Screening for Usage

Justin Hadad* and Kyle Woodward†

August 24, 2026

Abstract

In systems with significant wealth inequality, charging for a resource distorts efficient allocation across wealth types. We study a repeated allocation problem in which the principal wants to maximize resource usage, and agents vary in both their usage probability and their marginal value for money. In each period agents choose whether or not to make reservations, but may randomly cancel an allocated reservation. The principal may screen on usage probabilities using two practically-relevant instruments: monetary penalties and future eligibility restrictions (i.e., bans). We show that the usage-maximizing regulation has a simple structure: the principal never offers a refund for cancellation, and access restrictions are essentially deterministic. When available resource capacity is extremely small or extremely large, the optimal policy uses only prices, but at intermediate capacities the principal will implement eligibility restrictions. We show also that the benefit of bans relies on the existence of prices: a ban-only policy is never optimal.

1 Introduction

When designing systems to allocate scarce resources, policymakers often rely on monetary penalties to deter undesirable behavior. In settings where users with substantially different wealth levels compete, financial penalties can induce unintended distributional consequences: wealthy individuals may treat monetary costs as trivial, while the same penalties deter the poor. This inequality not only raises concerns with respect to fairness, but also undermines social surplus maximization.

We examine the challenge of regulation in a repeated market with inequality. To make our analysis concrete, we situate our approach in the domain of campsite reservations through Reserve California, the official campsite booking system of California State Parks. Reserve California manages roughly 9,000 campsites across just over 100 parks across California; each year, 1.25 million stay-nights yield nearly $100 million in revenue.[1] Despite its scale, Reserve California's campsite-booking system is widely criticized as broken: site reservations open six months in advance and

*University of Oxford; justin.hadad@balliol.ox.ac.uk

†Assign Group; kyle@assign.group

[1] All Reserve California figures are drawn from our analysis of Reserve California logs, obtained through FOIA. California State Parks's summary of park activity suggests 15,000 campsites across 280 parks, sheltering 6 million campers per year. We note that Reserve California does not manage reservations for all state-run campgrounds, including those which are first-come first-served.

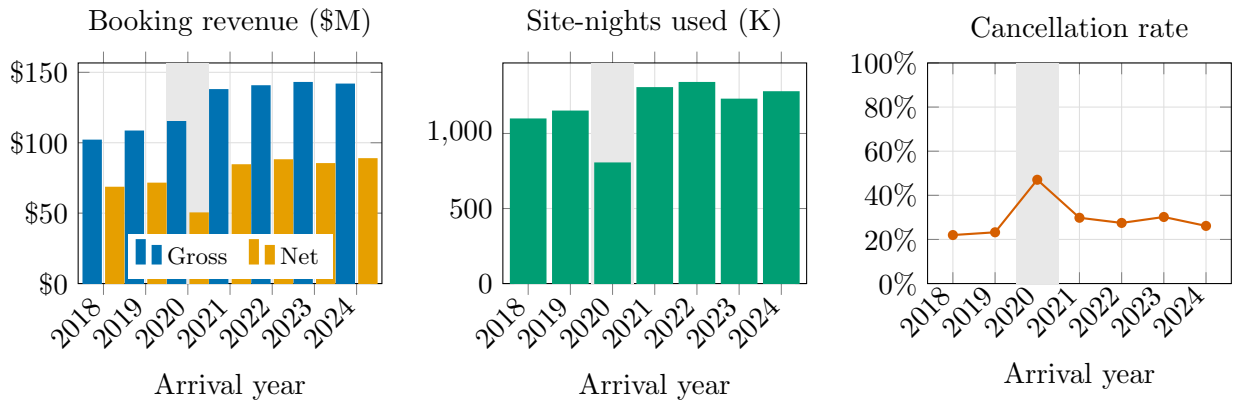


Figure 1: Time trends for Reserve California usage. (2020 highlighted due to COVID shock)

popular sites book in seconds. At the same time, many sites are left unused on any given night.[2] This apparent contradiction in campsite desirability was believed to be caused by the state's full-refund policy—all campsite cancellations were refundable except for an $8 change fee—and in 2023 the state passed AB 618 to address these issues. The bill took effect on July 1, 2026, and campsite cancellation is now penalized: cancellations within one week of arrival will forfeit the first night's fee; and cancellations within 24 hours of arrival will forfeit all fees paid.

At the same time, California is one of the most wealth-unequal states in the U.S.[3] Individuals with higher wealth generally have lower marginal utility for money, hence in the presence of extreme wealth inequality cancellation fees may hardly deter the rich, low-usage individuals while screening out poor individuals even when their usage probability is high. Thus while price screening might efficiently target high-usage prospective campers in an equal-wealth environment, it might screen adversely when facing wealthy and poor alike: wealthy agents book when they are unlikely to use the site, and poor agents do not book even when they are likely to use the site. A government which relies on traditional price-screening instruments to maximize the usage of its resource—in this case, campsites—might fail to bring about a socially optimal outcome.

AB 618's financial penalties were well-publicized (Thomas, 2026), but the bill also encoded a further enforcement mechanism. Individuals who renege on reservations three times in a one-year span may receive a one-year ban from Reserve California. In mechanism design contexts without wealth inequality, temporary bans cannot be justified above and beyond a standard price mechanism: restrictions on future payoffs can be mapped to a single threshold in value space, the same as a posted price.[4] In the presence of wealth inequality, however, it is ex ante unclear whether a

[2] "Unfortunately, our current outdated reservation system has led to a situation where many campsites are left empty," said California State Assemblymember Rebecca Bauer-Kahan (Office of Assemblymember Rebecca Bauer-Kahan, 2023).

[3] The U.S. Census Bureau's 2024 ACS found California to have the 4th highest level of wealth inequality of any U.S. state.

[4] At the core of this statement is the fact that any stationary distribution of campsite reservations in the presence of ban-based penalties can be weakly improved upon by a simple posted price. Since posted-price equilibria are time-invariant, it stands to reason that a regulator may prefer posted prices over bans in a dynamic environment.

designer interested in maximizing realized value would prefer to impose monetary or eligibility-based penalties. The potential value of equitable mechanisms based on future restrictions on consumption arises from the fact that, absent prices, agents have identical valuations for immediate consumption, and hence identical disutility from restrictions on future consumption. In the context of camping, a yearlong ban harms rich and poor alike. Shared value for camping implies that camping satisfies some properties of a numéraire good, unlike money which has wealth-dependent marginal value.

Ceteris paribus, screening on numéraire valuation is more socially efficient than screening on other valuations. However, screening using restrictions on future consumption comes at a cost: an agent who is removed from the market by a temporary ban cannot consume while under sanction, even if their consumption would improve resource utilization. For this reason it is not obvious ex ante whether relying on financial or restriction-based punishments yields higher realized surplus.

We evaluate these tradeoffs in a model in which infinitely-lived agents have fixed wealths but random per-period usage propensities. High-wealth agents have lower marginal utility for money, and low-wealth agents have higher marginal utility for money. All else equal, an agent with a high usage propensity will attempt to use the resource while an agent with a low usage propensity will not. Our principal has two design instruments: prices (and refunds), and temporary eligibility restrictions. As discussed above, prices distort the market away from efficient allocation as high-wealth agents book the resource even when their usage propensity is relatively low, but low-wealth agents book the resource only when the usage propensity is high. Eligibility restrictions distort away from efficiency by artificially shrinking the pool of consumers.

As a first step toward identifying whether price-based or ban-based mechanisms are preferred by a social planner interested in resource utilization, we begin our analysis by describing essential features of any optimal penalization scheme which uses these instruments. First, we show that in equilibrium the principal never offers refunds for cancellation (unlike California's recent policies). In our model full refunds turn payment into a usage-contingent payment, and the market price will either admit all high-wealth agents or all agents regardless of wealth. Since we show that, all else equal, usage increases as the spread between high- and low-wealth reservation thresholds decreases, refunds harm usage. Our analysis also handles the case of partial refunds. Second, in equilibrium, any eligibility restriction randomizes over at most two adjacent ban lengths. Fundamentally this is a consequence of Jensen's inequality, where restricting all cancelers' eligibility for an average amount of time keeps the eligible population larger than restricting a small fraction of cancelers' eligibilities for a long time.

We additionally devote attention to studying when a principal optimally rations; equivalently, when she overbooks her capacity and then randomizes allocation among agents who choose to book the resource. Rationing is a natural tool to overcome inequality: rather than target just high-wealth agents, the principal can elicit overbooking from some low-wealth agents and (more) high-wealth agents. Following a version of Simpson's paradox, when there are sufficiently many low-wealth agents in the system, rationing the combined mass of high-wealth agents and high-propensity, low-wealth agents increases average usage versus only targeting high-wealth agents. Underbooking,

meanwhile is never optimal when only prices are used, but may be optimal when bans are used. Note the threat of bans have two competing effects: while they increase usage conditional on booking, they also may decrease the steady-state eligible mass enough that expected usage decreases. Our analysis addresses these possibilities.

The core of our analysis identifies the nature of the usage-maximizing policy. Under the (optimal) design that includes no refunds and a nearly-deterministic ban length, the principal can choose some combination of a price and ban length. Our first main result shows that the effectiveness of a ban of any length goes to zero as capacity approaches extreme values. When C is small, the probability of being allocated in any given period is small enough that agents are not deterred by the threat of being made ineligible. Meanwhile, under large C, bans effectively deter agents, but are too costly for the principal: she prefers to keep the entire mass eligible and instead use prices to screen.

But bans can still be useful at intermediate capacities. We identify a simple sufficient condition for which incorporating a ban improves usage: If a policy which only uses prices optimally rations its capacity, then the principal can introduce a (small) ban and slightly decrease the price; this holds the booking mass constant but increases usage conditional on booking. However, the optimal deployment of bans is in conjunction with prices. We find that a ban-only policy is never optimal: introducing a (small) price always improves the principal's objective.

Conversely, when agents are patient, and inequality is not too small, policies that just use bans outperform policies that just use prices. This is because when the resource is scarce, financial penalties admit only high-wealth agents. Because bans screen independent of wealth, the effect of targeting the full population dominates the loss from holding potential consumers out of the market. We show in an example that "too small" and "sufficiently scarce" are not just limiting cases.

Our results offer prescriptive suggestions for when principals interested in maximizing resource usage in the presence of wealth inequality should impose monetary penalties versus eligibility-based penalties, and vice versa.

1.1 Related Literature

A series of papers study regulation in markets where value is only realized after transaction. The most closely related, Ma et al. (2018), analyzes a two-period resource allocation problem where a principal wants to maximize usage, and where agents realize their value for usage only after assignment. We also study contingent payment mechanisms, but we expand the scope in three ways: our repeated allocation environment lets use affect future access; we allow both monetary and nonmonetary instruments; and we focus on the differential impact of regulation under inequality. Other work investigates the question of revenue maximization in reservation systems. Non-refundable deposits are useful screening devices and overbooking can be useful under uncertainty in demand (Georgiadis and Tang, 2014; Lyu, 2024).

Beyond money, nonmarket mechanisms can maximize surplus when the objective differs from

revenue maximization. Whether monetary schemes dominate nonmarket mechanisms is subject to the distribution of the willingness to pay (Condorelli, 2013). Relatedly, money-burning and delays can make mechanisms incentive-compatible and increase surplus (Jackson and Sonnenschein, 2007; Burkett and Woodward, 2024; Patel and Urgun, 2024). Verdier and Reeling (2022) studies efficient dynamic allocation in an empirical model of rationed bear hunting permits in Michigan.

Our deliberate attention on policies that enforce eligibility restrictions is novel in this setting. Lundy et al. (2019) analyzes a setting of resource allocation via food banks where a designer can make agents pay debt in the form of future allocations, equivalent to a ban.

Our paper belongs to the *inequality-aware market-design* agenda which takes ethical institutional design seriously: "Policymakers, though, often attempt to maximize an objective function that evidently differs from either revenue or allocative efficiency" (Dworczak, 2024). A body of literature has emerged in this spirit and is instructive for the design of markets which exhibit wealth inequality (Akbarpour et al., 2024; Mäkimattila, 2025). Pathak et al. (2021) applies these principles to the equitable distribution of COVID-19 vaccines.

2 Model

Time is discrete and indexed by $t \in \{0, 1, 2, \ldots\}$. There is a unit mass of agents who are infinitely-lived and discount future payoffs by $\delta \in [0, 1)$. In each period, agents may book at most one reservation, and at most a mass $C \in (0, 1)$ of reservations can be allocated in any period. Agents share a common value $v > 0$ from *using* an allocated reservation; without loss of generality we fix $v = 1$ throughout. If an allocated reservation is not used, the agent obtains no consumption value from that allocation.

Agents differ in their marginal value for money. There are two wealth types $w \in \{w_L, w_H\}$ with $w_H \geq w_L > 0$; fraction $\mu_H \in [0, 1]$ of agents has wealth type w_H, and the remaining fraction $\mu_L := 1 - \mu_H$ has wealth type w_L. Each wealth level enters as the inverse marginal value for money; throughout, a payment of size x yields utility cost x/w.

To simplify expressions relating to wealth inequality, normalize the difference in the mass of wealth types and wealth levels as follows:

$$\mu_L = \frac{1}{1+\gamma}, \qquad \mu_H = \frac{\gamma}{1+\gamma}, \qquad \gamma \geq 0,$$

$$w_L = \frac{1}{1+\xi}, \qquad w_H = \frac{\xi}{1+\xi}, \qquad \xi \geq 1.$$

Under these normalizations γ captures changes in composition, while ξ indexes wealth dispersion; jointly, the pair (γ, ξ) characterize wealth inequality in the economy. As γ increases, the share of wealthy individuals increases, and as ξ increases wealth inequality grows. Throughout our analysis, we will make use of both the normalizations as well as the individual masses and wealth levels.

In each period, an agent has an *eligibility* as well as a private usage probability π_{it}. An agent is eligible in period t if she is allowed to book the resource in period t. Booking is a prerequisite for

allocation. Agent i's private usage probability $\pi_{it} \in [0,1]$ is drawn in each period from a cumulative distribution function F that is continuously differentiable on $(0,1)$ with density $f(\pi) > 0$. We assume f satisfies $0 < f(1) := \lim_{\pi \uparrow 1} f(\pi) < \infty$ and that f' is bounded on $(0,1)$. Across agents and across time, the draws $\{\pi_{it}\}$ are independent. If agent i is allocated a reservation in period t, then she uses it with probability π_{it} and cancels (equivalently, does not use) with complementary probability $1 - \pi_{it}$.

In each period, each eligible agent i makes a *booking decision* $b_{it} \in \{0,1\}$ with $b_{it} = 1$ corresponding to making a booking, and $b_{it} = 0$ corresponding to not. Let

$$A_t := \int b_{it} di$$

denote the mass of agents who book in period t.[5] Recall that an agent who is ineligible necessarily has $b_{it} = 0$. If $A_t \leq C$, then every booking is allocated a reservation. If $A_t > C$, then exactly mass C is allocated by symmetric random rationing among those who book. Let $y_{it} \in \{0,1\}$ indicate whether agent i is allocated a reservation in period t. Conditional on booking, the allocation probability is defined as

$$\beta_t := \Pr(y_{it} = 1 \mid b_{it} = 1) = \min\left\{1, \frac{C}{A_t}\right\}, \qquad \Pr(y_{it} = 1 \mid b_{it} = 0) = 0.$$

Finally, let $c_{it} \in \{0,1\}$ indicate cancellation,

$$\Pr(c_{it} = 1 \mid y_{it} = 1, \pi_{it}) = 1 - \pi_{it}, \qquad \Pr(c_{it} = 1 \mid y_{it} = 0, \pi_{it}) = 0.$$

2.1 Steady State and the Principal's Objective

The mechanisms we study induce stationary equilibria in which, for each wealth type $w \in \{w_L, w_H\}$, the booking decision is monotone in the privately observed usage probability π. Let z_{it} be a variable such that $z_{it} = 0$ when agent i is eligible to book in period t.[6] In any stationary equilibrium in our class, there exist type-specific cutoffs

$$\underline{\pi}_w \in [0,1] \quad \text{such that} \quad b_{it} = 1 \iff z_{it} = 0 \text{ and } \pi_{it} \geq \underline{\pi}_w \quad \text{for wealth type } w.$$

[5] Because draws of usage probabilities are independent across agents and across time, this definition is without loss.

[6] We define z_{it} rigorously in Appendix A.

For any cutoff $\underline{\pi} \in [0,1]$, the following are our primary sufficient statistics:

$$\begin{aligned}
\alpha(\underline{\pi}) &:= \Pr(\pi \geq \underline{\pi}) = 1 - F(\underline{\pi}) && \text{(booking mass)} \\
M(\underline{\pi}) &:= \mathbb{E}[\pi \mathbf{1}\{\pi \geq \underline{\pi}\}] = \int_{\underline{\pi}}^{1} x \, dF(x) && \text{(expected used mass by booking agents)} \\
m(\underline{\pi}) &:= \mathbb{E}[\pi \mid \pi \geq \underline{\pi}] = \frac{M(\underline{\pi})}{\alpha(\underline{\pi})} && \text{(average usage rate among booking agents).}
\end{aligned}$$

The function $m(\cdot)$ is the average usage probability among those who book given a specified booking threshold, and we assume that $m(1) = 1$. Observe that m is strictly increasing on $[0,1]$ and satisfies $m(0) = \mathbb{E}[\pi]$.

A policy may include access restrictions following cancellations, which can make some agents temporarily ineligible to book. In steady state, the mass of wealth-w agents who are eligible at the start of a period is constant over time. Denote this mass by $\mu_w \zeta_w$ where $\zeta_w \in (0,1]$ is the steady-state fraction of wealth-w agents who are eligible.

Fix a stationary equilibrium with type-specific booking cutoffs $(\underline{\pi}_H, \underline{\pi}_L)$ and eligibility masses (ζ_H, ζ_L). The steady-state mass of agents who book is then

$$A := \mu_H \zeta_H \alpha(\underline{\pi}_H) + \mu_L \zeta_L \alpha(\underline{\pi}_L). \tag{1}$$

Given A, the steady-state allocation probability conditional on booking is

$$\beta := \min\left\{1, \frac{C}{A}\right\}.$$

Thus $\beta = 1$ when $A \leq C$ and $\beta = C/A$ when $A > C$. When $A < C$, we say the policy *underbooks*; when $A > C$, we say the policy *overbooks*.

The principal's objective is to maximize *usage*: the expected mass of allocated reservations that are used in a period. In a steady state, expected usage per period is

$$U(\underline{\pi}_H, \underline{\pi}_L; \zeta_H, \zeta_L) := \left(\mu_H \zeta_H M(\underline{\pi}_H) + \mu_L \zeta_L M(\underline{\pi}_L)\right) \beta. \tag{2}$$

That is, $\mu_w \zeta_w M(\underline{\pi}_w)$ is the mass of *used* reservations that would arise from wealth type w if every booking were allocated, and β scales this by the probability that a booking is allocated under capacity rationing. We assume transfers do not enter the principal's objective.[7] Prices and refunds matter only through the incentives they induce, and access restrictions matter only through their effect on steady-state eligibility and booking behavior. The principal therefore chooses a feasible policy to maximize steady-state usage (2).

[7] In this analysis, we devote attention to mechanisms which exclusively affect usage; otherwise, the principal has bias toward price instruments.

2.2 Class of Mechanisms

At time $t = 0$, the principal commits to a policy from the class of mechanisms consisting of financial and eligibility-based penalties (defined below) and applies it uniformly over time. Fix a finite maximum ban length $T < \infty$. A policy is a triple

$$(p, r; \rho) \in \mathbb{R}_+ \times [0, p] \times \Delta(\{0, 1, \ldots, T\}),$$

where $p \geq 0$ is the posted price paid whenever a booking is allocated; $r \in [0, p]$ is the refund paid if an allocated booking is canceled; and ρ is a distribution over ban lengths $\tau \in \{0, 1, 2, \ldots T\}$, where $\tau = 0$ corresponds to no access restriction. See Appendix A for a formal specification of ban mechanics.

All together, a policy implements a steady-state outcome summarized by $(\underline{\pi}_H, \underline{\pi}_L; \zeta_H, \zeta_L)$.

2.3 Agent Behavior

We consider the agent's decision to book as a function of her flow utility $\hat{V}$. If allocated, the agent pays price p; if the agent cancels, she gets a refund r and faces a ban captured by an effective discount factor $\delta_\rho = \mathbb{E}_\rho[\delta^{\tau+1}] \in [0, \delta]$.[8] Given (π, w), her utility from booking (b) and from not booking (n) in a given period can be written

$$\hat{V}_w^b(\pi) = \beta\Big[\pi\Big((1-\delta)\Big(1 - \frac{p}{w}\Big) + \delta\hat{V}_w\Big) + (1-\pi)\Big(-(1-\delta)\frac{p-r}{w} + \delta_\rho\hat{V}_w\Big)\Big] + (1-\beta)\delta\hat{V}_w, \tag{3}$$

$$\hat{V}_w^n = \delta\hat{V}_w. \tag{4}$$

In (3), the agent is allocated the booking with probability β. If she uses the booking (which occurs with probability π) then she receives the flow payoff, which takes into account the price of booking and her continuation from not being banned; if she does not use the booking, then she is possibly granted a refund but suffers the continuation loss according to δ_ρ. In (4), the agent only incurs the continuation utility.

The booking cutoff $\underline{\pi}_w$ is defined by indifference between booking and not booking, thus is the value of π at which (3) is equal to (4) for a given wealth type. The booking cutoff for wealth type w is given by

$$\underline{\pi}_w = \frac{(\delta - \delta_\rho)\hat{V}_w + (1-\delta)\frac{p-r}{w}}{(\delta - \delta_\rho)\hat{V}_w + (1-\delta)\left(1 - \frac{r}{w}\right)}. \tag{5}$$

Note that if $p - r$ is held constant, then $\underline{\pi}_w$ is increasing in p: the less surplus is retained by the agent, the less likely they are to book. Relatedly, the lower the negative consequences of a temporary ban, the more willing the agent is to run the risk of punishment.

[8] Under our equilibrium restriction to steady state, the utility loss from being banned is constant.

3 Features of Optimal Policies

We begin our analysis by studying how the relevant instruments — prices (and refunds) and bans — are optimally deployed. Note that wealth inequality induces a tradeoff: while monetary punishments induce different marginal usage probabilities across wealth types, eligibility-based punishments remove agents from the market whose consumption would nevertheless improve social value.

We first show that refunds are the wrong way to encourage participation when facing wealth inequality; refunds neither ameliorate the distortion across wealth types nor screen on usage. Hence a principal can optimally set refunds to zero.[9]

Proposition 1. *The optimal policy does not use refunds.*

The presence of refunds turns prices into a use-contingent instrument, and therefore reduces the ability of prices to screen on usage probabilities. This is most clear at the extreme of full refunds, $r = p$. In this case the expected flow utility of an allocated booking is $\left(1 - \frac{p}{w}\right)\pi$, and agents will book as long as $\frac{p}{w} < 1$, independent of usage probability. Reducing refunds improves the screening ability of prices and therefore improves usage. Ironically, full refunds are perfectly fair; however, unlike eligibility-based penalties which retain an ability to screen on usage, issuing full refunds in this model is equivalent to uniformly rationing allocation across all agents — which, in general, is not usage-maximizing.

We now remark on how the principal sets the price p in light of there being no refunds. For a given level of inequality, it may be optimal for the principal to overbook her capacity — but in a price-only regime, it is never optimal to underbook.

Proposition 2. *Any optimal policy which only uses prices will never underbook. But an optimal policy which uses prices may overbook.*

By lowering the price in a regime without bans, the principal converts her unused capacity while affecting neither the usage probability of the units she would have allocated anyway, nor the probability that a booking is allocated. Thus she can locally increase bookings and raise expected usage. Hence underbooking is dominated.

Meanwhile, if bookings are dominated by the high-wealth type, then the principal may prefer to set a policy which incentivizes the marginal low-wealth type to book, even when the policy is such that the number of bookings already equals her capacity. In that case, rationing among all booking agents can increase principal surplus, and she overbooks. Indeed, ignore the possibility of eligibility restrictions, and consider the principal's objective of expected usage:

$$\beta\left[\mu_H M\left(\underline{\pi}_H\right) + \mu_L M\left(\underline{\pi}_L\right)\right].$$

[9] There may be an equivalent policy to the optimal one which uses refunds; this policy emerges either when only one wealth-type is served or when there is a full refund. In Appendix C.1, we show that in those cases, a smaller price paired with no refund emulates both cases.

When the mass of bookings is below C, expected usage is strictly decreasing in $\underline{\pi}_w$ for both wealth types. Once the mass of bookings exceeds C the mechanism employs rationing and the change of usage with respect to $\underline{\pi}_w$ for each w depends on the comparative relationship between the numerator and denominator (within β), both of which decrease as thresholds increases. Thus there may be an interior, overbooked optimal mechanism. Whether exact booking or overbooking is optimal depends on discrete comparison of local optima.

In the presence of eligibility-based punishment, reducing prices and thus decreasing thresholds for each wealth type has two simultaneous effects: more agents book, and more agents cancel. It is possible that lowering the price can decrease the steady-state mass of eligible agents $\zeta_w(0)$, thus underbooking, which increases the steady-state mass of agents who are eligible, can be preferred to eliciting at least C bookings. The effect of decreased steady-state mass is of second-order importance when capacity C is sufficiently small.[10]

We now focus on ban instruments. Consider that $\rho \in \Delta(\{0, ..., T\})$ defined in Section 2.2 gives the designer flexibility in how she distributes bans across periods. We find that the optimal ban length when used in any policy is structurally simple: if an agent cancels an allocated booking, they are punished with a ban randomized between $\tau^\star$ and $\tau^\star + 1$ periods, for some $\tau^\star$.

Proposition 3. *An optimal policy that uses a ban has ban length supported on at most two durations. If it is supported by two durations, these durations are adjacent.*

Observe that we can always hold the effective punishment constant while moving punishment forward, which does not affect the decision to book. But moving punishment forward will increase the steady-state mass of potential applicants, which is optimal for the principal. Since $\delta^{\tau+1}$ is strictly decreasing in τ, those two nonzero entries must lie on adjacent τ (or one if $\bar{\delta}/\delta = \delta^{\tau+1}$ exactly).

While we showed in Proposition 2 that underbooking is suboptimal for a principal who only uses prices to screen, the same is not true in general for a policy which uses bans. As we discussed previously, it may be the case that increasing thresholds, while increasing usage conditional on booking, may decrease the steady-state eligible mass enough that expected usage decreases.

Remark 1. *An optimal policy which uses bans may underbook or overbook.*

Propositions 1–3 and Remark 1 provide structure on how any optimal policy should be set: the policy consists exclusively of a price and/or ban length, and whether she underbooks, exactly books, or overbooks is determined by the instruments used in that policy and the market primitives. With only prices, weak overbooking is optimal, and the principal should ration only if inequality induces positive benefit from recruiting the marginal low-wealth type. When bans are included in the policy, whether underbooking is useful is determined by the steady-state eligible mass: lax punishments can induce more bookings which, when the thresholds are sufficiently low, can lead to more cancellations, which is suboptimal for the designer.

[10] In Lemma 2, we show that any optimal policy sends $\underline{\pi}_w \nearrow 1$ for each wealth type w, in which case the effect of increasing usage dominates.

4 Comparing Mechanisms

The preceding propositions reduce the policy choice to a price p and a relative ban impact, which we will write as

$$(\delta - \mathbb{E}_\rho[\delta^{\tau+1}]) =: \kappa.$$

We now analyze these two instruments, and consider under what conditions of the market the optimal policy can be simply described via just a price or a ban. We also will discuss how instruments we analyzed in Section 3 might be optimally deployed together. Let $(p; \kappa)$ denote the policy which the principal implements. The principal's problem is

$$\max_{p,\kappa} \left[\beta \left(\mu_H \zeta_H M(\underline{\pi}_H) + \mu_L \zeta_L M(\underline{\pi}_L)\right)\right]. \tag{6}$$

First, in light of the features described in Section 3, we observe that there is no such optimal policy where the threshold usage rate is made higher for the high-wealth type than for the low-wealth type.

Lemma 1. *For all policies $(p; \kappa)$, equilibrium usage thresholds satisfy $\underline{\pi}_H^\star \leq \underline{\pi}_L^\star$.*

Bans treat each wealth type symmetrically. So Lemma 1 essentially notes that there is no such price whose marginal impact on the high-wealth type is higher than for the low-wealth type. We thus preclude the possibility that, because the high-wealth type faces less effect of monetary punishments, bans have a larger differential effect on them. We show instead that prices have a consistent effect of screening more harshly on the low-wealth type.

We remark on how inequality impacts the effect of policies which use a single instrument. When the market has no inequality, bans have no utility for the principal; similarly, when the market can only address one of the two wealth types (namely the high-wealth type), the optimal policy will not deploy bans.

Corollary 1. *When $\xi = 1$ or $\gamma = 0$, every optimal policy uses only prices. Moreover, whenever optimal usage is achieved by serving only one wealth type, that outcome is implemented without bans.*

When there is perfect equality, price-only policies perfectly screen on usage rate: agents make a booking only when their expected usage compensates the monetary loss. In that market, the thresholds are the same across wealth types. And when only a single wealth type is served, the intuition which motivates the use of prices is the same: should a policy in this case include a ban, then the principal can implement a price which strictly increases the mass of agents who are eligible, and thus increase the threshold.

Away from these extreme cases, ban instruments can be useful in light of inequality. It follows from the logic of Corollary 1 that a price-only instrument becomes less effective as $\xi \nearrow \infty$, in which case a ban-only instrument increases in relative use for the principal.

Corollary 2. *Ban-only policies become weakly more competitive relative to price-only policies as wealth dispersion increases.*

Increasing wealth dispersion incentivizes the principal to use the screening instrument which has constant effect in wealth. Note Corollary 2 concerns dispersion only, and not composition γ. As $\gamma \nearrow \infty$, the policy is more likely to serve just the high-wealth type, in which case prices can dominate, as we note in Corollary 1. But as $\gamma \searrow 0$, the policy again serves just one type, the low-wealth type, and price-only policies dominate again.

We now present and discuss our first main result, which concerns the optimal policy under extreme capacities. There, policies which use only prices are optimal.

Theorem 1. *Fix $\delta < 1$, $\gamma > 0$, $\xi > 1$. Then there exists $0 \leq \underline{C}(\gamma, \xi) \leq \bar{C}(\gamma, \xi) < 1$ such that*

(i) if $C < \underline{C}(\gamma, \xi)$, every optimal mechanism uses no bans and serves only the high-wealth type;

(ii) if $C > \bar{C}(\gamma, \xi)$, every optimal outcome uses no bans, serves both wealth types, and exactly fills capacity.

First, consider the case of small C. Observe in general that the incentive generated by a ban is the value of the future booking opportunities which it deducts from an agent's continuation utility. When C is small, there are two possibilities: $\beta = C/A$ can be made very small, thus the threat of losing eligibility grows irrelevant as $C \searrow 0$, where an eligible agent is unlikely to receive an allocation anyway. Otherwise, the principal tries to fill capacity exactly and induce a (very) high cutoff. But a finite ban cannot sustain a cutoff arbitrarily close to 1 due to removing agents from the market.

A series of intermediate results under small C illuminate the first part of Theorem 1, and indeed highlight the tradeoffs in our analysis. When C is small, the optimal policy must induce booking thresholds which can be made arbitrarily close to 1; this is achieved by prices.

Lemma 2. *Let C be small. For any $\pi < 1$ the optimal policy induces $\underline{\pi}^\star_H, \underline{\pi}^\star_L > \pi$.*

Consider what happens if the optimal policy induces a booking threshold $\underline{\pi}^\star_w$ which does not go to 1 as C becomes small. Then expected usage relative to capacity remains strictly under 100%. But then consider an alternative policy which uses only prices; the principal can then expected usage to 100%, thus low booking thresholds are inconsistent with usage maximization under low available capacity.

In the regime of small C, weak overbooking is optimal for any policy. In this case, with less than C agents who book, the regulator can locally increase incentives and increase bookings without tightening the allocation constraint, and thereby raise expected usage.

Lemma 3. *Let C be small. Provided the equilibrium steady-state population mass is at least C, it is never optimal to elicit less than mass C of bookings.*

Lemma 3 enhances Remark 1 to a positive result. There is a sufficiently small capacity where rationing is dominated by inducing exactly C bookings; underbooking is suboptimal in the same way we noted in Proposition 2.

Lastly, we note that in environments with small C, there is a natural relationship between allocational inequality, defined as the difference between $\underline{\pi}_H$ and $\underline{\pi}_L$, and the principal's objective of maximum usage.

Lemma 4. *Let C be small. Holding the booking mass constant, expected usage increases as allocational inequality $|\underline{\pi}_H - \underline{\pi}_L|$ decreases.*

Lemma 3 and Lemma 4 illustrate the key tensions in our analysis. When there is wealth inequality, prices exclude too many low-wealth applicants while permitting too many high-wealth applicants. This cannot be optimal for the regulator, ceteris paribus, because Lemma 4 shows that the regulator in general prefers $\underline{\pi}_L = \underline{\pi}_H$. Second, although bans treat applicants equally across wealth types, bans exclude too many potential applicants (who are in ineligibility states). Because Lemma 3 says that the regulator generally desires at least C applications, removing applicant mass is equivalent to reducing the type threshold for application.

Lemma 4 normatively justifies the attention we pay to mechanisms which use only monetary punishment or only eligibility-based punishment. When the mechanism uses only monetary penalties, the expected duration of ineligibility upon cancellation is $\sum_{t=1}^{T} t\rho(t) = 0$, and the equilibrium mass of eligible agents $\zeta_w^\star(0)$ is maximized. On the other hand, when the mechanism uses only eligibility-based penalties, equations (3) and (5) imply that booking thresholds are equal across wealth types. That is, there is no allocational inequality. It follows that these mechanisms represent the key tradeoffs faced by the principal; we focus on single-instrument mechanisms in Theorems 2 and 3.

How the tradeoff regarding the distortions renders at small C is fundamentally different from at large C. At sufficiently small C, future eligibility has vanishing value and finite bans cannot induce high cutoffs; in other words, bans do not have much bite. At large C, removing agents from the market has too much bite for the *principal*. She opts instead for the inefficiency induced by unequal screening in lieu of removing any agents at all. The intuition is most clear when we take $C \nearrow 1$. There, every additional booking with $\pi > 0$ adds usage; the principal optimally serves all agents and the distortion between wealth types vanishes. Under a ban policy, the principal still faces no distortion across wealth types but now removes agents whose surplus increases her objective.

Theorem 1 identifies the optimal policy under extreme capacities; these policies consist exclusively of prices. So the market distortion induced by any length of ban serves the principal poorly when the threat of punishment is either ineffective (small C) or too consequential (large C). But there remains the possibility for mixed policies to be optimal (and even for ban-only policies to be optimal) for intermediate capacities. We address these possibilities now.

Our next main result identifies the region in which a price-only policy could be improved by incorporating a ban. We find a simple sufficient condition: Whenever the optimal price-only policy

overbooks, and the principal rations, the principal can improve her objective by decreasing the price slightly and using a ban.

Theorem 2. *If the best price-only policy overbooks, then bans are useful.*

First, observe that an optimal price-only policy would not overbook if it served only one wealth type; in that case, the principal could raise the price until the mass of bookings exactly equaled capacity. This would eliminate the booking agents with the lowest usage probabilities and strictly increase expected usage conditional on allocation. So when a policy overbooks, both wealth types have a threshold below 1. In this case, any optimality of overbooking is a symptom of the distortion created by monetary screening. By Lemma 1, high-wealth agents book at lower usage probabilities than low-wealth agents; given she overbooks, the principal could in theory raise the price until demand equaled capacity, but doing so raises the low-wealth cutoff more quickly. Rationing thus compensates for the unequal screening generated by prices.

We show that reducing prices and incorporating a ban improves the composition of the booking pool. The lower price disproportionately decreases the low-wealth threshold, while the small ban imposes a punishment which holds the total booking mass fixed. The resulting cutoff gap is therefore smaller than under the price-only policy.

While Theorem 2 defends the use of bans in conjunction with prices, we show now that any ban-only policy can be improved upon by some other policy which incorporates a positive price.

Theorem 3. *A ban-only policy is never optimal.*

For an agent, the effect of a ban is determined by its discounted duration $\kappa/(1-\delta)$; for the principal, its effect on the stationary eligible population is determined by the mean duration of the ban, which we describe in Proposition 3. Note that the incentive effect of a ban is weak relative to its cost:

$$\frac{\kappa}{1-\delta} < \sum_{t=1}^{T} t\rho(t).$$

So agents discount the future periods in which they may be banned, but each of those periods (deterministically) removes the agent from the principal's eligible population.

Under a pure-ban policy, each wealth type shares the same threshold, as well as the same eligible fraction. Introducing a positive price (and possibly adapting the ban length) perturbs these thresholds and eligibility fractions. This is to the benefit of the principal: it raises both booking cutoffs, but raises the low-wealth threshold more strongly; higher cutoffs eliminate bookings with the lowest usage probabilities, and also reduce cancellations, so fewer agents trigger bans and a larger fraction of the population remains eligible. Intuitively, because the principal values those future periods more than the discounted agent does, replacing a small amount of ban-based screening with contemporaneous price screening increases usage.[11]

[11] Recall that there is no optimal prescription for how a ban-only policy elicits bookings; it may underbook or overbook. We handle each case separately. In the underbooking regime, a sufficiently small price leaves $\beta = 1$, so introducing a price directly increases steady-state usage. At exact booking or overbooking, changing the price can also change total booking mass and hence the amount of rationing.

So while bans screen in a way that is neutral to wealth, their imposed punishment excludes agents for longer than agents value that punishment. A small price partially replaces this costly form of screening with a contemporary incentive, and the effect of the wealth distortion from the price is locally dominated by the gain from reducing reliance on discounted punishment. Indeed, this helps explain why an optimal policy may combine prices and bans.

Still, we do not preclude the possibility that a ban-only policy dominates a price-only policy. Indeed, we conclude with a positive result about ban-only policies: there is an intermediate region above small C identified in Theorem 1 where, given inequality is not too extreme and $\delta \nearrow 1$, the optimal ban-only policy generates higher usage than the optimal price-only policy.

Proposition 4. *Let $\mu_H < 1/2$. When C is small and δ is large, optimal eligibility-based punishment generates higher usage than optimal price-based mechanisms.*

For the preceding result, we assume that the maximum ban length T is sufficiently large relative to the fixed capacity C. As agents grow more patient, eligibility-based punishments become increasingly punitive in order to dissuade inefficient booking. In the patient limit agents remain ineligible with positive probability, and the eligible mass is lower under eligibility-based punishment than under monetary punishment. However, when C is small monetary punishments can only target high-wealth individuals. The tradeoff for the principal is whether to target lower-usage high-wealth individuals, or an endogenously smaller mass of all agents. As low-wealth agents become less competitive, eligibility-based punishments perform better because the unequal screening power of prices has an increasingly consequential impact on usage.

In the stationary equilibrium which occurs in the patient limit, the optimal ban-only policy makes half the population ineligible at all times. At $\mu_H < 1/2$, the fraction which the price-only policy addresses is smaller than the mass addressed by the ban-only policy, and so a ban-only policy dominates.

5 A Uniform Economy

We now present an example to highlight how the instruments of prices and bans compare under inequality, in the "patient limit" $\delta \to 1$.[12] For each fixed C, we let $\delta \nearrow 1$ and then allow the maximum ban length T to be sufficiently large to implement the limiting ban policy.[13] Let the distribution of usage probabilities be uniform: in each period, each eligible agent draws a private usage probability $\pi \sim \text{Unif}[0,1]$. In the uniform distribution, our sufficient statistics from Subsection 2.1 are analytically tractable:

$$\alpha(\underline{\pi}) := \Pr(\pi \geq \underline{\pi}) = 1 - \underline{\pi}, \qquad M(\underline{\pi}) := \mathbb{E}[\pi \, \mathbf{1}\{\pi \geq \underline{\pi}\}] = \int_{\underline{\pi}}^{1} x \, dx = \frac{1 - \underline{\pi}^2}{2},$$

[12] Steady-state equilibrium under monetary penalties is independent of patience. For equilibrium under eligibility restrictions we apply the tools developed in the proof of Proposition 4.

[13] Thus the results included here do not contradict Theorem 1. That result fixes $\delta < 1$ and a finite maximum ban length T.

5.1 Prices

Suppose the principal may use only monetary punishments, and wants to serve both wealth types.[14] From Proposition 1, the principal does not offer a refund upon cancellation. When the principal has access to prices alone, she can directly influence threshold usage rates as a function of the price p she sets. The threshold condition for wealth-type w reduces to $\pi \geq \underline{\pi}_w(p) := \frac{p}{w}$, thus under the wealth normalization, becomes

$$\underline{\pi}_L(p) = \frac{p}{w_L} = \min\{1, (1+\xi)p\}, \qquad \underline{\pi}_H(p) = \frac{p}{w_H} = \frac{1+\xi}{\xi}\, p.$$

Following Section 2, let $A(p)$ denote the mass of bookings and let $N(p)$ denote usage among those who book, each as a function of the price:

$$A(p) := \mu_L \alpha(\underline{\pi}_L(p)) + \mu_H \alpha(\underline{\pi}_H(p)), \qquad N(p) := \mu_L M(\underline{\pi}_L(p)) + \mu_H M(\underline{\pi}_H(p)).$$

We know from Proposition 2 that the principal allocates all of her capacity, so expected usage is given by

$$U^m(p) = \beta(p) N(p) = \frac{C N(p)}{A(p)} \tag{7}$$

where we used that $\beta(p) = \frac{C}{A(p)}$. With C fixed, maximizing $U^m(p)$ over $A(p) \geq C$ is equivalent to choosing p which satisfies $\beta(p) \leq 1$ and which maximizes (7). There are two candidates for the optimal price: the corner solution at which demand exactly meets supply, and the root which overbooks.[15] Thus we compare utility under the capacity-filling price p_C and the overbooking price p_{OB}.

We consider the corner solution first and derive price p_C. Setting $A(p) = C$ yields

$$p_C(\xi, \gamma; C) = \frac{\xi(1+\gamma)}{(1+\xi)(\xi+\gamma)}(1-C). \tag{8}$$

At p_C we have $\beta = 1$, so $U^m(p_C) = N(p_C)$ and hence the utility when capacity is *exactly filled* is

$$U^{m,\text{EF}}(\xi, \gamma; C) = \frac{1}{2}\Big(1 - \mathcal{K}(\xi, \gamma)(1-C)^2\Big), \qquad \mathcal{K}(\xi, \gamma) := \frac{(1+\gamma)(\gamma+\xi^2)}{(\gamma+\xi)^2}.$$

$\mathcal{K}$ is a measure of wealth inequality. Larger wealth dispersion ξ makes p/w a noisier proxy for π, which increases the inequality measure $\mathcal{K}$ and lowers welfare. We defer comparative statics on inequality to the end of this subsection.

[14] We discuss this latter condition at the end of this example.

[15] By this construction, it is possible that a root sets the price p such that capacity is underbooked, but this violates Lemma 3. Note that the overbooking root may not exist, in which case the corner solution is optimal.

The second candidate for the optimal price has $A(p) \geq C$ slack. Differentiating (7), the root is

$$p_{\mathrm{OB}}(\xi,\gamma) = \frac{\xi(1+\gamma)}{(1+\xi)(\xi+\gamma)}\left[1 - \frac{(\xi-1)\sqrt{\gamma}}{\sqrt{(1+\gamma)(\xi^2+\gamma)}}\right]. \tag{9}$$

Observe the relationship between (8) and (9): The overbooking price is a lower price, and thus overbooks, if $\frac{(\xi-1)\sqrt{\gamma}}{\sqrt{(1+\gamma)(\xi^2+\gamma)}} > C$. Let us define this critical capacity level as

$$C^\star(\xi,\gamma) := A\big(p_{\mathrm{OB}}(\xi,\gamma)\big) = \frac{(\xi-1)\sqrt{\gamma}}{\sqrt{(1+\gamma)(\xi^2+\gamma)}}. \tag{10}$$

Taking these together, the interior point p_{OB} is feasible if and only if $A(p_{\mathrm{OB}}) \geq C$, equivalently $C \leq C^\star(\xi,\gamma)$. We can then insert the overbooking price p_{OB} into (7) to yield utility

$$U^{m,\mathrm{OB}}(\xi,\gamma;C) = C \cdot \frac{N\left(p_{\mathrm{OB}}\right)}{A\left(p_{\mathrm{OB}}\right)} = C \cdot \frac{1}{1+C^\star(\xi,\gamma)} = \frac{C}{1+C^\star(\xi,\gamma)}.$$

Hence the value of $C^\star(\xi,\gamma)$ determines the optimal price and the subsequent utility level:

$$U^{m,\star}(\xi,\gamma;C) = \begin{cases} \dfrac{C}{1+C^\star(\xi,\gamma)} & \text{if } C < C^\star(\xi,\gamma), \\ \dfrac{1}{2}\Big(1 - \mathcal{K}(\xi,\gamma)\,(1-C)^2\Big) & \text{if } C \geq C^\star(\xi,\gamma). \end{cases} \tag{11}$$

The planner's pricing problem therefore becomes a trade-off between (i) using a high price p_C to exactly fill capacity, which avoids rationing but worsens selection distortions as ξ rises (captured by the inequality term $\mathcal{K}(\xi,\gamma)$), and (ii) using a lower price p_{OB} that deliberately overbooks so that rationing substitutes for precise screening when the price signal is too noisy.

We now consider comparative statics on (10).

Observation 1. *As wealth inequality decreases, $\xi \to 1$, exact booking becomes preferred to overbooking, $C^\star \to 0$.*

When there is no gap in wealth level the principal has no incentive to overbook, because the principal can target usage thresholds across wealth types using the same price. This argument remains valid when wealth inequality is small. Conversely, increasing levels of wealth inequality via ξ always makes the principal more likely to overbook.

Observation 2. *As the population distribution of wealth becomes imbalanced, $\gamma \to 0$ or $\gamma \to \infty$, exact booking becomes preferred to overbooking, $C^\star \to 0$.*

When there are few high-wealth types the principal cannot clear the booking market without setting a price low enough to attract at least some low-wealth types; hence exact-fill pricing always wins. Similarly, when there are so many high-usage, high-wealth types that the principal can clear the booking market while never attracting any low-wealth types, exact booking always wins. For

intermediate γ overbooking patterns are nonmonotone. As γ increases from 0 to ξ, more rich agents are willing to book and then cancel, and overbooking can recover the low-wealth types disincentivized from booking by a price which exactly clears the booking market. Once $\gamma > \xi$, the applicant pool becomes skewed toward the rich so the benefit of overbooking shrinks, hence $C^\star$ falls, on (ξ, ∞).

5.2 Bans

Absent money, wealth types behave symmetrically and use a common cutoff $\underline{\pi}$. The principal can implement any cutoff $\underline{\pi}$ subject to filling capacity.[16] Expected usage per period equals

$$U^b = \beta \cdot \zeta(0) \cdot M(\underline{\pi}) = \beta \cdot \zeta(0) \cdot \frac{1 - \underline{\pi}^2}{2}.$$

Under the benchmark, the principal chooses a ban policy so that capacity is exactly filled without rationing, so $A^b = C$ and hence $\beta = 1$. Let τ be the (deterministic) ban length. Then

$$\zeta(0) M(\underline{\pi}) = \frac{1 - \underline{\pi}^2}{2 + \tau(1 - \underline{\pi})^2},$$

which the agent chooses $\underline{\pi}$ to maximize, hence $\tau(1 - \underline{\pi})^2 = 2\underline{\pi}$. Plugging into $\zeta(0)$ gives $\zeta(0) = \frac{1}{1+\underline{\pi}}$. So for any cutoff $\underline{\pi}$ that is induced by some ban length τ, the implied steady-state eligible mass is $\zeta(0) = 1/(1 + \underline{\pi})$. Equivalently, the principal can implement any $\underline{\pi} \in (0, 1)$ by choosing $\tau = 2\underline{\pi}/(1 - \underline{\pi})^2$.

Apply the fact that capacity is filled exactly under the ban policy, so that $C = (1 - \underline{\pi})\zeta(0)$ implies

$$\frac{1 - \underline{\pi}}{1 + \underline{\pi}} = C \quad \to \quad \underline{\pi}^{b,\star}(C) = \frac{1 - C}{1 + C}, \quad \zeta^{b,\star}(0) = \frac{1}{1 + \underline{\pi}^{b,\star}(C)} = \frac{1 + C}{2}.$$

Expected usage per period equals allocated mass times expected usage probability conditional on booking:

$$U^b = C \cdot \mathbb{E}[\pi \mid \pi \geq \underline{\pi}] = C \cdot \frac{M(\underline{\pi})}{\alpha(\underline{\pi})} = C \cdot \frac{\frac{1-\underline{\pi}^2}{2}}{1 - \underline{\pi}} = C \cdot \frac{1 + \underline{\pi}}{2}. \tag{12}$$

And plugging $\underline{\pi}^{b,\star}(C) = \frac{1-C}{1+C}$ into (12) gives

$$U^{b,\star}(C) = C \cdot \frac{1 + \underline{\pi}^{b,\star}(C)}{2} = C \cdot \frac{1 + \frac{1-C}{1+C}}{2} = \frac{C}{1 + C}. \tag{13}$$

Comparing (11) and (13), we observe Theorem 2. Note

$$U^{b,\star}(C) > U^{m,\star}(\xi, \gamma; C) \iff C < C^\star(\xi, \gamma).$$

[16] Overbooking is never optimal in a mechanism which imposes only eligibility-based penalties. See the discussion in Lemma 7 for details.

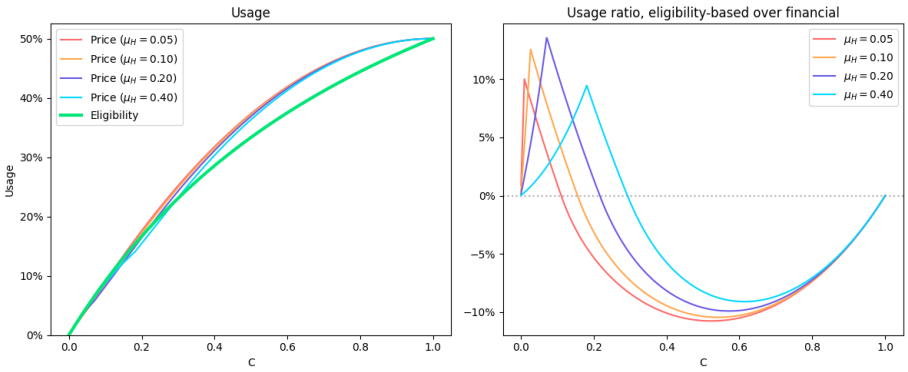


Figure 2: Performance of financial versus eligibility-based penalties. The left panel presents usage by mechanism, and the differences are marginal for low capacities C. The right panel presents relative usage $U^{b,\star}/U^{m,\star}$, and clearly illustrates Proposition 4. In both plots, $w_H/w_L = 2$.

Observation 3. *The ban-only policy is preferred to the price-only policy if and only if the price-only policy overbooks.*

Thus in our uniform example the level of inequality at which overbooking becomes optimal is the level of inequality at which bans begin to dominate the price regime. When inequality grows, a principal would prefer to either overbook and subsequently ration, or to threaten agents with a loss of access. Figure 2 illustrates this transition.

We now discuss our assertion that both wealth types are served. Recall that the optimal price-only policy may serve just one wealth type, in which case it is optimal for the principal to induce bookings exactly equal to her capacity. Observation 3 is necessary and sufficient here only because we elide this case. Indeed, without that restriction, the optimal price-only policy does not overbook but is still dominated by the optimal ban-only policy.

6 Conclusion

We study an empirically relevant game: a principal wants to allocate resources to a group of agents who vary in their value for money, but whose wealth levels affect the incentive properties of regulation. We motivate the use of access restrictions, especially under wealth inequality: the principal is better off threatening agents with temporary removal from the game than screening using prices. We highlight a tradeoff between a designer maintaining competition and equalizing punishment across the population.

We note that in practice, platform designers are not constrained to select between financial and eligibility-based punishments, and may select a more exotic mechanism. We leave this possibility unaddressed for two reasons. First, we design with an eye toward practical applications, where monetary penalties and temporary bans are common features of reservation markets.[17] Second, in

[17]Cf. our leading example of California state campsites. Additionally, the online restaurant reservation platform OpenTable states that the platform, "[...] shall have no liability for any charges made to the debit or credit card

the general mechanism design problem the principal may be able to leverage the history of each agent's actions to determine whether they are high- or low-wealth, and discriminate appropriately in steady-state. To the best of our knowledge, this does not correspond to the mechanisms observed in practice. Future work can study this environment where usage propensities persist over time, or are potentially correlated with wealth level — for instance, if wealthier agents have better outside options and thus smaller usage probabilities.

References

Mohammad Akbarpour, Piotr Dworczak, and Scott Duke Kominers. 2024. Redistributive allocation mechanisms. *Journal of Political Economy* 132, 6 (2024), 1831–1875.

Justin Burkett and Kyle Woodward. 2024. Intertemporal Allocation with Unknown Discounting. (2024).

Daniele Condorelli. 2013. Market and non-market mechanisms for the optimal allocation of scarce resources. *Games and Economic Behavior* 82 (November 2013), 582–591.

Piotr Dworczak. 2024. Inequality and Market Design. *ACM SIGecom Exchanges* 22, 1 (June 2024), 83–92. https://doi.org/10.1145/3699824.3699831

George Georgiadis and Christopher S. Tang. 2014. Optimal reservation policies and market segmentation. *International Journal of Production Economics* 154 (Aug. 2014), 81–99. https://doi.org/10.1016/j.ijpe.2014.04.021

Matthew O. Jackson and Hugo F. Sonnenschein. 2007. Overcoming Incentive Constraints by Linking Decisions. *Econometrica* 75, 1 (2007), 241–257.

Taylor Lundy, Alexander Wei, Hu Fu, Scott Duke Kominers, and Kevin Leyton-Brown. 2019. Allocation for Social Good: Auditing Mechanisms for Utility Maximization. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC '19)*. ACM, Phoenix, AZ, USA, 785–803. https://doi.org/10.1145/3328526.3329623

Qianjun Lyu. 2024. Optimal Refund Mechanism with Consumer Learning. https://doi.org/10.48550/arXiv.2404.14927 arXiv:2404.14927 [econ].

Hongyao Ma, Reshef Meir, David C. Parkes, and James Zou. 2018. Contingent Payment Mechanisms for Resource Utilization. https://doi.org/10.48550/arXiv.1607.06511 arXiv:1607.06511 [cs].

Mikael Mäkimattila. 2025. Inequality-Aware Regulation. (2025).

account for any failure to cancel your reservation in accordance with a Restaurant's cancellation policy ... [Additionally,] your Account will be suspended if you are a no-show for four reservations within a 12-month period." US (2026)

Office of Assemblymember Rebecca Bauer-Kahan. 2023. Governor Newsom Signs Assemblymember Bauer-Kahan's Legislation to Improve Access to Campsites. Press Release, California State Assembly, District 16.

Rohit Patel and Can Urgun. 2024. Costly Verification and Money Burning. (2024).

Parag A Pathak, Tayfun Sönmez, M Utku Ünver, and M Bumin Yenmez. 2021. Fair allocation of vaccines, ventilators and antiviral treatments: leaving no ethical value behind in health care rationing. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 785–786.

Gregory Thomas. 2026. Big Change Coming to California Campsite Reservations. *San Francisco Chronicle* (23 June 2026). https://www.sfchronicle.com/outdoors/article/california-campgrounds-cancellation-penalty-22309804.php

OpenTable US. 2026. OpenTable Terms of Use - OpenTable. https://www.opentable.com/c/legal/terms-and-conditions/. [Accessed 09-02-2026].

Valentin Verdier and Carson Reeling. 2022. Welfare effects of dynamic matching: An empirical analysis. *The Review of Economic Studies* 89, 2 (2022), 1008–1037.

A Modeling

A.1 Formal model of bans

We now consider the representation of a ban. We previously mentioned that agent i is eligible in period t if and only if $z_{it} = 0$; the variable $z_{it} \in \mathbb{Z}_+$ should be interpreted as the number of periods of ineligibility remaining at the start of period t for agent i. The state evolves according to

$$z_{i,t+1} = \begin{cases} z_{it} - 1, & \text{if } z_{it} > 0, \\ \tau_{it}, & \text{if } z_{it} = 0,\ y_{it} = 1,\ \text{and } c_{it} = 1,\ \text{where } \tau_{it} \sim \rho, \\ 0, & \text{otherwise.} \end{cases}$$

Given a policy $(p, r; \rho)$ and the induced booking cutoffs $(\underline{\pi}_H, \underline{\pi}_L)$, let $\zeta_{tw}(k)$ denote the period-t mass of wealth-w agents with ban state $z_{it} = k$,[18] and recall β_t is the probability that a booking is allocated. Then

$$\begin{aligned} \zeta_{t+1,w}(\tau) &= \zeta_{t,w}(\tau+1) + \zeta_{t,w}(0)\beta_t\big(\alpha(\underline{\pi}_w) - M(\underline{\pi}_w)\big)\rho(\tau), \qquad (\tau > 0), \\ \zeta_{t+1,w}(0) &= \zeta_{t,w}(1) + \zeta_{t,w}(0)\beta_t\big(\alpha(\underline{\pi}_w) - M(\underline{\pi}_w)\big)\rho(0) \\ &\quad + \zeta_{t,w}(0)\Big(\big(1 - \alpha(\underline{\pi}_w)\big) + \big(1 - \beta_t\big)\alpha(\underline{\pi}_w) + \beta_t M(\underline{\pi}_w)\Big). \end{aligned}$$

[18] We will use commas to separate subscripts when the notation may be unclear.

In steady state, $\zeta_{t+1,w}(\tau) = \zeta_{tw}(\tau)$ for all τ when $\underline{\pi}_w$ and β_t are determined by agent equilibrium behavior. As $\zeta_{tw}(\cdot)$ does not depend on t, we write the steady-state eligible fraction for wealth type w as $\zeta_w(0) \in (0, 1]$.

B Preliminary results

We introduce two new objects which we use throughout the Appendix.

$$H(\underline{\pi}) := \mathbb{E}\left[(1-\pi)\mathbf{1}\{\pi \geq \underline{\pi}\}\right] = \int_{\underline{\pi}}^{1} (1-x)dF(x) = \alpha(\underline{\pi}) - M(\underline{\pi}).$$

$$G(\underline{\pi}) := M(\underline{\pi}) - \underline{\pi}\alpha(\underline{\pi}) = \mathbb{E}\left[(\pi-\underline{\pi})\mathbf{1}\{\pi \geq \underline{\pi}\}\right] = \int_{\underline{\pi}}^{1} (x-\underline{\pi})dF(x).$$

$H(\underline{\pi})$ is the expected unused mass and $G(\underline{\pi})$ is the expected excess usage probability above the cutoff. Note the derivative $G'(\underline{\pi}) = -\alpha(\underline{\pi})$. We will also continue to use $\kappa = \delta - \delta_\rho$ which we introduced in the beginning of Section 2.2.

Given the stationary equilibrium, we can formally write the stationary flow utility of an agent of a given wealth type, and pin down the thresholds as a function of the policy.

Lemma 5. *Fix a policy $(p, r; \rho)$. The stationary flow utility of wealth type w is given by*

$$\widehat{V}_w = \frac{\beta(1-\delta)\left[M(\underline{\pi}_w) - \frac{p}{w}\alpha(\underline{\pi}_w) + \frac{r}{w}H(\underline{\pi}_w)\right]}{(1-\delta) + \beta\kappa H(\underline{\pi}_w)}. \tag{14}$$

The threshold is then pinned down by the policy: $\Phi_w(\underline{\pi}_w; p, r) = 0$ for any interior threshold of type w where

$$\Phi_w(\underline{\pi}; p, r) := (1-\delta)\left(\underline{\pi} - \frac{p}{w}\right) + (1-\delta)(1-\underline{\pi})\frac{r}{w} - \beta\kappa\left(1 - \frac{p}{w}\right)G(\underline{\pi}). \tag{15}$$

Proof. Observe the flow utility from booking is

$$\widehat{V}^b_w(\pi) = \beta\Bigg[\pi\left((1-\delta)\left(1-\frac{p}{w}\right) + \delta\widehat{V}_w\right) + (1-\pi)\left(-(1-\delta)\frac{p-r}{w} + \delta_\rho\widehat{V}_w\right)\Bigg] + (1-\beta)\delta\widehat{V}_w,$$

while the flow utility from not booking is $\widehat{V}^n_w = \delta\widehat{V}_w$. The agent books if and only if $\pi \geq \underline{\pi}_w$. The

stationary flow utility then equals $\hat{V}_w^b$ only when $\pi \geq \underline{\pi}_w$, thus

$$\begin{aligned}
\widehat{V}_w &= \int_0^{\underline{\pi}_w} \widehat{V}_w^n \, dF(\pi) + \int_{\underline{\pi}_w}^1 \widehat{V}_w^b(\pi) \, dF(\pi) \\
&= \delta \widehat{V}_w + \int_{\underline{\pi}_w}^1 \left[\widehat{V}_w^b(\pi) - \widehat{V}_w^n \right] dF(\pi) \\
&= \delta \widehat{V}_w + \beta \Bigg[(1-\delta) \left(M(\underline{\pi}_w) - \frac{p}{w} \alpha(\underline{\pi}_w) + \frac{r}{w} H(\underline{\pi}_w) \right) \\
&\qquad - \kappa \widehat{V}_w H(\underline{\pi}_w) \Bigg].
\end{aligned}$$

Rearranging for $\hat{V}_w^b$ gives (14). Meanwhile, the type at the threshold is indifferent between booking and not booking; $\widehat{V}_w^b(\pi) - \widehat{V}_w^n = 0$. So the following equality must hold:

$$(1 - \underline{\pi}_w) \kappa \widehat{V}_w = (1-\delta) \left[\underline{\pi}_w - \frac{p}{w} + (1 - \underline{\pi}_w) \frac{r}{w} \right]. \tag{16}$$

Multiplying the stationary equation by $1 - \underline{\pi}_w$ and substituting (16) gives

$$\begin{aligned}
&(1-\delta)(1-\underline{\pi}_w) \widehat{V}_w \\
&\quad = \beta(1-\delta) \Bigg\{ (1 - \underline{\pi}_w) \left[M(\underline{\pi}_w) - \frac{p}{w} \alpha(\underline{\pi}_w) + \frac{r}{w} H(\underline{\pi}_w) \right] \\
&\qquad - \left[\underline{\pi}_w - \frac{p}{w} + (1 - \underline{\pi}_w) \frac{r}{w} \right] H(\underline{\pi}_w) \Bigg\} \\
&\quad = \beta(1-\delta) \left(1 - \frac{p}{w} \right) G(\underline{\pi}_w).
\end{aligned}$$

Dividing by $1 - \delta$ gives $(1 - \underline{\pi}_w) \hat{V}_w = \beta \Big(1 - \frac{p}{w} \Big) G(\underline{\pi}_w)$, and plugging in for $\hat{V}_w$ in (14) gives (15). □

We also provide a closed form for the steady-state equilibrium mass of agents who are eligible.

Lemma 6. *The steady-state equilibrium mass of agents who are eligible to book (i.e., not in an access-restriction state) is*

$$\zeta_w^\star(0) = \left[1 + \sum_{t=1}^{T} \left(\alpha(\underline{\pi}_w) - M(\underline{\pi}_w) \right) \beta^\star t \rho(t) \right]^{-1}.$$

Proof. Let $\zeta_w^\star(z)$ be the steady-state mass of wealth-type w with state variable z, and let $\lambda_{wzz'}^\star$ be the steady-state probability that a wealth-w agent transitions from state variable z to state variable z'. Note that $\lambda_{wzz'} \neq 0$ only if $z = 0$ or $z = z' + 1$. Then $\zeta_w^\star(T) = \lambda_{w0T}^\star \zeta_w^\star(0)$, and $\zeta_w^\star(t) = \lambda_{w0t}^\star \zeta_w^\star(0) + \lambda_{w,t+1,t}^\star \zeta_w^\star(t+1)$ for all $t < T$. Since the state variable z counts down the time until the agent may book again, $\lambda_{w,t+1,t}^\star = 1$ for all w, t, and hence $\zeta_w^\star(t) = \lambda_{w0t}^\star \zeta_w^\star(0) + \zeta_w^\star(t+1)$.

Applying recursion gives

$$\zeta_w^\star(t) = \sum_{t'=t}^{T} \lambda_{w0t'}^\star \zeta_w^\star(0).$$

Since $\zeta_w^\star$ is a probability distribution, this implies that

$$\sum_{t=0}^{T}\sum_{t'=t}^{T} \lambda_{w0t'}^\star \zeta_w^\star(0) = 1 \quad \Longrightarrow \quad \zeta_w^\star(0) = \left[\sum_{t=0}^{T}\sum_{t'=t}^{T} \lambda_{w0t'}^\star\right]^{-1} = \left[\sum_{t=0}^{T} (t+1)\,\lambda_{w0t}^\star\right]^{-1}.$$

The transition probabilities $\lambda_{w0t}^\star$ are defined by the probability that an agent cancels and is punished with a t-period ban. We have already argued that fraction $(\alpha(\underline{\pi}_w) - M(\underline{\pi}_w))\beta^\star$ of wealth-w agents will be subject to some (possibly nondegenerate) punishment arising from cancellation. Of these, fraction $\rho(t)$ will face a t-period suspension. Then $\lambda_{w0t}^\star = (\alpha(\underline{\pi}_w) - M(\underline{\pi}_w))\beta^\star\rho(t)$ for $t > 0$, and $\lambda_{w00}^\star = 1 - (\alpha(\underline{\pi}_w) - M(\underline{\pi}_w))\beta^\star(1-\rho(0))$. Then we have

$$\zeta_w^\star(0) = \left[1 + \sum_{t=1}^{T} \left(\alpha\left(\underline{\pi}_w\right) - M\left(\underline{\pi}_w\right)\right)\beta^\star t\rho\left(t\right)\right]^{-1}.$$

□

C Proofs for Section 3

C.1 Optimal Prices and Refunds

Proof of Proposition 1. We will show that for every optimal policy $(p, r; \rho)$, there exists an optimal policy $(\widetilde{p}, 0; \rho)$. Thus, without loss of optimality, the principal may restrict attention to policies without refunds (although some refund-using policies may also be optimal). We first rewrite the cutoff condition. For any wealth type w that books with positive probability, necessarily $p < w$. Define

$$x_w := \frac{p-r}{w-r}.$$

The cutoff condition (15) factors as

$$\Phi_w(\underline{\pi}_w; p, r) = \left(1 - \frac{r}{w}\right)\left[(1-\delta)(\underline{\pi}_w - x_w) - \beta\kappa(1-x_w)G(\underline{\pi}_w)\right] = 0. \tag{17}$$

Intuitively, x_w is the normalized no-refund price that would induce the same cutoff for wealth type w, holding β and ρ fixed. Note the solution to the expression in brackets is strictly increasing in x_w.

First, suppose that both wealth types book and that there is a strictly positive refund that is not the full refund: $0 < r < p$. Then $p < w_L$, and so $x_H < x_L$, thus $\underline{\pi}_H < \underline{\pi}_L$. Note the two cutoffs can be varied locally and independently by varying (p, r). Now, fix ρ and hold the total booking

mass fixed, so β is also fixed. Use the following substitutions:

$$B_\rho := \sum_{t=1}^{T} t\rho(t), \qquad K := \beta B_\rho, \qquad \zeta(\pi) := \frac{1}{1+KH(\pi)}.$$

We will show that the changes in booking mass and expected usage are not proportional across wealth types. Write

$$\mathcal{A}(\pi) := \zeta(\pi)\alpha(\pi), \qquad \mathcal{N}(\pi) := \zeta(\pi)M(\pi).$$

Direct differentiation gives

$$\mathcal{A}'(\pi) = f(\pi)\zeta(\pi)^2 \left[KG(\pi) - 1\right], \tag{18}$$

$$\mathcal{N}'(\pi) = f(\pi)\zeta(\pi)^2 \left[KG(\pi) - \pi\right]. \tag{19}$$

Whenever $\mathcal{A}'(\pi) \neq 0$, write the proportional change as

$$R(\pi) := \frac{\mathcal{N}'(\pi)}{\mathcal{A}'(\pi)} = \frac{KG(\pi) - \pi}{KG(\pi) - 1}.$$

And note the proportional change is increasing:

$$R'(\pi) = \frac{1 + KH(\pi)}{\left[KG(\pi) - 1\right]^2} > 0.$$

Therefore $R(\underline{\pi}_H) \neq R(\underline{\pi}_L)$. Indeed, if the two cutoffs lie on the same side of the point satisfying $KG(\pi) = 1$, this follows from the strict monotonicity of R. If they lie on opposite sides, then

$$R(\underline{\pi}_H) > 1 > R(\underline{\pi}_L).$$

Finally, if $KG(\underline{\pi}_w) = 1$ for one wealth type, then $\mathcal{A}'(\underline{\pi}_w) = 0$, while

$$\mathcal{N}'(\underline{\pi}_w) = f(\underline{\pi}_w)\zeta(\underline{\pi}_w)^2(1 - \underline{\pi}_w) > 0,$$

so the same conclusion follows.

The total booking and potential-usage masses are

$$A = \mu_H \mathcal{A}(\underline{\pi}_H) + \mu_L \mathcal{A}(\underline{\pi}_L)$$

and

$$N = \mu_H \mathcal{N}(\underline{\pi}_H) + \mu_L \mathcal{N}(\underline{\pi}_L).$$

The preceding inequalities imply that the gradients of A and N with respect to $(\underline{\pi}_H, \underline{\pi}_L)$ are not proportional. Hence there exists a local perturbation of the two cutoffs that leaves A unchanged and strictly increases N. This perturbation can be implemented by a nearby policy. Since A, and

therefore β, is unchanged, realized usage $U = \beta N$ strictly increases. Thus no policy with $0 < r < p$ and both wealth types active can be optimal.

Now suppose both wealth types book and $p = r$. In this case, (17) becomes

$$\left(1 - \frac{p}{w}\right)\left[(1-\delta)\underline{\pi}_w - \beta\kappa G(\underline{\pi}_w)\right] = 0.$$

Because $p < w_L$, the first factor is strictly positive. The remaining expression is exactly the cutoff condition under the policy $(0, 0; \rho)$. The two policies therefore induce identical cutoffs, eligibility fractions, booking mass, allocation probability, and usage. Thus this full-refund policy has an outcome-equivalent no-refund implementation.

It remains to consider a policy under which only one wealth type books. The active type must be the high-wealth type. Let its cutoff be $\underline{\pi}_H$, and define its effective no-refund price by

$$\widehat{p} := w_H x_H = w_H \frac{p - r}{w_H - r}.$$

Fix the original allocation probability β and punishment distribution ρ. Under the no-refund policy $(\widehat{p}, 0; \rho)$, the high-wealth type has exactly the same cutoff and eligibility fraction as under the original policy. The low-wealth type may additionally enter, so the resulting booking mass is weakly greater than the original booking mass A.

As the no-refund price increases continuously from $\widehat{p}$ to w_H, the booking mass varies continuously from a value weakly above A to zero. Hence there exists some $\widetilde{p} \in [\widehat{p}, w_H]$ for which the no-refund policy $(\widetilde{p}, 0; \rho)$ induces exactly booking mass A. It therefore also induces the original allocation probability β.

Every wealth type that books under this new policy has a cutoff weakly above $\underline{\pi}_H$. Since $m(\pi)$ is strictly increasing, the potential-usage mass under the new policy satisfies

$$\begin{aligned}
\widetilde{N} &= \sum_{w \in \{H,L\}} \mu_w \widetilde{\zeta}_w \alpha(\widetilde{\pi}_w) m(\widetilde{\pi}_w) \\
&\geq m(\underline{\pi}_H) \sum_{w \in \{H,L\}} \mu_w \widetilde{\zeta}_w \alpha(\widetilde{\pi}_w) \\
&= A\, m(\underline{\pi}_H) = N.
\end{aligned}$$

Because the booking mass and hence β are unchanged, $\widetilde{U} = \beta \widetilde{N} \geq \beta N = U$. Thus every optimal policy serving only one wealth type also has an optimal no-refund implementation.

We have exhausted all cases. Therefore optimal usage can always be implemented by a policy with $r = 0$. □

Proof of Proposition 2. Suppose underbooking occurs. Then $\mu_H \zeta_H(0)\alpha(\underline{\pi}_H) + \mu_L \zeta_L(0)\alpha(\underline{\pi}_L) < C$. By slightly reducing prices the principal can increase $\zeta_w(0)\alpha(\underline{\pi}_w)$; since the base mechanism is underbooking, a marginal increase in booking mass does not affect the allocation probability $\beta = 1$. Thus reducing prices strictly increases expected usage. □

C.2 Optimal Bans

Proof of Proposition 3. The agent's decision depends on the punishment distribution ρ only through $\mathbb{E}_{\tau\sim\rho}[\delta^{\tau+1}]$.[19] We show that for any ρ, there is a $\rho^\star$ supported on at most two ban lengths such that $\mathbb{E}_{\tau\sim\rho}[\delta^{\tau+1}] = \mathbb{E}_{\tau\sim\rho^\star}[\delta^{\tau+1}]$ and $\zeta_{\rho 0}(\pi) \leq \zeta_{\rho^\star 0}(\pi)$. That is, the principal can move from punishment distribution ρ to punishment distribution $\rho^\star$ and increase the steady-state mass of eligible agents without affecting the agents' incentives.

From Lemma 6 in Appendix B, we write the steady state mass as

$$\zeta_{\rho 0}(\pi) = \left[1 + \sum_{\tau=1}^{T} (\alpha(\pi) - M(\pi))\,\beta^\star \tau \rho(\tau)\right]^{-1}$$

Since $(\alpha(\pi) - M(\pi))\beta^\star$ is constant in this argument, we write

$$\zeta_{\rho 0}(\pi) = \left[1 + K \sum_{\tau=1}^{T} \tau \rho(\tau)\right]^{-1}.$$

Then we solve

$$\max_{\rho'} \left[1 + K \sum_{\tau=1}^{T} \tau \rho'(\tau)\right]^{-1}, \text{ s.t. } \rho' \in \Delta_T \text{ and } \sum_{\tau=0}^{T} \delta^\tau \rho'(\tau) = D.$$

We replace the maximization of the inverse with the minimization of the non-inverse,

$$\min_{\tilde{\rho}} \sum_{\tau=1}^{T} \tau \tilde{\rho}(\tau), \text{ s.t. } \tilde{\rho} \in \Delta_T \text{ and } \sum_{\tau=0}^{T} \delta^{\tau+1} \tilde{\rho}(\tau) = D.$$

The Lagrangian of this problem with respect to $\tilde{\rho}(\tau)$ is

$$\tau - \lambda_\tau + \gamma + \delta^{\tau+1}\eta = 0,$$

where λ_τ is the multiplier on $\tilde{\rho}(\tau) \geq 0$, γ is the multiplier on $\sum_{\tau=0}^{T} \rho(\tau) = 1$, and η is the multiplier on $\mathbb{E}_{\tilde{\rho}}[\delta^{\tau+1}] = D$. If $\rho^\star(\tau) > 0$ then $\lambda_\tau = 0$, and hence $\tau + \delta^{\tau+1}\eta = -\gamma$ for all τ such that $\rho^\star(\tau) > 0$. Note that $\frac{d^2}{d\tau^2}[\tau + \delta^{\tau+1}\eta] = [\ln\delta]^2\delta^{\tau+1} > 0$, hence $\tau + \delta^{\tau+1}\eta$ is convex. It follows that there are at most two values $\{\tau_1^\star, \tau_2^\star\} \ni \tau$ such that $\rho^\star(\tau) > 0$. Additionally, since λ_τ is weakly positive convexity implies that $|\tau_2^\star - \tau_1^\star| \in \{0, 1\}$.

The result follows from the observation that with excess bookings, the principal can improve their utility by increasing the booking threshold. □

[19] The agent's utility also depends on the allocation probability β. In steady-state equilibrium this probability affects stage and continuation utilities equally, hence the effect can be ignored here.

D Proofs for Section 4

Proof of Lemma 1. We will show that $\frac{d\underline{\pi}_w}{d\frac{p}{w}} > 0$ for any interior cutoff, and then use the fact that $w_H \geq w_L$ to conclude that $\underline{\pi}_H \leq \underline{\pi}_L$. Note the optimal policy has $r = 0$ and $\kappa = \delta - \delta_\tau$ for some τ. Take the partials of (14) toward using the implicit function theorem:

$$\frac{\partial \Phi}{\partial \underline{\pi}_w}(\underline{\pi}_w; x) = (1-\delta) - \beta\kappa(1-x)G'(\underline{\pi}_w) = (1-\delta) + \beta\kappa(1-x)\left(1 - F(\underline{\pi}_w)\right) > 0.$$

$$\frac{\partial \Phi}{\partial x}(\underline{\pi}_w; x) = -(1-\delta) + \beta\kappa G(\underline{\pi}_w) < 0$$

To see that $\frac{\partial \Phi}{\partial x}(\underline{\pi}_w; x)$ is negative, observe that at any interior solution we have $\Phi(\underline{\pi}_w; x) = 0$, so

$$\beta\kappa G(\underline{\pi}_w) = (1-\delta)\frac{\underline{\pi}_w - x}{1-x} < (1-\delta).$$

Hence

$$\frac{d\underline{\pi}_w}{dx} = -\frac{\frac{\partial \Phi}{\partial x}}{\frac{\partial \Phi}{\partial \underline{\pi}_w}} > 0$$

at any interior solution, so $\underline{\pi}_w$ is strictly increasing in $x = p/w$. □

Proof of Corollary 1. Consider any optimal outcome covered by the statement, and let $\bar{w}$, $\bar{\mu}$, $\underline{\pi}$, and ζ denote the wealth, total mass, common cutoff, and eligible fraction of the agents it serves. Its booking and potential-usage masses are $A = \bar{\mu}\zeta\alpha(\underline{\pi})$ and $N = Am(\underline{\pi})$. Set $p_0 = \bar{w}\underline{\pi}$. Under the no-refund, no-ban price p_0, the originally served agents alone generate booking mass $\bar{\mu}\alpha(\underline{\pi}) \geq A$, whereas booking mass falls continuously to zero as the price rises to $\bar{w}$. Hence some $\widetilde{p} \in [p_0, \bar{w}]$ induces exactly A bookings. Every type that books at $\widetilde{p}$ has a cutoff at least $\underline{\pi}$, so its potential-usage mass is at least $Am(\underline{\pi}) = N$. Because booking mass is unchanged, the allocation probability is also unchanged. This gives an optimal price-only implementation. If the original policy used a nontrivial ban, then $\zeta < 1$, making the improvement strict and contradicting optimality. □

Proof of Corollary 2. Let $U_P^\star(\xi)$ and $U_B^\star$ denote maximal usage under price-only and ban-only policies, respectively. We will show that $U_B^\star - U_P^\star(\xi)$ is weakly increasing in wealth dispersion.

A ban-only policy contains no wealth-dependent monetary term, so its induced cutoffs, eligible fractions, and usage are independent of ξ; hence $U_B^\star$ is independent of ξ. Now fix $1 \leq \xi_1 < \xi_2$ and consider any price-only policy at ξ_2. Write its high- and low-wealth cutoffs as a and $b := \min\{\xi_2 a, 1\}$. By continuity and monotonicity of booking mass, there is a price at ξ_1 that produces the same booking mass with cutoffs a' and $b' := \min\{\xi_1 a', 1\}$ satisfying $a \leq a' \leq b' \leq b$. Equality of booking masses implies

$$\mu_H\big[F(a') - F(a)\big] = \mu_L\big[F(b) - F(b')\big].$$

The change in potential usage is therefore

$$\mu_L \int_{b'}^{b} x\, dF(x) - \mu_H \int_{a}^{a'} x\, dF(x) \geq 0,$$

because the two integrals contain equal probability mass and every usage probability in $[b', b]$ is weakly greater than every usage probability in $[a, a']$. Because booking mass is unchanged, the allocation probability is also unchanged. Thus every price-only outcome at ξ_2 is weakly improved upon by a price-only outcome at ξ_1. Taking the maximum over price-only policies gives $U_P^\star(\xi_1) \geq U_P^\star(\xi_2)$. Since $U_B^\star$ is constant in ξ, it follows that $U_B^\star - U_P^\star(\xi)$ is weakly increasing in ξ. □

D.1 Proof of Theorem 1

Proof of Lemma 2. Fix $\pi < 1$. Because $\underline{\pi}_w^\star \in [0,1]$ for all C, we may assume that $\underline{\pi}_w^\star \to P_w$ and $C \searrow 0$ for $w \in \{L, H\}$. If $P_L, P_H < 1$, then the principal can improve usage by increasing prices (this has the effect of increasing the expected usage rate of booking agents, $M(\underline{\pi}_w^\star)/\alpha(\underline{\pi}_w^\star)$, the reservation acceptance probability $\beta^\star$, and the mass of eligible agents $\zeta_w^\star(0)$). Then assume that w and w' are such that $P_{w'} = 1$ and $P_w < 1$. Expected usage along this sequence of equilibria is

$$U^\star = \mu_w M\left(\underline{\pi}_w^\star\right) \beta^\star \zeta_w^\star\left(0\right) + \mu_{w'} M\left(\underline{\pi}_{w'}^\star\right) \beta^\star \zeta_{w'}^\star\left(0\right).$$

By the assumption that $P_w < 1$, the mechanism employs rationing when C is sufficiently small. Then

$$\beta^\star = \left[\mu_w \alpha\left(\underline{\pi}_w^\star\right) \zeta_w^\star\left(0\right) + \mu_{w'} \alpha\left(\underline{\pi}_{w'}^\star\right) \zeta_{w'}^\star\left(0\right)\right]^{-1} C.$$

Consider an alternative mechanism which eliminates bans in favor of a price $p = w_H F^{-1}(1 - C/\mu_H)$. When C is small ($C < \mu_H$) this price is well defined, and such that $p > w_L$ so that only high-wealth agents will book. Expected usage in this mechanism is $U^p = \mu_H M \circ F^{-1}(1 - C/\mu_H)$.

It is then sufficient to compare $U^\star \gtrless U^p$ when C is small. That is,

$$\lim_{C \searrow 0} \frac{\mu_w M\left(\underline{\pi}_w^\star\right) \zeta_w^\star\left(0\right) + \mu_{w'} M\left(\underline{\pi}_{w'}^\star\right) \zeta_{w'}^\star\left(0\right)}{\mu_w \alpha\left(\underline{\pi}_w^\star\right) \zeta_w^\star\left(0\right) + \mu_{w'} \alpha\left(\underline{\pi}_{w'}^\star\right) \zeta_{w'}^\star\left(0\right)} \gtrless \lim_{C \searrow 0} \frac{1}{C} \mu_H M \circ F^{-1}\left(1 - \frac{C}{\mu_H}\right).$$

Since $P_w < 1$ and $P_{w'} = 1$, the left-hand side is $\mathbb{E}[\pi | \pi \geq \underline{\pi}_w] < 1$. The right-hand side is $\frac{d}{dx} M(F^{-1}(1-x))|_{x=0} = 1$. It follows that $U^p > U^\star$ when C is small, contradicting the assumed optimality of the mechanisms inducing $(\underline{\pi}_L^\star, \underline{\pi}_H^\star)$. □

Lemma 7 (Effect of Threshold on Booking Outcomes)**.** *When π is large, both $\zeta_0^\star(\cdot)\alpha(\cdot)$ and $\zeta_0^\star(\cdot)M(\cdot)$ are decreasing.*

Proof. From Lemma 6,

$$\frac{d}{d\pi} \zeta_0^\star\left(\pi\right) = -K\left(1 - \pi\right) \alpha'\left(\pi\right) \zeta_0^\star\left(\pi\right)^2, \quad K = \sum_{t=1}^{T} \beta^\star t \rho\left(t\right).$$

Then the derivative of $\zeta_0^\star(\cdot)\alpha(\cdot)$ is

$$\begin{aligned}\frac{d}{d\pi}\left[\zeta_0^\star(\pi)\,\alpha(\pi)\right] &= -K(1-\pi)\,\alpha'(\pi)\,\zeta_0^\star(\pi)^2\,\alpha(\pi)+\alpha'(\pi)\,\zeta_0^\star(\pi)\\ &\overset{\text{sgn}}{=} K(1-\pi)\,\alpha(\pi)-\zeta_0^\star(\pi)^{-1}\\ &= K(1-\pi)\,\alpha(\pi)-\left[1+K(\alpha(\pi)-M(\pi))\right]\\ &= K(M(\pi)-\pi\alpha(\pi))-1.\end{aligned}$$

As $\pi \nearrow 1$, $M(\pi)-\pi\alpha(\pi)\searrow 0$, establishing the first result.

Similarly, the derivative of $\zeta_0^\star(\cdot)M(\cdot)$ is

$$\begin{aligned}\frac{d}{d\pi}\left[\zeta_0^\star(\pi)\,M(\pi)\right] &= -K(1-\pi)\,\alpha'(\pi)\,\zeta_0^\star(\pi)^2\,M(\pi)+M'(\pi)\,\zeta_0^\star(\pi)\\ &\overset{\text{sgn}}{=} K(1-\pi)\,M(\pi)-\pi\zeta_0^\star(\pi)^{-1}\\ &= K(1-\pi)\,M(\pi)-\left[1+K(\alpha(\pi)-M(\pi))\right]\pi\\ &= K(M(\pi)-\pi\alpha(\pi))-\pi.\end{aligned}$$

As $\pi \nearrow 1$, $M(\pi)-\pi\alpha(\pi)\searrow 0$, establishing the second result. □

Proof of Lemma 3. By Lemma 7 we show that by slightly reducing prices the principal can increase $\zeta_w(0)\alpha(\underline{\pi}_w)$. The claim then follows exactly as in Proposition 2. □

Proof of Lemma 4. If the principal can directly manipulate the thresholds $\underline{\pi}_w$, the steady-state distribution of agents in state $z_i=0$ depends only on $\underline{\pi}_w$ and not directly on wealth. We can thus write $\zeta(\underline{\pi}_w)$ for $\zeta_w(0;\underline{\pi}_w)$. The total mass of bookings is

$$\mu_H\zeta(\underline{\pi}_H)\,\alpha(\underline{\pi}_H)+\mu_L\zeta(\underline{\pi}_L)\,\alpha(\underline{\pi}_L).$$

Then an infinitesimal change in $\underline{\pi}$ that leaves the total mass of bookings unchanged is such that

$$\mu_H\frac{d}{d\underline{\pi}_H}\left[\zeta(\underline{\pi}_H)\,\alpha(\underline{\pi}_H)\right]d\underline{\pi}_H+\mu_L\frac{d}{d\underline{\pi}_L}\left[\zeta(\underline{\pi}_L)\,\alpha(\underline{\pi}_L)\right]d\underline{\pi}_L=0.$$

For simplicity, write $d/d\underline{\pi}_w=\zeta_w'\alpha_w+\zeta\alpha_w'$. Then the above condition can be written as

$$\frac{\mu_H d\underline{\pi}_H}{\mu_L d\underline{\pi}_L}=-\frac{\zeta_L'\alpha_L+\zeta_L\alpha_L'}{\zeta_H'\alpha_H+\zeta_H\alpha_H'}.$$

We want to show that

$$\mu_H\frac{d}{d\underline{\pi}_H}\left[\zeta(\underline{\pi}_H)\,M(\underline{\pi}_H)\right]d\underline{\pi}_H+\mu_L\frac{d}{d\underline{\pi}_L}\left[\zeta(\underline{\pi}_L)\,M(\underline{\pi}_L)\right]d\underline{\pi}_L\geq 0.$$

This is equivalent to

$$\frac{\mu_H d\underline{\pi}_H}{\mu_L d\underline{\pi}_L} \geq -\frac{\zeta_L' M_L + \zeta_L M_L'}{\zeta_H' M_H + \zeta_H M_H'}.$$

Thus it is sufficient to show

$$\frac{\zeta_L' M_L + \zeta_L M_L'}{\zeta_H' M_H + \zeta_H M_H'} \geq \frac{\zeta_L' \alpha_L + \zeta_L \alpha_L'}{\zeta_H' \alpha_H + \zeta_H \alpha_H'}. \tag{20}$$

From Lemma 7 we have that both numerators and denominators are negative. Then it is sufficient to show

$$\frac{\zeta_L' M_L + \zeta_L M_L'}{\zeta_L' \alpha_L + \zeta_L \alpha_L'} \geq \frac{\zeta_H' M_H + \zeta_H M_H'}{\zeta_H' \alpha_H + \zeta_H \alpha_H'}.$$

Since $\underline{\pi}_L > \underline{\pi}_H$, for this it is in turn sufficient to show

$$\frac{d}{d\pi}\left[\frac{\zeta' M + \zeta M'}{\zeta' \alpha + \zeta \alpha'}\right] \geq 0.$$

Following the simplifications in Lemma 6, this is

$$\begin{aligned}
\frac{d}{d\pi}\left[\frac{-K(1-\pi)\alpha' M + \zeta^{-1} M'}{-K(1-\pi)\alpha'\alpha + \zeta^{-1}\alpha'}\right] &= \frac{d}{d\pi}\left[\frac{K(1-\pi)M - \pi\zeta^{-1}}{K(1-\pi)\alpha - \zeta^{-1}}\right] \\
&= \frac{d}{d\pi}\left[\frac{K(1-\pi)M - (1+K(\alpha-M))\pi}{K(1-\pi)\alpha - (1+K(\alpha-M))}\right] \\
&= \frac{d}{d\pi}\left[\frac{K(M-\pi\alpha) - \pi}{K(M-\pi\alpha) - 1}\right] \\
&= \frac{d}{d\pi}\left[1 + \frac{1-\pi}{K(M-\pi\alpha) - 1}\right] \\
&\overset{\text{sgn}}{=} -(K(M-\pi\alpha) - 1) - (1-\pi)K\left(M' - \alpha - \pi\alpha'\right) \\
&= 1 - K\left[(M-\pi\alpha) - (1-\pi)\alpha\right] \\
&= 1 - K(M-\alpha).
\end{aligned}$$

This is positive when π is large, establishing the desired result. □

Proof of Theorem 1, part (i). Suppose toward a contradiction that there exists a sequence $C_n \searrow 0$ and associated optimal policies under which the low-wealth type is served. Denote the induced objects along this sequence by a subscript n. By Proposition 1, we can restrict attention to an optimal policy that has $r_n = 0$. Lemma 2 implies

$$\underline{\pi}_{H,n} \to 1, \qquad \underline{\pi}_{L,n} \to 1.$$

Moreover, Lemma 3 implies that the policy elicits at least mass C_n of bookings. In particular, some agents book, and Lemma 1 implies that high-wealth agents book with positive probability. Hence $\underline{\pi}_{H,n} < 1$ and the high-wealth cutoff is interior.

Using $r_n = 0$ in (14), the stationary utility of the high-wealth type satisfies

$$0 \le \widehat{V}_{H,n} = \frac{\beta_n(1-\delta)\left[M(\underline{\pi}_{H,n}) - \frac{p_n}{w_H}\alpha(\underline{\pi}_{H,n})\right]}{(1-\delta) + \beta_n\kappa_n H(\underline{\pi}_{H,n})}$$
$$\le M(\underline{\pi}_{H,n}).$$

Since $\underline{\pi}_{H,n} \to 1$, it follows that $\widehat{V}_{H,n} \to 0$.

The interior cutoff condition (5) now becomes

$$\underline{\pi}_{H,n} = \frac{\kappa_n \widehat{V}_{H,n} + (1-\delta)\frac{p_n}{w_H}}{\kappa_n \widehat{V}_{H,n} + (1-\delta)}.$$

Because κ_n is bounded and $\widehat{V}_{H,n} \to 0$, we obtain $\underline{\pi}_{H,n} - \frac{p_n}{w_H} \to 0$. Together with $\underline{\pi}_{H,n} \to 1$, this implies $p_n \to w_H$. Since $w_H > w_L$, we therefore have $p_n > w_L$ for all sufficiently large n.

But when $p_n > w_L$, even a low-wealth agent with usage probability $\pi = 1$ strictly prefers not to book. Indeed, because such an agent never cancels, the incremental payoff from booking is

$$\widehat{V}^b_{L,n}(1) - \widehat{V}^n_{L,n} = \beta_n(1-\delta)\left(1 - \frac{p_n}{w_L}\right) < 0.$$

Thus no low-wealth agent books, contradicting. It follows that there exists $C_0 > 0$ such that every optimal policy serves only the high-wealth type whenever $C < C_0$.

It remains to show that such a policy cannot use a nontrivial ban. Consider a policy that serves only the high-wealth type and induces cutoff $\underline{\pi}_H$, eligible fraction ζ_H, and booking mass $A = \mu_H \zeta_H \alpha(\underline{\pi}_H)$. If the policy uses a nontrivial ban, then Lemma 6 gives

$$\zeta_H = \left[1 + \beta H(\underline{\pi}_H) \sum_{t=1}^{T} t\rho(t)\right]^{-1} < 1.$$

Choose $\widetilde{\pi}_H$ so that

$$\alpha(\widetilde{\pi}_H) = \zeta_H \alpha(\underline{\pi}_H).$$

Because $\zeta_H < 1$ and α is strictly decreasing, $\widetilde{\pi}_H > \underline{\pi}_H$. Now eliminate the ban and set $\widetilde{p} = w_H \widetilde{\pi}_H$. For sufficiently small C, we have $\widetilde{p} > w_L$, so the new policy continues to serve only the high-wealth type. It also induces the same booking mass:

$$\mu_H \alpha(\widetilde{\pi}_H) = \mu_H \zeta_H \alpha(\underline{\pi}_H) = A.$$

Consequently, the allocation probability β is unchanged. Expected usage under the new price-only

policy is

$$\begin{aligned}\widetilde{U} &= \beta\mu_H M(\widetilde{\pi}_H)\\ &= \beta\mu_H \alpha(\widetilde{\pi}_H) m(\widetilde{\pi}_H)\\ &= \beta\mu_H \zeta_H \alpha(\underline{\pi}_H) m(\widetilde{\pi}_H)\\ &> \beta\mu_H \zeta_H \alpha(\underline{\pi}_H) m(\underline{\pi}_H)\\ &= \beta\mu_H \zeta_H M(\underline{\pi}_H) = U,\end{aligned}$$

where the strict inequality follows from $\widetilde{\pi}_H > \underline{\pi}_H$ and the strict monotonicity of $m(\cdot)$. Hence any policy that serves only the high-wealth type and uses a nontrivial ban is strictly dominated by a price-only policy.

Finally, Lemma 3 rules out underbooking. A high-wealth-only price policy cannot optimally overbook either. If $A > C$, the principal can raise the price until $A = C$, thereby removing the high-wealth booking agents with the lowest usage probabilities and strictly increasing usage conditional on allocation. Therefore the optimal policy in fact satisfies $A = C$ and is implemented by

$$\underline{\pi}_H^\star = F^{-1}\left(1 - \frac{C}{\mu_H}\right), \qquad p^\star = w_H F^{-1}\left(1 - \frac{C}{\mu_H}\right).$$

Thus, for all sufficiently small C, every optimal policy uses no ban and serves only the high-wealth type. □

Proof of Theorem 1, part (ii). Throughout the proof, sums indexed by w are over $w \in \{H, L\}$. Use $B_\rho := \sum_{t=1}^T t\rho(t)$ again for the mean duration of the ban. Moreover,

$$\frac{\kappa_\rho}{1-\delta} = \sum_{t=1}^T \rho(t)\left(\delta + \delta^2 + \cdots + \delta^t\right) < \sum_{t=1}^T t\rho(t) = B_\rho \tag{21}$$

whenever ρ is nontrivial.

First, we characterize optimal outcomes as $C \nearrow 1$. Consider an arbitrary sequence $C_n \nearrow 1$ and an arbitrary optimal policy at each C_n. Denote its cutoffs, eligible fractions, allocation probability, booking mass, and usage by $\underline{\pi}_{w,n}$, $\zeta_{w,n}$, β_n, A_n, and $U_n^\star$, respectively. The policy with zero price and no ban makes every agent eligible and induces every agent to book, and therefore generates usage $C_n M(0)$. Optimality and feasibility consequently give

$$C_n M(0) \le U_n^\star = \beta_n \sum_w \mu_w \zeta_{w,n} M(\underline{\pi}_{w,n}) \le \sum_w \mu_w \zeta_{w,n} M(\underline{\pi}_{w,n}) \le M(0).$$

It follows that $U_n^\star \to M(0)$ and $\sum_w \mu_w \zeta_{w,n} M(\underline{\pi}_{w,n}) \to M(0)$. Because every μ_w is strictly positive and every term $M(0) - \zeta_{w,n} M(\underline{\pi}_{w,n})$ is nonnegative, we must have that

$$\underline{\pi}_{w,n} \to 0 \qquad \text{and} \qquad \zeta_{w,n} \to 1$$

for both wealth types. In particular, $A_n = \sum_w \mu_w \zeta_{w,n} \alpha(\underline{\pi}_{w,n}) \to 1$. Since $A_n \leq 1$, we also have $\beta_n = \min\{1, C_n/A_n\} \geq C_n$, and hence $\beta_n \to 1$. Further, from Lemma 6 we can write

$$\zeta_{w,n}^{-1} - 1 = \beta_n B_{\rho_n} H(\underline{\pi}_{w,n}).$$

The left-hand side converges to zero, whereas $\beta_n H(\underline{\pi}_{w,n}) \to H(0) > 0$. Therefore $B_{\rho_n} \to 0$.

We now rule out nontrivial bans for all sufficiently large capacities. Suppose otherwise; then there is a sequence $C_n \nearrow 1$ and an optimal policy at each C_n with $B_{\rho_n} > 0$. By Proposition 1, each such policy has an optimal no-refund counterpart $(p_n, 0; \rho_n)$ with the same punishment distribution. It is therefore enough to obtain a contradiction for these no-refund representatives.

With $r = 0$, the cutoff condition becomes

$$\underline{\pi}_w - \frac{p}{w} = \beta \frac{\kappa_\rho}{1-\delta} \left(1 - \frac{p}{w}\right) G(\underline{\pi}_w). \tag{22}$$

Equation (21) and that $B_{\rho_n} \to 0$ imply $\kappa_{\rho_n}/(1-\delta) \to 0$. Rearranging (22) yields

$$\frac{p_n}{w} = \frac{\underline{\pi}_{w,n} - \beta_n \frac{\kappa_{\rho_n}}{1-\delta} G(\underline{\pi}_{w,n})}{1 - \beta_n \frac{\kappa_{\rho_n}}{1-\delta} G(\underline{\pi}_{w,n})},$$

so $p_n \to 0$.

Suppress the subscript n and write $N := \sum_w \mu_w \zeta_w M(\underline{\pi}_w)$, so that realized usage is $U = \beta N$. For sufficiently large n, we have $p < w_L$, and (22) implies $\underline{\pi}_w \geq p/w$ for both wealth types.

Consider removing the ban while initially leaving the price unchanged. The resulting no-ban policy has booking and potential-usage masses

$$A^0(p) := \sum_w \mu_w \alpha\left(\frac{p}{w}\right), \qquad N^0(p) := \sum_w \mu_w M\left(\frac{p}{w}\right).$$

Because $B_\rho > 0$ and both types are active for sufficiently large n, we have $\zeta_w < 1$. Together with $\underline{\pi}_w \geq p/w$, this gives $A^0(p) > A$. Since A^0 is continuous and strictly decreasing near zero, there is a unique $\widetilde{p} > p$ such that

$$A^0(\widetilde{p}) = A. \tag{23}$$

Moreover, $A \to 1 = A^0(0)$ implies $\widetilde{p} \to 0$. The original policy and the new no-ban policy therefore have the same booking mass and hence the same allocation probability.

The potential-usage gain from removing the ban while holding the price at p satisfies

$$\begin{aligned} N^0(p) - N &= \sum_w \mu_w \left[M\left(\frac{p}{w}\right) - M(\underline{\pi}_w) + (1-\zeta_w) M(\underline{\pi}_w)\right] \\ &\geq \sum_w \mu_w (1-\zeta_w) M(\underline{\pi}_w), \end{aligned} \tag{24}$$

where the inequality follows because M is decreasing. By Lemma 6,

$$\frac{(1-\zeta_w)M(\underline{\pi}_w)}{B_\rho} = \frac{\beta H(\underline{\pi}_w)M(\underline{\pi}_w)}{1+\beta B_\rho H(\underline{\pi}_w)}.$$

Because $\beta \to 1$, $\underline{\pi}_w \to 0$, and $B_\rho \to 0$, the right-hand side converges to $H(0)M(0) > 0$. Hence, for both wealth types and all sufficiently large n,

$$(1-\zeta_w)M(\underline{\pi}_w) \geq \frac{H(0)M(0)}{2} B_\rho.$$

It follows from (24) that

$$N^0(p) - N \geq cB_\rho, \qquad c := \frac{H(0)M(0)}{2} > 0.$$

We next bound the additional booking mass created by removing the ban. Let $\Delta A := A^0(p) - A$. Since f' is bounded, f is bounded on $(0,1)$; write $\overline{f} := \sup_{\pi\in(0,1)} f(\pi) < \infty$. Equation (22) and (21) imply

$$0 \leq \underline{\pi}_w - \frac{p}{w} \leq B_\rho.$$

It follows that $\alpha(p/w) - \alpha(\underline{\pi}_w) \leq \overline{f} B_\rho$. Furthermore, $1 - \zeta_w \leq \beta B_\rho H(\underline{\pi}_w) \leq B_\rho$. Therefore,

$$\begin{aligned} 0 < \Delta A &= \sum_w \mu_w \left[\alpha\left(\frac{p}{w}\right) - \alpha(\underline{\pi}_w) + (1-\zeta_w)\alpha(\underline{\pi}_w)\right] \\ &\leq (\overline{f}+1)B_\rho. \end{aligned} \tag{25}$$

Write $K := \overline{f} + 1$.

Finally, consider the potential usage lost by raising the no-ban price from p to $\widetilde{p}$. For each wealth type,

$$M\left(\frac{p}{w}\right) - M\left(\frac{\widetilde{p}}{w}\right) = \int_{p/w}^{\widetilde{p}/w} x\, dF(x) \leq \frac{\widetilde{p}}{w_L}\left[\alpha\left(\frac{p}{w}\right) - \alpha\left(\frac{\widetilde{p}}{w}\right)\right].$$

Summing across wealth types and using (23) and (25) gives

$$\begin{aligned} N^0(p) - N^0(\widetilde{p}) &\leq \frac{\widetilde{p}}{w_L}\left[A^0(p) - A^0(\widetilde{p})\right] \\ &= \frac{\widetilde{p}}{w_L}\Delta A \leq \frac{K\widetilde{p}}{w_L} B_\rho. \end{aligned} \tag{26}$$

Because $\widetilde{p} \to 0$, we have $K\widetilde{p}/w_L < c/2$ for all sufficiently large n. Then

$$N^0(\widetilde{p}) - N \geq \left(c - \frac{K\widetilde{p}}{w_L}\right) B_\rho > \frac{c}{2} B_\rho > 0.$$

The allocation probability is unchanged, so the new policy strictly increases realized usage. This contradicts optimality. Because the no-refund replacement preserves ρ, the argument rules out a

nontrivial ban in every optimal policy for all sufficiently large C.

We next show that every optimal policy serves both wealth types. If one type is not served, the contribution of the remaining type to potential usage is at most $\max\{\mu_H, \mu_L\}M(0)$. The zero-price, no-ban policy instead generates $CM(0)$. Hence every optimal policy serves both types whenever $C > \max\{\mu_H, \mu_L\}$.

It remains to establish exact filling. At this point, we can use the case analysis in the proof of Proposition 1. Because both wealth types are active, no policy with $0 < r < p$ can be optimal. Thus, if an optimal policy uses a positive refund, it must satisfy $r = p$.

Suppose $r = p > 0$. Because the policy has no ban, its cutoff condition reduces to $\underline{\pi}_w = 0$ for both wealth types. Thus every agent books, so $A = 1$, $\beta = C$, and usage is $CM(0)$. This outcome is strictly suboptimal. Indeed, choose a sufficiently small no-refund price $q > 0$ such that

$$A^0(q) := \sum_w \mu_w \alpha\left(\frac{q}{w}\right) > C.$$

The resulting price-only policy overbooks and generates usage

$$U(q) = C\frac{N^0(q)}{A^0(q)}, \qquad N^0(q) := \sum_w \mu_w M\left(\frac{q}{w}\right).$$

Because $m(\cdot)$ is strictly increasing and $q/w > 0$ for both wealth types,

$$\begin{aligned} N^0(q) &= \sum_w \mu_w \alpha\left(\frac{q}{w}\right) m\left(\frac{q}{w}\right) \\ &> m(0) \sum_w \mu_w \alpha\left(\frac{q}{w}\right) = M(0)A^0(q). \end{aligned}$$

Consequently, $U(q) > CM(0)$, contradicting the optimality of the full-refund policy. Therefore every optimal policy at sufficiently large capacity has $r = 0$.

With no refund and no ban, the cutoffs are $\underline{\pi}_w = p/w$. Using the notation from before, because $A^0(0) = 1$ and $A^0(w_L) \le \mu_H < C$ for sufficiently large C, there is a unique $p_C \in (0, w_L)$ satisfying $A^0(p_C) = C$. Moreover, $p_C \to 0$ as $C \nearrow 1$.

An optimal policy cannot underbook. Indeed, underbooking means $A^0(p) < C = A^0(p_C)$ and hence $p > p_C$. Lowering the price from p to p_C strictly increases N^0, while both policies have allocation probability one.

Now suppose that an optimal policy overbooks. Then $p < p_C$, and its usage is

$$U(p) = C\frac{N^0(p)}{A^0(p)}.$$

For $q \in (0, p_C]$, direct differentiation gives

$$\frac{d}{dq}\left[\frac{N^0(q)}{A^0(q)}\right] = \frac{1}{A^0(q)^2}\sum_w \mu_w \frac{f(q/w)}{w}\left[N^0(q) - \frac{q}{w}A^0(q)\right]. \tag{27}$$

Choose $\varepsilon > 0$ sufficiently small that $M(\varepsilon) > \varepsilon$. Because $p_C \to 0$, for all sufficiently large C we have $p_C/w_L < \varepsilon$. Hence, for every $q \in [0, p_C]$,

$$N^0(q) \geq M(\varepsilon) \qquad \text{and} \qquad \frac{q}{w}A^0(q) \leq \varepsilon.$$

Every expression in brackets in (27) is therefore strictly positive. Thus $N^0(q)/A^0(q)$ is strictly increasing on $[0, p_C]$. Raising the price from p to p_C consequently strictly increases usage, contradicting optimality.

Neither underbooking nor overbooking can therefore be optimal. Every optimal policy at sufficiently large capacity satisfies $A^0(p) = C$ and $\beta = 1$. Taking $\bar{C} < 1$ larger than the finitely many thresholds identified above proves that whenever $C > \bar{C}$, every optimal outcome uses no ban, serves both wealth types, and exactly fills capacity. □

D.2 Proof of Theorem 2

Proof. Consider the optimal price-only policy with some price p^P; we will denote objects included in and induced by this policy by superscript P. Whenever an optimal price overbooks, it must serve both wealth types. The thresholds, booking mass, and allocation probability then satisfy

$$\underline{\pi}_H^P = \frac{p^P}{w_H}, \qquad \underline{\pi}_L^P = \frac{p^P}{w_L} = \xi\underline{\pi}_H^P,$$

$$A^P = \frac{\gamma}{1+\gamma}\alpha(\underline{\pi}_H^P) + \frac{1}{1+\gamma}\alpha(\underline{\pi}_L^P), \qquad \beta^P = \frac{C}{A^P} < 1.$$

Note from Lemma 1 we have $0 < \underline{\pi}_H^P < \underline{\pi}_L^P < 1$ and because there are no bans, $\zeta_H^P = \zeta_L^P = 1$. Under the price-only policy, the continuation value evaluates to

$$\widehat{V}_w^P = \beta^P\left[M(\underline{\pi}_w^P) - \underline{\pi}_w^P\alpha(\underline{\pi}_w^P)\right]. \tag{28}$$

Now, consider a small probability of a one-period ban: $\rho^\varepsilon(0) = 1 - \varepsilon$ and $\rho^\varepsilon(1) = \varepsilon$. Continue to set $r = 0$ and adjust the price to p^ε, with $p^0 = p^P$, so that the induced booking mass remains equal to A^P. Such a price adjustment exists locally.[20] Using superscript ε to denote the objects induced by this new policy, write

$$A(p^\varepsilon, \rho^\varepsilon) = A^\varepsilon = A^P$$

where the left-hand side $A(\cdot,\cdot)$ returns the booking mass; the price p^ε is that for which the booking

[20] This follows from the implicit function theorem; booking mass is locally strictly decreasing in price at the price-only policy.

mass, when paired with the small ban probability, equals that under the optimal price. Because the booking mass is held fixed, the allocation probability also remains fixed at $\beta^P = C/A^P$.

The price-only policy had continuation evaluated as in (28) because $\delta_{\rho^P} = \delta$. Under the ban we introduced, the effective discount rate becomes $\delta_{\rho^\varepsilon} = (1-\varepsilon)\delta + \varepsilon\delta^2$. We construct the compensating price change that keeps the total booking mass fixed, and differentiate usage along that fixed booking path. In the following, we use two derivatives which arise from the chain rule:

$$\left.\frac{d\zeta_w^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} = -\beta^P \left[\alpha(\underline{\pi}_w^P) - M(\underline{\pi}_w^P)\right],$$
$$\left.\frac{d\underline{\pi}_w^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} = \frac{1}{w}\left.\frac{dp^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} + \delta\beta^P(1-\underline{\pi}_w^P)\left[M(\underline{\pi}_w^P) - \underline{\pi}_w^P\alpha(\underline{\pi}_w^P)\right].$$

The steady-state mass equation follows from Lemma 6. Now we impose $A^\varepsilon = A^P$ and differentiate that constraint to find the compensating price:

$$\begin{aligned}
0 = \left.\frac{dA^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} &= \sum_{w\in\{H,L\}} \mu_w \left[\left.\frac{d\zeta_w^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} \alpha(\underline{\pi}_w^P) - f(\underline{\pi}_w^P)\left.\frac{d\underline{\pi}_w^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0}\right] \\
&= \sum_{w\in\{H,L\}} \mu_w \Bigg\{ -\beta^P\left[\alpha(\underline{\pi}_w^P) - M(\underline{\pi}_w^P)\right]\alpha(\underline{\pi}_w^P) \\
&\qquad - f(\underline{\pi}_w^P)\left[\frac{1}{w}\left.\frac{dp^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} + \delta\beta^P(1-\underline{\pi}_w^P)\left[M(\underline{\pi}_w^P) - \underline{\pi}_w^P\alpha(\underline{\pi}_w^P)\right]\right]\Bigg\} \\
&= -\beta^P \sum_{w\in\{H,L\}} \mu_w \Bigg\{ \left[\alpha(\underline{\pi}_w^P) - M(\underline{\pi}_w^P)\right]\alpha(\underline{\pi}_w^P) \\
&\qquad + \delta f(\underline{\pi}_w^P)(1-\underline{\pi}_w^P)\left[M(\underline{\pi}_w^P) - \underline{\pi}_w^P\alpha(\underline{\pi}_w^P)\right]\Bigg\} \\
&\quad - \left[\sum_{w\in\{H,L\}} \mu_w \frac{f(\underline{\pi}_w^P)}{w}\right]\left.\frac{dp^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0}.
\end{aligned} \tag{29}$$

This equation determines the necessary change in price. Note that every direct effect of the ban reduces mass: some agents are made ineligible, and it raises the cost of cancellation, which in turn increases the cutoff. So maintaining $A^\varepsilon = A^P$ requires the price to decrease. Indeed, we can rearrange (29) for the necessary compensation:

$$\left.\frac{dp^\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} = -\beta^P \frac{\sum_{w\in\{H,L\}} \mu_w \left\{\left[\alpha(\underline{\pi}_w^P) - M(\underline{\pi}_w^P)\right]\alpha(\underline{\pi}_w^P) + \delta f(\underline{\pi}_w^P)(1-\underline{\pi}_w^P)\left[M(\underline{\pi}_w^P) - \underline{\pi}_w^P\alpha(\underline{\pi}_w^P)\right]\right\}}{\sum_{w\in\{H,L\}} \mu_w \frac{f(\underline{\pi}_w^P)}{w}}.$$

We now consider the principal's objective. Because $A^\varepsilon = A^P$ we have $\beta^\varepsilon = \beta^P$, and expected

usage under the perturbation is

$$U^{\varepsilon} = \beta^P \sum_{w\in\{H,L\}} \mu_w \zeta_w^{\varepsilon} M(\underline{\pi}_w^{\varepsilon}).$$

The derivative evaluates to

$$\begin{aligned}\left.\frac{dU^{\varepsilon}}{d\varepsilon}\right|_{\varepsilon=0} = & -(\beta^P)^2 \sum_{w\in\{H,L\}} \mu_w \Bigg\{ \left[\alpha(\underline{\pi}_w^P) - M(\underline{\pi}_w^P)\right] M(\underline{\pi}_w^P) \\ & \qquad + \delta \underline{\pi}_w^P f(\underline{\pi}_w^P)(1-\underline{\pi}_w^P)\left[M(\underline{\pi}_w^P) - \underline{\pi}_w^P \alpha(\underline{\pi}_w^P)\right] \Bigg\} \\ & - \beta^P \left.\frac{dp^{\varepsilon}}{d\varepsilon}\right|_{\varepsilon=0} \sum_{w\in\{H,L\}} \mu_w \frac{\underline{\pi}_w^P f(\underline{\pi}_w^P)}{w}.\end{aligned}$$

There are two simultaneous effects: the first large sum is the direct effect of the ban, and the last line is the effect of the compensating price reduction. Observe that the price derivative enters the usage equation multiplied by $\sum_{w\in\{H,L\}} \mu_w \frac{\underline{\pi}_w^P f(\underline{\pi}_w^P)}{w}$, whereas it enters the booking equation multiplied by $\sum_{w\in\{H,L\}} \mu_w \frac{f(\underline{\pi}_w^P)}{w}$. Define their ratio as

$$\overline{\pi}^P := \frac{\displaystyle\sum_{w\in\{H,L\}} \mu_w \frac{\underline{\pi}_w^P f(\underline{\pi}_w^P)}{w}}{\displaystyle\sum_{w\in\{H,L\}} \mu_w \frac{f(\underline{\pi}_w^P)}{w}} = \frac{(\gamma/\xi)\underline{\pi}_H^P f(\underline{\pi}_H^P) + \underline{\pi}_L^P f(\underline{\pi}_L^P)}{(\gamma/\xi) f(\underline{\pi}_H^P) + f(\underline{\pi}_L^P)}.$$

Thus,

$$\sum_{w\in\{H,L\}} \mu_w \frac{\underline{\pi}_w^P f(\underline{\pi}_w^P)}{w} = \overline{\pi}^P \sum_{w\in\{H,L\}} \mu_w \frac{f(\underline{\pi}_w^P)}{w}.$$

This will help us eliminate the compensating price change from the derivative of usage. We can now write

$$\begin{aligned}\left.\frac{dU^{\varepsilon}}{d\varepsilon}\right|_{\varepsilon=0} = \frac{(\beta^P)^2}{1+\gamma} \Bigg\{ & \gamma\left[\alpha(\underline{\pi}_H^P) - M(\underline{\pi}_H^P)\right]\alpha(\underline{\pi}_H^P)\left[\overline{\pi}^P - m(\underline{\pi}_H^P)\right] \\ & + \left[\alpha(\underline{\pi}_L^P) - M(\underline{\pi}_L^P)\right]\alpha(\underline{\pi}_L^P)\left[\overline{\pi}^P - m(\underline{\pi}_L^P)\right] \\ & + \delta\gamma f(\underline{\pi}_H^P)(1-\underline{\pi}_H^P)\left[M(\underline{\pi}_H^P) - \underline{\pi}_H^P\alpha(\underline{\pi}_H^P)\right]\left[\overline{\pi}^P - \underline{\pi}_H^P\right] \\ & + \delta f(\underline{\pi}_L^P)(1-\underline{\pi}_L^P)\left[M(\underline{\pi}_L^P) - \underline{\pi}_L^P\alpha(\underline{\pi}_L^P)\right]\left[\overline{\pi}^P - \underline{\pi}_L^P\right] \Bigg\}. \qquad (30)\end{aligned}$$

It remains to sign the expression in braces. Since the price-only policy strictly overbooks, its objective is

$$U^P(p) = C\frac{\gamma M(p/w_H) + M(p/w_L)}{\gamma\alpha(p/w_H) + \alpha(p/w_L)}.$$

Because both types participate, p^P is an interior optimum. Its first-order condition implies

$$\begin{aligned}\overline{\pi}^P &= \frac{\gamma M(\underline{\pi}_H^P) + M(\underline{\pi}_L^P)}{\gamma\alpha(\underline{\pi}_H^P) + \alpha(\underline{\pi}_L^P)} \\ &= \frac{\gamma\alpha(\underline{\pi}_H^P)m(\underline{\pi}_H^P) + \alpha(\underline{\pi}_L^P)m(\underline{\pi}_L^P)}{\gamma\alpha(\underline{\pi}_H^P) + \alpha(\underline{\pi}_L^P)}.\end{aligned} \tag{31}$$

Using (31), the first two terms in braces in (30) equal

$$\begin{aligned}&\frac{\gamma\alpha(\underline{\pi}_H^P)\alpha(\underline{\pi}_L^P)}{\gamma\alpha(\underline{\pi}_H^P) + \alpha(\underline{\pi}_L^P)}\left[m(\underline{\pi}_L^P) - m(\underline{\pi}_H^P)\right] \\ &\qquad\times\left[\alpha(\underline{\pi}_H^P) - M(\underline{\pi}_H^P) - \alpha(\underline{\pi}_L^P) + M(\underline{\pi}_L^P)\right].\end{aligned} \tag{32}$$

This expression is strictly positive. Indeed, $m(\cdot)$ is strictly increasing, whereas

$$\alpha(\underline{\pi}) - M(\underline{\pi}) = \int_{\underline{\pi}}^1 (1-\pi)\,dF(\pi)$$

is strictly decreasing. Both differences in (32) are therefore positive because $\underline{\pi}_H^P < \underline{\pi}_L^P$.

The final two terms in braces in (30), after removing the positive factor δ, equal

$$\begin{aligned}&\frac{\gamma f(\underline{\pi}_H^P)f(\underline{\pi}_L^P)(\underline{\pi}_L^P - \underline{\pi}_H^P)}{(\gamma/\xi)f(\underline{\pi}_H^P) + f(\underline{\pi}_L^P)} \\ &\quad\times\left\{(1-\underline{\pi}_H^P)\left[M(\underline{\pi}_H^P) - \underline{\pi}_H^P\alpha(\underline{\pi}_H^P)\right] - \frac{1}{\xi}(1-\underline{\pi}_L^P)\left[M(\underline{\pi}_L^P) - \underline{\pi}_L^P\alpha(\underline{\pi}_L^P)\right]\right\}.\end{aligned} \tag{33}$$

This expression is also strictly positive; $M(\underline{\pi}) - \underline{\pi}\alpha(\underline{\pi}) = \int_{\underline{\pi}}^1(\pi - \underline{\pi})\,dF(\pi)$ is strictly decreasing in $\underline{\pi}$. And $1 - \underline{\pi}_H^P > 1 - \underline{\pi}_L^P > 0$, so it follows that the expression inside braces in (33) is strictly positive.

Therefore the first two terms in braces in (30) have a strictly positive sum, as do the final two terms whenever $\delta > 0$; when $\delta = 0$, the first sum remains strictly positive. Hence for every $\delta \in [0,1)$, we have that $\frac{dU^\varepsilon}{d\varepsilon}\big|_{\varepsilon=0} > 0$. Taking the preceding claims together, for every sufficiently small $\varepsilon > 0$, the perturbed policy continues to induce booking mass $A^P > C$, uses a nontrivial one-period ban, and generates strictly more usage than the optimal price-only policy. Thus bans are useful. □

D.3 Proof of Theorem 3

Proof. Let $(0, 0; \rho)$ assign positive probability to a positive ban length and induce a stationary equilibrium with $\underline{\pi}_w < 1$ for each wealth type. We will show that there exists a sequence of policies $(p^\varepsilon, 0; \rho^\varepsilon)$ such that $(p^\varepsilon, 0; \rho^\varepsilon) \to (0, 0; \rho)$ for $p^\varepsilon > 0$, and under price p^ε, where

$$U(p^\varepsilon, 0; \rho^\varepsilon) > U(0, 0; \rho)$$

for every sufficiently small $\varepsilon > 0$.

Let $B_\rho = \sum_{t=1}^{T} t\rho(t)$. Because $p = r = 0$ wealth does not enter the booking decision. Hence the pure-ban policy induces a common cutoff and the eligible fraction is common across both wealth types:

$$\underline{\pi}_H = \underline{\pi}_L = \underline{\pi}, \qquad \zeta_H = \zeta_L = \zeta = [1 + \beta B_\rho H(\underline{\pi})]^{-1}$$

where the eligible fraction is derived in Lemma 6. Recall the cutoff condition which we derived in Lemma 5; in the pure-ban policy, $p = 0$ so

$$(1 - \delta)\underline{\pi} = \beta\kappa G(\underline{\pi}). \tag{34}$$

Now consider that for any patience level $\delta < 1$, the mean duration of a nontrivial ban strictly exceeds its discounted incentive effect:

$$\begin{aligned} \frac{\kappa}{1-\delta} &= \sum_{t=1}^{T} \rho(t) \left(\delta + \delta^2 + \cdots + \delta^t\right) \\ &< \sum_{t=1}^{T} t\rho(t) = B_\rho. \end{aligned} \tag{35}$$

We noted in Remark 1 that any ban-only policy can underbook or overbook. So we have to check each regime separately and determine whether a positive price improves on the principal's objective. Observe the derivatives of booking mass and usage with respect to the threshold:

$$\frac{d}{d\underline{\pi}} \left[\zeta\alpha(\underline{\pi})\right] = \zeta^2 f(\underline{\pi}) \left[\beta B_\rho G(\underline{\pi}) - 1\right], \tag{36}$$

$$\begin{aligned} \frac{d}{d\underline{\pi}} \left[\zeta M(\underline{\pi})\right] &= \zeta^2 f(\underline{\pi}) \left[\beta B_\rho G(\underline{\pi}) - \underline{\pi}\right] \\ &= \beta\zeta^2 f(\underline{\pi}) G(\underline{\pi}) \left[B_\rho - \frac{\kappa}{1-\delta}\right] > 0. \end{aligned} \tag{37}$$

The second equality in (37) follows from the pure-ban cutoff condition (34).

By introducing a positive price, we no longer can treat each threshold as common across the wealth types. If β and ρ are held fixed, differentiating the general cutoff condition with respect to the price at $p = 0$ gives

$$\left[1 + \frac{\beta\kappa}{1-\delta}\alpha(\underline{\pi})\right] d\underline{\pi}_w = \frac{1 - \underline{\pi}}{w}\, dp.$$

It will be helpful to define the population-weighted average change in the cutoffs:

$$d\underline{\pi}^\mu = \frac{\gamma}{1+\gamma} d\underline{\pi}_H + \frac{1}{1+\gamma} d\underline{\pi}_L.$$

Then we can combine the preceding equations to determine the average change as a function of the

ban and change in price:

$$d\underline{\pi}^{\mu} = \frac{1-\underline{\pi}}{1+\frac{\beta\kappa}{1-\delta}\alpha(\underline{\pi})}\frac{1+\xi}{1+\gamma}\left(1+\frac{\gamma}{\xi}\right)dp > 0. \tag{38}$$

We now consider the three booking regimes.

Underbooking. Suppose $A < C$. A sufficiently small price leaves the policy underbooked, so $\beta = 1$ remains fixed. Keeping ρ unchanged, (37)–(38) give

$$dU = \zeta^2 f(\underline{\pi})\left[B_\rho G(\underline{\pi}) - \underline{\pi}\right] d\underline{\pi}^{\mu} > 0.$$

Exact booking. Suppose $A = C$, so $\beta = 1$. First, if $B_\rho G(\underline{\pi}) < 1$, then (36) shows that a small price moves the policy into underbooking, and the preceding argument applies.

Second, if $B_\rho G(\underline{\pi}) \geq 1$, then the price perturbation moves weakly toward overbooking. Let

$$N = \mu_H \zeta_H M(\underline{\pi}_H) + \mu_L \zeta_L M(\underline{\pi}_L).$$

At the initial symmetric outcome, $A = \zeta\alpha(\underline{\pi})$ and $N = \zeta M(\underline{\pi})$, and every local perturbation satisfies

$$dN - m(\underline{\pi})dA = \zeta f(\underline{\pi})\left[m(\underline{\pi}) - \underline{\pi}\right] d\underline{\pi}^{\mu}. \tag{39}$$

Accounting for the probability of rationing on the overbooking side gives

$$\left[1 + \frac{\beta\kappa}{1-\delta}\alpha(\underline{\pi}) + \underline{\pi}\frac{f(\underline{\pi})}{\alpha(\underline{\pi})}\left(\beta B_\rho G(\underline{\pi}) - 1\right)\right] d\underline{\pi}^{\mu} = (1-\underline{\pi})\frac{1+\xi}{1+\gamma}\left(1+\frac{\gamma}{\xi}\right)dp. \tag{40}$$

At $\beta = 1$ the coefficient on the left is strictly positive. Thus $d\underline{\pi}^{\mu} > 0$. Since $U = CN/A$ on the overbooking side, (39) implies

$$dU = \beta\zeta f(\underline{\pi})\left[m(\underline{\pi}) - \underline{\pi}\right] d\underline{\pi}^{\mu} > 0.$$

Overbooking. Suppose $A > C$, so $\beta = C/A < 1$. If $\beta B_\rho G(\underline{\pi}) \geq 1$, equation (40) again has a strictly positive coefficient. Therefore a small price, holding ρ fixed, raises $d\underline{\pi}^{\mu}$ and hence raises usage.

It remains to consider $\beta B_\rho G(\underline{\pi}) < 1$. Here, raising the price while holding ρ fixed need not improve usage. We instead introduce a price and simultaneously (slightly) weaken the ban. For s near 1, define

$$\rho^s(t) = s\rho(t) \quad \text{for } t \geq 1, \qquad \rho^s(0) = 1 - s\big(1-\rho(0)\big).$$

Then $B_{\rho^s} = sB_\rho$ and $\kappa_{\rho^s} = s\kappa$. Choose $s = s(p)$ so that A remains fixed. Since A is fixed, β is fixed as well. Differentiating the booking constraint gives

$$ds = \frac{f(\underline{\pi})\left[\beta B_\rho G(\underline{\pi}) - 1\right]}{\beta\alpha(\underline{\pi})H(\underline{\pi})B_\rho} d\underline{\pi}^{\mu}. \tag{41}$$

Combining this expression with the cutoff conditions gives

$$\left[1+\frac{\beta\kappa}{1-\delta}\alpha(\underline{\pi})-\frac{\underline{\pi}f(\underline{\pi})\left[\beta B_\rho G(\underline{\pi})-1\right]}{\beta\alpha(\underline{\pi})H(\underline{\pi})B_\rho}\right]d\underline{\pi}^\mu=(1-\underline{\pi})\frac{1+\xi}{1+\gamma}\left(1+\frac{\gamma}{\xi}\right)dp. \tag{42}$$

Because $\beta B_\rho G(\underline{\pi})-1<0$, the coefficient on the left is strictly positive. Hence $dp>0$ implies $d\underline{\pi}^\mu>0$. Equation (41) then gives $ds<0$, so the required adjustment indeed weakens the ban. And because A and β are fixed, (39) yields

$$dU=\beta\zeta f(\underline{\pi})\left[m(\underline{\pi})-\underline{\pi}\right]d\underline{\pi}^\mu>0.$$

Thus every nontrivial pure-ban policy with an interior cutoff is locally dominated by a policy with a strictly positive price and zero refund. □

Proof of Proposition 4. For the moment, ignore the agent's wealth type w. Let $\delta^\star$ be the agent's expected discount rate conditional on punishment, $\delta^\star=\sum_{t=0}^{T}\delta^{t+1}\rho(t)$. Recall that the agent's value from booking (b) or not (n) is[21]

$$\begin{aligned}V^b(\pi)&=(1-\pi)\,\delta^\star\hat{V}+\pi\left((1-\delta)+\delta\hat{V}\right),\\ V^n(\pi)&=\delta\hat{V}.\end{aligned}$$

As derived in (5) the indifferent type $\underline{\pi}$ is such that

$$V^b(\underline{\pi})=V^n(\underline{\pi})\quad\Longrightarrow\quad\underline{\pi}=\frac{(\delta-\delta^\star)\hat{V}}{(1-\delta)+(\delta-\delta^\star)\hat{V}}=\frac{\frac{\delta-\delta^\star}{1-\delta}\hat{V}}{1+\frac{\delta-\delta^\star}{1-\delta}\hat{V}}.$$

Ex ante, before the cancellation type π is realized, the agent's value is

$$\begin{aligned}\hat{V}&=\int_0^{\underline{\pi}}\delta\hat{V}dF(\pi)+\int_{\underline{\pi}}^{1}(1-\pi)\,\delta^\star\hat{V}+\pi\left((1-\delta)+\delta\hat{V}\right)dF(\pi)\\ &=(1-\alpha(\underline{\pi}))\,\delta\hat{V}+\alpha(\underline{\pi})\,\delta^\star\hat{V}+\left((1-\delta)+(\delta-\delta^\star)\hat{V}\right)M(\underline{\pi}).\end{aligned}$$

This gives

$$\left((1-\delta)+(\delta-\delta^\star)\left(\alpha(\underline{\pi})-M(\underline{\pi})\right)\right)\hat{V}=(1-\delta)\,M(\underline{\pi})\quad\Longrightarrow\quad\hat{V}=\frac{M(\underline{\pi})}{1+\frac{\delta-\delta^\star}{1-\delta}\left(\alpha(\underline{\pi})-M(\underline{\pi})\right)}.$$

It suffices now to show that when δ is large, there is some access-based punishment that improves the principal's utility over that obtained from a price-based mechanism. To see this, assume that there is $\eta\geq 0$ so that $\delta^\star=\delta^{\eta+1}$.[22] Then $\lim_{\delta\nearrow 1}(\delta-\delta^\star)/(1-\delta)=\eta$.

[21] Punishment occurs with probability $1-\pi$ because the construction of $\delta^\star$ incorporates the possibility of degenerate zero-period bans.

[22] This is a matter of letting T grow with δ, and assigning $\rho(t)$ as appropriate.

Because all involved terms are bounded, there is a subsequence of δ so that $M(\underline{\pi})$ and $\alpha(\underline{\pi})$ both converge. If $\lim_{\delta \nearrow 1} M(\underline{\pi}) = 0$, then $\lim_{\delta \nearrow 1} \hat{V} = 0$; this in turn implies that $\lim_{\delta \nearrow 1} \underline{\pi} = 0$ and hence $\lim_{\delta \nearrow 1} M(\underline{\pi}) > 0$, a contradiction. Then it must be that $\lim_{\delta \nearrow 1} M(\underline{\pi}) := M^\star(\underline{\pi}) > 0$ and, since $\alpha(\cdot) \geq M(\cdot)$, $\lim_{\delta \nearrow 1} \alpha(\underline{\pi}) := \alpha^\star(\underline{\pi}) > 0$.

We thus have, in the limit,

$$\hat{V} = \frac{M^\star(\underline{\pi})}{1 + \eta\left(\alpha^\star(\underline{\pi}) - M^\star(\underline{\pi})\right)}, \text{ and } \underline{\pi} = \frac{\eta \hat{V}}{1 + \eta \hat{V}}.$$

Together, these give

$$\underline{\pi}^\star = \frac{\eta M^\star(\underline{\pi}^\star)}{1 + \eta \alpha^\star(\underline{\pi}^\star)} \quad \Longrightarrow \quad \eta = \frac{\underline{\pi}}{M^\star(\underline{\pi}^\star) - \underline{\pi}^\star \alpha^\star(\underline{\pi})}. \tag{43}$$

It follows that for any threshold $\underline{\pi}^\star$ there is a η that achieves this threshold in the limit.

As a final step, we show that marginal increases of C from 0 result in higher realized value for the principal when they implement access-based punishments rather than prices. Note first that when C is small, price-based mechanisms serve only high-wealth agents. Then the optimal price-based wealth threshold $\underline{\pi}_H$ satisfies $1 - F(\underline{\pi}_H) = C/\mu_H$, and $\frac{d}{dC}\underline{\pi}_H = -[\mu_H dF(\underline{\pi}_H)]^{-1}$.[23]

Analyzing $\frac{d}{dC}\underline{\pi}^\star$ is more involved. For sake of approximation, assume that $\delta^\star = \delta^{\eta+1}$ represents η periods of access restriction, whether or not η is an integer; Proposition 3 establishes that this approximation is reasonable. Applying Lemma 6 gives

$$\zeta_w^\star(0) = \left[1 + \left(\alpha(\underline{\pi}) - M(\underline{\pi})\right)\eta\right]^{-1}.$$

Then the optimal threshold under access-based punishment satisfies

$$\begin{aligned} 1 - F(\underline{\pi}) = C/\zeta_w^\star(0) &= \left[1 + \left(\alpha(\underline{\pi}) - M(\underline{\pi})\right)\eta\right] C \\ &= \left[1 + \frac{\left(\alpha^\star(\underline{\pi}) - M^\star(\underline{\pi})\right)\pi}{M^\star(\underline{\pi}) - \underline{\pi}\alpha^\star(\underline{\pi})}\right] C = \underbrace{\left[\frac{\left(1 - \underline{\pi}^\star\right) M^\star(\underline{\pi}^\star)}{M^\star(\underline{\pi}^\star) - \underline{\pi}^\star \alpha^\star(\underline{\pi}^\star)}\right]}_{h(\underline{\pi}^\star)} C. \end{aligned}$$

We thus have

$$\frac{1 - F(\underline{\pi})}{1 - \underline{\pi}} = \frac{M(\underline{\pi})\, C}{M(\underline{\pi}) - \pi\alpha(\underline{\pi})}.$$

Note that the limit of the left-hand side, as $\underline{\pi} \nearrow 1$, is $dF(1)$. Then

$$dF(1) = \lim_{\underline{\pi} \nearrow 1} \frac{M(\underline{\pi})\, C}{M(\underline{\pi}) - \underline{\pi}\alpha(\underline{\pi})} = \lim_{\underline{\pi} \nearrow 1} \frac{M'(\underline{\pi})\, C + M(\underline{\pi})\frac{dC}{d\underline{\pi}}}{M'(\underline{\pi}) - \alpha(\underline{\pi}) - \underline{\pi}\alpha'(\underline{\pi})} = -\lim_{\underline{\pi} \nearrow 1} \frac{M'(\underline{\pi})\, C + M(\underline{\pi})\frac{dC}{d\underline{\pi}}}{\alpha(\underline{\pi})}.$$

[23] Lemma 7 establishes that there is no rationing in allocation when C is small.

We have already noted that $M(\underline{\pi})/\alpha(\underline{\pi}) = \mathbb{E}[\pi|\pi \geq \underline{\pi}]$. With $d^2F(\cdot)$ bounded, we also obtain

$$\lim_{\underline{\pi}\nearrow 1} \frac{M'(\underline{\pi})\,C}{\alpha(\underline{\pi})} = \lim_{\underline{\pi}\nearrow 1} \frac{M''(\underline{\pi})\,C + M'(\underline{\pi})\,\frac{dC}{d\underline{\pi}}}{\alpha'(\underline{\pi})} = \lim_{\underline{\pi}\nearrow 1} \frac{dC}{d\underline{\pi}}.$$

It follows that

$$dF(1) = -2\frac{dC}{d\underline{\pi}}.$$

By the inverse function theorem, $dC/d\pi = [d\underline{\pi}/dC]^{-1}$, giving

$$\left.\frac{d\underline{\pi}}{dC}\right|_{\underline{\pi}=1} = -\left[\frac{1}{2}dF(1)\right]^{-1} \gtrless -\left[\mu_H dF(1)\right]^{-1}.$$

Then $\frac{d\underline{\pi}}{dC}|_{C=0} > \frac{d\underline{\pi}_H}{dC}|_{C=0}$ when $\mu_H < \frac{1}{2}$, implying the desired result. □