Allocating heterogeneous goods through wait times and prices*

Filip Tokarski
Stanford GSB

April 12, 2025

Abstract

A designer allocates a stock of two kinds of goods to agents with heterogeneous preferences, aiming to maximize the agents' welfare. She can incentivize agents to choose goods in a socially optimal manner using two wasteful screening instruments: charging them money for the goods and requiring them to wait to receive them. I show that the optimal mechanism never requires agents to wait for either good and ensures that supply constraints hold by charging market-clearing prices.

1 Introduction

Affordable housing programs often offer a wide variety of units. In Amsterdam, public housing stock includes remote tower blocks as well as apartments in central, upscale neighbourhoods (Van Dijk, 2019). Among them, larger, better maintained, or more centrally located developments are more desired, and households wait much longer to get them. Similarly, wait times are higher for units whose rents are subsidized more heavily relative to market rates (Van Ommeren and Van der Vlist, 2016). Consequently, households participating in Amsterdam's housing program face a trade-off between expected wait times, rents, and their values for different locations or projects. While affordable housing programs around the world differ in terms of the exact mechanism or its implementation, the same core trade-off is often present. A designer interested in improving such programs must therefore understand how differences in wait times and rents make agents self-select into the available housing developments.

Similarly, healthcare systems face the challenge of allocating a limited supply of different treatments to patients with conditions that vary in severity and progression. To make these assignments more efficient, patients requesting scarce or costly therapies are often mandated to

*I am grateful to Joey Feffer, Federico Llarena, Sam Wycherley, Benjamin Brooks, Rafael Berriel, Andrzej Skrzypacz, Piotr Dworczak, Jacob Leshno, Michael Ostrovsky, Ilya Segal and Rebecca Diamond for their helpful comments and suggestions.

secure referrals or demonstrate that less invasive methods have been attempted previously. Indeed, the World Health Organization (2023) states that the referral system is "aiming to ensure patient access to specialist health care when needed, while maintaining resource efficiency". While referral requirements may reduce the burden on the healthcare system by ensuring that only patients with a high need pursue scarce treatment options, they can also delay the therapy itself. For instance, the American Physical Therapy Association claims that "direct access restrictions cause unnecessary delays for people who would benefit from treatment by a PT."[1]

I study the optimal design of the trade-off between values for different goods, wait times and payments by considering a stylized model where a welfarist designer chooses how to allocate two scarce goods amongst a mass of agents. Since supply is limited, the designer wants to encourage agents to select goods in a socially efficient manner; she can incentivize agents to do so by making those who receive certain goods wait or pay for them. Importantly, however, the designer cares only about the welfare of agents and does not want to extract revenue from them. Thus, both wait times and payments are essentially wasteful screening instruments. I find that even in this case, screening agents using only payments dominates any mechanism screening with wait times or combining the two devices. In optimum, all goods are allocated immediately upon arrival and the designer ensures supply constraints hold by charging prices for overdemanded goods.

The result can be understood through two complementary intuitions. First, ensuring that supply constraints hold using wait times disproportionately hurts high-value agents, while doing so with payments is equally costly to everyone. I illustrate this in a simple single-dimensional example where a unit mass of agents can either choose an abundant outside option or a scarce good. While the utilities of the marginal agent choosing the scarce good will be the same when either payments or wait times are used, the latter instrument deprives inframarginal takers of the scarce good of rents associated with their high values. By contrast, pricing the good preserves such rents, leading to higher welfare.

However, this intuition does not fully extend to the case of multiple goods. Indeed, in a multidimensional setting, the designer can use combinations of wait times and payments to make agents 'sort' into goods in rich ways, with some patterns of sorting requiring the use of wait times. A second intuition notices that the two instruments also differ qualitatively in *what* they screen on, and that prices always sort agents in a better way than do wait times—since the costs of waiting are associated with delaying the good's receipt, they can only extract information about agents' *relative* values for the offered goods, but not about the *intensity* of their need for either good. Payments, on the other hand, elicit *how much* agents value the distributed goods, allowing the designer to implement more efficient allocations.

From a technical perspective, my model can be seen as combining a wasteful screening device that enters an agent's utility *additively* with one that *multiplies the agent's type*. While wait times are the most common 'multiplicative' screening device, one could interpret it as the designer

[1] https://www.apta.org/advocacy/issues/direct-access-advocacy

degrading the good's quality à la Deneckere and McAfee (1996). In this case, the main result tells us that it is better to keep the quality of both goods high and price it in, rather than trying to entice more agents to choose one type of good by offering a cheaper and worse variety of it. Relatedly, the designer could be setting a multi-part tariff on a good which is itself scarce, but whose use does not generate substantial costs. She could then still require payments for usage in order to encourage agents to select one good over another.

My model is also an instance of a tractable multidimensional screening problem. By restricting attention to deterministic mechanisms, I am able to characterize the design of price and wait time menus for the two goods as interconnected single-dimensional screening problems. The interaction between them is summarized by a boundary in the type space that separates the sets of types who choose each good. The multidimensional problem can then be broken up into two stages: first, determining the optimal way to implement a given boundary, and second, solving an optimal control problem to select the optimal boundary among all implementable ones.

This paper relates closely to the work on costly screening and money-burning (e.g. Hartline and Roughgarden (2008), Condorelli (2012)). However, the literature focuses on cases where the designer only has access to a single wasteful screening device. An exception is Yang (2021) who considers a monopolist with both wasteful and non-wasteful instruments and characterizes cases where the wasteful one should not be used. In contrast, this paper considers a designer using two wasteful screening instruments. My comparison of screening devices—wait times and payments—also relates to the work of Akbarpour et al. (2023) who study when one screening device dominates another for a planner aiming to maximize a social welfare function. Unlike them, I allow the designer to *combine* instruments and show that screening with payments alone dominates any mechanism using both devices. Finally, my paper relates to a literature on waitlist design. While no paper has studied combining waitlists with payments in settings with heterogeneous goods, a substantial literature examines designing such waitlists without transfers. Ashlagi et al. (2024) demonstrate that allocative efficiency can be improved by coarsening agents' information about the qualities of allocated goods. Arnosti and Shi (2020) compare common non-monetary mechanisms in terms of targeting and match efficiency. Barzel (1974), Bloch and Cantala (2017), and Leshno (2022) observe that in environments with homogeneous waiting costs, wait times may 'act as prices', screening for agents with higher valuations. I refine this intuition by showing that the screening properties of wait times are impeded when the cost of waiting stems from delayed receipt—in those cases, wait times can only screen on agents' relative values for the offered goods.

2 Model

A designer distributes two types of goods, A and B. Their supplies are equal to $\mu_A, \mu_B > 0$, with $\mu_A + \mu_B \leq 1$. There is a unit mass of agents whose values for the two goods are given by a and b, respectively, with $(a, b) \in [0,1]^2$. Agents' values are distributed according to F with the following properties:

Assumption 1. *F has a continuous, full support pdf f on $[0,1]^2$. Moreover, agents' values for the two goods are independent, so there exist continuous, full support pdfs g, h on $[0,1]$ such that:*

$$f(a,b) = g(a) \cdot h(b).$$

Let G, H be their cdfs. I assume that their anti-hazard rates, $\frac{G(v)}{g(v)}$, $\frac{H(v)}{h(v)}$, are strictly increasing.

The designer chooses a menu of wait times and payments for each of the goods. That is, an agent can choose which good she wants to get and then choose a wait time and payment option from the relevant good's menu. She can also not participate, which gives her utility 0. When a type-(a,b) agent participates and receives good y, her utility is:

$$x \cdot a - p \text{ if } y = A,$$
$$x \cdot b - p \text{ if } y = B,$$

where $p \in \mathbb{R}_+$ is the price the agent pays and $x \in [0,1]$ is her discounting due to having had to wait for the good. The designer chooses the menus to maximize the total welfare of agents. Thus, by the Revelation Principle, we can reduce her problem to picking allocation rules for payments, $p : [0,1]^2 \to \mathbb{R}_+$, discounting, $x : [0,1]^2 \to [0,1]$, and goods, $y : [0,1]^2 \to \{\varnothing, A, B\}$, to maximize:

$$W = \int U[a,b,(p,t,y)(a,b)] \, \mathrm{d}F(a,b), \tag{W}$$

subject to (IC) and (IR) constraints, and the supply constraint (S):

$$\text{for all } (a,b),(a',b') \in [0,1]^2, \quad U[a,b,(p,t,y)(a,b)] \geq U[a,b,(p,t,y)(a',b')], \tag{IC}$$

$$\text{for all } (a,b) \in [0,1]^2, \quad U[a,b,(p,t,y)(a,b)] \geq 0, \tag{IR}$$

$$\int \mathbb{1}_{y(a,b)=A} \, \mathrm{d}F(a,b) \leq \mu_A, \quad \int \mathbb{1}_{y(a,b)=B} \, \mathrm{d}F(a,b) \leq \mu_B. \tag{S}$$

Here $U[a,b,(p,x,y)(a',b')]$ denotes the utility type (a,b) gets from reporting (a',b') in the mechanism (p,x,y). I also impose the following technical restriction on admissible mechanisms:

Assumption 2. *The designer is restricted to discounting rules $x : [0,1]^2 \to [0,1]$ that are piecewise continuously differentiable in each dimension of the type.*

I call a mechanism (p,x,y) satisfying (IC),(IR), (S) and Assumption 2 *feasible*.

A few features of the model are worth discussing. First, I assume that the designer does not value the revenue she receives from the mechanism. This might correspond to an environment where payments are not monetary, but represent a wasteful ordeal such as form-filling or travelling to a distant office. Alternatively, it may capture the nature of social programs whose participants are significantly poorer than the average taxpayer, so a redistributive designer would

not want to use them to collect revenue.[2] In practice, we might expect the designer to consider transfers wasteful, but still have *some* value for the collected revenue. For government programs, this could be the case when rebating it to participants is possible but costly due to bureaucratic inefficiency or because distributing cash lacks the screening benefits of in-kind transfers—when the designer hands out a free (or subsidized) inferior good, only relatively poor agents will want to participate as wealthier ones can afford higher-quality alternatives. Thus, the subsidy is automatically targeted to those who need it most (Besley and Coate, 1991). As soon as the designer hands out cash, such targeting disappears as money is desired by everyone, regardless of wealth. Note, however, that the extreme assumption that revenue is completely wasted works *against* the main result that the designer only wants to screen agents using money. If the designer considered transfers neutral or attributed some weight in $(0, 1)$ to them, she would have even more reason to rely on payments.

Second, my model does not permit the designer to use lotteries between the two goods. While this assumption is restrictive, it renders an otherwise unwieldy model tractable. As I explain in the following section, the fact that the designer chooses two separate menus of payments and wait times, one for each good, lets me represent mechanisms as boundaries splitting the type-space into three distinct regions. This additional structure allows me to find the best mechanism by optimizing over the space of implementable boundaries. Moreover, restricting the designer to wait times and payments, without the use of lotteries, makes the comparison between these two common screening instruments clearer.

Finally, in many settings wait times are not an explicit design choice but arise naturally as byproducts of the system's equilibrium dynamics—public housing waitlists, for instance, develop endogenously to clear markets when rental prices are too low to do so. Nonetheless, endogenously arising differences in wait times continue to exhibit screening properties described by the model. Moreover, even in these cases, we can view waitlists as a consequence of design choices, and thus, indirectly, as designed objects. For instance, the imbalance between wait times for different kinds of housing can be influenced by suitably adjusting their rents. By raising the subsidy for the less desired unit and increasing the price of the popular one, the designer can bring the lengths of their waitlists closer together, and in doing so diminish the 'screening role' of wait times in the program. Relatedly, one can interpret the model as a dynamic one in which flows of goods and agents arrive over time. The designer then operates two waitlists, letting arriving agents choose which waitlist to join and whether they want to pay extra to reduce their wait time in it. The model then corresponds to a patient designer choosing menus of such pay-to-skip options to maximize total welfare in the waitlists' steady states. The supply constraints (S) ensure that the mass of agents assigned to either waitlist in the representative

[2] While the model does not explicitly account for wealth differences or heterogeneous welfare weights among agents, this can be viewed as an approximation of a scenario where such differences exist but are relatively small *between* participants compared to the gap between participants and the average taxpayer. This is especially likely when the designer allocates inferior goods, such as public housing in undesirable areas. The designer's welfare-weighted objective can then be approximated by a constant weight on all participants, which is distinct from that on revenue, representing her welfare weight for the average taxpayer.

steady-state period does not exceed the mass of the corresponding good that arrives in it.

3 Feasible mechanisms

Let us first describe the properties of feasible mechanisms. It will be convenient to characterize them in terms of good-specific indirect utilities $U_A, U_B : [0,1] \to \mathbb{R}_+$, defined as follows:

$$U_A(a) = \max_{(a',b')} \left\{ x(a',b') \cdot a - p(a',b') : \quad y(a',b') \in \{A, \varnothing\} \right\}, \tag{1}$$

$$U_B(b) = \max_{(a',b')} \left\{ x(a',b') \cdot b - p(a',b') : \quad y(a',b') \in \{B, \varnothing\} \right\}. \tag{2}$$

Note that U_A and U_B are convex and increasing. Intuitively, $U_A(a)$ and $U_B(b)$ represent the highest utility type (a,b) could get from selecting some wait time and payment option for the A- and the B-goods, respectively (or not participating). Then agents for whom $U_A(a) < U_B(b)$ choose good A and those from whom $U_A(a) > U_B(b)$ choose good B.

Good-specific indirect utilities depend only on one dimension of the type—an agent's value for good B does not affect her choice of wait time and payment option if she chooses good A. We can thus write each type's discounting x purely as a function of the relevant component of the type. To that end, define $x_A, x_B : [0,1] \to [0,1]$ as:

$$U_A'(a) := x_A(a), \quad U_B'(b) := x_B(b).^{3}$$

Indeed, by the envelope theorem (Milgrom and Segal, 2002), $x(a,b)$ equals to $x_A(a)$ and $x_B(b)$ for almost all types who get A and B, respectively.

We will now use U_A and U_B to describe agents' choices of goods. To that end, let us also define a mechanism's *lowest participating values* as follows:

$$\underline{a} = \sup\{a : \ U_A(a) = 0\}, \quad \underline{b} = \sup\{b : \ U_B(b) = 0\}.^{4} \tag{3}$$

Definition 1. *Let a* ***boundary*** *be a function $z : [\underline{a}, \overline{a}] \to [\underline{b}, \overline{b}]$ that is continuous, strictly increasing and satisfies $\overline{a} \leq 1$ and $\overline{b} \leq 1$, with one of them holding with equality.*

Lemma 1. *All types pointwise below the lowest participating values do not get either good, that is $y(a,b) = \varnothing$ for all $(a,b) < (\underline{a}, \underline{b})$.*[5]

Suppose a positive mass of agents gets either good. Then the good choices of types $(a,b) > (\underline{a}, \underline{b})$ are characterized by some boundary $z : [\underline{a}, \overline{a}] \to [\underline{b}, \overline{b}]$. A type $(a,b) > (\underline{a}, \underline{b})$ gets good A if (a,b) is below

[3] We also assume that x_A and x_B are left-continuous. This pins down the values of these functions in places where U_A, U_B are not differentiable.

[4] Since the mechanism offers a non-participation option, we always have $U_B(0) = U_A(0) = 0$. Thus, these suprema are well-defined.

[5] When comparing vectors, I will use $\geq$ and $>$ for pointwise comparisons.

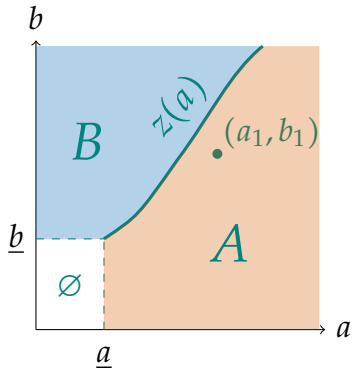

Figure 1: Types below the boundary (orange) choose good A and types above it (blue) choose good B.

the boundary z, that is, if $z(a) > b$, and gets good B if (a, b) is above the boundary z, that is, if $z(a) < b$. Moreover, types at the boundary are indifferent between both goods, thus:

$$U_A(a) = U_B(z(a)) \quad \textit{for all} \quad a \in [\underline{a}, \overline{a}]. \tag{I}$$

I then say that the mechanism ***implements*** *the boundary z.*

When goods A and B are only given out at positive prices, types with sufficiently low values for both of them, i.e. $(a, b) < (\underline{a}, \underline{b})$, do not participate. To understand the good choices of participating types, consider some (a_1, b_1) choosing good A (Figure 1). Then any type (a, b) with $a > a_1$ and $b < b_1$ will also choose good A—since she values the A-good even more than (a_1, b_1) and values the B-good even less, all the payment and wait time options for good B are strictly less attractive to her than they were to (a_1, b_1). We can now notice that the types who are indifferent between their best options for either good lie on an increasing curve z originating from $(\underline{a}, \underline{b})$. By the above logic, all types below this curve choose A and pick some payment and wait time option from its menu, while types above it choose B.

We can therefore think of our multidimensional mechanism design problems as two single-dimensional problems connected endogenously through the boundary z. While agents on its either side effectively face one-dimensional problems, making one of the goods more attractive invites more types to switch to it, effectively deforming the boundary.

Note also that despite the boundary $z(a)$ being written as a function of a, the setting is symmetric with respect to both goods. Thus, any results about $z(a)$ also apply to its inverse, $z^{-1}(b)$. This observation will be analytically useful: throughout, we will encounter properties that must hold either for the boundary $z(a)$ or for its inverse. However, the symmetry of the setting w.r.t. the two goods guarantees that assuming them for $z(a)$ is without loss.

4 Optimal mechanism

I show that the planner's optimal mechanism is extremely simple:

Theorem 1. *The optimal mechanism implements the efficient allocation of goods, and allocates both of them without waiting. It posts a separate price for each good. The prices are chosen so that the whole supply of both goods is allocated.*

While the designer could incentivize agents to select the socially optimal good by using a combination of wait times and payments, Theorem 1 says she should only use the latter. That is, payments dominate wait times as a screening device even when the designer has zero value for revenue. Moreover, the optimal price mechanism guarantees that the allocation of both goods is efficient, in the sense that it maximizes the aggregate value of the goods to recipients.

More abstractly, one can understand the result as saying that wasteful screening devices that enter an agent's utility *additively* are superior to those that *multiply the agent's type*. Consequently, one can also interpret $x < 1$ as damage or a usage restriction that the designer imposes on the good. In this case, Theorem 1 tells us that the designer would never want to damage either good in order to incentivize fewer agents to choose it.

I present two complementary intuitions behind Theorem 1. The first one explains why an analogous result would hold in a simple, one-dimensional version of the model where only one good is scarce. In such an environment, any mechanism that satisfies the supply constraint must deter low-value agents by imposing some burden on all agents taking A, and by doing so enforce the right type cutoff for choosing it. However, when this burden is imposed through wait times, it hurts inframarginal agents with high values for A even more strongly than it hurts the cutoff type it is meant to deter. Payments, on the other hand, are 'equally costly' to everyone, thus leaving more rents to high value types.

In the above example, however, the set of agents who choose good A is described by a simple cutoff. By contrast, in a multidimensional model the designer can use different combinations of wait times and payments to assign goods to agents in complex ways. The second intuition therefore highlights why screening with wait times leads types to choose goods less efficiently than does screening with money alone—since wait times multiply agents' values for goods, they screen on *relative* preferences for them. Prices, however, enable participants to express absolute valuations, which is what fundamentally matters for allocative efficiency.

4.1 Intuition 1: wait times are more costly to inframarginal agents

Consider a simplified model where a unit mass of agents has types a distributed according to G with full support on $[0,1]$. The designer has a mass $\mu_A \in (0,1)$ of good A and an unlimited supply of good B. All agents value good B at $b \in (0, G^{-1}(\mu_A))$, that is, strictly more than μ_A agents prefer good A to good B. For simplicity, we assume the designer can ask agents to

wait for the scarce A-good, but not the B-good. Thus, the designer chooses an allocation rule $y : [0,1] \to \{A, B\}$, a discounting rule $x_A : [0,1] \to [0,1]$, and a payment rule $p : [0,1] \to \mathbb{R}_+$ to maximize total welfare, subject to IC, IR and supply constraints analogous to those in the main model.

Proposition 1. *The optimal mechanism in the one-dimensional model offers good B for free and posts a price for good A, which she allocates without waiting. The whole available supply of A is allocated.*

Proof. Since good B is always allocated without waiting, all agents receiving it must by paying the same price p_B for it. Note also that any feasible mechanism must allocate good A to at most $\mu_A \in (0,1)$ agents. A single-crossing argument then tells us there exists some $\underline{a} \in [0,1]$ such that:

$$y(a,b) \begin{cases} = A, & \text{if } a > \underline{a}, \\ = B, & \text{if } a < \underline{a}. \end{cases}$$

Let $\underline{a}^* \in (0,1)$ be the cutoff for which the supply constraint binds. By Myerson's Lemma (Myerson, 1981) we can then reduce the problem to choosing some $p_B \geq 0$, $\underline{a} \in [\underline{a}^*, 1]$ and an increasing $x_A : (\underline{a}, 1] \to [0,1]$. Total welfare then becomes:

$$W = G(\underline{a}) \cdot (b - p_B) + \int_{\underline{a}}^{1} U(v)\, dG(v) \quad \text{where} \quad U(a) = b - p_B + \int_{\underline{a}}^{a} x_A(v) dv.$$

Total welfare therefore decreases in p_B and increases pointwise in $x_A(a)$. Thus, the optimal single-good mechanism features $p_B = 0$, $\underline{a} = \underline{a}^*$, and $x_A(a) = 1$ for all $a > \underline{a}^*$. Note that the payment for types above $\underline{a}$ must be constant and equal to some p_A. □

Intuitively, every feasible mechanism must allocate the A-good to agents with types above some cutoff $\underline{a}$. This cutoff must then be enforced by making good A costly enough so that type $\underline{a}$ is indifferent between A and the outside option B. Let us compare two ways of enforcing such a cutoff. First, the designer could impose a wait on anyone requesting good A, i.e. decrease x_A to the point where the cutoff agent is indifferent:

$$b = x_a \cdot \underline{a}.$$

The designer could also charge a price p_A enforcing the same cutoff:

$$b = \underline{a} - p_A.$$

While both of these mechanisms hurt the cutoff type $\underline{a}$ equally, the wait time mechanism is more costly to *inframarginal* types $a > \underline{a}$ who choose good A. Payments, on the other hand, are equally costly to everybody. Thus, the payment mechanism leaves more surplus to high-value takers of A, leading to higher welfare (Figure 2).

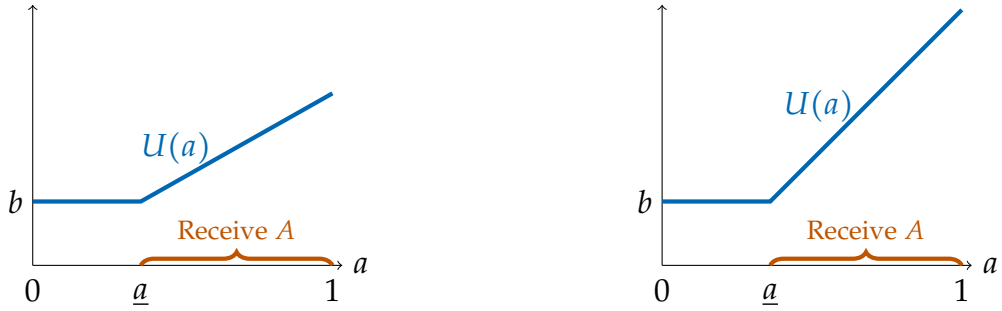

Figure 2: Indirect utilities $U(a) = b + \int_{\underline{a}}^{a} x_A(v)dv$ for the wait time mechanism (left) and the price mechanism (right).

4.2 Intuition 2: wait times screen on relative values, prices screen on absolute values

In a two-dimensional model, being able to combine wait times and payments gives the deisgner more freedom to 'sort' agents into the two goods. I now explain why wait times screen agents in a qualitatively worse way than do payments, and thus why the designer would not benefit from incorporating them into the mechanism. To illustrate this, I consider two examples: one where $\mu_A + \mu_B = 1$ and the designer does not use payments, and one where the designer uses only payments and achieves the efficient allocation of goods.

Proposition 2. *Suppose $\mu_A + \mu_B = 1$ and consider a mechanism that allocates both goods without using money. Then there exists $k \in (0, \infty)$ such that all types with $a/b > k$ take good A and all types with $a/b < k$ take good B.*

Proof. When $p(a, b) \equiv 0$, all types choosing either good must have the same discounting x_A, x_B. The boundary z for this mechanism must then satisfy the boundary indifference condition:

$$a \cdot x_A = z(a) \cdot x_B \quad \text{for all} \quad a \in [0, \overline{a}]. \tag{I}$$

Thus, by Lemma 1, all types for whom $a/b > k = x_A/x_B$ choose good A, and all types for whom $a/b < k$ choose good B. □

Under such a mechanism both goods are handed out for free, and so all applicants with non-zero value for either good will want one. Moreover, in the absence of payments, the 'market will clear' based on wait times—if good A is overdemanded, that is, preferred by more agents than the mass of this good—the wait time for it will be longer. This will in turn deter some agents who prefer good A from choosing it and encourage them to take B instead. Importantly, however, wait times can only screen agents based on their *relative values* for the two goods, that is, the ratio a/b. Graphically, this corresponds to the boundary z partitioning the type space along a ray originating from zero (Figure 3). The slope of the boundary reflects the ratio a/b for

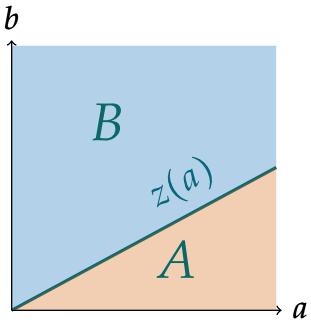

Figure 3: Without money, agents choose goods based on a/b.

which agents are indifferent between the two goods with those wait times. This slope is pinned down by the relative supply of the two goods.

Thus, when we screen agents with wait time, the designer will not be able to distinguish between two agents whose value ratios a/b are equal but whose *absolute* values differ. However, the designer cares about these two agents' allocations to different extents. If the former agents' values for both goods are higher, it is more important to give her the good she prefers. The first-best mechanism would therefore distort the assignments to agents whose values for both goods are low (by giving them the less demanded one) and leave the overdemanded good to those whose absolute value for it is large. Screening with payments alone can accomplish exactly that. Indeed, a payment-only mechanism can achieve the efficient allocation, that is, maximize:

$$\int \mathbb{1}_{y(a,b)=A} \cdot a + \mathbb{1}_{y(a,b)=B} \cdot b \; \mathrm{d}F(a,b), \tag{E}$$

subject to the supply constraint (S).

Proposition 3. *A mechanism that posts a single price for each good and allocates the whole supply of both goods maximizes allocative efficiency* (E).

Proof. Consider a linear relaxation of the problem of maximizing allocative efficiency, namely the problem of choosing $q_A, q_B : [0,1]^2 \to [0,1]$ to maximize:

$$\int q_A(a,b) \cdot a + q_B(a,b) \cdot b \; \mathrm{d}F(a,b),$$

subject to:

$$\int q_A(a,b)\,\mathrm{d}F(a,b) \le \mu_A, \quad \int q_B(a,b)\,\mathrm{d}F(a,b) \le \mu_B, \tag{4}$$

$$q_A(a,b) + q_B(a,b) \le 1 \;\text{ for every } (a,b) \in [0,1]^2. \tag{5}$$

Since $\mu_A + \mu_B \le 1$ and a unit mass of types has positive values for both goods, both supply constraints (4) will holds with equality. The objective and constraints are linear so the solution

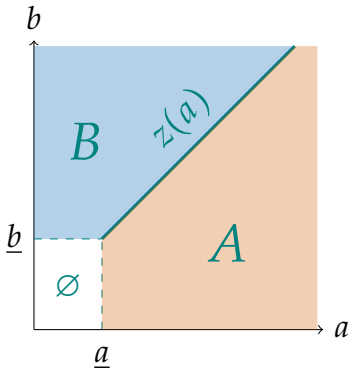

Figure 4: The boundary corresponding to the two-posted-price mechanism.

exists and must also maximize:

$$\int q_A(a,b)\cdot(a-\eta_A)+\ q_B(a,b)\cdot(b-\eta_B)\ \mathrm{d}F(a,b), \tag{6}$$

subject to (5) for some multipliers $\eta_A, \eta_B \geq 0$. Note also that $\eta_A, \eta_B < 1$. Otherwise, the maximizer of (6) would not allocate one of the goods at all, and we know that supply constraints must hold with equality. Now, notice that q_A, q_B maximize (6) if and only if they satisfy the following almost everywhere:

$$q_A(a,b)=\begin{cases}1, & \text{if } a-\eta_A>\max\{0,\, b-\eta_B\},\\ 0, & \text{otherwise,}\end{cases}, \quad q_B(a,b)=\begin{cases}1, & \text{if } b-\eta_B>\max\{0,\, a-\eta_A\},\\ 0, & \text{otherwise.}\end{cases}$$

Such an allocation is implemented by a mechanism with no wait times and two posted prices equal to η_A, η_B. □

The proposition is a standard example of the efficiency of prices. However, the mechanism implementing the efficient allocation might require agents to make large payments, and so a designer who cares about agents' welfare might still find this form of screening very costly.

Graphically, the two-posted-price mechanism corresponds to a linear boundary $z(a)$ with slope 1 originating from the point $(\underline{a}, \underline{b})$ (Figure 4). This can be seen from the boundary indifference condition (I). Differentiating it gives:

$$x_A(a) = x_B(z(a))\cdot z'(a). \tag{DI}$$

Under this mechanism both goods come with no discounting, so $x_A, x_B = 1$, implying $z'(a) = 1$. Moreover, the lowest participating values $\underline{a}$ and $\underline{b}$ have to equal to the two goods' prices.

5 Proof of Theorem 1

I first use Lemma 1 to characterize feasible mechanisms in terms of their corresponding discounting rules x_A, x_B and the boundary z they implement. The proof of this and other facts and lemmas can be found in the Appendix.

Lemma 2. *A mechanism is feasible if and only if its corresponding x_A, x_B are piecewise continuously differentiable, weakly increasing and implement a boundary z satisfying the following properties:*

1. *The supply constraint (S′) holds:*

$$\int_{\underline{a}}^{1} \int_{0}^{z(\min[a,\bar{a}])} f(a,v)dv\, da \leq \mu_A, \quad \int_{\underline{b}}^{1} \int_{0}^{z^{-1}(\min[b,\bar{b}])} f(v,b)dv\, db \leq \mu_B. \tag{S′}$$

2. *z is piecewise twice continuously differentiable on $(\underline{a}, \bar{a})$.*

3. *z has strictly positive and finite left- and right-derivatives at all $a \in (\underline{a}, \bar{a})$, and a strictly positive and finite left-derivative at $\bar{a}$.*

The result changes the way we express supply constraints. Rather than look at good allocations $y(a,b)$ directly, it takes advantage of the fact that types who get good A (B) are below (above) the boundary. It then measures the masses of agents getting either good by integrating over agents below and above z (Figure 5). Also, Assumption 2 guarantees that the boundaries implemented by feasible mechanisms are well-behaved, satisfying properties 2 and 3.

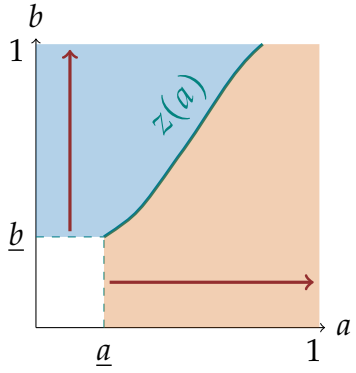

Figure 5: The supply condition (S′) ensures that the probability masses below the boundary (orange) and above it (blue) are at most μ_A and μ_B, respectively. The red arrows mark the directions of integration in the left-hand-sides in (S′).

Lemma 1 also lets us rewrite total welfare (W) in terms of good-specific indirect utilities U_A, U_B and their associated boundary z:

$$W[z] := \int_{\underline{a}}^{1} \int_{0}^{z(\min[a,\bar{a}])} f(a,v)dv \cdot U_A(a)\, da \;+\; \int_{\underline{b}}^{1} \int_{0}^{z^{-1}(\min[b,\bar{b}])} f(v,b)dv \cdot U_B(b)\, db. \tag{W′}$$

The rest of the argument can be broken down into two stages: in the first stage, I consider all the mechanisms that implement a particular boundary z and find the one that does so optimally. Then, in the second stage, I consider mechanisms optimally implementing different boundaries z and look for the boundary z^* that maximizes (W′). I use optimal control tools to show that the optimal boundary z^* has to be linear, and thus offer a single wait time and price option for each good. Finally, I show that the optimal mechanism allocates the whole available supply of both goods, and that its corresponding boundary has slope 1. This in turn means the mechanism does not require wait times for either good.

5.1 Optimally implementing a fixed boundary

We now fix any boundary z and look for a mechanism that optimally implements it. By Lemma 2, discounting rules x_A, x_B for all feasible mechanisms are weakly increasing. Recall also that any mechanism implementing z satisfies boundary indifference (I):

$$U_A(a) = U_B(z(a)) \quad \text{for all} \quad a \in [\underline{a}, \overline{a}]. \tag{I}$$

Differentiating (I) tells us that x_A, x_B have to satisfy the following condition almost everywhere:

$$x_A(a) = x_B(z(a)) \cdot z'(a). \tag{DI}$$

Now, consider some type (a', b') on the boundary and suppose that the boundary is convex on some interval $[a'', a']$. Then $z'(a)$ would be increasing on that interval, and would satisfy:

$$\frac{x_A(a)}{x_B(z(a))} = z'(a).$$

Since $x_B(z(a))$ is non-decreasing on this interval, it means that $x_A(a)$ has to be strictly increasing on it, and thus distorted downwards from $x_A(a')$ below a'. More generally, in order to make a boundary curve up or down somewhere, a mechanism must induce a separating allocation of discounting on at least one side of the boundary there. However, inducing such separation is costly, as it requires imposing wasteful wait times on agents with lower types. Note, however, that altering the shape of the boundary could, in principle, be beneficial despite this cost, as it might entice agents to choose between the A- and B-goods in a more socially efficient way. Nevertheless, more separation cannot be better *conditional on implementing the same boundary*. Consequently, a fixed boundary is implemented most efficiently by a mechanism that separates agents' allocations as little as possible while still satisfying boundary indifference (I). This observation is formalized in the following Lemma:

Definition 2. *Let a closed interval be a* ***convex (concave) region*** *of z if z is convex (concave) on it.*

Lemma 3. *Fix any boundary satisfying properties* 1 – 3 *of Lemma* 2*. Then the mechanism optimally implementing z exists, is unique, and satisfies the following properties:*

1. *On concave regions,* $x_A(a)$ *is constant and* $x_B(z(a)) \propto 1/z'(a)$.
2. *On convex regions,* $x_B(z(a))$ *is constant and* $x_A(a) \propto z'(a)$.
3. *At least one of* $x_A(a)$ *and* $x_B(z(a))$ *is continuous at every* $a \in (\underline{a}, \overline{a}]$.
4. *If* $\overline{a} < 1$, x_A *is constant on* $(\overline{a}, 1]$. *If* $\overline{b} < 1$, x_B *is constant on* $(\overline{b}, 1]$.
5. $\max[x_A(1), x_B(1)] = 1$.

Intuitively, minimizing separation means that at most one of $x_A(a), x_B(z(a))$ is strictly increasing on any region of the boundary. If both of them were strictly increasing somewhere, we could keep raising them pointwise in a (DI)-preserving manner until one of the monotonicity constraints started binding. Wherever the boundary is convex, the monotoniciy constraint on x_B will bind, wherever it is concave, that on x_A will (Figure 6).

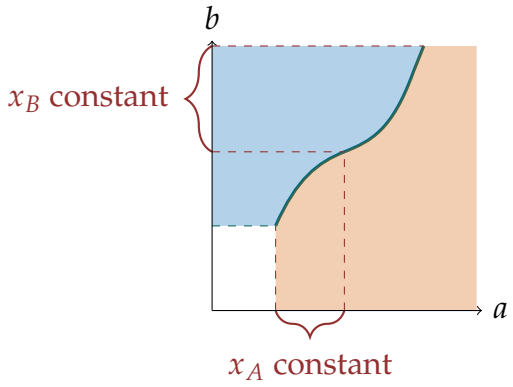

Figure 6: x_B (x_A) is constant where the boundary z is convex (concave).

5.2 Choosing the optimal boundary

Having pinned down the optimal way to implement a given z, we can turn to searching for the best boundary among all optimally implemented ones. Let z^* be optimal boundary and $\underline{a}^*, \underline{b}^*$ be the lowest participating values associated with it.

Proposition 4. *The optimal boundary* $z^* : [\underline{a}^*, \overline{a}^*] \to [\underline{b}^*, \overline{b}^*]$ *is linear.*

I prove this proposition by considering optimal control problems of choosing a boundary on a part of a convex/concave region. I show that no boundary with strictly convex/concave parts or kinks can satisfy the necessary optimality conditions, and thus that the optimal boundary has to be linear. This observation greatly simplifies the search for the optimal mechanism. Indeed, a linear boundary corresponds to a mechanism offering only two wait time and price options: one for good A and one for good B. Knowing this additional structure lets me show that the

designer always wants to allocate the whole supply of both goods. Intuitively, if she were to discard some of one good's supply, she could do better by simply lowering its price and letting demand for it increase. This intuition underlies the proof of the following lemma:

Lemma 4. *The optimal mechanism allocates the whole supply of both goods.*

The final step of the proof requires showing that the optimal linear boundary indeed has a slope of 1. Intuitively, this is the only slope which can be implemented by not requiring any agents to wait. To see this, recall that the following condition has to hold for any boundary with slope s:

$$x_A(a) = x_B(x(a)) \cdot s.$$

Lemma 3 then tells us that the optimal implementation of the boundary features constant x_A, x_B, with the higher one being equal to 1. Thus, when $s \geq 1$, we have $x_A = 1$ and $x_B = 1/s$. While setting a slope of $s > 1$ would require introducing wait times for one of the goods, it would also lower its price. Still, the intuitions from Section 4 suggest this trade-off should be resolved in favour of not inflicting delays:

Lemma 5. *The optimal boundary z^* is linear with a slope of* 1.

Thus, the optimal mechanism allocates the whole supply of both goods, and does so without delays. To make (IC) hold, it must then charge the same prices p_A, p_B to all agents receiving a given good. By Proposition 3, this mechanism also implements the efficinet allocation.

6 Conclusion

When a welfarist designer allocates two scarce goods among agents with heterogeneous valuations, screening with prices strictly dominates screening with wait times even if the designer has no value for the collected revenue. Intuitively, requiring agents to wait disproportionately penalizes high-value types, while charging money imposes a uniform cost across all takers and thus preserves more surplus. Moreover, wait times can only screen on relative preferences for the two goods, whereas prices allow the designer to elicit absolute valuations; thus, they tend to implement more efficient allocations. Despite the model's stylized nature, these two intuitions provide more general lessons about the nature of prices and wait times as screening instruments.

While completely eliminating wait times is impractical in many contexts—waitlists for housing inevitably appear due to supply and demand fluctuations—the core observations of the paper remain relevant even then. For instance, public housing programs frequently feature extreme disparities in wait times between different developments, effectively enticing participants to choose between them based on willingness to wait. The policymaker could mitigate such disparities by adjusting the units' rents or subsidies, thereby reducing the reliance on waiting as a screening device.

References

AKBARPOUR, M., P. DWORCZAK, AND F. YANG (2023): "Comparison of Screening Devices," in *Proceedings of the 24th ACM Conference on Economics and Computation*, 60–60.

ARNOSTI, N. AND P. SHI (2020): "Design of lotteries and wait-lists for affordable housing allocation," *Management Science*, 66, 2291–2307.

ASHLAGI, I., F. MONACHOU, AND A. NIKZAD (2024): "Optimal Allocation via Waitlists: Simplicity Through Information Design," *The Review of Economic Studies*, rdae013.

BARZEL, Y. (1974): "A theory of rationing by waiting," *The Journal of Law and Economics*, 17, 73–95.

BESLEY, T. AND S. COATE (1991): "Public Provision of Private Goods and the Redistribution of Income," *The American Economic Review*, 81, 979–984.

BLOCH, F. AND D. CANTALA (2017): "Dynamic Assignment of Objects to Queuing Agents," *American Economic Journal: Microeconomics*, 9, 88–122.

CONDORELLI, D. (2012): "What money can't buy: Efficient mechanism design with costly signals," *Games and Economic Behavior*, 75, 613–624.

DENECKERE, R. J. AND P. R. MCAFEE (1996): "Damaged goods," *Journal of Economics & Management Strategy*, 5, 149–174.

HARTLINE, J. D. AND T. ROUGHGARDEN (2008): "Optimal mechanism design and money burning," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 75–84.

LESHNO, J. D. (2022): "Dynamic matching in overloaded waiting lists," *American Economic Review*, 112, 3876–3910.

MILGROM, P. AND I. SEGAL (2002): "Envelope theorems for arbitrary choice sets," *Econometrica*, 70, 583–601.

MYERSON, R. B. (1981): "Optimal Auction Design," *Mathematics of Operations Research*, 6, 58–73.

NEUSTADT, L. W. (1976): *Optimization: A Theory of Necessary Conditions*, Princeton University Press.

SEIERSTAD, A. AND K. SYDSAETER (1986): *Optimal control theory with economic applications*, Elsevier North-Holland, Inc.

VAN DIJK, W. (2019): "The socio-economic consequences of housing assistance," *University of Chicago Kenneth C. Griffin Department of Economics job market paper, 0–46 i–xi*, 36.

VAN OMMEREN, J. N. AND A. J. VAN DER VLIST (2016): "Households' willingness to pay for public housing," *Journal of Urban Economics*, 92, 91–105.

WORLD HEALTH ORGANIZATION (2023): "High-value referrals: learning from challenges and opportunities of the COVID-19 pandemic: concept paper," *High-value referrals: learning from challenges and opportunities of the COVID-19 pandemic: concept paper*.

YANG, F. (2021): "Costly multidimensional screening," *arXiv preprint arXiv:2109.00487*.

A Omitted proofs

The following fact and corollary will be useful throughout.

Fact 1. *Let f be differentiable on $(0,\gamma)$ and $\lim_{x\to 0^+} f'(x) = l \in \mathbb{R}$. Then $f'_+(0) = l$.*

Proof. For $x \in (0,\gamma)$, we have $f(x) - f(0) = \int_0^x f'(t)dt$. We therefore want to show that:

$$\lim_{x\to 0^+} \frac{f(x)-f(0)}{x} = \lim_{x\to 0^+} \frac{1}{x}\int_0^x f'(t)dt = l \quad \Leftrightarrow \quad \lim_{x\to 0^+} \frac{1}{x}\int_0^x \big(f'(t)-l\big)\,dt = 0.$$

We do it by showing that for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x \in (0,\delta)$ we have:

$$\left|\frac{1}{x}\int_0^x \big(f'(t)-l\big)\,dt\right| < \epsilon.$$

Fix any such ϵ. Since $f'(t) \to l$, there exists $\delta > 0$ such that for all $t \in (0,\delta)$ we have:

$$|f'(t) - l| < \epsilon.$$

Take any $x \in (0,\delta)$, integrate the above inequality over $(0,x)$ and divide through by x. Then:

$$\frac{1}{x}\int_0^x |f'(t) - l|dt < \epsilon.$$

The triangle inequality then gives:

$$\frac{1}{x}\left|\int_0^x \big(f'(t)-l\big)\,dt\right| \le \frac{1}{x}\int_0^x |f'(t)-l|dt < \epsilon.$$

□

Corollary 1. *Let f be differentiable on $(-\gamma,0)$ and $(0,\gamma)$ and such that:*

$$\lim_{x\to 0^+} f'(x) = \lim_{x\to 0^-} f'(x) = l \in \mathbb{R}.$$

Then $f'(0)$ exists and is equal to l.

A.1 Proof of Lemma 1

Suppose some type $(a,b) < (\underline{a},\underline{b})$ could weakly benefit from requesting either good. Then some type $(a+\epsilon, b+\epsilon) < (\underline{a},\underline{b})$, for $\epsilon > 0$ small enough, would strictly benefit from it, so one of $U_A(a+\epsilon)$ and $U_B(b+\epsilon)$ would have to be strictly above zero. Since U_A, U_B are increasing, this contradicts the definition of $(\underline{a},\underline{b})$. Thus, $y(a,b) = \varnothing$ for all $(a,b) < (\underline{a},\underline{b})$.

Analogously, all types for whom $a > \underline{a}$ or $b > \underline{b}$ strictly benefit from choosing one of the goods. Moreover, a positive mass of types gets either good, and thus $\underline{a},\underline{b} < 1$. Let us now identify the set of types $(a,b) \geq (\underline{a},\underline{b})$ who are indifferent between the two goods, that is, for whom $U_A(a) = U_B(b)$. This is the case for $(\underline{a},\underline{b})$ by construction. Recall also that U_A, U_B are continuous and strictly increasing on $[\underline{a},1]$ and $[\underline{b},1]$, respectively. Therefore, all indifferent types must lie on a continuous and strictly increasing curve originating from $(\underline{a},\underline{b})$. Let $z(a) : [\underline{a},\overline{a}] \to \mathbb{R}$ describe this curve; notice that either $\overline{a} = 1$ or $z(\overline{a}) = 1$. By construction, any type $(a, z(a)) > (\underline{a},\underline{b})$ is indifferent between her best options for both goods. Then, by the standard single-crossing argument, any type $(a', z(a))$ with $a' > a$ strictly prefers to choose good A. Analogously, any type (a, b') with $b' > z(a)$ strictly prefers to choose good B.

A.2 Proof of Lemma 2

($\Rightarrow$). Fix any increasing and piecewise continuously differentiable x_A, x_B that implement a boundary satisfying properties 1 – 3. I will construct a feasible mechanism that corresponds to them. Define $\tilde{U}_A, \tilde{U}_B$ such that:

$$\tilde{U}_A(a) = \int_0^a x_A(v)da, \quad \tilde{U}_B(b) = \int_0^b x_B(v)db.$$

I will show that $\tilde{U}_A, \tilde{U}_B$ are the good-specific indirect utilities for the following mechanism, and that the mechanism is feasible:

$$p(a,b) = \begin{cases} a \cdot x_A(a) - \tilde{U}_A(a), & \text{if } \tilde{U}_A(a) \geq \tilde{U}_B(b), \\ b \cdot x_B(b) - \tilde{U}_B(b), & \text{if } \tilde{U}_B(b) > \tilde{U}_A(a), \end{cases}$$

$$x(a,b) = \begin{cases} x_A(a), & \text{if } \tilde{U}_A(a) \geq \tilde{U}_B(b), \\ x_B(b), & \text{if } \tilde{U}_B(b) > \tilde{U}_A(a), \end{cases}$$

$$y(a,b) = \begin{cases} \varnothing, & \text{if } (a,b) \leq (\underline{a},\underline{b}), \\ A, & \text{if } a > \underline{a} \text{ and } \tilde{U}_A(a) \geq \tilde{U}_B(b), \\ B, & \text{if } b > \underline{b} \text{ and } \tilde{U}_B(b) > \tilde{U}_A(a). \end{cases}$$

A standard envelope argument verifies that, under this payment rule, no (a,b) wants to misreport to (a',b') for which $y(a',b') = y(a,b)$. That is, conditional on choosing the good she was assigned, (a,b) prefers her assigned wait time and payment option. Then $\tilde{U}_A, \tilde{U}_B$ are indeed

the good-specific indirect utilities for this mechanism because:

$$U_A(a) = a \cdot x_A(a) - \big(a \cdot x_A(a) - \tilde{U}_A(a)\big) = \tilde{U}_A(a),$$
$$U_B(b) = b \cdot x_B(b) - \big(b \cdot x_B(b) - \tilde{U}_B(b)\big) = \tilde{U}_B(b).$$

Verifying that (IC) holds therefore only requires checking that no (a,b) wants to misreport to (a',b') for which $y(a',b') \neq y(a,b)$. But since $\tilde{U}_A(a)$ and $\tilde{U}_B(b)$ are the best utilities (a,b) can get from either good, this is true by the construction of $y(a,b)$. Note also that (IR) must hold as both good-specific indirect utilities are positive everywhere.

It therefore remains to check the supply condition (S). By Lemma 1:

$$\int \mathbb{1}_{y(a,b)=A} \, \mathrm{d}F(a,b) = \int_{\underline{a}}^{1} \int_{0}^{z(\min[a,\bar{a}])} f(a,v)\, dv\, da \leq \mu_A,$$
$$\int \mathbb{1}_{y(a,b)=B} \, \mathrm{d}F(a,b) = \int_{\underline{b}}^{1} \int_{0}^{z^{-1}(\min[b,\bar{b}])} f(v,b)\, dv\, db \leq \mu_B,$$

where the inequalities hold by property 1. Thus, the supply constraint (S) must hold.

($\Leftarrow$). Fix any feasible mechanism (p,x,y). Then its U_A and U_B must be convex, and so x_A, x_B must be weakly increasing. Furthermore, by Lemma 1, types $(a,0)$ where $a > \underline{a}$ choose good A and types $(0,b)$ where $b > \underline{b}$ choose good B. Then the following holds almost everywhere:

$$x_A(a) = x(a,0) \;\text{ for }\; a \in [\underline{a},\bar{a}], \quad x_B(b) = x(0,b) \;\text{ for }\; b \in [z(\underline{a}), z(\bar{a})]. \tag{7}$$

Since x is piecewise continuously differentiable, so are x_A and x_B.

I now show the boundary z satisfies property 2. Integrating (7) tells us that $\tilde{U}_A$ and $\tilde{U}_B$ are piecewise twice continuously differentiable. Now, by Lemma 1, z satisfies the boundary indifference condition:

$$U_A(a) = U_B(z(a)) \quad \text{for all} \quad a \in [\underline{a},\bar{a}]. \tag{I}$$

U_B is strictly increasing on $[z(\underline{a}), z(\bar{a})]$, so it is invertible there:

$$U_B^{-1}(U_A(a)) = z(a) \quad \text{for all} \quad a \in [\underline{a},\bar{a}].$$

Since U_A and U_B were piecewise twice continuously differentiable on $[\underline{a},\bar{a}]$ and $[z(\underline{a}), z(\bar{a})]$, respectively, z is piecewise twice continuously differentiable on $(\underline{a},\bar{a})$.

I now prove z satisfies property 3. I show that left-derivatives are positive and finite for all $a \in (\underline{a},\bar{a}]$ (the argument for right-derivatives on $(\underline{a},\bar{a})$ is analogous). Since z is piecewise twice continuously differentiable, it suffices to check the finitely many points where it is not differentiable. Consider any such $\hat{a}$. Then z is differentiable in some neighbourhood $(\hat{a}-\epsilon, \hat{a})$ for $\epsilon > 0$ small enough. On that interval, differentiating (I) gives:

$$x_A(a) = x_B(z(a)) \cdot z'(a). \tag{DI}$$

Since $x_B(z(a)) > 0$ on this interval, we have:

$$z'(a) = \frac{x_A(a)}{x_B(z(a))}.$$

Note that $x_A(a) \uparrow x_A(a^-) > 0$ and $x_B(z(a)) \uparrow x_B(z(a)^-) > 0$ as $a \uparrow \hat{a}$. Thus, we also know that $z'(a) \uparrow x_A(a^-)/x_B(z(a)^-)$. Fact 1 then completes the proof.

Finally, let us show that (S′) holds. Recall $U_A(a) > U_B(b)$ for almost all agents for whom $y(a,b) = A$, so by Lemma 1:

$$\int_{\underline{a}}^{1} \int_{0}^{z(\min[a,\bar{a}])} f(a,v)dv\, da = \int \mathbb{1}_{y(a,b)=A}\, \mathrm{d}F(a,b) \leq \mu_A,$$

where the inequality holds by (S). An analogous expression also holds for B.

A.3 Proof of Lemma 3

In what follows I prove properties 1, 3, and 5; property 2 is symmetric to property 1. The proof of property 4 is analogous to that of property 2. I later show that a mechanism satisfying properties 1 – 5 exists for every implementable boundary and is unique.

Fact 2 (Property 1)**.** *On concave regions, $x_A(a)$ is constant and $x_B(z(a)) \propto 1/z'(a)$.*

Proof. Let x_A, x_B be some discounting rules implementing z and consider any concave region $[\underline{v}, \overline{v}]$. I will propose feasible $\tilde{x}_A, \tilde{x}_B$ that implement z and improve upon x_A, x_B; the improvement is strict if the statement of the fact did not hold for the original x_A, x_B.

Define $\xi : [\underline{v}, \overline{v}] \to \mathbb{R}$ such that:

$$\xi(v) = \begin{cases} 0, & \text{if } v = \underline{a}, \\ \dfrac{x_A(\overline{v})}{x_A(v)}, & \text{if } v \neq \underline{a}. \end{cases}$$

x_A is increasing, so $\xi(v)$ is decreasing and ≥ 1 for $v > \underline{a}$. Let $\tilde{x}_A, \tilde{x}_B$ be discounting rules s.t.:

$$\tilde{x}_A(a) = \begin{cases} x_A(a) \cdot \xi(\underline{v}) & \text{if } a \leq \underline{v}, \\ x_A(a) \cdot \xi(a) & \text{if } \underline{v} < a \leq \overline{v}, \\ x_A(a) & \text{if } a > \overline{v}, \end{cases} \quad \tilde{x}_B(b) = \begin{cases} x_B(b) \cdot \xi(z^{-1}(\underline{v})) & \text{if } b \leq z(\underline{v}), \\ x_B(b) \cdot \xi(z^{-1}(a)) & \text{if } z(\underline{v}) < b \leq z(\overline{v}), \\ x_B(b) & \text{if } b > z(\overline{v}). \end{cases}$$

I will denote their corresponding good-specific indirect utilities by $\tilde{U}_A, \tilde{U}_B$. The remainder of the proof consists of showing that this alternative mechanism implements the boundary z, that it is feasible, and that it improves upon the original mechanism.

Implementing the boundary. It suffices to show that $\tilde{U}_A, \tilde{U}_B$ satisfy the boundary indifference condition:

$$\tilde{U}_A(a) = \tilde{U}_B(z(a)) \quad \text{for all} \quad a \in [\underline{a}, \overline{a}]. \tag{I′}$$

First, recall that (I) for x_A, x_B implies (DI) almost everywhere:

$$x_A(a) = x_B(z(a)) \cdot z'(a). \tag{DI}$$

Now, consider (DI) on $a \in [0, \underline{a}]$ and multiply both sides of (DI) by $\xi(\underline{v})$ to obtain:

$$\underbrace{x_A(a) \cdot \xi(\underline{v})}_{=\tilde{x}_A(a)} = \underbrace{x_B(z(a)) \cdot \xi(\underline{v})}_{=\tilde{x}_B(z(a))} \cdot z'(a). \tag{DI'}$$

Integrating the above equation over $[0, \underline{a}]$ then gives $\tilde{U}_A(a) = \tilde{U}_B(z(a))$ on this interval.

Let us now show that (I') holds on $[\underline{v}, \overline{v}]$. By the above, we know that:

$$\tilde{U}_A(\underline{v}) = \tilde{U}_B(z(\underline{v})). \tag{8}$$

Take (DI) for $a \in [\underline{v}, \overline{v}]$ and multiply both sides by $\xi(a)$:

$$\underbrace{x_A(a) \cdot \xi(a)}_{=\tilde{x}_A(a)} = \underbrace{x_B(z(a)) \cdot \xi(a)}_{=\tilde{x}_B(z(a))} \cdot z'(a).$$

Fix any $a \in [\underline{v}, \overline{v}]$ and integrate the above equation over $[\underline{v}, a]$. This gives:

$$\tilde{U}_A(a) - \tilde{U}_A(\underline{v}) = \tilde{U}_B(z(a)) - \tilde{U}_B(z(\underline{v})).$$

Summing it with (8) then yields (I') for any $a \in [\underline{v}, \overline{v}]$.

Let us then show (I') holds on $[\overline{v}, \overline{a}]$ as well. By the above, we know that:

$$\tilde{U}_A(\overline{v}) = \tilde{U}_B(z(\overline{v})). \tag{9}$$

Recall that $\tilde{x}_A(a) = x_A(a)$ and $\tilde{x}_B(z(a)) = x_B(z(a))$ for $a \in [\overline{v}, \overline{a}]$. Take any a in this interval. Then integrating both sides of (DI) over $[\overline{v}, a]$ and summing it with (9) yields (I') for any $a \in [\overline{v}, \overline{a}]$. □

Feasibility. Since the original mechanism also implemented z, Lemma 2 tells us it remains to show that $\tilde{x}_A, \tilde{x}_B$ are piecewise continuously differentiable and weakly increasing. Note $\tilde{x}_A, \tilde{x}_B$ inherit piecewise continuous differentiability from x_A, x_B. They are also weakly increasing on $[\underline{a}, \underline{v}]$ and $[\underline{b}, z(\underline{v})]$, respectively, as $\tilde{x}_A, \tilde{x}_B$ are constructed by rescaling weakly increasing x_A, x_B by $\xi(\underline{v}) > 0$. On $(\underline{v}, \overline{v}]$, we have:

$$\tilde{x}_A(a) = x_A(a) \cdot \xi(a) = x_A(a) \cdot \frac{x_A(\overline{v})}{x_A(a)} = x_A(\overline{v}),$$

and thus $\tilde{x}_A$ is constant on this interval. Moreover, the 'pasting' of x_A at $\underline{v}$ preserves monotonicity as $x_A(\underline{v}) = x_A(\underline{v}^+)$. Now, on $(z(\underline{v}), z(\overline{v})]$ we have:

$$\tilde{x}_B(z(a)) = x_B(z(a)) \cdot \xi(z^{-1}(z(a))) = x_B(z(a)) \cdot \frac{x_A(\overline{v})}{x_A(a)}.$$

However, Fact 1, (DI), and the left-continuity of x_A tell us that $x_B(z(a)) = x_A(a)/z'_-(a)$, giving:

$$\tilde{x}_B(z(a)) = \frac{x_A(\overline{v})}{z'_-(a)}.$$

Here $z'_-(a)$ is the left derivative of $z(a)$, which exists because z is concave on this interval. Then the LHS is increasing as $z'_-(a)$ is positive and decreasing on the concave region $[\underline{v}, \overline{v}]$. Moreover, the 'pasting' of $x_B(b)$ at $z(\underline{v})$ preserves monotonicity as:

$$\tilde{x}_B(z(\underline{v})) = x_B(z(\underline{v})) \cdot \frac{x_A(\overline{v})}{x_A(\underline{a})} = \tilde{x}_B(z(\underline{v})^+).$$

Let us finally consider $(\overline{v}, \overline{a}]$ and $(z(\overline{v}), z(\overline{a})]$. Since $\tilde{x}_A, \tilde{x}_B$ agree with x_A, x_B there, they will be weakly increasing. Moreover, the pastings at $\overline{v}$ and $z(\overline{v})$ preserve monotonicity because:

$$\tilde{x}_A(\overline{v}) = x_A(\overline{v}^+), \quad \tilde{x}_B(z(\overline{v})) = x_B(z(\overline{v})^+).$$

Improvement. Note $\tilde{x}_A, \tilde{x}_B$ are pointwise higher than x_A, x_B and so $\tilde{U}_A, \tilde{U}_B$ are pointwise higher than U_A, U_B. Since total welfare (W′) features all good-specific indirect utilities with strictly positive weights, the proposed mechanism improves upon the originial one. If the original mechanism did not satisfy property 1, $\tilde{U}_A, \tilde{U}_B$ would be strictly higher than U_A, U_B on some interval, and so the improvement would be strict.

Fact 3 (Property 3). *At least one of $x_A(a)$ and $x_B(z(a))$ is continuous at every $a \in (\underline{a}, \overline{a}]$.*

Proof. Suppose there is $\hat{a} \in (\underline{a}, \overline{a})$ where neither $x_A(a)$ nor $x_B(z(a))$ are continuous; I will construct $\tilde{x}_A, \tilde{x}_B$ that implement z and strictly improve upon the original allocation. Define:

$$\chi := \max\left[\frac{x_A(\hat{a}^+)}{x_A(\hat{a})}, \frac{x_B(z(\hat{a})^+)}{x_B(z(\hat{a}))}\right], \tag{10}$$

where these right-limits exist because x_A, x_B are both increasing. Since both x_A, x_B are left-continuous and discontinuous at $\hat{a}$ and $z(\hat{a})$, respectively, we have $\chi > 1$.

Assume without loss that $\frac{x_A(\hat{a}^+)}{x_A(\hat{a})} \leq \frac{x_B(z(\hat{a})^+)}{x_B(z(\hat{a}))}$ and consider the proposed improvement:

$$\tilde{x}_A(a) = \begin{cases} x_A(a) \cdot \chi, & \text{if } a \leq \hat{a}, \\ x_A(a), & \text{if } a > \hat{a}, \end{cases} \qquad \tilde{x}_B(b) = \begin{cases} x_B(b) \cdot \chi, & \text{if } b \leq z(\hat{a}), \\ x_B(b), & \text{if } b > z(\hat{a}). \end{cases}$$

I will now verify that a mechanism with these discounting rules implements the boundary z, that it is feasible, and that it strictly improves upon the original one.

Implementing the boundary. It suffices to show that $\tilde{U}_A, \tilde{U}_B$ satisfy the boundary indifference condition (I′). We know that U_A, U_B satisfy (I). Since $\tilde{U}_A, \tilde{U}_B$ coincide with U_A, U_B for $a \geq \hat{a}$, (I′) holds there too. Moreover, integrating $\tilde{x}_A, \tilde{x}_B$ tells us that for $a < \hat{a}$:

$$\tilde{U}_A(a) = U_A(a) \cdot \chi, \quad \tilde{U}_B(z(a)) = U_B(z(a)) \cdot \chi,$$

meaning that (I′) holds there too.

Feasibility. It suffices to show that $\tilde{x}_A, \tilde{x}_B$ are piecewise continuously differentiable and weakly increasing. Note $\tilde{x}_A, \tilde{x}_B$ inherit piecewise continuous differentiability from x_A, x_B. Let us now show x_A, x_B are weakly increasing on $[\underline{a}, \hat{a}]$ and $[z(\underline{a}), z(\hat{a})]$, respectively. There, $\tilde{x}_A$ and $\tilde{x}_B$ are constructed by rescaling x_A and x_B by a positive constant, and thus are increasing. They are also trivially increasing on $(\hat{a}, \overline{a}]$ and $(z(\hat{a}), z(\underline{a})]$, as $\tilde{x}_A, \tilde{x}_B$ and x_A, x_B coincide there. It therefore remains to check the pasting points $\hat{a}, z(\hat{a})$. For $\hat{a}$, we verify this as follows:

$$\tilde{x}(\hat{a}) = x_A(\hat{a}) \cdot \chi = x_A(\hat{a}) \cdot \frac{x_A(\hat{a}^+)}{x_A(\hat{a})} = x_A(\hat{a}^+).$$

For $z(\hat{a})$, we have:

$$\tilde{x}_B(z(\hat{a})) = x_B(z(\hat{a})) \cdot \chi = x_B(z(\hat{a})) \cdot \frac{x_A(\hat{a}^+)}{x_A(\hat{a})}.$$

However, recall we assumed that $\frac{x_A(\hat{a}^+)}{x_A(\hat{a})} \leq \frac{x_B(z(\hat{a})^+)}{x_B(z(\hat{a}))}$, and thus:

$$\tilde{x}_B(z(\hat{a})) \cdot \chi = x_B(z(\hat{a})) \cdot \frac{x_A(\hat{a}^+)}{x_A(\hat{a})} \leq x_B(z(\hat{a})) \cdot \frac{x_B(z(\hat{a})^+)}{x_B(z(\hat{a}))} = x_B(z(\hat{a})^+).$$

Improvement. The argument is analogous to that in the proof of Fact 3. □

Fact 4 (Property 5)**.** $\max[x_A(1), x_B(1)] = 1$.

Proof. Suppose $\max[x_A(1), x_B(1)] < 1$ and define:

$$\chi = \frac{1}{\max[x_A(1), x_B(1)]}, \quad \tilde{x}_A(a) = x_A(a) \cdot \chi, \quad \tilde{x}_B(b) = x_B(b) \cdot \chi.$$

Note $\tilde{x}_A, \tilde{x}_B$ inherit piecewise continuous differentiability and monotonicity from x_A, x_B and still implement z as rescaling them does not affect (I). They are both a.e. strictly above x_A, x_B and thus, analogously to the preceding arguments, give higher total welfare than x_A, x_B. □

It remains to show that the mechanism satisfying properties 1 – 5 exists for every boundary and is unique. Let us start with uniqueness. By Lemma 2, every such boundary can be partitioned into finitely many concave and convex regions. Let $[\underline{v}_1, \overline{v}_1]$ be the last such region and fix some strictly positive allocation of x_A at point $\overline{v}_1$. Then properties 1 – 3 ensure that $x_A(\overline{v}_1)$ pins down discounting allocations everywhere on $(\underline{a}, \overline{v}_1]$ and $(\underline{b}, z(\overline{v}_1))]$. To see why, consider first the last region $[\underline{v}_1, \overline{v}_1]$. Then the allocations on $(\underline{v}_1, \overline{v}_1]$ and $(z(\underline{v}_1), z(\overline{v}_1))]$ are pinned down by $x_A(\overline{v}_1)$, the left-continuity of x_A, x_B at $\overline{v}_1$ and $z(\overline{v}_1)$, and property 1 or 2 (depending on whether it is a convex or concave region). If $\underline{v}_1 = \underline{a}$, we are done. Otherwise, let $[\underline{v}_2, \underline{v}_1]$ be the next concave/convex region. The allocation on $(\underline{v}_2, \underline{v}_1]$ is then pinned down by the right limits $x_A(\underline{v}_1^+), x_B(z(\underline{v}_1)^+)$, properties 1 or 2, and the 'pasting' property 3 which requires that one of x_A or x_B be continuous at $\underline{v}_1$ or $z(\underline{v}_1)$. Which one of x_A and x_B is pasted smoothly is pinned down by the monotonicity requirement. An inductive argument then extends this reasoning to all the convex/concave intervals.

Thus, $x_A(\overline{v}_1)$ pins down the allocation everywhere on $(\underline{a}, \overline{v}_1]$ and $(\underline{b}, z(\overline{v}_1)]$. Then, by property 4 and the 'pasting' property 3, the allocations $x_A(\overline{v}_1)$ and $x_B(z(\overline{v}_1))$ also pin down the discounting rules on $(\overline{a}, 1]$ or $(\overline{b}, 1]$, whenever such regions exist. Moreover, condition 5 tells us that the largest allocation of x_A, x_B has to equal 1. This pins down the scale of the constructed allocation, and thus pins down $x_A(\underline{v}_1)$ itself. Consequently, properties 1 – 5 pin down the optimal mechanism uniquely for every z.

Finally, the existence of such a mechanism is guaranteed by the properties of z given by Lemma 2. Indeed, consider any concave/convex region $[\underline{v}, \overline{v}]$ and fix $x_A(\overline{v}) > 0$. Since the left-derivative $z'_-(\overline{v})$ exists and is strictly positive, there exist allocation rules x_A, x_B that satisfy properties 1 or 2 on it (depending on whether the region is convex or concave). Moreover, if $\underline{v} > \underline{a}$, the existence of a strictly positive right-derivative $z'_+(\underline{v})$ ensures that $x_A(\underline{v}), x_B(z(\underline{v}))$ satisfying these properties are strictly positive. Consequently, the mechanism described above is indeed associated with admissible allocation rules x_A, x_B.

A.4 Proof of Proposition 4

A.4.1 The optimal boundary is piecewise linear. I first prove that the optimal boundary z^* cannot have strictly concave regions (by symmetry, the same then applies to convex regions). I begin by showing that z^* has to solve the following optimal control problem on every closed interval where it is concave and twice continuously differentiable:

Problem 1. *Choose the control $u : [\underline{v}, \overline{v}] \to \mathbb{R}_-$ and state variables $z, y, q : [\underline{v}, \overline{v}] \to \mathbb{R}$ to maximize:*

$$-\int_{\underline{v}}^{\overline{v}} G(a)H(z(a))\,da, \tag{11}$$

subject to the following laws of motion:

$$z'(v) = y(v), \quad y'(v) = u(v), \quad q'(v) = g(v) \cdot H(z(v)),$$

and the following end-point constraints:

$$z(\underline{v}) = z^*(\underline{v}), \quad z(\overline{v}) = z^*(\overline{v}), \tag{12}$$

$$y(\underline{v}) = z^{*\prime}_+(\underline{v}), \quad y(\overline{v}) = z^{*\prime}_-(\overline{v}), \tag{13}$$

$$q(\underline{v}) = 0, \quad q(\overline{v}) = \int_{\underline{v}}^{\overline{v}} g(v) \cdot H(z^*(v))\,dv. \tag{14}$$

The states z and y correspond to the boundary and its derivative, the control u corresponds to its second derivative, and q is introduced to capture the supply constraint.

Lemma 6. *Let $[\underline{v}, \overline{v}]$ be a concave region such that z^* is twice continuously differentiable on it. Then z^* has to solve Problem 1 on $[\underline{v}, \overline{v}]$.*

Proof. First, note that the optimal boundary z^* is absolutely continuous on $[\underline{v}, \overline{v}]$ as it is continuously differentiable on this interval. Since it is also concave there, it is admissible in Problem

1. Now, consider any $z : [\underline{v}, \overline{v}] \to \mathbb{R}$ that is admissible in Problem 1 and define:

$$\tilde{z}(v) = \begin{cases} z(v) & \text{if } v \in [\underline{v}, \overline{v}], \\ z^*(v) & \text{elsewhere.} \end{cases}$$

Note that $\tilde{z}$ satisfies properties 1 – 3 of Lemma 2 and thus, by Lemma 3, is implementable.

Therefore, for z^* to be optimal in the original problem, it must give higher total welfare (W′) than any such $\tilde{z}$. In what follows I show that this is only true if z^* solves Problem 1. I do so by showing that for all z that are admissible in Problem 1, total welfare (W′) is identically equal to (20) up to an affine transformation. This argument relies on the following corollary which is given by a construction analogous to those in the proof of Lemma 3:

Corollary 2. *Let $z_1, z_2 : [\underline{a}, \overline{a}] \to \mathbb{R}$ be implementable boundaries and suppose $[\underline{v}, \overline{v}]$ is a concave region for both z_1 and z_2. Suppose further that:*

$$z_1(a) = z_2(a) \quad \textit{for all} \quad a \notin (\underline{v}, \overline{v}),$$

and that:

$$z'_{1+}(\underline{v}) = z'_{2+}(\underline{v}), \quad z'_{1-}(\underline{v}) = z'_{2-}(\underline{v}).$$

Then the good-specific indirect utilities optimally implementing z_1 and z_2 agree everywhere except $(\underline{v}, \overline{v})$ and $(z_1(\underline{v}), z_1(\overline{v}))$.

Intuitively, Corollary 2 says that perturbing a boundary inside an interval where it is concave does not affect the indirect utilities optimally implementing it outside this interval. Thus, Corollary 2 tells us that total welfare depends on z only through the following terms:

$$\int_{\underline{v}}^{\overline{v}} U_A(a) \cdot g(a) \cdot H(z(a)) da + \int_{z(\underline{v})}^{z(\overline{v})} U_B(b) \cdot G(z^{-1}(b)) \cdot h(b) db. \tag{15}$$

A change of variables lets us rewrite the latter term as follows:

$$\begin{aligned} \int_{z(\underline{v})}^{z(\overline{v})} U_B(b) \cdot G(z^{-1}(b)) h(b)\, db &= \int_{z^{-1}(z(\underline{v}))}^{z^{-1}(z(\overline{v}))} \underbrace{U_B(z(a))}_{:= U_A(a)} \cdot z'(a) \cdot G(z^{-1}(z(a))) h(z(a))\, da \\ &= \int_{\underline{v}}^{\overline{v}} U_A(a) \cdot z'(a) \cdot G(a) h(z(a)) da. \end{aligned}$$

Thus, (15) becomes:

$$\int_{\underline{v}}^{\overline{v}} U_A(a) \left[g(a) H(z(a)) + z'(a) \cdot G(a) h(z(a)) \right] da.$$

Integrating by parts yields:

$$U_A(\overline{v}) G(\overline{v}) H(z(\overline{v})) - U_A(\underline{v}) G(\underline{v}) H(z(\underline{v})) - \int_{\underline{v}}^{\overline{v}} U'_A(a) G(a) H(z(a))\, da.$$

However, $\overline{v}, \underline{v}, z(\overline{v}), z(\underline{v})$ are fixed for all z admissible in Problem 1 and $U_A(\underline{v}), U_A(\overline{v})$ are fixed by Corollary 2. Thus, for such z, (W′) is identically equal the following, up to a constant:

$$-\int_{\underline{v}}^{\overline{v}} x_A(a)G(a)H(z(a))\,da. \tag{16}$$

Finally, by Lemma 3, x_A is constant on the concave interval $[\underline{v}, \overline{v}]$. Moreover, its value there is pinned down by $z'_+(\underline{v})$, which is fixed for all admissible z. Thus, total welfare (W′) is identically equal to (11) up to an affine transformation. □

We can now show that z^* is piecewise linear. Suppose towards a contradiction that $z^{*\prime\prime}(\hat{v}) < 0$ for some $\hat{v}$. Since $z^{*\prime\prime}$ is piecewise continuous, there must be some interval $[\underline{v}, \overline{v}]$ around $\hat{v}$ such that $z^{*\prime\prime}(v) < 0$ on it. Consider Problem 1 for that interval. As shown, z^* restricted to $[\underline{v}, \overline{v}]$ must be the optimal z for that problem. Let $(z^*, y^*, q^*, u^*, \xi, \phi, \eta)$ be optimal the collection of states, controls and costates associated with z^*. The Hamiltonian for this problem is:

$$\mathcal{H} = -G(a)H(z^*(a)) + \mu(a) \cdot g(a)H(z^*(a)) + \xi(a) \cdot y(a) + \phi(a) \cdot u(a), \tag{17}$$

where $\mu(a)$ is the costate on q, ξ is the costate on z and ϕ is the costate on y. By the Maximum Principle, we then have:

$$\mu'(a) = 0.$$

Since $\mu(a)$ is constant, I will simply write it as μ. Moreover, we have:

$$\xi'(a) = -\big(-G(a) + \mu \cdot g(a)\big)h(z^*(a)) = \big(G(a) - \mu \cdot g(a)\big)h(z^*(a)), \tag{18}$$

and:

$$\phi'(a) = -\xi(a), \tag{19}$$

giving:

$$\phi''(a) = -\xi'(a) = -\big(G(a) - \mu \cdot g(a)\big)h(z^*(a)).$$

The Maximum principle further tells us that controls $u^*(v) < 0$ must maximize the Hamiltonian everywhere in $(\underline{v}, \overline{v})$. However, the Hamiltonian depends on the control linearly and so the optimal control can be interior only if $\phi(v) = 0$ on $(\underline{v}, \overline{v})$. In particular, this means that $\phi''(a)$ has to be zero in that region. Since $h(z(a)) > 0$, this gives:

$$0 = -G(a) + \mu \cdot g(a) \quad \Rightarrow \quad \frac{G(a)}{g(a)} = \mu,$$

which cannot hold since G/g is strictly increasing. Thus, $z^{*\prime\prime}(a) = 0$ wherever z^* is twice-differentiable. Since it was piecewise twice continuously differentiable, it follows that z^* is piecewise linear.

A.4.2 The optimal boundary is linear. We know the optimal boundary z^* is piecewise linear, so it is also absolutely continuous on all of $[\underline{a}, \overline{a}]$. This lets us apply a similar optimal control method on the initial convex/concave region $[\underline{a}, \underline{v}]$. Assume without loss that it is concave.

Problem 2. *Choose the control $u : [\underline{a}, \overline{v}] \to \mathbb{R}_-$, state variables $z, y, q : [\underline{a}, \overline{v}] \to \mathbb{R}$, a number of jumps $n \in \mathbb{N}$, jump locations and jump sizes, $a_i \in [\underline{a}, \overline{v}]$ and $v_i \in \mathbb{R}_-$ for $i \in \{1, \dots, n\}$ to maximize:*

$$-\int_{\underline{a}}^{\overline{v}} G(a)H(z(a))\,da. \tag{20}$$

subject to the following laws of motion:

$$z'(v) = y(v), \quad y'(v) = u(v), \quad q'(v) = g(v) \cdot H(z(v)),$$

the jump function for every i:

$$v_i = y_+(a_i) - y_-(a_i),$$

and the following end-point constraints:

$$z(\underline{a}) = z^*(\underline{a}), \quad z(\overline{v}) = z^*(\overline{v}), \tag{21}$$

$$y(\underline{a}) \text{ free}, \quad y(\overline{v}) = z_-^{*\prime}(\overline{v}), \tag{22}$$

$$q(\underline{a}) = 0, \quad q(\overline{v}) = \int_{\underline{a}}^{\overline{v}} g(v) \cdot H(z^*(v))\,dv. \tag{23}$$

We then get the following lemma whose proof is analogous to that of Lemma 6:

Lemma 7. *Let $[\underline{a}, \overline{v}]$ be a concave interval of z^*. Then the optimal boundary z^* on this interval has to solve Problem 2.*

Take $\overline{v}$ such that $[\underline{a}, \overline{v}]$ is the largest concave interval starting with $\underline{a}$. Since z^* is piecewise linear, the largest initial concave interval either covers all of $[\underline{a}, \overline{a}]$, or consists of at least two linear pieces. In what follows, I show that the solution to Problem 2 cannot have jumps, and thus that the latter case cannot happen. This in turn proves that z^* is linear.

Let us now analyze the necessary conditions for z^* restricted to $[\underline{a}, \overline{v}]$ to solve Problem 2. The Hamiltonian and costate equations for this problem are the same as for Problem 1, and given by (17), (18) and (19), and μ, the costate for q, is also constant. However, the initial value of y is now free, so its costate at the beginning of the interval is zero (see Neustadt (1976), p. 234).

$$\phi(\underline{a}) = 0. \tag{24}$$

Now, by the Maximum Principle with jumps (see Seierstad and Sydsaeter (1986), Theorem 7, p. 196-197) we know that:

1. $\phi(\cdot)$ is continuous and differentiable except possibly at jump points,
2. $\phi(a^*) = 0$ when a^* is a jump point,
3. At all a where there is no jump, $\phi(a) \geq 0$.

I now show that ϕ is twice continuously differentiable on $(\underline{a}, \overline{v})$. For a other than jump points this follows because:

$$\phi'(a) = -\xi(\underline{a}) - \int_{\underline{a}}^{a} \big(G(t) - \mu \cdot g(t)\big)\, h(z(t))dt. \tag{25}$$

Now, let $a^* \in (\underline{a}, \overline{v})$ be a jump point. Then (25) holds on some open neighborhoods to the left and right of a^*. We know that $\phi'(a)$ is differentiable there, with:

$$\phi''(a) = -\big(G(a) - \mu \cdot g(a)\big)\, h(z(a)),$$

which, just like $\phi'(a)$, inherits continuity from g, h and G. Moreover, we see that:

$$\lim_{a \to a^{*+}} \phi'(a) = \lim_{a \to a^{*-}} \phi'(a), \quad \lim_{a \to a^{*+}} \phi''(a) = \lim_{a \to a^{*-}} \phi''(a),$$

where these limits are finite. Corollary 1 then tells us that $\phi'(a^*), \phi''(a^*)$ also exist and equal to these limits, and thus that ϕ is indeed twice continuously differentiable on $(\underline{a}, \overline{v})$.

Thus, if there is an interior jump at $a^* \in (\underline{a}, \overline{v})$, we must have $\phi(a^*) = 0$ there. Since we also know that $\phi(a) \geq 0$ outside of jump points, we must therefore have $\phi'(a^*) = 0$ and $\phi''(a^*) \geq 0$ there. I will show this cannot happen. Note that:

$$\begin{aligned} \phi''(a) &= -\big(G(a) - \mu\, g(a)\big)\, h\big(z(a)\big) \\ &= \big(\mu\, g(a) - G(a)\big)\, h\big(z(a)\big) \\ &= \left(\mu - \frac{G(a)}{g(a)}\right) h\big(z(a)\big) \cdot g(a). \end{aligned}$$

Recall that $G(a)/g(a)$ is strictly increasing by Assumption 1. Thus, $\phi''(a)$ is either strictly negative everywhere on $(\underline{a}, \overline{v})$ or positive until some $\tilde{a} \in (\underline{a}, \overline{v})$ and then negative forever after. In the former case, $\phi''(a) < 0$ for all $a \in (\underline{a}, \overline{v})$, and so $\phi''(a^*) \geq 0$ can never hold for an interior a^*. In the latter case, there exists some $\tilde{a} \in (\underline{a}, \overline{v})$ such that:

$$\phi''(a) \begin{cases} > 0, & \text{if } a < \tilde{a}, \\ = 0, & \text{if } a = \tilde{a}, \\ < 0, & \text{if } a > \tilde{a}. \end{cases}$$

Now, I show that for every $a \in (\underline{a}, \tilde{a}]$ we have $\phi'(a) > 0$. For suppose $\phi'(a) \leq 0$ for some $a \in (\underline{v}, \tilde{a}]$. Then, since $\phi''(a) > 0$ on $(\underline{a}, \tilde{a})$, it must be that $\phi'(a) < 0$ everywhere on $(\underline{a}, \tilde{a})$. But since $\phi(\underline{a}) = 0$, this would mean that for all $a \in (\underline{a}, \tilde{a}]$:

$$\phi(a) = \underbrace{\phi(\underline{a})}_{=0} + \int_{\underline{a}}^{a} \underbrace{\phi'(t)}_{<0}\, dt < 0,$$

which cannot be as $\phi(a) \geq 0$ on the whole interval by the Maximum Principle with jumps.

Thus, $\phi'(a) > 0$ for $a \in (\underline{a}, \tilde{a}]$ and $\phi''(a) < 0$ for all $a \in (\underline{a}, \tilde{a}]$. Therefore there is no $a^* \in (\underline{a}, \overline{a})$ for which $\phi'(a^*) = 0$ and $\phi''(a^*) \geq 0$. This in turn means that z^* consists of a single linear piece on $[\underline{a}, \overline{v}]$, which completes the proof.

A.5 Proof of Lemma 4

I first show that the optimal mechanism allocates both goods. I do so by finding the optimal mechanism allocating only one good and showing that it can be improved by also allocating some of the other one.

Consider a mechanism allocating only A. Any feasible mechanism must allocate it to at most $\mu_A \in (0,1)$ agents. A single-crossing argument then tells us there exists some $\underline{a} \in [0,1]$ such that:

$$y(a,b) \begin{cases} = A, & \text{if } a > \underline{a}, \\ = \varnothing, & \text{if } a < \underline{a}. \end{cases}$$

Let $\underline{a}^* \in (0,1)$ be the cutoff for which the supply constraint binds. By Myerson's Lemma we can then reduce the problem to choosing some $p_B \geq 0$, $\underline{a} \in [\underline{a}^*, 1]$ and an increasing $x_A : (\underline{a}, 1] \to [0,1]$. Total welfare then becomes:

$$W = \int_{\underline{a}}^{1} U_A(v)\, dG(a) \quad \text{where} \quad U_A(a) = \int_0^a x_A(v) da,$$

which increases pointwise in $x_A(a)$. Thus, the optimal single-good mechanism features $\underline{a} = \underline{a}^*$ and $x_A(a) = 1$ for all $a > \underline{a}^*$. Note that the mechanism can be implemented by offering only $x_A = 1$ at the price of $\underline{a}^*$.

Now, augment this mechanism by also offering good B with zero wait time at price $1 - \epsilon$, for $\epsilon > 0$. I first show the new mechanism is strictly better, and then that it is feasible for ϵ sufficiently small. Adding this option cannot reduce welfare, so we must just show that it strictly benefits a positive mass of types. But all types with $a < \underline{a}^*$ and $b > 1 - \epsilon$ would be getting nothing (and hence utility zero) under the old mechanism and get strictly positive utility now, so the new mechanism is indeed a strict improvement.

To see why the mechanism is feasible, note that good B will be taken only by agents with $b \geq 1 - \epsilon$, of whom there are $1 - H(1 - \epsilon)$. By the continuity of H, this mass approaches zero as $\epsilon \to 0$, and thus is below μ_B for ϵ small enough. Since only agents with $a \geq \underline{a}^*$ get the A-good, the supply constraints (S) hold.

Thus, the optimal mechanism allocates strictly positive amounts of both goods and, by Lemma 1, has a 'boundary structure'. I now show that the optimal mechanism allocates the whole supply of both goods. By Lemma 4, we can restrict attention to mechanisms with a linear boundary. Such mechanisms offer only two options: good A with discounting x_A at price p_A and good B with discounting x_B at price p_B. Now, suppose one of the supply constraints (S) is slack for such a mechanism; assume without loss this is the case for good A.

Since the mechanism allocates a strictly positive amount of good A but its supply constraint is slack, the price of good A has to be interior: $p_A \in (0,1)$. Now, consider an alternative mecha-

nism offering x_A, x_B for prices $p_A - \epsilon$ and p_B, respectively. This mechanism improves the utilities of all agents, and strictly so for the positive mass of agents who chose good A under the original mechanism. It therefore suffices to show the alternative mechanism is feasible for $\epsilon > 0$ sufficiently small. Note that the mass of agents who take A under the new mechanism is:

$$\int \mathbb{1}_{x_A a - (p_A - \epsilon) > \max[0,\, x_B b - p_B]} \, dF(a,b),$$

since the set of indifferent agents is zero-mass. Moreover:

$$\lim_{\epsilon \to 0} \int \mathbb{1}_{x_A a - (p_A - \epsilon) > \max[0,\, x_B b - p_B]} \, dF(a,b) = \int \mathbb{1}_{x_A a - p_A > \max[0,\, x_B b - p_B]} \, dF(a,b),$$

which is the mass of agents who got good A under the original mechanism. Since the supply constraint (S) for good A was slack, it remains slack for the alternative one when ϵ is sufficiently small. Similarly, reducing the price for good A can only relax the supply constraint for good B, and thus (S) is satisfied for ϵ small enough.

A.6 Proof of Lemma 5

Let $z_s : [\underline{a}_s, \underline{b}_s] \to [\overline{a}_s, \overline{b}_s]$ denote a linear boundary with slope s, that is:

$$z_s(a) = \underline{b}_s + s \cdot (a - \underline{a}_s) \quad \text{for } a \in [\underline{a}_s, \overline{a}_s],$$

where $\underline{a}_s$ and $\underline{b}_s$ are chosen so that the supply constraints (S) bind for both goods (Figure 7a). Since the type distribution is atomless, such a boundary z_s exists for any $s > 0$. Let us also define an *extended* s-sloped boundary $\tilde{z}_s$ as follows:

$$\tilde{z}_s(a) = \begin{cases} 0, & \text{if } a < \underline{a}_s, \\ z_s, & \text{if } a \in [\underline{a}_s, \overline{a}_s], \\ 1, & \text{if } a > \overline{a}_s. \end{cases}$$

That is, $\tilde{z}_s$ equals to z_s on the latter's domain, takes value zero below it and takes value 1 above it (Figure 7b). Note that for every $s, s' > 0$ such that $s > s'$, the boundaries g_s and $g_{s'}$ cross exactly once, with $g_{s'}$ crossing from above (Figure 8). This in turn implies that $\overline{a}_s$ is decreasing in s and $\overline{b}_s$ is increasing in s. Moreover, the lowest participating values of the two boundaries must also satisfy $\underline{a}_{s'} \leq \underline{a}_s$ and $\underline{b}_{s'} \geq \underline{b}_s$. Otherwise, both supply conditions in (S) could not hold with equality for both boundaries. Finally, since the distribution F is atomless, $\underline{b}_s, \underline{a}_s$ and $\overline{b}_s, \overline{a}_s$ must be differentiable functions of s.

Before proving that the optimal boundary has slope 1, I show a useful auxiliary fact:

Fact 5. *$Q(s)$, defined below, is strictly decreasing in s for $s \geq 1$:*

$$Q(s) := \int_0^1 G(a) H(\tilde{z}_s(a)) \, da.$$

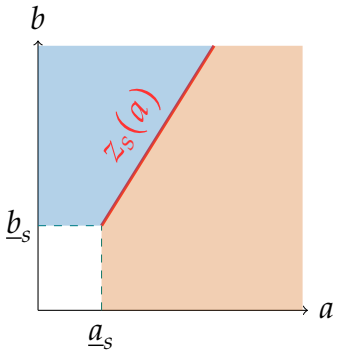

Figure 7a: s-sloped boundary z_s.

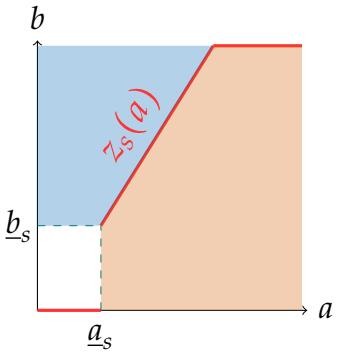

Figure 7b: s-sloped extended boundary $\tilde{z}_s$.

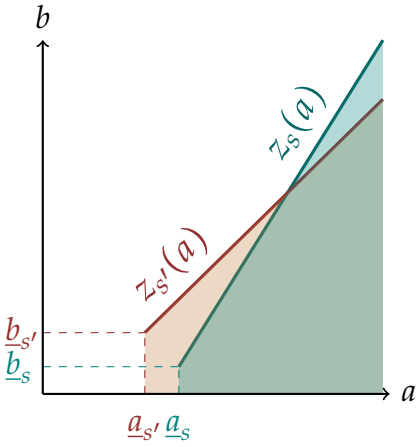

Figure 8: Let $s > s'$. Then the s'-sloped boundary $z_{s'}$ crosses the s-sloped boundary z_s once, and from above.

Proof. Fix some $s_1, s_2 \geq 1$ such that $s_1 > s_2$ and consider the difference:

$$\begin{aligned} Q(s_1) - Q(s_2) &= \int_0^1 G(a)H(\tilde{z}_{s_1}(a))\,da - \int_0^1 G(a)H(\tilde{z}_{s_2}(a))\,da \\ &= \int_0^1 \frac{G(a)}{g(a)}\Big[g(a)\left(H(\tilde{z}_{s_1}(a)) - H(\tilde{z}_{s_2}(a))\right)\Big]da. \end{aligned}$$

Since $s_1 > s_2$, $\tilde{z}_{s_2}$ crosses $\tilde{z}_{s_1}$ only once, and from above. Let a^* be their crossing point. We can then write the LHS as:

$$\int_0^{a^*} \frac{G(a)}{g(a)} \underbrace{g(a)\Big(H(\tilde{z}_{s_1}(a)) - H(\tilde{z}_{s_2}(a))\Big)}_{\leq 0}\,da + \int_{a^*}^1 \frac{G(a)}{g(a)} \underbrace{g(a)\Big(H(\tilde{z}_{s_1}(a)) - H(\tilde{z}_{s_2}(a))\Big)}_{\geq 0}\,da,$$

where the inequalities hold strictly in some neighborhood of a^*. Since $G(a)/g(a)$ is strictly increasing by Assumption 1, we obtain the following bound:

$$
\begin{aligned}
Q(s_1) - Q(s_2) \;>\; & \frac{G(a^*)}{g(a^*)} \int_0^{a^*} g(a)\Big(H(\tilde{z}_{s_1}(a)) - H(\tilde{z}_{s_2}(a))\Big)\, da \\
& + \frac{G(a^*)}{g(a^*)} \int_{a^*}^1 g(a)\Big(H(\tilde{z}_{s_1}(a)) - H(\tilde{z}_{s_2}(a))\Big)\, da \\
= \; & \frac{G(a^*)}{g(a^*)} \left(\int_0^1 g(a) H(\tilde{z}_{s_1}(a))\, da - \int_0^1 g(a) H(\tilde{z}_{s_2}(a))\, da \right) \;=\; 0.
\end{aligned}
$$

The difference is zero because (S) held with equality for both boundaries. □

I now show that all boundaries with slope $s > 1$ are dominated by some other boundary with a lower slope. Then, by symmetry, all boundaries with slope $s < 1$ will be dominated by some boundary with a higher slope. Consequently, $s = 1$ will be uniquely optimal. Denote by x_A^s, x_B^s the constant discounting levels which optimally implement z_s. Recall also that if $s > 1$, $x_A^s = 1$ and $x_B^s = 1/s$. Fix any boundary z_{s_1} with $s_1 > 1$ and recall that either $\bar{a}_{s_1} = 1$ or $\bar{b}_{s_1} = 1$. Consider two cases.

Case 1: $\bar{a}_{s_1} \neq 1$. We can use (W′) to write total welfare as follows:

$$
W[z_s] = \int_{\underline{a}_s}^1 \int_0^{z_s(\min[a,\bar{a}_s])} f(a,v)dv \cdot U_A(a)\; da \;+\; \int_{\underline{b}_s}^1 \int_0^{z_s^{-1}(\min[b,\bar{b}_s])} f(v,b)dv \cdot U_B(b)\; db. \quad \text{(W′)}
$$

Since $\bar{a}_s \neq 1$, we must have $\bar{b}_s = 1$ and so $W[z_s]$ becomes:

$$
\begin{aligned}
W[z_s] = & \int_{\underline{a}_s}^{\bar{a}_s} U_A(a)\, g(a) H\big(z_s(a)\big)\, da + \int_{z_s(\underline{a}_s)}^{z_s(\bar{a}_s)} U_B(b)\, G\big(z_s^{-1}(b)\big) h(b)\, db \\
& + \int_{\bar{a}_s}^1 U_A(a)\, g(a) H(1)\, da.
\end{aligned}
$$

A change of variables yields:

$$
W[z_s] = \underbrace{\int_{\underline{a}_s}^{\bar{a}_s} U_A(a) \cdot \big(g(a)H(z_s(a)) + z_s'(a) \cdot G(a) h(z_s(a))\big)\, da}_{:=K} + \underbrace{\int_{\bar{a}_s}^1 U_A(a) \cdot g(a) H(1) da}_{:=L}.
$$

Let us now integrate K and L by parts. This gives:

$$
K = U_A(\bar{a}_s) \cdot G(\bar{a}_s) H(z(\bar{a}_s)) - U_A(\underline{a}_s) \cdot G(\underline{a}_s) H(z_s(\underline{a}_s)) - \int_{\underline{a}_s}^{\bar{a}_s} U_A'(a) \cdot G(a) H(z_s(a))\, da.
$$

Recall that $U_A(\underline{a}_s) = 0$, $U'_A(a) = x^s_A = 1$, and $H(z_s(\overline{a}_s)) = H(\overline{b}_s) = H(1) = 1$. Thus, we get:

$$K = U_A(\overline{a}_s) \cdot G(\overline{a}_s) - \int_{\underline{a}_s}^{\overline{a}_s} G(a)H(z_s(a))\, da.$$

Integrating L by parts gives:

$$L = U_A(1) - U_A(\overline{a}_s)G(\overline{a}_s) - \int_{\overline{a}_s}^{1} G(a) \cdot H(1)\, da.$$

Summing K and L, we get:

$$\begin{aligned} W[z_s] &= U_A(1) - \int_{\underline{a}_s}^{\overline{a}_s} G(a)H(z_s(a))\, da - \int_{\overline{a}_s}^{1} G(a)H(1)\, da \\ &= 1 - \underline{a} - \int_{\underline{a}_s}^{\overline{a}_s} G(a)H(z_s(a))\, da - \int_{\overline{a}_s}^{1} G(a)H(1)\, da, \end{aligned}$$

where the equality follows since $U_A(1) = \int_{\underline{a}}^{1} x_A da = 1 - \underline{a}$ by the envelope theorem. We can now express $W[z_s]$ more concisely by writing it using the extended boundary $\tilde{z}_s$:

$$W[z_s] = 1 - \underline{a}_s - \int_0^1 G(a)H(\tilde{z}_s(a))\, da = 1 - \underline{a} - Q(s).$$

Now, since $\overline{a}_{s_1} \neq 1$ and $\overline{b}_s, \overline{a}_s$ change continuously in s, there exists $s_2 \in (s_1, 1)$ such that $\overline{a}_{s_2} \neq 1$. Then also $\overline{b}_{s_2} \neq 1$ and thus we can apply the formula derived above to both z_{s_1} and z_{s_2}:

$$W[z_{s_1}] = 1 - \underline{a}_{s_1} - Q(s_1), \quad W[z_{s_2}] = 1 - \underline{a}_{s_2} - Q(s_2).$$

It thus suffices to show that $W[z_{s_1}] < W[z_{s_2}]$. However, $s_1 > s_2$, so $\underline{a}_{s_1} > \underline{a}_{s_2}$. Moreover, Fact 5 tells us that $Q(s_1) > Q(s_2)$, which completes the proof.

Case 2: $a_{s_1} = 1$. Analogously to the previous case, we can write total welfare as:

$$W[z_s] = \underbrace{\int_{\underline{a}_s}^{1} U_A(a) \cdot \big(g(a)H(z_s(a)) + z'_s(a) \cdot G(a)h(z_s(a))\big)\, da}_{:=K} + \underbrace{\int_{\overline{b}_s}^{1} U_B(b) \cdot G(1)h(b) db}_{:=L}.$$

Let us now integrate K and L by parts. This gives:

$$K = U_A(1) \cdot G(1)H(z_s(1)) - U_A(\underline{a}_s) \cdot G(\underline{a}_s)H(z(\underline{a}_s)) - \int_{\underline{a}_s}^{1} U'_A(a) \cdot G(a)H(z_s(a))\, da.$$

Recall that $U_A(\underline{a}_s) = 0$, $U_A'(a) = x_A^s = 1$, $U_A(1) = U_B(\overline{b}_s)$ and $H(z(1)) = H(\overline{b}_s)$. Thus, we get:

$$K = U_B(\overline{b}_s) \cdot H(\overline{b}_s) - \int_{\underline{a}_s}^{1} G(a)H(z_s(a))\, da = U_B(\overline{b}_s) \cdot H(\overline{b}_s) - Q(s).$$

Now, integrating L by parts gives:

$$L = U_B(1) - U_B(\overline{b})H(\overline{b}) - \int_{\overline{b}}^{1} U_B'(b)H(b)\, db,$$

but since $U_B'(b) = x_B^s = 1/s$ and $U_B(1) = \frac{1}{s}(1 - \underline{b}_s)$, we have:

$$L = \frac{1}{s}(1 - \underline{b}_s) - U_B(\overline{b}_s)H(\overline{b}_s) - \frac{1}{s}\int_{\overline{b}_s}^{1} H(b)\, db.$$

Summing K and L, we get:

$$W[z_s] = \frac{1}{s}(1 - \underline{b}_s) - \frac{1}{s}\int_{\overline{b}_s}^{1} H(b)\, db - Q(s). \tag{26}$$

Now, take any $s_2 \in [1, s_1)$. Since is $\overline{a}_s$ decreasing in s and $\overline{a}_{s_1} = 1$, this means that $\overline{a}_{s_2} = 1$ as well. Therefore, (26) describes total welfare under both z_{s_1} and z_{s_2}.

We now show that $W[z_{s_2}] > W[z_{s_1}]$. Indeed, $Q(s_2) < Q(s_1)$ by Fact 5. It therefore suffices to show that $M(s)$, defined below, is decreasing in s:

$$M(s) := \frac{1}{s}(1 - \underline{b}_s) - \frac{1}{s}\int_{\overline{b}_s}^{1} H(b)\, db.$$

Recall that $\underline{b}_s, \overline{b}_s$ change differentiably in s, and so it suffices to show that $M'(s) < 0$. Note:

$$M'(s) = -\frac{1}{s^2}\left(1 - \underline{b}_s - \int_{\overline{b}_s}^{1} H(b)\, db\right) - \frac{1}{s}\frac{d}{ds}\left(\underline{b}_s + \int_{\overline{b}_s}^{1} H(b)\, db\right),$$

so it suffices to show that:

$$-\frac{1}{s}\left(1 - \underline{b}_s - \int_{\overline{b}_s}^{1} H(b_s)\, db\right) < \underline{b}_s' - \overline{b}_s'(s) \cdot H(\overline{b}_s).$$

Recall that $\overline{b}_s' \geq 0$, so we can strengthen the inequality to:

$$-\frac{1}{s}\left(1 - \underline{b}_s - \int_{\overline{b}_s}^{1} H(b)\, db\right) < \underline{b}_s' - \overline{b}_s'(s).$$

Furthermore, we have $\overline{b}_s = \underline{b}_s + s \cdot (1 - \underline{a}_s)$, so:

$$\begin{aligned}\overline{b}'_s &= 1 - \underline{a}_s + \underline{b}'_s - s \cdot \underline{a}'_s \\ &\leq 1 - \underline{a}_s + \underline{b}'_s,\end{aligned}$$

giving an even stronger inequality:

$$-\frac{1}{s}\left(1 - \underline{b}_s - \int_{\overline{b}_s}^{1} H(b)\, db\right) < \underline{b}'_s - (1 - \underline{a}_s + \underline{b}'_s) = -(1 - \underline{a}_s).$$

Now, since $s \cdot (1 - \underline{a}_s) = (\overline{b}_s - \underline{b}_s)$, the above inequality is equivalent to:

$$\begin{aligned}-\left(1 - \underline{b}_s - \int_{\overline{b}_s}^{1} H(b)\, db\right) &< -\left(\overline{b}_s - \underline{b}_s\right), \\ \underline{b}_s + \int_{\overline{b}_s}^{1} H(b)\, db + \overline{b}_s - \underline{b}_s &< 1, \\ \int_{\overline{b}_s}^{1} H(b)\, db + \overline{b}_s &< 1.\end{aligned}$$

which holds because H is a cdf.