x\right)^2},$$

where

$$g(x) = \left(1-\frac{1}{\alpha}\right)x^{2+\frac{1}{\alpha}} + \frac{n+1}{\alpha}x^{1+\frac{1}{\alpha}} - n\left(1+\frac{1}{\alpha}\right)x^{\frac{1}{\alpha}} - (x-n)^2.$$

Note that the sign of the derivative is determined by $g(x)$, since the denominator is positive for $1 \le x \le n$, that is,

$$\operatorname{sgn} f'(x;\alpha,n) = \operatorname{sgn} g(x). \tag{A.1}$$

We will show that g is strictly increasing over $[1,n]$. To this end, we have

$$g'(x) = x^{\frac{1}{\alpha}-1}q(x) + 2(n-x),$$

where

$$q(x) = \left(2+\frac{1}{\alpha}\right)\left(1-\frac{1}{\alpha}\right)x^2 + \left(1+\frac{1}{\alpha}\right)\left(\frac{n+1}{\alpha}\right)x - \frac{n}{\alpha}\left(1+\frac{1}{\alpha}\right).$$

Since we are interested in the domain $[1,n]$, it suffices to show that $q(x) > 0$ over it. For $\alpha > 1$, q is a convex quadratic, with its minimizer being equal to

$$-\frac{\left(1+\frac{1}{\alpha}\right)\left(\frac{n+1}{\alpha}\right)}{2\left(2+\frac{1}{\alpha}\right)\left(1-\frac{1}{\alpha}\right)} < 0.$$

Hence, $q(x) \ge q(1)$ for $x \in [1,n]$. Similarly, for $\alpha < 1$, q is a concave quadratic, and as such, for $x \in [1,n]$ we have $q(x) \ge \min\{q(1), q(n)\}$. For $\alpha = 1$, $q(x) = 2(n+1)x - 2n$, which is positive for $x \ge 1$. Then, $q(x) > 0$ in $[1,n]$ for all $\alpha > 0$, if and only if $q(1) > 0$ and $q(n) > 0$. Note that for $r = 1$, we get $q(1) = 2$ and

$q(n) = 2n^2$, and

$$\frac{dq(1)}{dr} = 2 > 0, \quad \frac{dq(n)}{dr} = 2n^2 > 0,$$

which demonstrates that $q(1)$ and $q(n)$ are positive. Furthermore,

$$g(n) = n^{1+\frac{1}{\alpha}}(n-1) > 0.$$

Using the above, the fact that g is continuous and strictly increasing over $[1, n]$ and (A.1), we deduce that if $g(1) < 0$, there exists a unique $m \in (1, n)$ such that

$$\operatorname{sgn} f'(x; \alpha, n) \begin{cases} < 0, & \text{if } 1 \leq x < m, \\ > 0, & \text{if } m < x \leq n. \end{cases}$$

Similarly, if $g(1) \geq 0$, f is strictly increasing for $1 \leq x \leq n$. It follows that $-f$ is unimodal.

(b) Let $\theta_n = n^{\frac{\alpha}{\alpha+1}}$. Using the mean value Theorem, for every $n \geq 2$, there exists a $\psi_n \in [\theta_n, \xi^\star]$ (or $[\xi^\star, \theta_n]$, depending on if $\theta_n \leq \xi^\star$), such that

$$f(\theta_n; \alpha, n) = f(\xi^\star; \alpha, n) + f'(\psi_n; \alpha, n)(\theta_n - \xi^\star),$$

or, equivalently,

$$\frac{f(\xi^\star; \alpha, n)}{f(\theta_n; \alpha, n)} = 1 - \frac{f'(\psi_n; \alpha, n)(\theta_n - \xi^\star)}{f(\theta_n; \alpha, n)}.$$

We will show that, for a sufficiently small $\epsilon > 0$

(I.) $f'(\psi_n; \alpha, n) = O\left(n^{-\frac{\min\{1,\alpha\}+2\alpha}{\alpha+1}+2\epsilon}\right)$,

(II.) $\theta_n - \xi^\star = O\left(n^{\frac{\alpha}{\alpha+1}+\epsilon}\right)$,

(III.) $f(\theta_n; \alpha, n) = \Theta\left(n^{-\frac{\alpha}{\alpha+1}}\right)$.

Using the above facts, it is easy to see that

$$\frac{f(\xi^\star; \alpha, n)}{f(\theta_n; \alpha, n)} = 1 - \frac{f'(\psi_n; \alpha, n)(\theta_n - \xi^\star)}{f(\theta_n; \alpha, n)} = 1 - O\left(n^{-\frac{\min\{1,\alpha\}}{\alpha+1}+3\epsilon}\right) \to 1,$$

and thus $f(\xi^\star;\alpha,n) = \Theta\left(n^{-\frac{\alpha}{\alpha+1}}\right)$.

(I.) We first show that for any sufficiently large n,

$$n^{\frac{\alpha}{\alpha+1}-\epsilon} \leq \xi^\star \leq n^{\frac{\alpha}{\alpha+1}+\epsilon}. \tag{A.2}$$

By part (a), $\xi^\star$ is the unique root of g in the interval $[1,n]$. Moreover, g is strictly increasing. The dominant term of

$$\begin{aligned} g\left(n^{\frac{\alpha}{\alpha+1}-\epsilon}\right) = \left(1-\frac{1}{\alpha}\right) n^{\left(2+\frac{1}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}-\epsilon\right)} + \frac{1}{\alpha} n^{1-\frac{\alpha+1}{\alpha}\epsilon} + \frac{1}{\alpha} n^{2-\frac{\alpha+1}{\alpha}\epsilon} \\ - \left(1+\frac{1}{\alpha}\right) n^{1+\frac{1}{\alpha+1}-\frac{1}{\alpha}\epsilon} - n^2 - n^{\frac{2\alpha}{\alpha+1}-2\epsilon} + 2n^{1+\frac{\alpha}{\alpha+1}-\epsilon}, \end{aligned}$$

is $-n^2$, and hence, for sufficiently large n we have $g\left(n^{\frac{\alpha}{\alpha+1}-\epsilon}\right) < 0$. Similarly, the dominant term of $g\left(n^{\frac{\alpha}{\alpha+1}+\epsilon}\right)$ is $\frac{1}{\alpha}n^{2+\frac{\alpha+1}{\alpha}\epsilon}$, and for sufficiently large n we have $g\left(n^{\frac{\alpha}{\alpha+1}+\epsilon}\right) > 0$. The claim then follows. Using the above bound, for sufficiently large n, we also get that $\psi_n \geq n^{\frac{\alpha}{\alpha+1}-\epsilon}$. We now provide a bound for the denominator of $f'(\psi_n;\alpha,n)$. In particular, for sufficiently large n, we get that for $x \leq n^{\frac{\alpha}{\alpha+1}+\epsilon}$,

$$\frac{d}{dx}\left(x^{1+\frac{1}{\alpha}} + nx - x^2\right) = \left(1+\frac{1}{\alpha}\right) x^{\frac{1}{\alpha}} + n - 2x > 0,$$

which shows that the denominator is strictly increasing. Hence, using the lower bound on ψ_n,

$$\begin{aligned} \frac{1}{\left(\psi_n^{1+\frac{1}{\alpha}} + n\psi_n - \psi_n^2\right)^2} &\leq \frac{1}{\left(n^{\left(\frac{\alpha}{\alpha+1}-\epsilon\right)\left(1+\frac{1}{\alpha}\right)} + n^{1+\frac{\alpha}{\alpha+1}-\epsilon} - n^{\frac{2\alpha}{\alpha+1}-2\epsilon}\right)^2} \\ &\leq \frac{n^{-2-\frac{2\alpha}{\alpha+1}+2\epsilon}}{\left(n^{-\frac{\alpha}{\alpha+1}-\frac{1}{\alpha}\epsilon} + 1 - n^{-\frac{1}{\alpha+1}}\right)^2} = O\left(n^{-2-\frac{2\alpha}{\alpha+1}+2\epsilon}\right). \end{aligned}$$

We now provide a bound for the numerator. Since g is strictly increasing and

$\xi^\star$ is a root, we get

$$\begin{aligned}
|g(\psi_n)| &\le |g(\theta_n)| \\
&= \left| \left(1 - \frac{1}{\alpha}\right) \alpha^{\frac{2\alpha+1}{\alpha+1}} n^{-\frac{1}{\alpha+1}+2} + n - \right. \\
&\qquad \left. - \left(1 + \frac{1}{\alpha}\right) \alpha^{\frac{1}{\alpha+1}} n^{-\frac{\alpha}{\alpha+1}+2} - \alpha^{\frac{2\alpha}{\alpha+1}} n^{-\frac{2}{\alpha+1}+2} + 2\alpha^{\frac{\alpha}{\alpha+1}} n^{-\frac{1}{\alpha+1}+2} \right| \\
&= O\left(n^{-\frac{\min\{1,\alpha\}}{\alpha+1}+2}\right).
\end{aligned}$$

If we combine the above results, we get $f'(\psi_n; \alpha, n) = O\left(n^{-\frac{\min\{1,\alpha\}+2\alpha}{\alpha+1}+2\epsilon}\right)$.

(II.) Follows from (A.2).

(III.) We have

$$\begin{aligned}
f(\theta_n; \alpha, n) &= \frac{n + n - n^{\frac{\alpha}{\alpha+1}}}{n + n^{1+\frac{\alpha}{\alpha+1}} - n^{\frac{2\alpha}{\alpha+1}}} \\
&= \frac{n\left(2 - n^{-\frac{1}{\alpha+1}}\right)}{n^{1+\frac{\alpha}{\alpha+1}}\left(n^{-\frac{\alpha}{\alpha+1}} + 1 - n^{-\frac{1}{\alpha+1}}\right)} = \Theta\left(n^{-\frac{\alpha}{\alpha+1}}\right).
\end{aligned}$$

□

Proposition 4. *There exists a point $z \in \mathbf{R}^n$ that attains the minimum of (4.41), for which*

$$z_1 = \ldots = z_{n-1} = 1.$$

Proof. For $\alpha = 1$, problem (4.41) is writen as

$$\begin{array}{ll}
\text{minimize} & \dfrac{1}{n}\left(\dfrac{z_n}{z_1} + \dfrac{z_n}{z_2} + \ldots + \dfrac{z_n}{z_{n-1}} + 1\right) \\
\text{subject to} & \dfrac{1}{n} \le z_n \le z_{n-1} \le \ldots \le z_1 \le 1.
\end{array}$$

If z is an optimal solution of the above, then clearly $z_1 = \ldots = z_{n-1} = 1$.

We now deal with the case of $\alpha \neq 1$. We first show that if z is an optimal solution of (4.41), then $z_1 = \ldots = z_{n-1}$. We analyze the cases $0 < \alpha < 1$ and $\alpha > 1$ separately.

For $0 < \alpha < 1$, the function $z_1^{1-\alpha} + \ldots + z_{n-1}^{1-\alpha}$ is strictly concave, and the function $z_1^{-\alpha} + \ldots + z_{n-1}^{-\alpha}$ is strictly convex. If z is an optimal solution of (4.41) for which $z_1 = \ldots = z_{n-1}$ is violated, we construct a point $\bar{z} \in \mathbf{R}^n$, such that its first $n-1$

components are all equal to the mean of $z_1, \dots, z_{n-1}$ and $\bar{z}_n = z_n$. We show that $\bar{z}$ is feasible for (4.41) and it achieves a strictly lower objective value compared to z, a contradiction. Note that by strict concavity/ convexity we get

$$\bar{z}_1^{1-\alpha} + \ldots + \bar{z}_{n-1}^{1-\alpha} > z_1^{1-\alpha} + \ldots + z_{n-1}^{1-\alpha},$$

and

$$\bar{z}_1^{-\alpha} + \ldots + \bar{z}_{n-1}^{-\alpha} < z_1^{-\alpha} + \ldots + z_{n-1}^{-\alpha},$$

respectively. For feasibility, $0 \le \bar{z}_n \le \ldots \le \bar{z}_1 \le 1$ is immediate and

$$\bar{z}_n^{-\alpha} = z_n^{-\alpha} \le z_1^{1-\alpha} + \ldots + z_{n-1}^{1-\alpha} + z_n^{1-\alpha} < \bar{z}_1^{1-\alpha} + \ldots + \bar{z}_{n-1}^{1-\alpha} + \bar{z}_n^{1-\alpha}.$$

Finally, compared to z, if we evaluate the objective of (4.41) at $\bar{z}$, the numerator strictly decreases and the denominator strictly increases, hence the objective value strictly decreases.

For $\alpha > 1$, let z be an optimal solution of (4.41) for which $z_{j+1} < z_j$ for some $j = 1, \dots, n-2$. We similarly construct a feasible point $\bar{z}$ for (4.41) that achieves a strictly lower objective value than z. Let $\bar{z}_i = z_i$ for all $i \neq j, j+1$, $\bar{z}_j = z_j - \epsilon$ and $\bar{z}_{j+1} = z_{j+1} + \delta\epsilon$, where $\epsilon > 0$ and

$$\delta = \frac{z_j^{-\alpha} - \mu}{z_{j+1}^{-\alpha}}, \qquad \mu \in \left(0, z_j^{-\alpha}\left(\frac{z_j - z_{j+1}}{z_j}\right)\right).$$

For small enough ϵ, $0 \le \bar{z}_n \le \ldots \le \bar{z}_1 \le 1$ is immediate. Using a first order Taylor series expansion,

$$\begin{aligned}\bar{z}_j^{1-\alpha} + \bar{z}_{j+1}^{1-\alpha} &= z_j^{1-\alpha} + z_{j+1}^{1-\alpha} + (z_j^{-\alpha} - \delta z_{j+1}^{-\alpha})(\alpha - 1)\epsilon + O(\epsilon^2) \\ &> z_j^{1-\alpha} + z_{j+1}^{1-\alpha}\end{aligned}$$

for small enough ϵ, since $z_j^{-\alpha} > \delta z_{j+1}^{-\alpha} \Leftrightarrow \mu > 0$. As a result,

$$\bar{z}_1^{1-\alpha} + \ldots + \bar{z}_{n-1}^{1-\alpha} + \bar{z}_n^{1-\alpha} > z_1^{1-\alpha} + \ldots + z_{n-1}^{1-\alpha} + z_n^{1-\alpha},$$

and $\bar{z}$ is feasible. Moreover, the denominator of the objective strictly increases. Thus it suffices to show that the numerator decreases. To this end, we have

$$\begin{aligned}\bar{z}_j^{-\alpha} + \bar{z}_{j+1}^{-\alpha} &= z_j^{-\alpha} + z_{j+1}^{-\alpha} + (z_j^{-\alpha-1} - \delta z_{j+1}^{-\alpha-1})\alpha\epsilon + O(\epsilon^2) \\ &< z_j^{-\alpha} + z_{j+1}^{-\alpha}\end{aligned}$$

for small enough ϵ, since $z_j^{-\alpha-1} < \delta z_{j+1}^{-\alpha-1} \Leftrightarrow \mu < z_j^{-\alpha}\left(\frac{z_j - z_{j+1}}{z_j}\right)$.

Since for every optimal solution of (4.41), we have $z_1 = \ldots = z_{n-1}$, problem (4.41) can be writen equivalently as

$$\begin{aligned} \text{minimize} \quad & g(z_1, z_2) = \frac{(n-1)z_1^{-\alpha}z_2 + z_2^{1-\alpha}}{(n-1)z_1^{1-\alpha} + z_2^{1-\alpha}} \\ \text{subject to} \quad & 0 \le z_2 \le z_1 \le 1 \\ & z_2^{-\alpha} \le (n-1)z_1^{1-\alpha} + z_2^{1-\alpha}. \end{aligned} \tag{A.3}$$

It suffices to show that there exists an optimal solution z of (A.3) for which $z_1 = 1$.

Let z be an optimal solution of (A.3).

If $0 < \alpha < 1$, assume that $z_1 < 1$. Then, increase z_1 by a small enough amount such that it remains less than 1. The quantity $z_1^{1-\alpha}$ increases, so the new point we get is feasible. Also, the quantity $z_1^{-\alpha}$ decreases. Hence, the new point is feasible and achieves a strictly lower objective value, a contradiction.

If $\alpha > 1$, the point z lies on the boundary of the feasible set or is a stationary point of the objective. Suppose that z is not a stationary point, *i.e.*, $\nabla g(z_1, z_2) \neq 0$. If $z_1 = z_2$, the objective evaluates to 1 for any such z, so we can assume $z_1 = 1$. We next rule out the possibility of z lying on the $z_2^{-\alpha} = (n-1)z_1^{1-\alpha} + z_2^{1-\alpha}$ boundary with $z_1 < 1$. Suppose that it does. We will demonstrate that we can always find a feasible direction along which the objective decreases. We have

$$\begin{aligned} \frac{\partial g}{\partial z_1} &= \frac{(n-1)z_1^{-\alpha}z_2}{\left((n-1)z_1^{1-\alpha} + z_2^{1-\alpha}\right)^2}\left(-(n-1)z_1^{-\alpha} - \alpha z_1^{-1}z_2^{1-\alpha} + (\alpha-1)z_2^{-\alpha}\right), \\ \frac{\partial g}{\partial z_2} &= -\frac{z_1}{z_2}\frac{\partial g}{\partial z_1}. \end{aligned}$$

Note that we assumed that $\nabla g(z) \neq 0$, hence $\frac{\partial g}{\partial z_1}(z) \neq 0$. Suppose that $\frac{\partial g}{\partial z_1}(z) > 0$. Then, $(1, \delta)$ is a direction along which the objective decreases, for large enough $\delta > 0$, since

$$\frac{\partial g}{\partial z_1}(z) + \delta \frac{\partial g}{\partial z_2}(z) = \frac{\partial g}{\partial z_1}(z) \left(1 - \delta \frac{z_1}{z_2}\right) < 0.$$

It is also a feasible direction, since for $\epsilon > 0$ small enough, $0 \leq z_2 + \delta\epsilon \leq z_1 + \epsilon \leq 1$, and is also a direction along which $(n-1)z_1^{1-\alpha} + z_2^{1-\alpha} + z_2^{-\alpha}$ increases, since

$$\begin{aligned}(n-1)z_1^{-\alpha} + \delta\left((1-\alpha)z_2^{-\alpha} + \alpha z_2^{-\alpha-1}\right) &= (1-\alpha)(z_2^{-\alpha} - z_2^{1-\alpha}) + \\ &\quad + \delta\left((1-\alpha)z_2^{-\alpha} + \alpha z_2^{-\alpha-1}\right) \\ &= z_2^{-\alpha}((1-\alpha)(1-z_2) + \delta\left(\frac{a}{z_2} - (\alpha-1)\right) \\ &> 0\end{aligned}$$

for large enough δ. Similarly, if $\frac{\partial g}{\partial z_1}(z) < 0$, one can show that $(1, \delta)$ is again a feasible direction along which the objective decreases, for

$$\frac{(\alpha-1)(1-z_2)z_2}{\alpha - (\alpha-1)z_2} < \delta < \frac{z_2}{z_1},$$

if one can select such δ. Otherwise, one can show that $(-1, -\delta)$ is a feasible direction along which the objective decreases, for

$$\frac{z_2}{z_1} < \delta < \frac{(\alpha-1)(1-z_2)z_2}{\alpha - (\alpha-1)z_2}.$$

We have thus established that if z is not a stationary point, then there also exists an optimal solution for which $z_1 = 1$. We next show that the same holds true if z is a stationary point.

Suppose that z is a stationary point, *i.e.*, $\nabla g(z_1, z_2) = 0$. Then, we have

$$(n-1)z_1^{1-\alpha} + \alpha z_2^{1-\alpha} - (\alpha-1)z_1 z_2^{-\alpha} = 0.$$

Using the above, the objective evaluates to

$$g(z_1, z_2) = \frac{\alpha}{\alpha - 1} \frac{z_2}{z_1}.$$

Moreover, if $z_1 = \lambda z_2$ for some $\lambda \geq 1$, the stationarity condition yields

$$(n-1)\lambda^{1-\alpha} - (\alpha - 1)\lambda + \alpha = 0,$$

an equation that has a unique solution in $[1, \infty)$. Let $\bar{\lambda}$ be the solution. Then, the problem (A.3) constrained on the stationary points of its objective can be expressed as

$$\begin{array}{ll} \text{minimize} & \frac{\alpha}{\alpha-1}\frac{z_2}{z_1} \\ \text{subject to} & z_1 = \bar{\lambda} z_2, \quad z_1 \leq 1 \\ & z_2^{-\alpha} \leq (n-1) z_1^{1-\alpha} + z_2^{1-\alpha}, \end{array}$$

or, equivalently,

$$\begin{array}{ll} \text{minimize} & \frac{\alpha}{\alpha-1}\frac{1}{\bar{\lambda}} \\ \text{subject to} & z_1 = \bar{\lambda} z_2 \\ & \frac{1}{(\alpha-1)(\bar{\lambda}-1)} \leq z_2 \leq \frac{1}{\bar{\lambda}}. \end{array}$$

In case the above problem is feasible, we pick $z_2 = \frac{1}{\bar{\lambda}}$, and $z_1 = 1$ and the proof is complete. □

Proposition 5. *Consider a resource allocation problem with n players, $n \geq 2$. Let the utility set, denoted by $U \subset \mathbf{R}^n$, be compact and convex. If the players have equal maximum achievable utilities (greater than zero),*

$$\text{POF}\,(U; 1) \leq 1 - \frac{2\sqrt{n} - 1}{n}. \quad \textit{(price of proportional fairness)}$$

Let $\{\alpha_k \in \mathbf{R} \mid k \in \mathbf{N}\}$ be a sequence such that $\alpha_k \to \infty$ and $\alpha_k \geq 1$, $\forall k$. Then,

$$\limsup_{k\to\infty} \text{POF}\,(U; \alpha_k) \leq 1 - \frac{4n}{(n+1)^2}. \quad \textit{(price of max-min fairness)}$$

Proof. Let f be defined as in Proposition 3. Using Theorem 2 for $\alpha = 1$ we get

$$\begin{aligned}
\operatorname{POF}(U;1) &\leq 1 - \min_{x\in[1,n]} f(x;1,n) \\
&= 1 - \min_{x\in[1,n]} \frac{x^2+n-x}{nx} \\
&= 1 - \frac{2\sqrt{n}-1}{n}.
\end{aligned}$$

Similarly, for any $k \in \mathbf{N}$ and $\alpha = \alpha_k$

$$\operatorname{POF}(U;\alpha_k) \leq 1 - \min_{x\in[1,n]} f(x;\alpha_k,n),$$

which implies that

$$\begin{aligned}
\limsup_{k\to\infty} \operatorname{POF}(U;\alpha_k) &\leq \limsup_{k\to\infty} \left(1 - \min_{x\in[1,n]} f(x;\alpha_k,n)\right) \\
&\leq 1 - \liminf_{k\to\infty} \min_{x\in[1,n]} f(x;\alpha_k,n). \qquad \text{(A.4)}
\end{aligned}$$

Consider the set of (real-valued) functions $\{f(\,.\,;\alpha_k,n)\,|\,k\in\mathbf{N}\}$ defined over the compact set $[1,n]$. We show that the set is equicontinuous, and that the closure of the set $\{f(x;\alpha_k,n)\,|\,k\in\mathbf{N}\}$ is bounded for any $x\in[1,n]$. Boundedness follows since $0 \leq f(x;\alpha,n) \leq 1$ for any $\alpha > 0$ and $x \in [1,n]$. The set of functions $\{f(\,.\,;\alpha_k,n)\,|\,k\in\mathbf{N}\}$ shares the same Lipschitz constant, as for any $k\in\mathbf{N}$, $\alpha_k \geq 1$ and $x\in[1,n]$ we have

$$\begin{aligned}
|f'(x;\alpha_k,n)| &= \left|\frac{\left(1-\frac{1}{\alpha_k}\right)x^{2+\frac{1}{\alpha_k}} + \frac{n+1}{\alpha_k}x^{1+\frac{1}{\alpha_k}} - n\left(1+\frac{1}{\alpha_k}\right)x^{\frac{1}{\alpha_k}} - (x-n)^2}{\left(x^{1+\frac{1}{\alpha_k}}+(n-x)x\right)^2}\right| \\
&\leq \left|\left(1-\frac{1}{\alpha_k}\right)x^{2+\frac{1}{\alpha_k}} + \frac{n+1}{\alpha_k}x^{1+\frac{1}{\alpha_k}} - n\left(1+\frac{1}{\alpha_k}\right)x^{\frac{1}{\alpha_k}} - (x-n)^2\right| \\
&\leq \left(1-\frac{1}{\alpha_k}\right)x^{2+\frac{1}{\alpha_k}} + \frac{n+1}{\alpha_k}x^{1+\frac{1}{\alpha_k}} + n\left(1+\frac{1}{\alpha_k}\right)x^{\frac{1}{\alpha_k}} + (x-n)^2 \\
&\leq n^3 + (n+1)n^2 + 2n^2 + n^2 = 2(n^3+2n^2).
\end{aligned}$$

As a result, the set of functions $\{f(\,.\,;\alpha_k,n)\,|\,k\in\mathbf{N}\}$ is equicontinuous.

Using the above result,

$$\lim_{k\to\infty} \min_{x\in[1,n]} f(x;\alpha_k,n) = \min_{x\in[1,n]} \lim_{k\to\infty} f(x;\alpha_k,n).$$

Thus, (A.4) yields

$$\begin{aligned}\limsup_{k\to\infty} \mathrm{POF}\,(U;\alpha_k) &\leq 1 - \liminf_{k\to\infty} \min_{x\in[1,n]} f(x;\alpha_k,n) \\ &= 1 - \min_{x\in[1,n]} \lim_{k\to\infty} f(x;\alpha_k,n) \\ &= 1 - \min_{x\in[1,n]} \lim_{k\to\infty} \frac{x^{1+\frac{1}{\alpha_k}} + n - x}{x^{1+\frac{1}{\alpha_k}} + (n-x)x} \\ &= 1 - \min_{x\in[1,n]} \frac{n}{x + (n-x)x} \\ &= 1 - \frac{4n}{(n+1)^2}. \qquad \square\end{aligned}$$

Appendix B

A Model for Air Traffic Flow Management

The following is a model for air traffic flow management due to [12]. Consider a set of flights, $\mathscr{F} = \{1, \ldots, F\}$, that are operated by airlines over a (discretized) time period in a network of airports, utilizing a capacitated airspace that is divided into sectors. Let $\mathscr{F}_a \subset \mathscr{F}$ be the set of flights operated by airline $a \in \mathscr{A}$, where $\mathscr{A} = \{1, \ldots, A\}$ is the set of airlines. Similarly, $\mathscr{T} = \{1, \ldots, T\}$ is the set of time steps, $\mathscr{K} = \{1, \ldots, K\}$ the set of airports, and $\mathscr{J} = \{1, \ldots, J\}$ the set of sectors. Flights that are continued are included in a set of pairs, $\mathscr{C} = \{(f', f) : f' \text{is continued by flight} \, f\}$. The model input data, the main decision variables, and a description of the feasibility set are described below:

Data.

$$
\begin{aligned}
N_f &= \text{number of sectors in flight } f\text{'s path,}\\
P(f,i) &= \begin{cases} \text{the departure airport, if } i = 1,\\ \text{the } (i-1)\text{th sector in flight } f\text{'s path, if } 1 < i < N_f,\\ \text{the arrival airport, if } i = N_f, \end{cases}\\
P_f &= (P(f,i) : 1 \leq i \leq N_f),\\
D_k(t) &= \text{departure capacity of airport } k \text{ at time } t,\\
A_k(t) &= \text{arrival capacity of airport } k \text{ at time } t,\\
S_j(t) &= \text{capacity of sector } j \text{ at time } t,\\
d_f &= \text{scheduled departure time of flight } f,\\
r_f &= \text{scheduled arrival time of flight } f,\\
s_f &= \text{turnaround time of an airplane after flight } f,\\
l_{fj} &= \text{number of time steps that flight } f \text{ must spend in sector } j,\\
T_f^j &= \text{set of feasible time steps for flight } f \text{ to arrive to sector } j\ ,\\
\underline{T}_f^j &= \text{first time step in the set } T_f^j\text{, and}\\
\bar{T}_f^j &= \text{last time step in the set } T_f^j.
\end{aligned}
$$

Decision Variables.

$$
w_{ft}^j = \begin{cases} 1, & \text{if flight } f \text{ arrives at sector } j \text{ by time step } t,\\ 0, & \text{otherwise.} \end{cases}
$$

Feasibility Set. The variable w is feasible if it satisfies the constraints:

$$
\begin{aligned}
&\textstyle\sum_{f:P(f,1)=k}(w_{ft}^k - w_{f,t-1}^k) \leq D_k(t) \;\; \forall k \in \mathscr{K}, t \in \mathscr{T},\\
&\textstyle\sum_{f:P(f,N_f)=k}(w_{ft}^k - w_{f,t-1}^k) \leq A_k(t) \;\; \forall k \in \mathscr{K}, t \in \mathscr{T},\\
&\textstyle\sum_{f:P(f,i)=j,P(f,i+1)=j',i<N_f}(w_{ft}^j - w_{ft}^{j'}) \leq S_j(t) \;\; \forall j \in \mathscr{J}, t \in \mathscr{T},\\
&w_{f,t+l_{fj}}^{j'} - w_{ft}^j \leq 0 \;\; \forall f \in \mathscr{F}, t \in T_f^j, j = P(f,i), j' = P(f,i+1), i < N_f,\\
&w_{ft}^k - w_{f,t-s_f}^k \leq 0 \;\; \forall (f', f) \in \mathscr{C}, t \in T_f^k, k = P(f,i) = P(f', N_f),\\
&w_{ft}^j - w_{f,t-1}^j \geq 0 \;\; \forall f \in \mathscr{F}, j \in P_f, t \in T_f^j,\\
&w_{ft}^j \in \{0,1\} \;\; \forall f \in \mathscr{F}, j \in P_f, t \in T_f^j.
\end{aligned}
$$

The constraints correspond to capacity constraints for airports and sectors, connectivity between sectors and airports, and connectivity in time (for more details, see [12]).

Bibliography

[1] J. Ahn and J. C. Hornberger. Involving patients in the cadaveric kidney transplant allocation process: A decision-theoretic perspective. *Management Science*, 42(5):pp. 629–641, 1996.

[2] M. Armony and A. R. Ward. Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, 58(3):624–637, MayJune 2010.

[3] K. J. Arrow. Aspects of the theory of risk-bearing. Helsinki, 1965.

[4] ATA, 2008.

[5] A. B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263, 1970.

[6] C. Barnhart, D. Bertsimas, C. Caramanis, and D. Fearing. Equitable and efficient coordination in traffic flow management. *Submitted for publication*, 2009.

[7] N. Barr. *The Economics of the Welfare State*. Weidenfeld and Nicolson, 1987.

[8] A. Bergson. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics*, 52:310–334, 1938.

[9] D. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, 1987.

[10] D. Bertsimas and S. Gupta. A proposal for network air traffic flow management incorporating fairness and airline collaboration. *Operations Research, submitted for publication*, 2010.

[11] D. Bertsimas, I.C. Paschalidis, and J.N. Tsitsiklis. Optimization of multiclass queuing networks: polyhedral and nonlinear characterizations of achievable performance. *Annals of Applied Probability*, 4(1):43–75, 1994.

[12] D. Bertsimas and S. S. Patterson. The air traffic flow management problem with enroute capacities. *Operations Research*, 46(3):406–422, 1998.

[13] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[14] D. Bisias, A. Lo, and J. Watkins. Estimating the nih efficient frontier. *In preparation*, 2010.

[15] T. Bonald and L. Massoulié. Impact of fairness on internet performance. *SIGMETRICS Perform. Eval. Rev.*, 29(1):82–91, 2001.

[16] M. Butler and H. P. Williams. Fairness versus efficiency in charging for the use of common facilities. *The Journal of the Operational Research Society*, 53(12):1324–1329, 2002.

[17] D. Callahan and A.A. Wasunna. *Medicine and the market: equity v. choice*. The Johns Hopkins University Press, 2006.

[18] D. Chakrabarty, G. Goel, V. V. Vazirani, L. Wang, and C. Yu. Some computational and game-theoretic issues in Nash and nonsymmetric bargaining games. *Working Paper*, 2009.

[19] C. W. Chan, V. F. Farias, N. Bambos, and G. J. Escobar. Maximizing throughput of hospital intensive care units with patient readmissions. *Submitted for Publication*, 2009.

[20] C.W. Chan, M. Armony, and N. Bambos. Fairness of heterogeneous queues through cone scheduling. *Proceedings Manufacturing & Service Operations Management Conference*, 2010.

[21] J.R. Correa, A.S. Schulz, and N.E. Stier-Moses. Fast, fair, and efficient flows in networks. *Operations Research*, 55(2):215–225, March-April 2007.

[22] T. H. Cui, J. S. Raju, and Z. J. Zhang. Fairness and channel coordination. *Management Science*, 53(8):1303–1314, August 2007.

[23] I. David and U. Yechiali. A time-dependent stopping problem with application to live organ transplants. *Operations Research*, 33(3):pp. 491–504, 1985.

[24] I. David and U. Yechiali. One-attribute sequential assignment match processes in discrete time. *Operations Research*, 43(5):pp. 879–884, 1995.

[25] DHHS. Organ Procurement and Transplantation Network Final Rule. 2000. Electronic Code of Federal Regulations, Title 42–Public Health, Chapter I–Public Health Service, Department of Health and Human Services, Subchapter K–Health Resources Development, Part 121.

[26] E. Gelenbe and L. Mitrani. *Analysis and Synthesis of Computer Systems*. Academic, London, 1980.

[27] J. C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *The Journal of Political Economy*, 63(4):309–321, 1955.

[28] D. H. Howard. Why do transplant surgeons turn down organs?: A model of the accept/reject decision. *Journal of Health Economics*, 21(6):957 – 969, 2002.

[29] E. Kalai and M. Smorodinsky. Other solutions to Nash's bargaining problem. *Econometrica*, 43:510–18, 1975.

[30] L. Kaplow and S. Shavell. *Fairness versus welfare.* Harvard University Press, 2002.

[31] F. P. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, 1997.

[32] KPSAM. Kidney-Pancreas Simulated Allocation Model. 2008. Arbor Research Collaborative for Health. Scientific Registry of Transplant Recipients.

[33] T. Lan, D. Kao, M. Chiang, and A. Sabharwal. An axiomatic theory of fairness in network resource allocation. *Working paper*, 2009. http://arxiv.org/pdf/0906.0557.

[34] T. Lensberg. Stability and the Nash solution. *Journal of Economic Theory*, 45:330–341, 1988.

[35] H. Luss. On equitable resource allocation problems: a lexicographic minimax approach. *Operations Research*, 47(3):361–378, 1999.

[36] A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory.* Oxford University Press, 1995.

[37] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.*, 8(5):556–567, 2000.

[38] J. Nash. The bargaining problem. *Econometrica*, 18:155–62, 1950.

[39] S. Norman. Update on the development of a new kidney transplant allocation system. *Dial Transpl*, 38:400–406, 2009.

[40] ODADK. Organ Distribution: Allocation of Deceased kidneys. 3(5):1–13, June 2010. United Network for Organ Sharing Policies.

[41] A. R. Odoni and L. Bianco. *Flow Control of Congested Networks*, chapter The Flow Management Problem in Air Traffic Control. Springer-Verlag, Berlin, 1987.

[42] W. Ogryczak, M. Pioro, and A. Tomaszewski. Telecommunications network design and max-min optimization problem. *Journal of Telecommunications and Information Technology*, pages 43–56, 2005.

[43] A.O. Ojo, F.K Port, R.A. Wolfe, E.A. Mauger, L. Williams, and D.P. Berling. Comparative mortality risks of chronic dialysis and cadaveric transplantation in black end-stage renal disease patients. *Am J Kidney Dis*, 24(1):59–64, 1994.

[44] OPTNKTC. Report of the OPTN/UNOS Kidney Transplantation Committee to the Board of Directors. 2007. September 17-18, Los Angeles, California.

[45] OPTNKTC. Report of the OPTN/UNOS Kidney Transplantation Committee to the Board of Directors. 2008. February 20-21, Orlando, Florida.

[46] M. V. Pauly. Avoiding side effects in implementing health insurance reform. *N Engl J Med*, 362:671–673, 2010.

[47] F.K. Port, R.A. Wolfe, E.A. Mauger, D.P. Berling, and K. Jiang. Comparison of survival probabilities for dialysis patients versus cadaveric renal transplant recipients. *JAMA*, 270(11):1339–1343, 1993.

[48] J. W. Pratt. Risk aversion in the small and large. *Econometrica*, 32:122–136, 1964.

[49] B. Radunovic and J.-Y. Le Boudec. A unified framework for max-min and min-max fairness with applications. *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, 40(2):1061–1070, 2002.

[50] J. Rawls. *A Theory of Justice.* Harvard University Press, Cambridge, Mass., 1971.

[51] D. B. Resnick. Setting biomedical research priorities in the 21st century. *American Medical Association Journal of Ethics*, 5(7), July 2003.

[52] RFI. Kidney Allocation Concepts: Request for Information. 2008. OPTN/UNOS Kidney Transplantation Committee.

[53] R. Righter. A resource allocation problem in a random environment. *Operations Research*, 37(2):pp. 329–338, 1989.

[54] A. Roth. *Axiomatic Models of Bargaining.* Berlin and New York: Springer Verlag, 1979.

[55] R. J. Ruth, L. Wyszewianski, and G. Herline. Kidney transplantation: A simulation model for examining demand and supply. *Management Science*, 31(5):pp. 515–526, 1985.

[56] P. Samuelson. *Foundations of Economic Analysis.* Harvard University Press, Cambridge, Mass., 1947.

[57] P. Schnuelle, D. Lorenz, M. Trede, and F.J. Van Der Woude. Impact of renal cadaveric transplantation on survival in end-stage renal failure: Evidence for reduced mortality risk compared with hemodialysis during long-term follow-up. *J Am Soc Nephrol*, 9:2135–2141, 1998.

[58] A. Sen and J. E. Foster. *On Economic Inequality.* Oxford University Press, 1997.

[59] M. Shreedhar and G. Varghese. Efficient fair queueing using deficit round-robin. *IEEE/ACM Trans. Netw.*, 4(3):375–385, 1996.

[60] X. Su and S. A. Zenios. Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing & Service Operations Management*, 6(4):280–301, 2004.

[61] X. Su and S. A. Zenios. Patient choice in kidney allocation: A sequential stochastic assignment model. *Operations Research*, 53(3):443–445, May-June 2005.

[62] X. Su and S. A. Zenios. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Science*, 52(11):1647–1660, November 2006.

[63] X. Su and S.A. Zenios. Patient choice in kidney allocation: The role of the queueing discipline. *MANUFACTURING SERVICE OPERATIONS MANAGEMENT*, 6(4):280–301, 2004.

[64] M. Suthanthiran and T.B. Strom. Renal transplantation. *N Engl J Med*, page 331:365, 1994.

[65] M. D. Swenson. Scarcity in the intensive care unit: Principles of justice for rationing icu beds. *The American Journal of Medicine*, 92(5):551 – 555, 1992.

[66] K. Talluri and G. van Ryzin. An analysis of bid-price controls for network revenue management. *Management Science*, 44(11):1577–1593, November 1998.

[67] A. Tang, J. Wang, and S. H. Low. Counter-intuitive throughput behaviors in networks under end-to-end control. *IEEE/ACM Trans. Netw.*, 14(2):355–368, March 2006.

[68] P. Tsoucas. The region of achievable performance in a model of klimov. Research report rc16543, IBM T.J. Watson Research Center, Yorktown Heights, NY, 1991.

[69] UNOS. United Network for Organ Sharing. 2010. `http://www.unos.org/`.

[70] USRDS. U.S. Renal data system, annual data report: Atlas of chronic kidney disease and end-stage renal disease in the united states. 2009. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

[71] T. Vossen, M. Ball, and R. Hoffman. A general approach to equity in traffic flow management and its application to mitigating exemption bias in ground delay programs. *Air Traffic Control Quarterly*, 11:277–292, 2003.

[72] A. Wagstaff. Qalys and the equity-efficiency trade-off. *Journal of Health Economics*, 10(1):21 – 41, 1991.

[73] L. Waisanen, R.A. Wolfe, R.M. Merion, K. McCullough, and A. Rodgers. Simulating the allocation of organs for transplantation. *Health Care Management Science*, 7(4):331–338, 2004.

[74] R.A. Wolfe, K.P. McCullough, and A.B. Leichtman. Predictability of survival models for waiting list and transplant patients: Calculating LYFT. *Am J Transplant*, 9(7):1523–1527, 2009.

[75] R.A. Wolfe, K.P. McCullough, D.E. Schaubel, J.D. Kalbfleisch, S. Murray, M.D. Stegall, and A.B. Leichtman. 2007 SRTR report on the state of transplantation: Calculating life years from transplant (LYFT): Methods for kidney and kidney-pancreas candidates. *Am J Transplant*, 8(2):997–1011, 2008.

[76] Y. Wu, C. H. Loch, and L. Van der Heyden. A model of fair process and its limits. *Manufacturing & Service Operations Management*, 10(4):637–653, 2008.

[77] H. Peyton Young. *Equity: In Theory and Practice.* Princeton University Press, 1995.

[78] S. A. Zenios. Models for kidney allocation. In Margaret Brandeau, Franois Sainfort, and William Pierskalla, editors, *Operations Research and Health Care*, volume 70 of *International Series in Operations Research and Management Science*, pages 537–554. Springer New York, 2005.

[79] S. A. Zenios, G. M. Chertow, and L. M. Wein. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research*, 48(4):pp. 549–569, 2000.