Optimal Waiting Times and Assignments in Healthcare Provision*

Mark Braverman
Princeton University
mbraverm@cs.princeton.edu

Jing Chen
Institute for Advanced Study, Princeton and Stony Brook University
jingchen@math.ias.edu

Sampath Kannan
University of Pennsylvania
kannan@cis.upenn.edu

February 17, 2013

Abstract

We investigate computational and mechanism design aspects of optimal scarce resource allocation, where the primary rationing mechanism is through waiting times. Specifically we consider the problem of allocating medical treatments to a population of patients. Each patient has demand for exactly one unit of treatment, and can choose to be treated in one of k hospitals, $H_1, \ldots, H_k$. Different hospitals have different costs, which are fully paid by a third party —the "payer"— and do not accrue to the patients. The payer has a fixed budget B and can only cover a limited number of treatments in the more expensive hospitals. Access to over-demanded hospitals is rationed through waiting times: each hospital H_i will have waiting time w_i. In equilibrium, each patient will choose his most preferred hospital given his intrinsic preferences and the waiting times. The payer thus computes the waiting times and the number of treatments authorized for each hospital, so that in equilibrium the budget constraint is satisfied and the social welfare is maximized.

We show that even if the patients' preferences are known to the payer, the task of optimizing social welfare in equilibrium subject to the budget constraint is NP-hard, which gives strong evidence that such a computation requires at least exponential time and is beyond the ability of the payer's computation power. We also show that, with constant number of hospitals, if the budget constraint can be relaxed from B to $(1+\epsilon)B$ for an arbitrarily small constant ϵ, then the original optimum under budget B can be approximated very efficiently.

We further investigate the optimization problem over a much larger class of mechanisms that contains the equilibrium ones as special cases. In the setting with two hospitals, we show that under a natural assumption on the patients' preference profiles, optimal welfare is in fact attained by the randomized assignment mechanism, which allocates patients to the hospitals at random subject to the budget constraint, but avoids waiting times.

Finally, we discuss potential policy implications of our results, as well as follow-up directions and open problems.

Keywords: Healthcare, Mechanism design, Unit demand, Budget constraint, Waiting times

*The authors would like to thank Itai Ashlagi for pointing us to important references. The first author is supported by the Alfred P. Sloan Fellowship, an NSF CAREER award (CCF-1149888), NSF Award CCF-1215990, and a Turing Centenary Fellowship. The second author is supported by the Zurich Financial Services and NSF grant CCF-0832797. The third author is supported by NSF Award CCF-1137084. Part of this research was done when the third author was visiting Princeton University.

1 Introduction

In this paper we initiate the study of computational mechanism design in the context of optimal healthcare provision. Specifically, we consider the setting where waiting times, and not payments, are used to allocate scarce care resources among patients. Waiting times in healthcare provision is an important topic of public debate worldwide. For example, it has a central role in the ongoing debate surrounding the Patient Protection and Affordable Care Act ("Obamacare") in the United States. In a large number of countries with public health coverage financing, including Australia, Canada, Spain, and the United Kingdom, procedures such as elective surgery are rationed by waiting [22, 9]. While in the public perception waiting times are often associated with poor resource management, in the economics literature it is well-understood that queues of consumers will form whenever a good is priced below the good's perceived value, as long as supply is scarce [4, 17, 12] – independently of the ultimate distribution mechanism. In particular, waiting times in this context are dictated by economic incentive constraints and not by stochastic fluctuations as in classical queuing theory. Therefore, whenever "correct" monetary pricing is impossible or undesirable, waiting times should be incorporated explicitly into the allocation models.

There are two main distinct features in the study of resource allocation with waiting times. Firstly, if money is still involved, the setting leads to two non-interchangeable "currencies" of money and waiting time which feature in the design. This complicates the design problem, both conceptually and computationally. As we will see from our main results, even if one keeps money and waiting time separate, the fact that they cannot be "traded" for each other (thus reducing the setup to one currency) makes the problem much more difficult. Secondly, unlike monetary transfers, nobody benefits from one's waiting time, and thus waiting times represent a net loss in utility. In this regard the situation is similar to money-burning mechanisms [10], subject to the important caveat that time burnt is not interchangeable with money.

We focus on a setting where a population of patients is seeking a single health-care procedure (such as a surgery) from one of k hospitals. The procedure is fully financed by a third party —a "payer", e.g., the government or an insurer— and thus the only cost directly incurred by patients is the waiting time. We assume that the waiting time is the same for all patients treated in the same hospital. We also assume that waiting times are stable and known to the patients before they join one of the queues[1]. Each patient P_j has an estimate v_{ij} on his utility from being treated in hospital H_i if he is treated immediately. The patient is unrestricted in his choice of hospital. Thus, in equilibrium, he chooses the hospital that maximizes his utility given the waiting times w_i. Similarly to [8] we assume patients have quasi-linear utilities with respect to waiting time: $u_{ij} = v_{ij} - w_i$. The primary reason for this choice is that it is the most natural way to ensure that patients are treated equally by welfare-optimizing mechanisms. Since, as the mechanism designers, we do not have full access to the u_{ij}'s of individual patients but can observe waiting times, our welfare-loss due to waiting will just be the sum of all the waiting times in the system[2].

We assume that the payer has a fixed budget B it is willing to spend on providing the procedure to the entire patient population. The payer carries all the burden for financing the procedure. The cost of treating a patient at hospital H_i is c_i. Without loss of generality, we assume that the payer has enough budget to treat all patients in *some* hospital. This can always be achieved by adding a dummy hospital H_0 with zero cost, zero waiting time, and zero valuation by all patients, representing the option of not getting any treatment. The payer cannot deny care to any individual patient, however, it can decide how many treatment slots λ_i it is willing to fund in hospital H_i[3].

[1] This happens for example, if a patient signs up for a procedure at a fixed date a few months down the road.

[2] We can relax this assumption to allow utility functions of the form $u_{ij} = v_{ij} - U(w_i)$, where $U(w)$ is a function (common to all patients) that maps waiting time w to disutility caused by waiting w time units.

[3] It is natural to think of λ_i as a rate (e.g. in slots per year) at which the payer is willing to fund treatment in H_i.

The only constraints on the payer are the budget constraint and the requirement that the total number of available treatment slots is at least as high as the number of patients. Subject to the budget constraint, the goal of the payer is to maximize the aggregate welfare of the patients.

Patients are free to choose their hospitals. Therefore, once the capacities λ_i have been fixed, patients' hospital choices and waiting times will be such that the system is in equilibrium. In fact, given the capacities, the situation becomes quite similar to the setting of multi-item unit-demand auctions, which is very well understood [6, 7] (where patients are "buying" treatment using waiting time). In particular, among all equilibrium waiting times and assignments, there is a patient-optimal one that simultaneously gives each patient the best utility attainable by this patient in *any* equilibrium solution, as well as the lowest total waiting time. Therefore, we will assume that this is the equilibrium that gets realized when the capacities are fixed. We study and try to optimize the total welfare in this equilibrium.

Given that waiting times are a "dead weight" in terms of welfare, it is very natural to ask whether they can be avoided or reduced via a different allocation mechanism altogether. If monetary payments are not allowed, and patients are free to choose their hospitals, then the equilibrium discussed above is the only one possible. What if the payer has sufficient control over the patients that it can tell them where to receive their treatment, or otherwise restrict their options[4]? The simplest such mechanism would be a randomized assignment of patients to available slots. In such assignment, we benefit from zero waiting times. On the downside, we incur an efficiency loss: patients may not end up in the hospitals they prefer. How does this randomized assignment mechanism compare to the mechanism where patients are given a free choice and waiting times are used as a rationing tool? The answer to this question depends on the preference profiles of the patients. Informally speaking, if patients have strong *and diverse* preferences on where to be treated, then the free-choice equilibrium mechanism is better, since efficiency gains due to better allocation offset the inefficiency caused by waiting. At the other extreme, if all patients have similar preferences, then no efficiencies are to be gained from patients' choice, and randomized assignment mechanisms are superior. We further investigate this question in the case of two hospitals.

1.1 Main results

Finding the optimal equilibrium assignment

We first consider the setting where the payer decides the capacities for different hospitals, and the assignments and waiting times that are realized are based on the (patient-optimal) equilibrium solution to the resulting unit-demand problem. While the problem of finding the equilibrium given the capacities is tractable, it turns out that finding capacities that maximize patients' total utility in equilibrium is NP-hard[5]:

THEOREM 1. *Finding optimal equilibrium assignments is NP-hard.*

The hardness result motivates one to ask whether one can efficiently approximate the welfare of the optimal assignment. In fact, we show that if we relax the budget constraint to $(1+\epsilon)B$ with an arbitrarily small constant ϵ, we can achieve at least as much welfare as the best B-budget equilibrium assignment, using an algorithm whose running time is exponential in the number of hospitals, but is polynomial in the number of patients if the number of hospitals is constant:

[4]Possible "soft" mechanisms for doing this are discussed below.

[5]The notion of NP-hardness is closely related to the most fundamental conjecture in Computer Science, that is, $NP \neq P$. Showing that a computation problem is NP-hard is the "standard" way of establishing intractability of that problem, and gives strong evidence that to solve it requires time at least exponential in the input size and is beyond the ability of today's computation power. See [23, 5, 19] for introductions to NP-hardness and its possible relevance for Economics.

THEOREM 2. (rephrased) *Let k be the number of hospitals and m the number of patients. There is an algorithm that runs in time $O\left((\log_{1+\epsilon} m)^k \cdot m^4\right)$ and outputs capacities $\{\lambda_i\}_{i=1}^k$ for the hospitals and the corresponding equilibrium assignment such that, the total cost is $\sum_{i=1}^k c_i\lambda_i \leq (1+\epsilon)B$, and the social welfare is at least as high as that of the optimal equilibrium assignment with budget B.*

It remains an interesting open problem whether there is a welfare approximation scheme that does not exceed the budget. Also, it is unknown whether there is an approximation scheme that is polynomial in k.

When is the randomized assignment optimal?

We next turn our attention to the expanded setting where we are not limited to mechanisms that produce equilibrium solutions. The two "extreme" mechanisms are the equilibrium mechanism discussed above, and the randomized assignment mechanism (which assigns patients at random to available slots). In addition, there is an infinite number of various lotteries in-between these extremes, where the patient is presented with a choice over distributions over hospitals, along with an expected waiting time associated with each distribution. Intuitively, if there are no extreme variations among the patients' preferences, the randomized assignment should outperform other mechanisms, since it avoids the dead-weight cost of waiting times. We give further evidence that suggests that, welfare-wise, randomized assignment may be superior by analyzing the case when $k = 2$, i.e. when there are two hospitals.

In the two-hospital setting, with hospitals H_0 and H_1 such that the cost $c_0 < c_1$, we assume without loss of generality that no patient choosing hospital H_0 is facing any waiting time. Thus patients who prefer H_0 over H_1 will always choose H_0. We can therefore exclude them from consideration, and focus on patients who prefer H_1 over H_0. Further, in this case, λ_0 and λ_1 are completely determined by the budget constraint, and the only problem is to design a mechanism for allocating patients who prefer H_1 to the λ_1 available slots. Each patient x who prefers H_1 is associated with a value $v(x)$ representing how much time x is willing to wait to get treatment in H_1 instead of H_0. We assume a continuous distribution of patients, index them by the $[0, 1]$ interval, and rename them so that $v(x)$ is a non-decreasing function. Thus, for example, $v(0.5)$ represents the median time that patients preferring H_1 are willing to wait to be treated there. We prove:

THEOREM 3. (rephrased) *If $v(x)$ is concave, then no lotteries offered to the patients can beat the randomized assignment mechanism.*

Here by a lottery we mean a set of options each consisting of a probability of getting treated in H_1 (as opposed to H_0), and the waiting time there. We assume that patients are risk-neutral. This shows that for a broad class of preferences, the randomized assignment is welfare-maximizing even when waiting times are an option available to the payer. As a special case, this shows that randomized assignment has better welfare than the optimal equilibrium assignment. It would be interesting to find an analogous sufficient condition for three or more hospitals.

1.2 Discussion and open problems

In this paper we consider two separate issues. The first one is how to use waiting times to optimally allocate treatments in equilibrium when the payer faces budget constraints. The second one is whether it may be beneficial to do away with the equilibrium requirements (if possible).

While finding the optimal equilibrium assignment is NP-hard, our approximation result suggests that this problem might not be as difficult in practice. In many cases the number of treatment facilities involved is fairly small, making running time exponential in k feasible. Even better, in some cases the "hospitals" are actually treatment alternatives that vary in cost and in waiting time (e.g. physiotherapy is cheaper and more readily accessible than knee replacement), in which case

k may be as low as 2. For the general case where k can be big, it would be interesting to explore restrictions on the valuation matrix of the patients, such as an upper bound on the rank, that would make the exact optimization efficient.

Besides money-burning mechanisms [10], equilibrium allocation with waiting times has a strong connection to unit-demand auctions [6, 1], making the step of computing the optimal equilibrium waiting times based on capacities and patients' preferences a computationally feasible task. One natural question is whether this clean connection can be used in dynamic setting to show that the system will remain in the patient-optimal equilibrium as the population's preferences slowly shift over time. Another related question is whether it is possible to approximate optimal welfare in equilibrium if the payer only knows the approximate distribution of patient types in the population.

The study of waiting times as a rationing mechanism is closely related to the study of ordeal mechanisms [2], where other tools (e.g. excessive bureaucracy) are used in place of waiting times to reduce demand to the supply level[6]. These may be used in settings where queues are not an option such as school choice. Developing algorithmic mechanism design tools for these settings is a very interesting direction of study.

Our second result looks beyond equilibrium solutions. We give evidence that equilibrium solutions are in fact dominated in many cases. One immediate implication is that giving the payer power to restrict choice may in fact improve overall welfare. While this is perhaps not surprising, choice restriction may be very difficult or politically infeasible to implement in practice, due to the fact that patients have an inherent preference for choice [20].

There are important indirect ways, however, in which a payer (especially a government payer) may influence choice. One of them is through release (or non-release) of quality of care information about providers. The topic of quality of care information is an important one, both in theory and in practice. In the United States, for example, Medicare has started to publicly release hospital performance information as part of its pay-for-performance push [13]. The effect performance reporting has on *provider* incentives has been the subject of much study and discussion [21, 16]. It has even been suggested that it would be possible to manipulate reported quality metrics in a way that would force the provider to exert first-best quality and cost effort [18]. To the best of our knowledge, there has been no work on the effect of quality reporting on patient choice.

Inasmuch as quality information influences patients' choices, it may actually cause harm in the context of allocation mechanisms with waiting times. Consider a scenario where there are two hospitals, a good one H_g and a bad one H_b. All patients prefer the good hospital over the bad by the same amount, but they do not know which is which. As a result, both hospitals will receive half the patients, and waiting time will be zero. If the payer reveals that H_g is the good hospital through its quality of care disclosure, then all patients will prefer H_g over H_b by the same amount Δ. Unless H_g has enough slots for everybody, a queue of length Δ will form, making everyone worse-off than when they were ignorant. In effect, before the quality disclosure, uninformed patients implemented the randomized assignment – through free choice. Once the quality information was disclosed, the game moved to the equilibrium solution.

Our results indicate that in some cases a population of more informed patients will experience higher waiting times and lower overall utility than uninformed patients. This suggests an unfortunate potential side effect of information disclosure in cases where allocation is done by waiting times. Such a side effect deserves further study since, at the moment, quality information release is regarded as an absolute good. Understanding optimal information structures of information released to the patients in terms of overall welfare (as well as provider-side incentives) is an important and interesting direction of study.

[6]Note that in medicine not all ordeals are necessarily dead-weight loss. For example, the famous (and highly-demanded) Shouldice hernia clinic in Ontario, Canada requires its patients to lose weight before being admitted for a surgery [11], most clinics do not place such a requirement.

2 Our Model

We focus on the provision of a single healthcare service to a population of patients. Several hospitals provide the same service at different costs, and each patient has his own valuation for being served at each hospital immediately. A payer, which is taken to be the government from now on, covers the costs for all patients, so a patient's utility does not depend on the costs. But due to its budget constraint the government cannot simply let each patient choose his favorite hospital. Thus it uses waiting times to provide incentives for the patients to go to different hospitals. Among all choices of waiting times that meet the budget constraint, the government is interested in finding the optimal one that maximizes social welfare, that is, the total utility of the patients. More precisely,

- The set of *hospitals* is $\{H_1, \ldots, H_k\}$.
- For each $i \in [k]$, the *cost* of H_i serving one patient is $c_i \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ is the set of non-negative integers.
- The set of *patients* is $\{P_1, \ldots, P_m\}$.
- For each $i \in [k]$ and $j \in [m]$, the *value* of patient P_j for hospital H_i is $v_{ij} \in \mathbb{Z}^+$.
- An *assignment* of the patients to the hospitals is a triple (λ, w, h), where $\lambda = (\lambda_1, \ldots, \lambda_k) \in \{1, \ldots, m\}^k$ with $\sum_{i\in[k]} \lambda_i = m$ is the *capacity vector* of the hospitals, $w = (w_1, \ldots, w_k) \in (\mathbb{Z}^+)^k$ is the *waiting time vector*, and $h : [m] \to [k]$ is the *assignment function* such that $|h^{-1}(i)| = \lambda_i$ for each $i \in [k]$.

  According to such an assignment, patient P_j will receive the service at hospital $H_{h(j)}$ after waiting time $w_{h(j)}$.
- A patient P_j's *utility* under assignment (λ, w, h) is $u_j(\lambda, w, h) \triangleq v_{h(j)j} - w_{h(j)}$, that is, quasi-linear in the waiting time.

  The *social welfare* of this assignment is $SW(\lambda, w, h) \triangleq \sum_{j\in[m]} u_j(\lambda, w, h)$.
- The government has *budget* $B \in \mathbb{Z}^+$, and an assignment (λ, w, h) is *feasible* if $\sum_{i\in[k]} \lambda_i \cdot c_i \le B$.

  For the problem to be interesting, we assume that $mc_{\min} < B < mc_{\max}$, where $c_{\min}$ and $c_{\max}$ are respectively the minimum and the maximum cost of the hospitals.

Remark 1. *The hospitals' costs, the patients' valuations, and the waiting times are assumed to be integers without loss of generality. As long as they have finite description, we can always choose proper units so that all of them are integers.*

Remark 2. *The capacity λ_i of hospital H_i is the maximum number of patients that H_i is* allowed *to serve in a time unit, say a month. It may be different from H_i's real capacity, namely, the maximum number of patients that H_i is able to serve. The latter is assumed to be large enough. But due to the government's budget constraint, some hospitals may not be saturated. Accordingly, the w_i's do not directly depend on the λ_i's or the real capacities as in classical queuing theory. Instead they are indirectly used as a way of controlling the patients' incentives for going to different hospitals.*

Since in reality the government may not be able or willing to force a patient to go to a hospital assigned to him, it must ensure that wherever it wants that patient to go is indeed the best hospital for him, given the waiting times. Accordingly, we have the following definition.

Definition 1. *Assignment* (λ, w, h) *is an* equilibrium assignment *if: (1) it is feasible, (2) for each* $j \in [m]$ *we have* $u_j(\lambda, w, h) \geq 0$*, and (3) for each* $j \in [m]$ *and* $i \in [k]$ *we have*

$$u_j(\lambda, w, h) \geq v_{ij} - w_i.$$

Assignment (λ, w, h) *is an* optimal equilibrium assignment *if: (1) it is an equilibrium assignment, and (2) for any other equilibrium assignment* (λ', w', h')*,*

$$SW(\lambda, w, h) \geq SW(\lambda', w', h').$$

The social welfare of optimal equilibrium assignments is denoted by SW_{OEA}*.*

As we are interested in the (existence and) computation of optimal equilibrium assignments, we assume that the government has precise knowledge about the cost of each hospital. We may also assume that the government knows each patient's valuation for each hospital, but we do not need it. In fact, it is enough for the government to know the "distribution" of the k-dimensional valuation vectors of the patients, namely, the fraction of the patients having each particular valuation vector. Once it computes λ and w in the optimal solution, the assignment function h will be automatically implemented by the patients going to their favorite hospitals[7], and the government need not know where each patient is going. Notice that it is not sufficient for the government to know the distribution of the valuations for each single hospital, since the correlations between valuations for different hospitals will affect the outcome.

3 The Computational Complexity of Optimal Equilibrium Assignments

We begin with three observations about our model, as a warm-up.

The first observation is that, if the patients have unanimous preferences, namely, $v_{ij} = v_{ij'}$ for each $i \in [k]$ and each j, $j' \in [m]$, then no equilibrium assignment can improve the social welfare of the following trivial one: order the hospitals according to the patients' valuations decreasingly, find the first hospital H_i such that $mc_i \leq B$, and assign all patients to H_i with $w_i = 0$ and $w_{i'} = \max_{i'' \in [k]} v_{i''1}$ for any $i' \neq i$. Indeed, for any equilibrium assignment (λ, w, h) we have $v_{h(j)j} - w_{h(j)} = v_{h(j')j} - w_{h(j')}$ for each $j, j' \in [m]$. Letting $i^* = \text{argmin}_{i:h^{-1}(i)\neq\emptyset} c_i$, λ' be such that $\lambda'_{i^*} = m$ and $\lambda'_i = 0$ for all other i, h' be such that $h'(j) = i^*$ for all j, we have that (λ', w, h') is another equilibrium assignment with the same social welfare as (λ, w, h). Thus it suffices to look for an optimal equilibrium assignment that sends all patients to the same hospital. This is also intuitive: if the patients are all the same, then at equilibrium the government must make them equally happy, and it can do so by treating them in the same way.

Another observation is that, even if the government only cares about meeting the budget constraint in expectation, and is allowed to assign each patient to several hospitals probabilistically (with the total probability summing up to 1), the optimal social welfare it can get in expectation will just be the same as the optimal one obtained by deterministic assignments. This is so because, at equilibrium, all the hospitals to which a patient P_j is assigned with positive probability must yield the same utility for him. Thus assigning P_j deterministically to the one with the smallest cost leads to another equilibrium assignment with the same social welfare and still meeting the budget constraint. Accordingly, to maximize social welfare it suffices to consider only deterministic assignments.

[7] If there are more than one favorite hospitals for some patients, then we need to assume that ties are broken in a proper way so that the budget constraint is satisfied.

Finally, for any capacity vector λ meeting the budget constraint, the problem of finding equilibrium assignments with λ reduces to finding equilibrium prices in multi-item unit-demand auctions, which has been well studied —see, e.g., [6, 1, 3, 7]. More precisely, each hospital H_i corresponds to λ_i identical goods $H_{i1}, \ldots, H_{i\lambda_i}$, each patient P_j corresponds to a buyer, and the value of P_j for each good H_{ir} is v_{ij}. Each patient only wants a single good (i.e., being served once). It is well known that such an auction always has a set of equilibrium prices, which can be found by the Hungarian method [15]. It is then intuitive that such prices will correspond to the waiting times of the hospitals, and the allocation of the goods will correspond to the assignment function. The only caution is that, the goods may have different prices, but for a hospital to have a well-defined waiting time the prices of its corresponding goods must be all the same. Fortunately from the details of the Hungarian method one can see that this is indeed the case. Therefore for each capacity vector λ (whether it meets the budget constraint or not), there exists an equilibrium assignment with λ. In fact the optimal assignment with λ can be computed efficiently, following the result of [1]. We shall get back to this in Section 4.

Although equilibrium assignments can be efficiently found given the capacity vector λ, the problem of deciding the "correct" λ makes optimal equilibrium assignments hard to find, even in very special cases. (Notice that it does not suffice to solve a multi-item auction instance where each hospital corresponds to m identical goods, as the allocation may lead to an assignment violating the budget constraint.)

Theorem 1. *Finding optimal equilibrium assignments is NP-hard.*

Proof. The proof is by reduction. Namely, we show that if we can solve the optimal equilibrium assignment problem then we can solve another NP-hard problem in about the same time, concluding that the former is also NP-hard.

The reduction is from the classical knapsack problem [14], which is well known to be NP-hard. In this problem there are k items, $a_1, \ldots, a_k$, and each a_i has value v_i and cost c_i. We are also given a budget B, and the goal is to select a subset of items so as to maximize their total value while keeping their total cost less than or equal to B.

We can transform this problem to an assignment problem with $k+1$ hospitals and k patients. Each hospital H_i with $i \leq k$ has cost c_i, and each patient P_i has value v_i for H_i and 0 for all others. Hospital H_{k+1} has cost 0 and is valued 0 by all patients. The government has budget B.

Given an equilibrium assignment (λ, w, h) for our problem, we can construct a solution to the knapsack problem with total value equal to $SW(\lambda, w, h)$ —the set $A = \{i : h(i) = i\}$ is such a solution. Indeed, without loss of generality we can assume $h(i) = k+1$ whenever $h(i) \neq i$. By the definition of equilibrium assignments, we have $w_{k+1} = 0$, $w_i = v_i$ if $h(i) = k+1$, and $w_i = 0$ otherwise. Thus $SW(\lambda, w, h) = \sum_{i \in A} v_i$, which is the total value of A in the knapsack problem. As the total cost of (λ, w, h) is $\sum_{i \in A} c_i \leq B$, the set A meets the budget constraint in the knapsack problem.

It is easy to see that the other direction is also true, that is, given a solution $A \subseteq [k]$ to the knapsack problem, we can construct an equilibrium assignment (λ, w, h) whose social welfare equals the total value of A.

Accordingly, an optimal equilibrium assignment in our problem corresponds to an optimal solution to the knapsack problem. □

Remark 3. *The NP-hardness of the knapsack problem comes from the need for integrality. Its fractional version can be easily solved using a greedy bang-per-buck approach. But this is not the case in our problem. Indeed, as we have noted, given a fractional equilibrium assignment we can construct a deterministic equilibrium assignment with the same social welfare. Thus for our problem the fractional version is as hard as the integral version.*

4 Approximating Optimal Equilibrium Assignments with Arbitrarily Small Deficit

Although the optimization problem is hard when both the numbers of patients and hospitals are large, in practice we expect the number of hospitals to be small. Therefore it makes sense to assume that this number is a constant and try to solve the problem efficiently in this case.

Our first observation is that optimal equilibrium assignments can be found in time $O(m^k \text{poly}(m,k))$. This is so because, there are at most m^k possible assignment functions $h : [m] \to [k]$. For each h and the corresponding capacity vector λ which can be inferred from h, we can find the best waiting time vector w using the following linear programming (or prove that no feasible waiting time vector exists):

$$\begin{aligned} & \min_{w} \sum_{i\in[k]} w_i\lambda_i \\ \text{s.t.} \quad & \forall j \in [m], i \in [k], v_{h(j)j} - w_{h(j)} \geq v_{ij} - w_i, \\ & \sum_{i\in[k]} c_i\lambda_i \leq B. \end{aligned}$$

We then choose the assignment (λ, w, h) maximizing social welfare.

Given the above observation, we are interested in replacing the m^k part with a better bound. As we shall show, if the government is willing to violate its budget constraint by an arbitrarily small fraction, then the problem can be solved much more efficiently.

Definition 2. *Let ϵ be a positive constant. An assignment (λ, w, h) is an* equilibrium assignment with ϵ-deficit *if it is an equilibrium assignment with the feasibility condition replaced by the following condition: $\sum_{i\in[k]} \lambda_i c_i \leq (1+\epsilon)B$.*

We shall construct an algorithm that, in time $O(\log^k_{1+\epsilon} m \cdot (1+\epsilon)^3 m^4)$, finds an equilibrium assignment with ϵ-deficit whose social welfare is at least SW_{OEA}, the social welfare of the optimal equilibrium assignments.

4.1 A useful result in multi-unit auctions

Our algorithm uses that of [1] for multi-unit auctions as a black box, therefore we first recall their result (while using our notation to help establish the connection with our results).

Definition 3. *A* multi-unit auction*, or simply an* auction *in this paper, is a triple (g, m, v), where the set of goods is $\{1, 2, \ldots, g\}$, the set of bidders is $\{1, 2, \ldots, m\}$, and v is the* valuation matrix*, that is, a $g \times m$ matrix of non-negative integers. Each v_{ij} denotes the valuation of bidder j for good i.*

Given an auction (g, m, v), a matching *is a triple (u, p, μ), where $u = (u_1, \ldots, u_m) \in (\mathbb{Z}^+)^m$ is the* utility vector*, $p = (p_1, \ldots, p_g) \in (\mathbb{Z}^+)^g$ is the* price vector*, and $\mu \subseteq [g] \times [m]$ is a set of good-bidder pairs such that no bidder and no good occur in more than one pair. Bidders and goods that do not appear in any pair in μ are* unmatched*.*

Definition 4. *Given an auction (g, m, v), a matching (u, p, μ) is* weakly feasible *if for each $(i, j) \in \mu$ we have $u_j = v_{ij} - p_i$, and for each unmatched bidder j we have $u_j = 0$.*

A matching (u, p, μ) is feasible *if it is weakly feasible and for each unmatched good i we have $p_i = 0$.*

A matching (u, p, μ) is stable *if for each $(i, j) \in [g] \times [m]$ we have $u_j \geq v_{ij} - p_i$.*

A matching (u^*, p^*, μ^*) is bidder-optimal *if: (1) it is stable and feasible, and (2) for every matching (u, p, μ) that is stable and weakly feasible, and for every bidder j, we have $u^*_j \geq u_j$.*

In [1] the authors construct an algorithm, StableMatch, which given an auction (g, m, v), outputs a bidder-optimal matching (u^*, p^*, μ^*) in time $O(mg^3)$.

Notice that the original definitions in [1] have for each good-bidder pair a reserve price and a maximum price. In our model we do not need them, so the definitions above are more succinct than the original ones. In fact, as pointed out by [1], with maximum prices, there may be no bidder-optimal matching. But without them such a matching always exists, as shown by [6].

Notice also that [1] does not distinguish between weak feasibility and feasibility. But it is easy to see that their algorithm and its analysis still apply under our definitions. We shall use this difference in analyzing our algorithm.

Next we establish two properties for the matching (u^*, p^*, μ^*) output by StableMatch.

- *Property 1.* If $g \geq m$, then without loss of generality we can assume that (u^*, p^*, μ^*) has no unmatched bidder.

  Indeed, if there exists an unmatched bidder j, then there must exist an unmatched good i (since $g \geq m$). Since (u^*, p^*, μ^*) is bidder-optimal, we have $u_j^* = 0$, $p_i^* = 0$, and $u_j^* \geq v_{ij} - p_i^*$. Thus we have $v_{ij} = 0$, and the matching $(u^*, p^*, \mu^* \cup \{(i, j)\})$ is another bidder-optimal matching.

- *Property 2.* If two goods i, i' are identical, namely, $v_{ij} = v_{i'j}$ for each bidder j, then $p_i^* = p_{i'}^*$.

  Indeed, if both goods are unmatched then $p_i^* = p_{i'}^* = 0$. Otherwise, say $(i, j) \in \mu^*$. By definition, $u_j^* = v_{ij} - p_i^* \geq v_{i'j} - p_{i'}^*$. As $v_{ij} = v_{i'j}$, we have $p_i^* \leq p_{i'}^*$. If i' is unmatched then $p_{i'}^* = 0$, implying $p_i^* = 0$. If $(i', j') \in \mu^*$ then similarly we have $p_{i'}^* \leq p_i^*$, and thus $p_i^* = p_{i'}^*$ again.

4.2 Our algorithm for approximating optimal equilibrium assignments

Now we are ready to construct our algorithm for approximating optimal equilibrium assignments. The algorithm takes as input the number of patients m, the number of hospitals k, the hospitals' costs $c_1, \ldots, c_k$, the patients' valuations v_{ij}'s for the hospitals, the budget B, and a small constant $\epsilon > 0$. Letting (λ, w, h) be an optimal equilibrium assignment, the algorithm works by guessing λ, constructing a multi-unit auction based on the guessed vector, computing the bidder-optimal matching using StableMatch, and extracting the waiting time vector and the assignment function from the matching.

More precisely, let $L \triangleq \lceil \log_{1+\epsilon} m \rceil$, $C_0 \triangleq 0$, and $C_\ell \triangleq \lfloor (1+\epsilon)^\ell \rfloor$ for each $\ell = 1, \ldots, L$. The algorithm examines all the vectors $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_k) \in \{C_0, C_1, \ldots, C_L\}^k$ one by one, say lexicographically.

If $\sum_{i \in [k]} \hat{\lambda}_i \notin [m, (1+\epsilon)m]$ or $\sum_{i \in [k]} \hat{\lambda}_i c_i > (1+\epsilon)B$, the algorithm disregards this vector and moves to the next. Otherwise it constructs an auction $(g, m, \hat{v})$ as follows. The set of patients corresponds to the set of bidders; each hospital H_i corresponds to $\hat{\lambda}_i$ copies of identical goods $H_{i1}, \ldots, H_{i\hat{\lambda}_i}$, thus $g = \sum_{i \in [k]} \hat{\lambda}_i$; the valuation matrix $\hat{v}$ has rows indexed by $\{ir : i \in [k], r \in [\hat{\lambda}_i]\}$, columns indexed by $[m]$, and for each $j \in [m]$, $i \in [k]$, and $r \in [\hat{\lambda}_i]$, $\hat{v}_{ir,j} = v_{ij}$.

The algorithm then runs StableMatch with input $(g, m, \hat{v})$ to generate the bidder-optimal matching (u^*, p^*, μ^*), and extracts the waiting time vector $\hat{w}$ and the assignment function $\hat{h}$ as follows. For each hospital H_i, let $\hat{w}_i = p_{i1}^*$. For each patient P_j, let H_{ir} be the unique good to which P_j is matched (by Property 1 in Section 4.1 such a good always exists) according to μ^*, and let $\hat{h}(j) = i$. The triple $(\hat{\lambda}, \hat{w}, \hat{h})$ may not be an assignment as $\sum_{i \in [k]} \hat{\lambda}_i$ may be larger than m, but there is a unique capacity vector $\hat{\lambda}'$ such that $(\hat{\lambda}', \hat{w}, \hat{h})$ is an assignment.

The algorithm computes the social welfare of the assignment $(\hat{\lambda}', \hat{w}, \hat{h})$ for each $\hat{\lambda}$ that is not disregarded, and output the assignment (λ^*, w^*, h^*) with the maximum social welfare.

We prove the following theorem.

Theorem 2. *Our algorithm runs in time $O(\log^k_{1+\epsilon} m \cdot (1+\epsilon)^3 m^4)$, and outputs an equilibrium assignment with ϵ-deficit, (λ^*, w^*, h^*), such that $SW(\lambda^*, w^*, h^*) \geq SW_{OEA}$.*

Proof. The running time of the algorithm can be immediately seen. Indeed, if a vector $\hat{\lambda}$ is not disregarded, then it takes $O(mg) = O((1+\epsilon)m^2)$ time to construct the auction as $g \in [m, (1+\epsilon)m]$, $O(mg^3) = O((1+\epsilon)^3 m^4)$ time to run StableMatch, and $O(m)$ time to extract the assignment. Accordingly, it takes $O((1+\epsilon)m^2 + (1+\epsilon)^3 m^4 + m) = O((1+\epsilon)^3 m^4)$ time to examine a single vector $\hat{\lambda}$, and there are $O(\log^k_{1+\epsilon} m)$ vectors in total.

The remaining part of the theorem follows from the two lemmas below.

Lemma 1. *(λ^*, w^*, h^*) is an equilibrium assignment with ϵ-deficit.*

Proof. In fact, we show that for each vector $\hat{\lambda}$ that is not disregarded, the extracted assignment $(\hat{\lambda}', \hat{w}, \hat{h})$ is an equilibrium assignment with ϵ-deficit. To see why this is true, first notice that $\sum_{i\in[k]} \hat{\lambda}_i c_i \leq (1+\epsilon)B$ by the construction of the algorithm, thus

$$\sum_{i\in[k]} \hat{\lambda}'_i c_i \leq \sum_{i\in[k]} \hat{\lambda}_i c_i \leq (1+\epsilon)B. \tag{1}$$

Second, for each $j \in [m]$, letting $H_{\hat{h}(j)r}$ be the good matched to P_j according to μ^*, we have

$$u_j(\hat{\lambda}', \hat{w}, \hat{h}) = v_{\hat{h}(j)j} - \hat{w}_{\hat{h}(j)} = \hat{v}_{\hat{h}(j)r,j} - p^*_{\hat{h}(j)1} = \hat{v}_{\hat{h}(j)r,j} - p^*_{\hat{h}(j)r} = u^*_j \geq 0, \tag{2}$$

where the third equality is because of Property 2 in Section 4.1 (in particular, $H_{\hat{h}(j)1}$ and $H_{\hat{h}(j)r}$ are identical goods, and $p^*_{\hat{h}(j)1} = p^*_{\hat{h}(j)r}$), and the other equalities/inequality are by definition.

Third, since (u^*, p^*, μ^*) is a bidder-optimal matching for auction $(g, m, \hat{v})$, we have that for each $j \in [m]$, $i \in [k]$, and $r \in [\hat{\lambda}_i]$,

$$u^*_j \geq \hat{v}_{ir,j} - p^*_{ir} = v_{ij} - p^*_{i1} = v_{ij} - \hat{w}_i,$$

and thus

$$u_j(\hat{\lambda}', \hat{w}, \hat{h}) = u^*_j \geq v_{ij} - \hat{w}_i. \tag{3}$$

Equations 1, 2, and 3 together imply that every $(\hat{\lambda}', \hat{w}, \hat{h})$ is an equilibrium assignment with ϵ-deficit, and so is (λ^*, w^*, h^*). □

Lemma 2. *$SW(\lambda^*, w^*, h^*) \geq SW_{OEA}$.*

Proof. To see why this is true, arbitrarily fix an optimal equilibrium assignment (λ, w, h). Notice that for each hospital H_i, there exists a "good guess" $\hat{\lambda}_i \in \{C_0, \ldots, C_L\}$ such that

$$\lambda_i \leq \hat{\lambda}_i \leq (1+\epsilon)\lambda_i.$$

Since λ satisfies $\sum_{i\in[k]} \lambda_i = m$ and $\sum_{i\in[k]} \lambda_i c_i \leq B$, the vector $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ satisfies

$$\sum_{i\in[k]} \hat{\lambda}_i \in [m, (1+\epsilon)m] \quad \text{and} \quad \sum_{i\in[k]} \hat{\lambda}_i c_i \leq (1+\epsilon)B.$$

Thus it won't be disregarded by the algorithm. Let $(g, m, \hat{v})$ be the auction constructed from $\hat{\lambda}$, (u^*, p^*, μ^*) the output of StableMatch under input $(g, m, \hat{v})$, and $(\hat{\lambda}', \hat{w}, \hat{h})$ the assignment extracted from (u^*, p^*, μ^*). Following the same reasoning as in Equation 2, we have that for each $j \in [m]$, $u_j(\hat{\lambda}', \hat{w}, \hat{h}) = u_j^*$. Thus

$$SW(\hat{\lambda}', \hat{w}, \hat{h}) = \sum_{j \in [m]} u_j^*. \tag{4}$$

From (λ, w, h), we construct a matching (u, p, μ) for the auction $(g, m, \hat{v})$ as follows. For each bidder j, we have $u_j = v_{h(j)j} - w_{h(j)}$; for each good H_{ir} with $i \in [k]$ and $r \in [\hat{\lambda}_i]$, we have $p_{ir} = w_i$; and for each hospital H_i, letting $j_1 \leq j_2 \leq \cdots \leq j_{\lambda_i}$ be the patients assigned to H_i by h, we have $\mu = \{(j_r, ir) : i \in [k], r \in [\lambda_i]\}$.

It is easy to verify that the so constructed (u, p, μ) is stable and weakly feasible, thus by the optimality of u^* we have that for each $j \in [m]$,

$$u_j^* \geq u_j. \tag{5}$$

Moreover, for the same reason as Equation 4, we have

$$SW(\lambda, w, h) = \sum_{j \in [m]} u_j. \tag{6}$$

Equations 4, 5, and 6 together imply

$$SW(\hat{\lambda}', \hat{w}, \hat{h}) \geq SW(\lambda, w, h) = SW_{OEA}$$

as we want to show. □

In sum, Theorem 2 holds. □

5 The Optimality of the Randomized Assignment

Following our discussion in the Introduction, we consider the case of two hospitals, a "good" one H_1 and a "bad" one H_0. The patients are indexed by the interval $[0, 1]$, and each patient x has valuation $v(x) \geq 0$ for H_1 and 0 for H_0. We assume that the patients have been renamed and normalized, so that $v(x)$ is non-decreasing and $v(0) = 0$. Since the number of patients is infinite, we talk about the *cost density* $c_i(x)$ of each hospital, rather than the cost for serving a single patient. Without loss of generality, $c_1(x) \equiv 1$ and $c_0(x) \equiv 0$. The government has budget $B \in (0, 1)$, meaning that at most a B fraction of the patients can be served at H_1. The government's goal is to maximize the expected social welfare subject to the requirement that the budget constraint is satisfied in expectation.

In the randomized assignment, the government assigns each patient to H_1 with probability p and waiting time 0. The budget constraint gives

$$\int_0^1 p c_1(x) dx = p = B,$$

and the corresponding social welfare, denoted by SW_r, is

$$SW_r = \int_0^1 p v(x) dx = B \int_0^1 v(x) dx. \tag{7}$$

Below we compare this social welfare with that of lotteries.

Definition 5. *A* contract *is a pair (p, w), where $p \in [0,1]$ is the probability of assigning a patient to H_1, and $w \geq 0$ is the waiting time for that patient at H_1.*

A lottery *consists of a set of contracts, denoted by the domain $D \subseteq [0,1]$ of the probabilities, and the waiting time function $w(p)$ defined over D.*

Given a contract $C = (p, w)$ for patient x, the expected utility of x is

$$u(x, C) = p \cdot (v(x) - w).$$

Given a lottery $L = (D, w(p))$, each patient x chooses the contract $C(x) = (p(x), w(p(x)))$ maximizing his expected utility. Namely, for each $p \in D$,

$$u(x, C(x)) \geq u(x, (p, w(p))).$$

If there are more than one values of p that maximize the expected utility of x, we assume that $p(x)$ is the smallest one, so that the cost of serving patient x is minimized. Notice that $p(x)$ depends on x only indirectly, via the function $v(x)$: indeed, $p(x) = p(x')$ whenever $v(x) = v(x')$. Thus we can write $p(x)$ as $p(v(x))$.

As an example, the randomized assignment is a lottery with $D = \{B\}$ and $w(B) = 0$.[8] As another example, any equilibrium assignment is also a lottery, with $D = [0,1]$ and $w(p)$ always equal to the waiting time of H_1 specified by the equilibrium. Indeed, for every patient x, the contract maximizing his expected utility is to go to the hospital assigned by the equilibrium with probability 1.

Without loss of generality, we assume that D is a subinterval of $[0,1]$, denoted by $[a, b]$. Indeed, if a patient can choose between $(p_1, w(p_1))$ and $(p_2, w(p_2))$ according to the lottery, then by using a "mixed strategy" he can choose to be assigned to H_1 with any probability $p = \alpha p_1 + (1-\alpha)p_2$ with $\alpha \in [0,1]$, and corresponding expected waiting time $\alpha p_1 w(p_1) + (1-\alpha)p_2 w(p_2)$.

Also without loss of generality, we assume that the patients' expected waiting time function $p \cdot w(p)$ is convex, and thus differentiable almost everywhere. Indeed, for any contracts $C_1 = (p_1, w(p_1))$, $C_2 = (p_2, w(p_2))$, and $C = (p, w(p))$ with $p = \alpha p_1 + (1-\alpha)p_2$ for some $\alpha \in [0,1]$, if $p \cdot w(p) > \alpha p_1 w(p_1) + (1-\alpha)p_2 w(p_2)$, then a patient is always better off by mixing between C_1 and C_2 instead of choosing C. Thus we may simply assume that $p \cdot w(p) \leq \alpha p_1 w(p_1) + (1-\alpha)p_2 w(p_2)$.[9]

The social welfare and the budget constraint are naturally defined for lotteries, as follows.

Definition 6. *Given a lottery $L = ([a,b], w(p))$ and the contracts $(p(x), w(p(x)))$ chosen by the patients $x \in [0,1]$, letting $u(x) \triangleq u(x, (p(x), w(p(x))))$, the* social welfare *of L, denoted by SW_L, is*

$$SW_L = \int_0^1 u(x)dx.$$

Lottery L is feasible *if the budget constraint is satisfied, namely, $\int_0^1 p(x)dx = B$.*

Notice that we require a feasible lottery to use up all the budget. This is again without any loss of generality, since our theorem below implies that any lottery with cost $B' < B$ is beaten by the randomized assignment with budget B', and thus by the one with budget B.

We assume that the expected waiting time function $pw(p)$ is piece-wise twice differentiable in p. Notice that, although assuming twice differentiability of $pw(p)$ over the whole domain is too much, assuming it piece-wisely is quite natural. For example, the government may use different $w(p)$'s

[8] In general D can be a proper subset of $[0,1]$, as the government may not offer the whole interval $[0,1]$ for the patients to choose from.

[9] Notice that $w(p)$ itself may not be convex.

for different intervals of p, but inside each interval it uses a smooth $w(p)$. Both the randomized assignment and equilibrium assignments trivially satisfy this assumption.

The following theorem shows that, when the distribution of the patients' valuations accumulates toward the higher-value side, the randomized assignment is optimal compared with any lottery. Since equilibrium assignments are special cases of lotteries, the randomized assignment is optimal compared with them as well.

Theorem 3. *For any concave valuation function $v(x)$ and any feasible lottery $L = ([a, b], w(p))$, we have $SW_r \geq SW_L$.*

Proof. As the choice of $p(x)$ maximizes the utility of x, for any $\Delta > 0$ patient x prefers contract $C(x) = (p(x), w(p(x)))$ to contract $C(x + \Delta) = (p(x + \Delta), w(p(x + \Delta)))$, and patient $x + \Delta$ prefers $C(x + \Delta)$ to C. That is,

$$u(x) = p(x)[v(x) - w(p(x))] \geq p(x + \Delta)[v(x) - w(p(x + \Delta))],$$

and

$$u(x + \Delta) = p(x + \Delta)[v(x + \Delta) - w(p(x + \Delta))] \geq p(x)[v(x + \Delta) - w(p(x))].$$

Accordingly,

$$v(x) \cdot \Delta p(x) \leq \Delta(p(x) \cdot w(p(x))), \quad \text{and} \quad v(x + \Delta) \cdot \Delta p(x) \geq \Delta(p(x) \cdot w(p(x))). \tag{8}$$

As $pw(p)$ is piece-wise twice differentiable, all the differential equations and statements made in this paragraph hold piece-wisely, and we shall not mention the piece-wiseness again and again. To begin with, letting $\Delta \to 0$ in Equation 8, we have (with variable x omitted for conciseness)

$$v = \frac{d(pw(p))}{dp}, \tag{9}$$

where the function on the right-hand side is well defined and differentiable in p. As $p(v)$ is the inverse of Equation 9, it is differentiable in v. As $v(x)$ is concave, it is differentiable in x almost everywhere. Thus $p(x) = p(v(x))$ is differentiable in x. Accordingly, we have

$$\begin{aligned} du(x) &= dp \cdot (v - w) + p \cdot (dv - dw) = p \cdot dv + v \cdot dp - (w \cdot dp + p \cdot dw) \\ &= p \cdot dv + v \cdot dp - d(p \cdot w) = p \cdot dv + v \cdot dp - v \cdot dp = p \cdot dv. \end{aligned} \tag{10}$$

(Notice that $p(v)$ and $p(x)$ may not be continuous functions, but we only need them to be "nice" piece-wisely.)

Now putting all the pieces together and integrating both sides of Equation 10 over the whole domain, we have

$$u(x) = \int_0^{v(x)} p(\hat{v}) d\hat{v}. \tag{11}$$

As $v(x)$ is non-decreasing and concave, we have that $v'(x) \geq 0$ and $v'(x)$ is non-increasing. If there exists $x < 1$ such that $v'(x) = 0$, then let x_0 be the smallest number with $v'(x_0) = 0$; otherwise (i.e., $v(x)$ is strictly increasing) let $x_0 = 1$. We have that $v(x)$ is strictly increasing on $[0, x_0]$ and constant on $[x_0, 1]$. Let $v_0 = v(x_0)$. Following Equation 11 the social welfare of lottery

L is

$$\begin{aligned}
SW_L &= \int_0^1 u(x)dx = \int_0^1 \int_0^{v(x)} p(\hat{v})d\hat{v}dx = \int_0^{x_0} \int_0^{v(x)} p(\hat{v})d\hat{v}dx + \int_{x_0}^1 \int_0^{v_0} p(\hat{v})d\hat{v}dx \\
&= \int_0^{v_0} \left(p(\hat{v}) \int_{v^{-1}(\hat{v})}^{x_0} dx \right) d\hat{v} + \int_0^{v_0} \left(p(\hat{v}) \int_{x_0}^1 dx \right) d\hat{v} \\
&= \int_0^{v_0} p(\hat{v}) \cdot (x_0 - v^{-1}(\hat{v}))d\hat{v} + \int_0^{v_0} p(\hat{v}) \cdot (1 - x_0)d\hat{v} \\
&= \int_0^{x_0} p(x)(x_0 - x)v'(x)dx + \int_0^{x_0} p(x)(1 - x_0)v'(x)dx \\
&= \int_0^{x_0} p(x)(1 - x)v'(x)dx.
\end{aligned}$$

Similarly, the social welfare of the randomized assignment can be written as

$$\begin{aligned}
SW_r &= \int_0^1 Bv(x)dx = \int_0^1 \int_0^{v(x)} Bdvdx = \int_0^{x_0} \int_0^{v(x)} Bdvdx + \int_{x_0}^1 \int_0^{v_0} Bdvdx \\
&= \int_0^{v_0} \int_{v^{-1}(\hat{v})}^{x_0} Bdxd\hat{v} + \int_0^{v_0} \int_{x_0}^1 Bdxd\hat{v} = \int_0^{v_0} B(x_0 - v^{-1}(\hat{v}))d\hat{v} + \int_0^{v_0} B(1 - x_0)d\hat{v} \\
&= \int_0^{x_0} B(x_0 - x)v'(x)dx + \int_0^{x_0} B(1 - x_0)v'(x)dx = \int_0^{x_0} B(1 - x)v'(x)dx.
\end{aligned}$$

To prove $SW_r - SW_L \geq 0$, below we first show that $p(x)$ is non-decreasing. To do so, again notice that $p(x)$ maximizes the expected utility of x. Thus for any two patients $x_1 < x_2$, we have

$$u(x_1) = p(x_1)(v(x_1) - w(p(x_1))) \geq p(x_2)(v(x_1) - w(p(x_2)))$$

and

$$u(x_2) = p(x_2)(v(x_2) - w(p(x_2))) \geq p(x_1)(v(x_2) - w(p(x_1))).$$

Thus $p(x_2)(v(x_2) - v(x_1)) \geq p(x_1)(v(x_2) - v(x_1))$. If $v(x_2) = v(x_1)$ then $p(x_2) = p(x_1)$ (as we already said, $p(x)$ only depends on $v(x)$), otherwise $p(x_2) \geq p(x_1)$. That is, the function $p(x)$ is non-decreasing.

As L is feasible, we have $\int_0^1 p(x)dx = B$. Since $v(x)$ is constant on $[x_0, 1]$, so is $p(x)$. Therefore $p(x_0) \geq B$. Accordingly, there exists $x_B \in [0, x_0]$ such that $p(x) \leq B$ for all $x < x_B$, and $p(x) \geq B$ for all $x \geq x_B$. Thus we have

$$\begin{aligned}
SW_r - SW_L &= \int_0^{x_0} (B - p(x))(1 - x)v'(x)dx \\
&= \int_0^{x_B} (B - p(x))(1 - x)v'(x)dx + \int_{x_B}^{x_0} (B - p(x))(1 - x)v'(x)dx.
\end{aligned}$$

Notice that the value of $p(x_B)$ does not affect the value of the integration, thus without loss of generality we assume $p(x_B) = B$.

Again because $v'(x)$ is non-negative and non-increasing, for any $x \leq x_B$, we have $(1-x)v'(x) \geq (1 - x_B)v'(x_B) \geq 0$. Because $B - p(x) \geq 0$ for all $x \leq x_B$, we have

$$(B - p(x))(1 - x)v'(x) \geq (B - p(x))(1 - x_B)v'(x_B).$$

Similarly, for any $x \geq x_B$, we have $0 \leq (1 - x)v'(x) \leq (1 - x_B)v'(x_B)$ and $B - p(x) \leq 0$, which again implies

$$(B - p(x))(1 - x)v'(x) \geq (B - p(x))(1 - x_B)v'(x_B).$$

Thus

$$\begin{aligned} SW_r - SW_L &\geq \int_0^{x_B} (B - p(x))(1 - x_B)v'(x_B)dx + \int_{x_B}^{x_0} (B - p(x))(1 - x_B)v'(x_B)dx \\ &= (1 - x_B)v'(x_B) \int_0^{x_0} (B - p(x))dx. \end{aligned}$$

Following the budget constraint we have

$$\int_0^1 p(x)dx = \int_0^{x_0} p(x)dx + p(x_0)(1 - x_0) = B = \int_0^{x_0} Bdx + B(1 - x_0),$$

and thus

$$\int_0^{x_0} (B - p(x))dx = (p(x_0) - B)(1 - x_0).$$

Therefore

$$SW_r - SW_L \geq (1 - x_B)v'(x_B)(p(x_0) - B)(1 - x_0) \geq 0,$$

where the second inequality is because $x_B \leq 1$, $v'(x_B) \geq 0$, $p(x_0) \geq B$, and $x_0 \leq 1$.

In sum, no feasible lottery can generate more social welfare than the randomized assignment, and Theorem 3 holds. □

Remark 4. *Notice that the analysis above holds as long as $(1 - x)v'(x)$ is non-increasing. Thus the randomized assignment is optimal compared with any lottery even for some convex valuation function, such as $v(x) = e^x$. It would be interesting to fully characterize the condition under which the randomized assignment is optimal.*

References

[1] Aggarwal, Gagan, S. Muthukrishnan, Dávid Pál, and Martin Pál (2009), "General auction mechanism for search advertising." In *Proceedings of the 18th international conference on World Wide Web*, WWW '09, 241–250, ACM Press, New York, NY.

[2] Alatas, V., A. Banerjee, R. Hanna, B.A. Olken, R. Purnamasari, and M. Wai-Poi (2012), "Ordeal mechanisms in targeting: Theory and evidence from a field experiment in indonesia." Technical report, Mimeo, MIT.

[3] Ashlagi, Itai, Mark Braverman, and Avinatan Hassidim (2009), "Ascending unit demand auctions with budget limits." Working paper.

[4] Barzel, Y. (1974), "A theory of rationing by waiting." *Journal of Law and Economics*, 73–95.

[5] Bergemann, Dirk and Joan Feigenbaum (2008), "Economics and computation." Yale ECON 425/563 and CPSC 455/555, lecture notes.

[6] Demange, Gabrielle, David Gale, and Marilda Sotomayor (1986), "Multi-item auctions." *Journal of Political Economy*, 94, 863–872.

[7] Easley, David and Jon Kleinberg (2010), *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, chapter 10. Cambridge University Press.

[8] Gravelle, H. and L. Siciliani (2008), "Is waiting-time prioritisation welfare improving?" *Health Economics*, 17, 167–184.

[9] Gravelle, Hugh and Luigi Siciliani (2008), "Optimal quality, waits and charges in health insurance." *Journal of Health Economics*, 27, 663–674.

[10] Hartline, J.D. and T. Roughgarden (2008), "Optimal mechanism design and money burning." In *Proceedings of the 40th annual ACM symposium on Theory of computing*, 75–84, ACM.

[11] Heskett, James (2003), "Shouldice hospital limited." Harvard Business School.

[12] Iversen, T. (1993), "A theory of hospital waiting lists." *Journal of Health Economics*, 12, 55–71.

[13] Kahn, C.N., T. Ault, H. Isenstein, L. Potetz, and S. Van Gelder (2006), "Snapshot of hospital quality reporting and pay-for-performance under medicare." *Health Affairs*, 25, 148–162.

[14] Karp, Richard M. (1972), "Reducibility among combinatorial problems." In *Complexity of Computer Computations* (Raymond E. Miller and James W. Thatcher, eds.), 85–103.

[15] Kuhn, H. W. (1955), "The hungarian method for the assignment problem." *Naval Research Logistics Quarterly*, 2, 83–97.

[16] Lindenauer, P.K., D. Remus, S. Roman, M.B. Rothberg, E.M. Benjamin, A. Ma, and D.W. Bratzler (2007), "Public reporting and pay for performance in hospital quality improvement." *New England Journal of Medicine*, 356, 486–496.

[17] Lindsay, C.M. and B. Feigenbaum (1984), "Rationing by waiting lists." *The American Economic Review*, 404–417.

[18] Ma, C.A. and H.Y. Mak (2012), "Information disclosure and the equivalence of prospective payment and cost reimbursement." Technical report, Boston University-Department of Economics.

[19] Nisan, Noam, Tim Roughgarden, Éva Tardos, and Vijay Vazirani, eds. (2007), *Algorithmic Game Theory*. Cambridge University Press.

[20] Rosén, P., A. Anell, and C. Hjortsberg (2001), "Patient views on choice and participation in primary health care." *Health policy*, 55, 121–128.

[21] Rosenthal, M.B., R. Fernandopulle, H.S.R. Song, and B. Landon (2004), "Paying for quality: providers' incentives for quality improvement." *Health Affairs*, 23, 127–141.

[22] Siciliani, L. and J. Hurst (2005), "Tackling excessive waiting times for elective surgery: a comparative analysis of policies in 12 oecd countries." *Health policy*, 72, 201–215.

[23] Sipser, Michael (2012), *Introduction to the Theory of Computation*, 3 edition. Course Technology.