Non-linear Pricing with Maximum Demand *

Bing Liu†

October 15, 2025

PRELIMINARY AND INCOMPLETE

Abstract

The economics literature on non-linear pricing has offered significant insights into real-world practices of second-degree price discrimination, with a central conclusion that optimal strategies involve differentiated quantity bundles with a quantity discount. However, this prediction is in contrast with the prevalence of quantity premiums in utility markets, parcel shipping, and cloud storage, as well as the widespread adoption of unlimited options, subscription, and buffet pricing models.

This paper starts from the practical observation that consumers have a maximum demand, a finite quantity beyond which their marginal utility diminishes to zero, and demonstrates that accounting for the correlation between consumers' maximum demand and per-unit value, quantity premiums and unlimited options can emerge as features of the optimal non-linear pricing strategy.

In the case when the consumer's maximum demand and per-unit value are negatively correlated, the consumer preferences violates the ubiquitously held single-crossing assumption in the non-linear pricing and mechanism design literature. We provide a method to solve the problem. The optimal mechanism involves full surplus extraction, buffet pricing and non-monotonic allocation. Extending beyond the maximum demand model, we show these feature remain in the optimal mechanism for consumer preferences that violate single-crossing in a more general form. In this context, we demonstrate a case of monopolistic *over*-provision to illustrate how relaxing the single-crossing assumption can reverse standard economic intuitions.

We show that a quantity premium (discount) is optimal when per-unit value and the maximum demand are perfectly positively (negatively) correlated and that extends to the cases when per-unit value and the maximum demand are stochastically related.

*.

†Department of Economics, Stanford University. Email: bingliu@stanford.edu.

Keywords:
JEL-Classification:

1 Introduction

The economics literature on non-linear pricing has been successful in predicting and prescribing real-world practices of second-degree price discrimination. Quantity discounts such as "buy 2 get 1 free" or "1 for $4 and 3 for $10" are ubiquitous, seemingly confirming one of the key results in Maskin and Riley (1984), a seminal contribution to the theory of non-linear pricing:

> Quantity discounts are always optimal for buyers at the upper tail of the distribution. Invoking some further fairly mild restrictions, we can establish the considerably stronger result that quantity discounts are everywhere optimal.

This suggests that optimal non-linear pricing almost universally takes the form of quantity discounts.

Yet in practice, many prominent pricing schemes do not fit this prediction. A salient example is utility pricing. In Redwood City, California, residential water is priced in tiers, with the per-unit rate increasing at higher levels of total consumption[1]. Electricity pricing under PG&E's tiered rate plan follows a similar pattern: $0.40 per kWh up to a baseline allocation, and $0.50 per kWh thereafter. Analogous "quantity premium" appear elsewhere: Dropbox charges $11.99 per month for 2TB of storage but $19.99 for 3TB,[2] and DHL increases its worldwide document shipping rate by $1 for the first 0.5kg above 0.5kg, but by $5 for the next 0.5kg.

Even in markets where quantity discounts are observed, standard models struggle to account for the prevalence of the "unlimited" option and subscription models. Mint Mobile, for instance, offers 5GB for $20 and 15GB for $25—consistent with discounting—but also an "unlimited" option for $30. Ski resorts frequently offer day passes, multi-day bundles, and unlimited season pass; gyms commonly sell single drop-in classes, multi-class packages, and unlimited monthly membership; streaming platforms such as Netflix and HBO Max rely primarily on subscriptions that give users access to the whole catalogue for a fixed monthly fee.

Why then, if sellers can use second-degree price discrimination through menus of quantities, do we so often observe pricing schemes that feature an unlimited option, or even have only the unlimited option?

Markets that exhibit a quantity premium typically share a common feature: consumers have a maximum demand, a finite quantity beyond which their marginal utility drops to

[1] https://www.redwoodcity.org/departments/public-works/water/rates/current-rates

[2] By contrast, linear pricing would imply roughly $17.99 for 3TB.

zero. In utilities, for example, households derive values from water and electricity consumption only up to a certain quantity; beyond that, additional units provide no further benefit. The same holds for cloud storage or parcel shipping, where demand is naturally capped. Moreover, in many of these markets, consumers with a higher willingness to pay per unit also tend to have higher maximum demands. Households with more appliances not only consume more electricity but may also value each kilowatt-hour more. Professionals working with large datasets place greater per-unit value on storage than, say, a researcher who only needs to archive a few papers.

Maximum demand also plays a key role in consumer preferences in markets where 'unlimited option' and subscription-based pricing prevail. Consider ski passes: avid skiers want to maximize their rides over a season, while occasional skiers value only a handful of trips. Similarly, Netflix subscribers differ both in the total hours they can devote to viewing and in the intensity of their per-hour valuations: those with limited leisure may prize a few hours of high-quality entertainment more than those with ample time.

In this paper, we introduce a maximum-demand constraint into consumer preferences and show that this unifying framework provides an optimality-based rationale for both quantity premiums and the prevalence of unlimited options. In our model, each consumer is characterized by a two-dimensional type: a per-unit value, reflecting marginal utility, and a maximum demand, beyond which additional units yield no value. This can be interpreted as a linear approximation of preferences that vary both in marginal utility and in the rate at which marginal utility declines. We show that positive affiliation between per-unit value and maximum demand leads the optimal menu to feature quantity premiums, whereas negative affiliation (under further conditions) yields quantity discounts. Depending on the relationship of per-unit value and maximum demand, the optimal menu may also include some "unlimited" options or take the form of a pure subscription.

Adding maximum demand introduces two important deviations from the standard model. First, if the per-unit value and the maximum demand are positively affiliated, it is no longer true that under the optimal pricing, marginal utility of consumers with higher consumptions is lower than the marginal utility of consumers with lower consumption – the crucial condition for the standard quantity discount result in Maskin and Riley (1984). Second, if the per-unit value and the maximum demand are negatively affiliated, the order of utilities and marginal utilities among consumers is no longer preserved across all quantities, unlike in the standard model where single-crossing ensures such ordering. Figure 1 provides a visual illustration on how consumer preferences differ in our model

and a standard model.

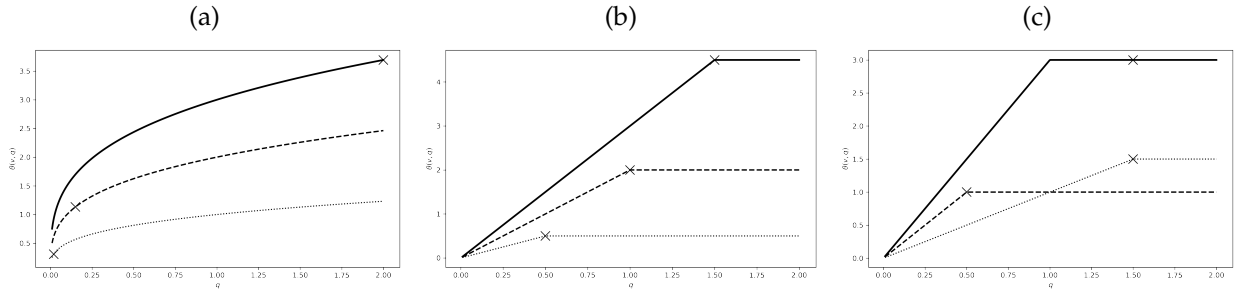


Figure 1: Visualization of consumer utility curves $v(\theta, q)$ of different types θ as a function of quantity q. Panel (a) shows an example from a standard model. Panel (b) and (c) are examples of consumer utility curves with maximum demand, when the per-unit value and maximum demand are positively and negatively affiliated, respectively. The allocations under optimal pricing to each type of consumer is indicated by a black cross in all three panels. Comparing (a) and (b), marginal utility of consumers with higher consumptions is lower than the marginal utility of consumers with lower consumption in (a) while the opposite occurs in (b). Comparing (a) and (c), consumers with higher type have both higher level of utility and marginal utility in (a) while it is no longer true in (c).

In the case of perfectly positive correlation (see Figure 1 (b)), the optimal mechanism features a cut-off per-unit value – all consumers with per-unit value higher than this threshold receive their full maximum demand. It is important to observe that consumers with higher maximum demand also have higher marginal utility (per-unit value). To ensure incentive compatibility, the marginal price must therefore increase with quantity, leading to a quantity premium. In utilities, if households with larger maximum demand also place greater value on each unit of electricity or water, then optimal pricing involves charging higher per-unit rates in higher consumption tiers— consistent with what we observe in practice. A similar rationale extends to markets such as cloud storage and parcel shipping.

In the case of a perfectly negative correlation (see Figure 1 (c)), consumers with higher marginal utility (per-unit value) have lower maximum demand. The single-crossing condition no longer holds. As a result, the usual approach of applying the envelope theorem does not apply. Solving the problem involves two steps. We first approximate the maximum-demand function by simple (step) functions—finite linear combinations of indicator functions—and, by taking the limit as the approximations refine, characterize a key feature of the optimal mechanism: quantity discount. Conditional on the quantity-discount schedule, we apply the standard principle of maximizing the seller's extraction of consumer surplus subject to incentive-compatibility and individual rational-

ity constraints. In this particular setting, it implies designing a payment function that is non-increasing in the consumer's per-unit value to ensure incentive compatibility, while capturing as much surplus as possible. The segment of the payment function that decreases in consumer's per-unit value corresponds to the a segment of the consumer's maximum willingness to pay (per-unit value $\times$ maximum demand) that decreases with their per-unit value. In fact, the strictly decreasing segments of the payment function is always exactly equal the consumer's maximum willingness to pay. In these regions, full surplus extraction is possible: consumers with higher maximum willingness to pay have lower per-unit value and therefore no incentive to deviate to a lower quantity at a higher per-unit price, while consumers with lower maximum willingness to pay have no incentive to choose a higher quantity which implies a higher aggregate payment. Conversely, when a consumer's maximum willingness to pay is non-decreasing in per-unit value, the payment function is a constant. In these regions, the incentive constraint binds. Any payment function that increases in per-unit value is not incentive compatible because that implies a higher aggregate price for a lower quantity. The resulting optimal pricing is a hybrid of *all-you-can-pay* (the strictly decreasing pricing segments) and *all-you-can-eat* (the constant pricing segments). This hybrid structure closely mirrors real world pricing schedules, such as tiered data plans from telecom companies such as Mint Mobile and various passes for ski resorts, theme parks, and gym classes. If consumer's maximum willingness to pay is non-decreasing for all consumer's per-unit values, the optimal mechanism charges a fixed fee – all consumers who pay the fixed fee can consumer any quantity they want. This directly corresponds to the all-you-can-eat pricing seen in buffets and the flat-rate subscription models prevalent on platforms such Netflix and Amazon Prime.

This contrasts with the standard mechanism-design approach, where one typically derives the optimal allocation first and then recovers payments via the envelope theorem. In our model the order is reversed: we first pin down the optimal payment schedule and then derive the allocation that implements those payments. Crucially, the optimal allocation need not be monotone in a consumer's per-unit valuation; it can be non-monotonic. An important implication is that in a monopoly market where consumers' maximum demand and per-unit valuation are negatively correlated, intermediate types may be excluded from the market even though lower- and higher-valuation consumers are served.

Beyond the maximum-demand model, we show that all-you-can-pay and all-you-can-eat schemes can still be optimal for more general preferences that violate the single-crossing condition. Crucially, these violations can flip standard economic intuition: instead of the monopolist under-providing, the inefficiency of monopoly may be over-provision.

The observation of a quantity premium (discount) when per-unit value and the maximum demand are perfectly positively (negatively) correlated extends to the cases when per-unit value and the maximum demand are stochastically related. When the maximum demand and per-unit value are positively affiliated, it is more likely for a consumer with higher maximum demand to have a higher per-unit value, the optimal pricing features quantity premium similar to the case with perfect positive correlation. Because of the quantity premium, consumers with a maximum demand equal to q and a per-unit value higher than the per-unit price of the q-unit bundle may not purchase the q-unit bundle – their surplus might be maximized by purchasing a lower quantity bundle with a lower per-unit price. Thus, different from the perfect correlation case, consumers may purchase a positive quantity that is below their maximum demand. This also creates a linearizing effect on the price schedule: for a q-unit bundle, instead of setting it to a very low price so as to selling it to a consumer with maximum demand equal to q, it is optimal for the seller to set a slightly higher price since the seller can also offer this bundle a consumer with a maximum demand higher than q but couldn't afford the higher quantity bundle, creating a profitable substitution. When the maximum demand and per-unit value are negatively affiliated, the revenue benefits from this substitution persists. So it is not automatically the case that a quantity discount is optimal. Once we rule out this substitution (Assumption 3), quantity discount is optimal.

This paper builds upon the extensive literature on non-linear pricing under the single-crossing assumption, including seminal works such as Mussa and Rosen (1978), Maskin and Riley (1984), and Goldman et al. (1984). Comprehensive reviews of this literature can be found in Wilson (1993) and Armstrong (2016).

Our work also relates to studies that explore the implications of violating the single-crossing assumption, such as Araujo and Moreira (2010), Araujo et al. (2015), Schottmüller (2015), Chen et al. (2022), and Kwak (2022). With the exception of Kwak (2022), which identifies conditions under which non-single-crossing problems can be reduced to those satisfying single-crossing, these papers introduce alternative conditions, such as double-crossing (Chen et al. (2022)), inversely U-shaped decision functions (Araujo et al. (2015)), and monotone decision functions (Schottmüller (2015)), to ensure global incentive compatibility alongside local incentive compatibility. In contrast, our paper adopts a practical approach, acknowledging that consumer preferences in real-world markets often violate single-crossing. By incorporating a maximum demand that negatively correlates with per-unit value—a crude approximation of real-world, non-single-crossing preferences—our model predicts optimal non-linear pricing strategies that align with the prevalence of unlimited options and subscription models. Furthermore, we demonstrate how

consumer preferences with non-single-crossing in more general forms can alter standard economic intuitions. Regarding quantity premiums, our findings relate to Liu (2025), which identifies quantity premiums as a general feature in linear utility models, of which ours is a special case.

The paper is organized as follows. In Section 2, we introduce the model. In Section 3, we discuss the case when the two dimensions are perfectly correlated. In Section 4, we extend the result to when the two dimensions are stochastically related.

2 Model

Setup There is a seller and a consumer. The seller can supply a homogeneous good at a constant marginal cost of $c \geq 0$. For most of the paper, we let $c = 0$ to concentrate on the effect of consumer's preferences. The consumer has a two-dimensional type, $\theta = (v, d)$, where $v \in [\underline{v}, \overline{v}]$ is the consumer's per-unit valuation of the good and $d \in [\underline{d}, \overline{d}]$ is the consumer's maximum demand. For a given type (v, d), an allocation q and a payment T, the consumer's payoff is

$$v \min\{q, d\} - T.$$

So v is the consumer's marginal utility for all units $q \leq d$ and 0 is the consumer's marginal utility for all units $q > d$. (v, d) is private information to the consumer and follows the joint distribution F with density $f(\cdot)$. We assume F is common knowledge and satisfies Assumption 1.[3]

Assumption 1. *The conditional distribution $F(\cdot; d)$ satisfies that*

$$vf(v; d) - (1 - F(v; d))$$

is weakly increasing in v.

Let $\phi(v; d) = v - \frac{1-F(v;d)}{f(v;d)}$ denote the conditional virtual type. We let $\hat{v}(d)$ be such that $\phi(\hat{v}(d); d) = 0$.

[3] Assumption 1 is closely related to the standard Myerson regularity condition, which requires the virtual value $\phi(v; d)$ to be increasing. These conditions are not nested: there exist distributions (e.g., Beta(1,3)) that satisfies Assumption 1 but not Myerson regularity. For some examples where Assumption 1 holds: the uniform distribution, symmetric or increasing Beta distributions, and many truncated normal distributions. It generally fails for the exponential and log-normal distributions over their full supports, though it holds over truncated intervals.

Seller's problem It is without loss to consider the seller offering a menu of quantity bundles $(q, t(q))_{q\in[0,\bar{d}]}$ to the consumer, with $t(q)$ being the per-unit price when the consumer purchases the q-unit bundle. The seller's objective is to maximize profit

$$\mathbb{E}[(t(q(\theta)) - c)\, q(\theta)]$$

subject to the consumer maximizing their own payoff, that is, for each θ

$$q(\theta) \in \arg\max_{q} \quad v \min\{d, q\} - t(q)q$$

Note that the average price function $t(\cdot)$ fully determines the seller's profit. So we denote $\pi(t) := \mathbb{E}[(t(q(\theta)) - c)\, q(\theta)]$ where $q(\theta) \in \arg\max_q \quad v\min\{d,q\} - t(q)q$.

Correlation structure between v and d. We consider two types of correlations between v and d: when v and d are positively affiliated and when v and d are negatively affiliated.

Definition 1. *v and d are positively (respectively, negatively) affiliated if for all $d > d'$, $\frac{f(v;d)}{f(v;d')} \geq \frac{f(v';d)}{f(v';d')}$ (respectively, $\leq$) for all $v > v'$.*

Quantity premium/discount.

Definition 2. *The optimal pricing exhibits quantity premium (respectively, discount) if there exists a solution $t(\cdot)$ to the seller's problem such that $t(q)$ is non-decreasing (respectively, non-increasing) in q.*

Intuitively, a quantity premium arises when larger purchases are charged a higher per-unit price, whereas a quantity discount occurs when larger quantities are priced more favorably on a per-unit basis. The following lemma is useful and known:

Lemma 1. *If v and d are positively (respectively, negatively) affiliated*

1. *$\hat{v}(d)$ is non-decreasing (non-increasing) in d,*
2. *$F(\cdot; d) \geq_{FOSD} F(\cdot; d')$ for all $d > d'$.*

3 Perfect correlation

We begin with the special case in which maximum demand and per-unit value are perfectly correlated. A consumer's type is $\theta = (v, d(v))$ for some monotone function $d(\cdot)$.

This reduces the problem to a single-dimensional screening problem. We take the direct revelation approach to the problem. Let $q : [\underline{v}, \overline{v}] \to [0, \overline{d}]$ and $T : [\underline{v}, \overline{v}] \to \mathcal{R}$ be the allocation and total payment rule. The seller's problem is to choose (q, T) to maximize profit

$$\mathbb{E}[T(v) - cq(v)]$$

subject to (IC)

$$v \min\{d(v), q(v)\} - T(v) \geq v \min\{d(v), q(v')\} - T(v'), \quad \forall v, v' \in [\underline{v}, \overline{v}]$$

and (IR)

$$\theta \min\{d(v), q(v)\} - T(v) \geq 0, \quad \forall v \in [\underline{v}, \overline{v}]$$

We take $c = 0$ to focus on the consumer side of the effect.

3.1 Increasing maximum demand

When $d(\cdot)$ is increasing, consumers with a higher type—represented here by a higher per-unit value v—have a greater marginal utility across all quantities. Panel (a) of Figure 2 provides a visualization of an example of these consumer utilities. Importantly, the single-crossing preferences assumption remains satisfied. Consequently, the problem can be solved using the standard envelope theorem approach. We formally present the resulting optimal mechanism in Proposition 1. Panel (b) of Figure 2 illustrates the optimal price schedule as a function of quantity q when $v \sim U[0, 3]$ and $d(v) = 0.5v$.

Proposition 1. *When $d(\cdot)$ is increasing, the optimal mechanism has the allocation rule that uses same cut-off type in a standard optimal auction, $\hat{v}$. In the optimal mechanism, for all $v < \hat{v}$, $q(v) = 0$; for all $v \geq \hat{v}$, $q(v) = d(v)$ and $T(v) = vd(v) - \int_{\hat{v}}^{v} d(x)dx$.*

See Appendix A.1 for the proof.

The optimal mechanism features a cut-off type, $\hat{v}$. This cut-off is, perhaps surprisingly, identical to that found in a standard optimal auction when all consumers have the same maximum demand. The reason for this is that the seller's optimization problem remains linear with respect to the consumer's allocation. Therefore, the core principle of mechanism design persists: the seller allocates a consumer their maximum demand $d(v)$ if and only if that consumer's virtual value is non-negative, that is, whenever $v \geq \hat{v}$.

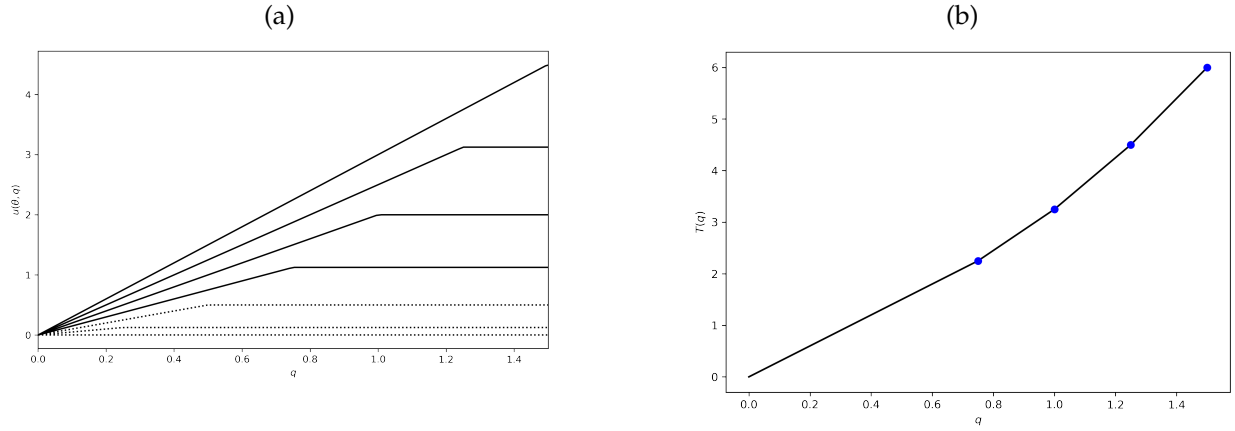


Figure 2: Panel (a): visualization of consumer utility curves $u(\theta, q)$ of different types θ as a function of quantity q when the maximum demand is increasing. Here $v \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$ and $d(v) = 0.5v$. Consumers of higher type have higher marginal utilities for all quantities. Solid lines represent consumer types that are allocated with a positive quantity under the optimal pricing. Panel (b): the optimal price schedule for Panel (a). The cutoff type is $v = 1.5$. The gradient of the optimal price schedule at $q \geq 0.75$ corresponds to the marginal utility of the consumer whose maximum demand is q.

Given $v > \hat{v}$, one can calculate the per-unit price the type $(v, d(v))$ pays

$$t(v) = \frac{T(v)}{d(v)} = \underline{v} + \int_{\underline{v}}^{v} 1 - \frac{d(x)}{d(v)} dx$$

For $v' > v \geq \hat{v}$,

$$\begin{aligned} t(v') = \frac{T(v')}{d(v')} &= \underline{v} + \int_{\underline{v}}^{v} \underbrace{1 - \frac{d(x)}{d(v')}}_{> 1 - \frac{d(x)}{d(v)}} dx + \underbrace{\int_{v}^{\underline{v}} 1 - \frac{d(x)}{d(v')} dx}_{\geq 0} \\ &\geq t(v). \end{aligned}$$

Therefore, the optimal mechanism exhibits quantity premium.

Corollary 1. *If $d(\cdot)$ is increasing, then the optimal mechanism exhibits quantity premium.*

A useful way to understand the intuition behind the quantity premium result is to consider how the *marginal* price at each quantity separates consumers of different types. Let $T(q)$ be any pricing schedule that maps quantities to aggregate prices. For a consumer that is already purchasing q units, it is optimal for them to purchase $q + \Delta$ units only if the marginal price at q, $T(q + \Delta) - T(q)$, is less than their marginal utility, in this case v, at q. Thus, the marginal price at q essentially acts as the cut-off type that determines which

consumers find it optimal to consume at least q units. The set of marginal utilities at each q is $\{v : d(v) \geq q\}$. Let F_q denote the distribution of the marginal utilities at q[4]. Crucially, F_q monotone likelihood ratio dominates $F_{q'}$ for any $q > q'$. It then follows that the optimal cut-off type, that is the marginal price, for each quantity q is increasing, explaining the quantity premium.

At a higher level, the emergence of a quantity premium reflects the fact that more information about a consumer's type is revealed when maximum demand is positively affiliated with per-unit valuation. Corollary 2 is a direct implication of Proposition 1

Corollary 2. *If $d(\cdot)$ is increasing, the per-unit price in the optimal mechanism is higher than $\hat{v}$.*

Application: effect of demand reduction. Type-dependent maximum demand may result from an exogenous demand reduction. For instance, consider a government means-tested program that provides the good at no cost, or at a substantially reduced price, to lower-type consumers. If the resulting residual demand, that the consumer's effective maximum demand in the monopoly market, is increasing in type, then Corollary 2 implies that customers face a higher per-unit price following the demand reduction.

3.2 Decreasing maximum demand

When the maximum demand $d(v)$ is decreasing in per-unit value v, the consumer's preference no longer satisfies the single-crossing condition. Figure 3 provides a visual comparison of the consumer utility curves that satisfy the single-crossing condition and the consumer utility curves corresponding to a decreasing maximum demand. In Figure 3(b), neither utilities nor marginal utilities are consistently ordered among different types across quantities. The ranking of marginal utilities changes at $q = 0.5$ and again at $q = 1$ while the ranking of total utilities changes at $q = 1$. In such cases, no monotone ordering of consumer "types" is possible. When the single-crossing condition is satisfied, higher types always optimally choose higher quantities than the lower types, regardless of the price schedule. With decreasing maximum demand, this monotonicity breaks down. Example 1 provides a simple illustration.

Example 1. *Consider two consumers, $1, 2$, with per-unit values $v_1 = 1 < v_2 = 2$ and a simple linear price schedule $T(q) = 0.8q$. then their preferences satisfy the single-crossing condition: consumer 1 can be regarded as the lower type and consumer 2 as the higher type. Given the price schedule, it is optimal for consumer 1 to purchase 1 unit. It is also optimal for consumer 2 to*

[4]Note that this is not true under the assumption in Maskin and Riley (1984).

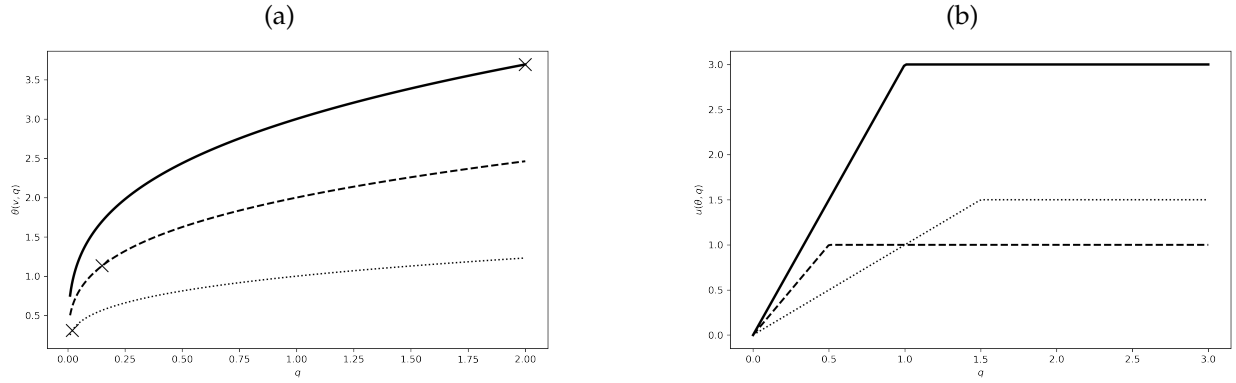


Figure 3: Panel (a): visualization of consumer utility curves $u(\theta, q)$ of different types θ as a function of quantity q that satisfies the single-crossing condition. Consumers of higher type have higher utilities and marginal utilities for all quantities. Panel (b): visualization of consumer utility curves $u(\theta, q)$ of different types θ as a function of quantity q when the maximum demand decreases in per-unit value. The order of consumer utility is not the same across different quantities: the consumer with utility represented by dashed curve than the consumer with utility represented by the dotted curve for $q < 1$ but the order swaps once $q > 1$. The order of consumer marginal utility is not the same across different quantities: the solid curve consumer has higher marginal utility than the dotted curve consumer for $q < 1$ but the order swaps once $q > 1$.

purchase 1 unit. In fact, since the marginal utility of 2 *is always higher than* 1 *for all quantities, consumer 2 would always purchase a higher quantity than consumer 1. When consumer 1 has a higher maximum demand than consumer 2 (see. Figure 4 Panel (b) where* $d(v_1) = 1 > d(v_2) = 0.5$*), these preferences no longer satisfy the single-crossing condition and the two consumers can no longer be ranked. Under the same linear pricing* $T = 0.8q$*, consumer 1 would purchase 1 unit and consumer 2 would purchase 0.5 units. Notably, the optimal pricing for the two types shown in Panel (a) is any menu with the two options of* q *units at the price of* q *and* 1 *unit at the price of* $2 - q$*, for any* $q \in [0, 1]$*. This optimal pricing generates a revenue of* 2 *for the seller. In contrast, the optimal pricing for the two types shown in Panel (b) takes the form of a fixed fee of* 1 *and the consumer can consume any quantity they want. Interestingly, this optimal pricing also generates a revenue of* 2 *for the seller, even though clearly there is a demand reduction for consumer 2 compared to the case in Panel (a).*

When solving a standard problem where the consumer preference satisfies the single-crossing condition, incentive compatibility requires the price of the high quantity to be low enough such that a higher type gets weakly higher payoff choosing the high quantity than the low quantity. The single-crossing property further ensures that only the downward incentive compatibility constraints are binding. Hence, leaving sufficient surplus to

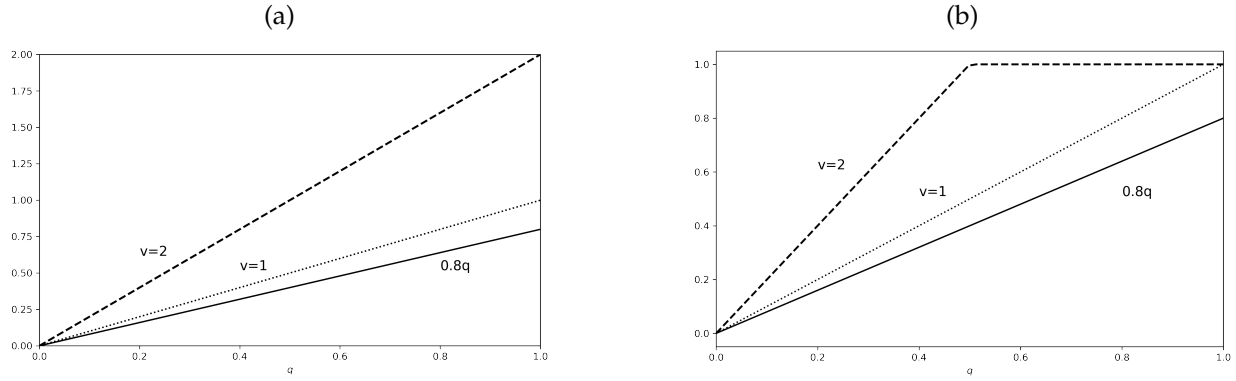


Figure 4: Dashed and dotted lines are consumer utility curves of different types, for $v = 2$ and $v = 1$ respectively. Solid line is the price schedule as a function of quantity where the per-unit price is 0.8. Panel (a): two consumers have the same maximum demand. Panel (b): consumer 1 has higher maximum demand than consumer 2.

higher types at higher quantities is sufficient for incentive constraints. One does not need to worry about the lower types deviating to a higher quantity bundle.

With decreasing maximum demand, however, two important complications arise. First, a type with lower marginal utility at a lower quantity may have higher marginal utility at a higher quantity, so satisfying the downward incentive constraint alone is *no longer* sufficient; such a type may profitably deviate to a higher-quantity bundle. Second, the downward incentive constraint is relevant only if the quantity intended for the lower type (lower v) is below the maximum demand of the higher type (higher v). If instead the quantity intended for the lower type exceeds the maximum demand of the higher type, then—given that higher quantities are associated with higher prices—the higher type would never choose that allocation, and the usual constraint of leaving sufficient surplus to higher types no longer applies. Example 2 and Figure 5 illustrates this in more detail.

Example 2. *Consider two consumers,* $1, 2$*, with per-unit values* $v_1 = 1.5 < v_2 = 2$ *and maximum demands* $d(v_1) = 1 > d(v_2) = 0.5$*. For simplicity, suppose the seller is restricted to offering a menu with two possible quantities,* $q \in 0.5, 1$*. Using the 'standard' approach (Figure 5), the bundle with* $q = 0.5$ *is assigned to the lower type (*v_1*). The highest feasible price is the utility that type* v_1 *obtains from consuming* 0.5 *units, namely* $0.5 \times 1.5 = 0.75$*. If the higher type,* v_2*, consumes this bundle* $(0.5, 0.75)$*, their payoff is* 0.25*. To ensure incentive compatibility, the price for the* $q = 1$ *bundle must therefore be at most* $\min 1, 0.5 \times 2 - 0.25 = 0.75$*. This yields the menu* $(0.5, 0.75), (1, 0.75)$*. However, this menu is* neither optimal nor incentive compatible*. The lower type (*v_1*) will optimally choose* $(1, 0.75)$ *since it provides higher utility. The optimal menu*

consists the two bundles: $(0.5, 1)$ *and* $(1, 1.5)$*, with type* v_1 *optimally chooses* $(1, 1.5)$ *and type* v_2 *optimally chooses* $(0.5, 1)$*. Note that the optimal menu also achieves* full surplus extraction.

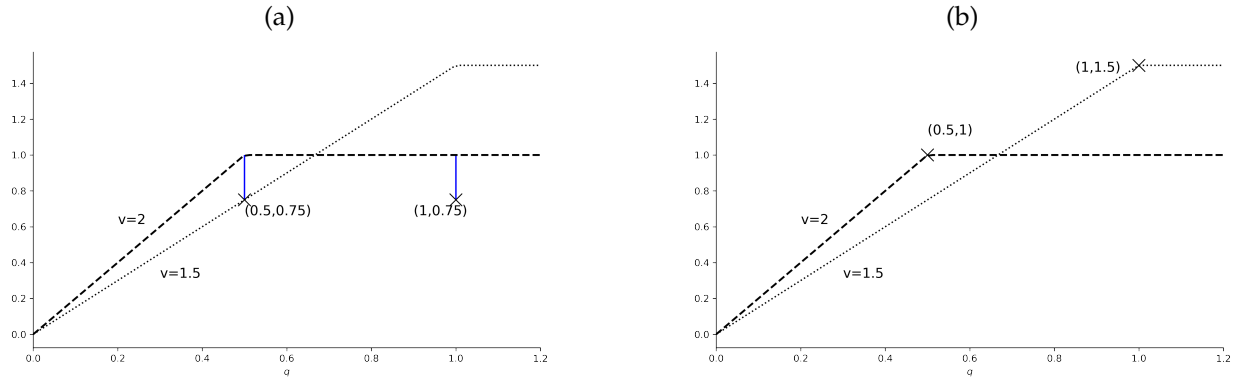


Figure 5: Dashed and dotted lines are consumer utility curves of different types, for $v = 2$ and $v = 1.5$ respectively. "X" markers represent potential bundles. Panel (a): (Not really)'incentive compatible' bundles based on the intuition from the standard model. Panel (b): optimal bundles.

We characterize the optimal pricing in two steps. First, we approximate the maximum demand function $d(\cdot)$ with n-step functions $d^n(\cdot)$. We show that the optimal pricing when the maximum demand is the step function $d^n(\cdot)$ exhibits quantity discounts. this quantity-discount structure is preserved in the limit as the approximation step size approaches zero and $d^n(\cdot) \to d(\cdot)$. Second, using the established quantity-discount property, we derive the detailed form of the optimal allocation and the associated payment rule. We outline the key intuitions in the main text and provide the formal proof in the appendix.

For the intuition of the quantity discount, consider the following maximum demand function that is a step function and non-increasing in the per-unit value $v \in [\underline{v}, \overline{v}]$

$$d(v) = \begin{cases} d_1, & v \in [\underline{v}, v^1) \\ d_2, & v \in [v^1, \overline{v}] \end{cases}$$

for some $0 < d_2 < d_1$. For cleaner presentation, let $U(v, T(v), q(v)$ denote the payoff of a consumer with per-unit value v given the allocation $q(v)$ and the payment $T(v)$. The incentive constraints of this problem can be decomposed into two parts:

- (Within-max demand IC): for all $v, v' \in [\underline{v}, v_d)$ and for all $v, v' \in [v_d, \overline{v}]$,

$$U(v, T(v), q(v)) \geq U(v, T(v'), q(v'))$$

- (Cross-max demand IC): for all $v \in [\underline{v}, v_d)$ and $v' \in [v_d, \overline{v}]$

$$U(v, T(v), q(v)) \geq U(v, T(v'), q(v'))$$

and for all $v' \in [\underline{v}, v_d)$ and $v \in [v_d, \overline{v}]$

$$U(v, T(v), q(v)) \geq U(v, T(v'), q(v')).$$

The two sets of IC constraints can be further reduced. First, the within-max demand IC can be replaced by a local IC constraint by appealing to the envelope theorem. If the within-max demand IC is sufficient, then the optimal mechanism would have two cut-off types v_1 and v_2 such that

$$v_2 - \frac{1 - F(v_2)}{f(v_2)} = 0, \quad v_1 - \frac{1 - F(v^1)}{f(v_1)} = 0.$$

Second, the cross-max demand IC can be replaced by a simple constraint: the aggregate payment $T(\cdot)$ is non-decreasing in the allocation $q(\cdot)$. The solution to the this problem has a simple form.

If $d_2 \cdot v_2 \leq d_1 \cdot v_1$, the cross-max demand IC is not binding. The optimal mechanism is to offer d_1 units at a total payment of $d_1 v_1$ and d_2 units at a total payment $d_2 v_2$. The per-unit price for the larger bundle $d1$ is $v1$, which is lower than the per-unit price $v2$ for the smaller bundle $d2$ — i.e., the optimal menu exhibits a quantity discount. If $d_2 \cdot v_2 > d_1 \cdot v_1$, the cross-max demand IC binds. It is no longer optimal to set a different price for d_1 units and d_2 units. The optimal solution is to set a fixed fee $T \in \arg\max_{T'} T' \cdot P(vd(v) \geq T')$, and consumers can consume their maximum demand after paying the fixed fee T. This is effectively a quantity-discount mechanism because higher-demand types obtain more units for the same fee.

When we increase the number of steps in the maximum demand function to n steps such that $d(v) = d_i$ for $v \in [v^{i-1}, v^i)$ for each $i \in \{1, \ldots, n-1\}$, the optimal mechanism has the same qualitative structure as in the two-step case. The cut-off types are now v_i's where for each i

$$v_i - \frac{F(v^i) - F(v_i)}{f(v_i)} = 0.$$

Note that $v_i < v_{i-1}$ for all i. For any $d_{i-1} > d_i$, there are two possibilities: either the cross-max demand IC is not binding, then the optimal per-uni prices for d_{i-1} and d_i are

$v_i < v_{i-1}$; or the cross-max demand IC binds, the optimal mechanism is a fixed fee for both $d_{i-1} > d_i$. In both scenarios, the optimal menu displays quantity discounts: per-unit prices do not increase with quantity.

As $n \to \infty$, $d^n(\cdot) \to d(\cdot)$ uniformly. The quantity discount feature remains in the solution. With linear utility and a non-increasing marginal price schedule, individual demand is bang-bang: a consumer with type v either purchases $d(v)$ or 0.

This bang-bang structure in allocation in turn pins down the shape of payments. Since the maximum demand $d(v)$ is decreasing in v and the served types, consumers who receive a positive amount, receive their maximum demand $d(v)$, the allocation $q(v)$ among all the served types is decreasing in v. Consequently, the payment function among served types, $T(v)$, is also decreasing in v.

Given this knowledge on the shape of the payment function, we consider an auxiliary problem: maximize expected revenue (i.e., captured consumer surplus) over all non-increasing payment functions, subject only to individual rationality. That is,

$$\overline{T}(v) \in \arg \max_{\hat{T}(\cdot) \text{ non-increasing}} \int_{\hat{T}(v) \le v \cdot d(v)} \hat{T}(v)\, dF(v) \tag{1}$$

Given the solution $\overline{T}(v)$ to the auxiliary problem, the candidate payment function is

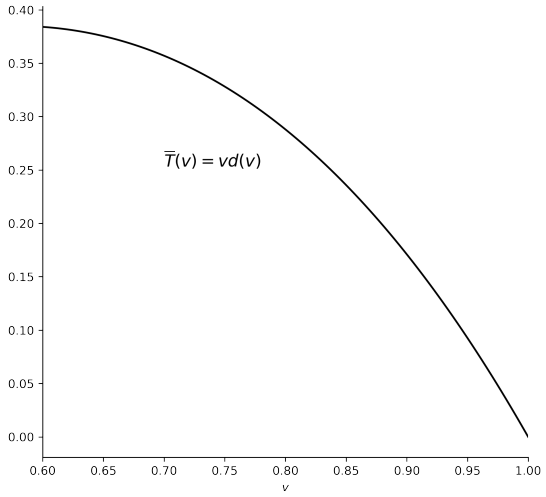
$$T(v) = \begin{cases} \overline{T}(v), & \overline{T}(v) \le v \cdot d(v) \\ 0, & \overline{T}(v) > v \cdot d(v). \end{cases}$$

Equivalently, the candidate payment function is the largest (in the F-integral sense) function that lies below the envelope $v \cdot d(v)$ and is non-increasing for its non-zero segments.

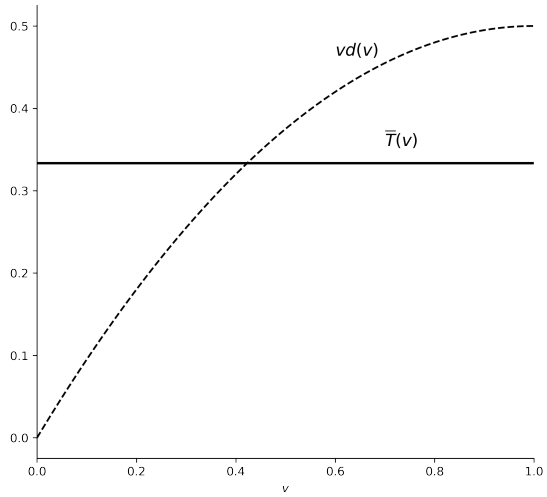
Two observations help characterize the solution to the auxiliary problem $\overline{T}(v)$ and the candidate payment function $T(v)$. First, if $v \cdot d(v)$ is non-increasing, then the $\overline{T}(v)$ equals $v \cdot d(v)$. Second, if $v \cdot d(v)$ is increasing, then the $\overline{T}(v)$ is a constant function. In fact, together with the segments where $T(v) = 0$, these are the only three kinds of segments of $T(v)$: 1) $T(v) = v \cdot d(v)$ for some non-increasing segments of $v \cdot d(v)$, 2) $T(v)$ is a constant for some increasing or non-monotone segments of $v \cdot d(v)$, and 3) $T(v) = 0$. To see why this is the case, consider any non-increasing function $h(v)$ that lies below $v \cdot d(v)$ for some interval $[v_1, v_2]$. Then one can construct another function $\overline{h}(v)$ that consists of only 1) and 2) and satisfies that $\overline{h}(v) \ge h(v)$ for all $v \in [v_1, v_2]$. Figure 6 provides visualizations of different shapes of consumer's maximum willingness to pay $vd(v)$ and its associated candidate payment function $\overline{T}(v)$.

In contrast to the standard mechanism design approach—where we usually first find

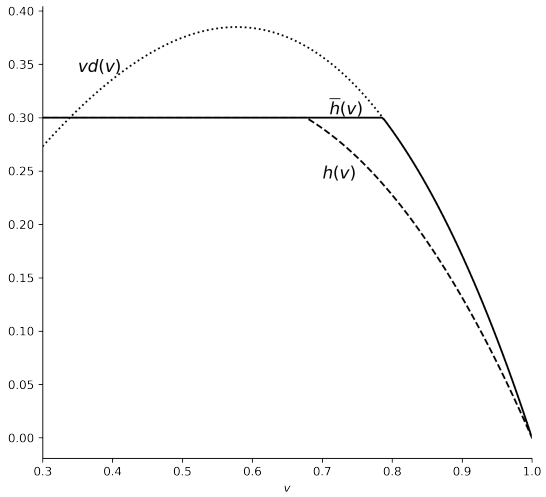
(a) Full surplus extraction



(b) Subscription/buffet pricing



(c)



(d) Tiered pricing with an 'unlimited' option

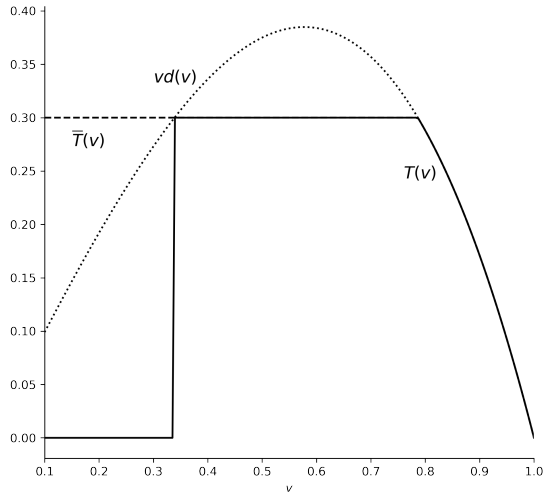


Figure 6: Panel (a): decreasing $vd(v)$ and $\overline{T}(v) = vd(v)$. Panel (b): increasing $vd(v)$ and $\overline{T}(v)$ is a constant. Panel (c): augmenting a decreasing function $h(v)$ (dashed) to $\hat{h}(v)$ (solid) that consists of a constant and a segment equal to $vd(v)$(dotted). Panel (d): given a non-monotone consumer maximum willingness to pay $vd(v)$ (dotted), $\overline{T}(v)$ (dashed) is the solution to (1) and $T(v)$ (solid) is the optimal payment function.

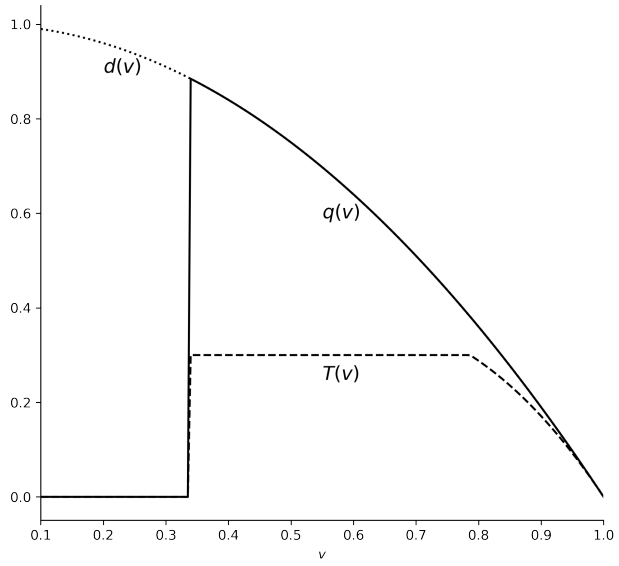


Figure 7: Allocation (solid line $q(v)$) determined by the candidate payment function (dashed line $T(v)$). The dotted line $d(v)$ is the maximum demand from the example in Figure 6 (d), and the dashed line $T(v)$ is the associated candidate payment function. The allocation coincides with the maximum demand whenever $T(v) > 0$ and equals zero whenever $T(v) = 0$.

the optimal allocation and then use incentive compatibility to write down the payments—here we first have the candidate payment $T(v)$, then the candidate allocation function is determined by $T(v)$.

$$q(v) = \begin{cases} d(v), & T(v) > 0 \\ 0, & T(v) = 0. \end{cases}$$

Essentially, consumers with types v that is paying a positive amount, i.e., $T(v) > 0$, receive their maximum demand $d(v)$ while others receive 0.

Now it remains to verify that the candidate allocation and payment functions satisfy global incentive compatibility. First note the price schedule $(q(v), T(v))_v$ features quantity discount: on the segments where $T(v)$ is a constant, the per-unit price $\frac{T(v)}{q(v)}$ is lower for higher $q(v)$; on the segments where $T(v) = vd(v)$, the per-unit price for a quantity $d(v)$ is v, which decreases in $d(v)$. Given the quantity discount, each consumer optimally purchase their maximum demand conditional on that they can afford it: there is no incentive to purchase more since it implies higher payments and zero marginal utility and there is no incentive to purchase less since it implies higher per-unit price.

We state the optimal mechanism in Proposition 2. See Appendix A.2 for its proof.

Proposition 2. *When $d(\cdot)$ is decreasing, the optimal mechanism consists of a sequence of $(v_r)_{r\geq 0}$, with payment and allocation satisfies one of three conditions*

1. *(all-you-can-pay) $q(v) = d(v)$ and $T(v) = vd(v)$ for all $v \in [v_r, v_{r+1}]$. $vd(v)$ is non-increasing in $v \in [v_r, v_{r+1}]$.*
2. *(all-you-can-eat) $q(v) = d(v)$ and $T(v) = v_r d(v_r)$ for all $v \in [v_r, v_{r+1}]$. $vd(v)$ is increasing or non-monotonic in $v \in [v_r, v_{r+1}]$.*
3. *(non-serving) $q(v) = 0$ and $T(v) = 0$ for $v \in [v_r, v_{r+1}]$.*

We now illustrate the economic intuitions behind these three regions and discuss how these might generalize.

All-you-can-pay/full surplus extraction In the all-you-can-pay regions, $vd(v)$ is non-increasing. It is feasible and optimal to perfectly separate all the different types. Consumers consume their maximum demand $d(v)$ and pay their *full surplus* $vd(v)$. Here, not only the order of the marginal utilities are not preserved across quantities, the order of the *utilities* are also not preserved across quantities. The change in the order of the utilities makes the screening costless and full surplus extraction possible. Example 3, which is a continuation of Example 2, illustrates this point. Within this region, for each quantity q, there exists a unique type that attains the highest utility for q. By pricing that quantity q exactly at this maximum value, the seller ensures that only this specific type can afford the bundle. Applying this principle across all quantities in the region leaves each consumer type with exactly one affordable option, perfectly screening them and preventing any profitable deviation. Consequently, the individual rationality constraint binds for every type, leaving them with zero net utility.

Example 3. *Consider two consumers in Example 2, $1, 2$, with per-unit values $v_1 = 1.5 < v_2 = 2$ and maximum demands $d(v_1) = 1 > d(v_2) = 0.5$. The preference of these two consumers satisfies the condition that $vd(v)$ decreases in v: $v_1 d(v_1) > v_2 d(v_2)$. The ranking of consumer valuations changes with the quantity offered. At $q = 0.5$, consumer 2 has the higher willingness to pay (utility of 1, versus 0.75 for consumer 1). At $q = 1$, the roles reverse: consumer 1 now has the higher willingness to pay (utility of 1.5, versus 1 for consumer 2). The optimal menu is $\{(0.5, 1), (1, 1.5)\}$. Each consumer has exactly one affordable bundle in the menu and has their surplus fully extracted by the seller. Figure 8 provides a visualization.*

Within our model, the existence of a unique type that attains the highest utility for each q is sufficient for all-you-can-pay mechanism to be optimal. This can be generalized to

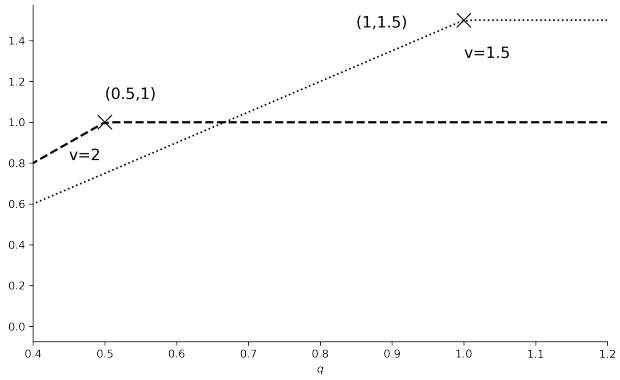


Figure 8: Utility curves for the two consumers in Example 2(dotted line for consumer q and dashed line for consumer 2). The optimal menu is $\{(0.5, 1), (1, 1.5)\}$. Each consumer has exactly one affordable bundle in the menu.

other consumer preferences but with additional conditions on how 'separable' the types are. Figure 9 illustrates when all-you-can-pay mechanism is and is not optimal for more general consumer preferences.

Application: monopoly over-provision The usual economic intuition is that the inefficiency of a monopoly market comes from the under-provision. That intuition, however, depends on consumer preferences satisfying the single-crossing condition. When preferences do not satisfy single-crossing, a monopoly's inefficiency can instead arise from over-provision. Consider the two consumers in Example 2, $1, 2$, with per-unit values $v_1 = 1.5 < v_2 = 2$ and maximum demands $d(v_1) = 1 > d(v_2) = 0.5$. Let the incremental cost of increasing production from 1 unit to 1.5 units be $c > 0$. The maximum social welfare that can be obtained from 1.5 units of the good is $0.5 \times 2 + 1 \times 1.5 = 2.5$; while the maximum social welfare from 1 unit is $0.5 \times 2 + 0.5 \times 1.5 = 1.75$. Thus supplying an extra 0.5 units raises social welfare by $2.5 - 1.75 = 0.75$. By contrast, the monopoly revenue rises by more: the monopoly revenue at 1.5 units is $0.5 \times 2 + 1 \times 1.5 = 2.5$, whereas the monopoly revenue at 1 unit is $0.75 \times 2 = 1.5$, so the monopolist's revenue gain from increasing output is 1! For $0.75 < c < 1$, the socially efficient provision is 1 unit but a profit-maximizing monopoly will supply 1.5 units.

All-you-can-eat In the all-you-can-eat regions, $vd(v)$ is not monotonically decreasing. n sharp contrast to the "all-you-can-pay" region, it is no longer optimal to separate types. Instead, the optimal mechanism pools all types by offering a single bundle. This is driven by a fundamental conflict: the types with the highest total utility are now the same types with the lowest marginal utility. If there is a price differential for a larger quantity, con-

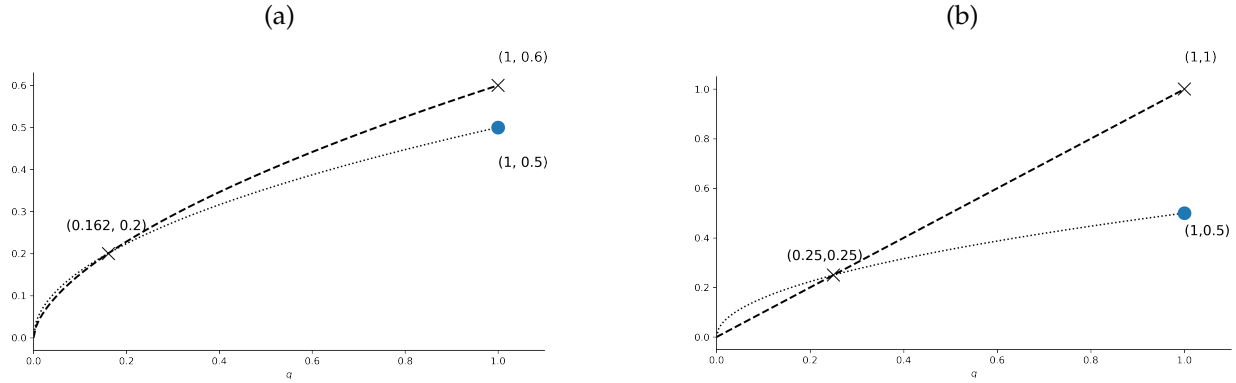


Figure 9: For more general consumer preference, a change of order of utilities for different types is not sufficient for all-you-can-pay menus to be optimal. For it to be optimal, the types must be sufficiently 'separable': the utility curve for the types who have higher utilities at lower quantities has to flatten out quickly at higher quantities compared to the types with higher utilities at higher quantities. Consider a utility function for each quantity q, $u(\theta, q) = \theta q^{\theta}$, where type θ represents the consumer's elasticity. We examine two scenarios. Panel (a): the two types are $\theta \in \{0.5, 0.6\}$. The all-you-can-pay menu that perfectly separates the two types is $\{(0.162, 0.2), (1, 0.6)\}$ which yields a revenue of 0.8. However, the optimal menu is to not separate the two types. A single bundle $\{(1, 0.5)\}$ yields a revenue of 1. Panel (b): the two types are $\theta \in \{0.5, 1\}$. The all-you-can-pay menu is $\{(0.25, 0.25), (1, 1)\}$. It yields a revenue of 1.25 and is the maximum revenue that the seller could get. The alternative menu is $\{(1, 0.5)\}$ that yields a revenue of 1.

sumers with low marginal utility are less willing to pay for additional quantity, so they will naturally choose the smaller bundle; however, in this region, these are precisely the consumers with the highest overall willingness to pay. Any price differential thus creates a perverse incentive, causing the seller's most valuable customers to purchase the cheaper option. This mechanism works against revenue generation. The only way to resolve this is to eliminate the incentive to self-select, which means setting the price differential to zero. Consequently, offering a single all-you-can-eat bundle is the optimal strategy. Example 4 illustrates this intuition.

Example 4. *Consider two consumers,* $1, 2$*, with per-unit values* $v_1 = 0.8 < v_2 = 2$ *and maximum demands* $d(v_1) = 1 > d(v_2) = 0.5$*. The preference of these two consumers is such that* $vd(v)$ *increases in* v*:* $v_1 d(v_1) < v_2 d(v_2)$*. Consumer 2 has higher willingness to pay for all quantity while consumer 1 has higher marginal utility for* $q > 0.5$*. The separating menu* $\{(0.5, 0.4), (1, 0.8)\}$ *will lead consumer 2 to purchase* 0.5 *units and a total revenue of* 1.2*. On the other hand, setting a fixed fee of* 0.8 *and allow consumer to consume any quantity will lead to a revenue of* 1.6*. Figure 10 provides a visualization.*

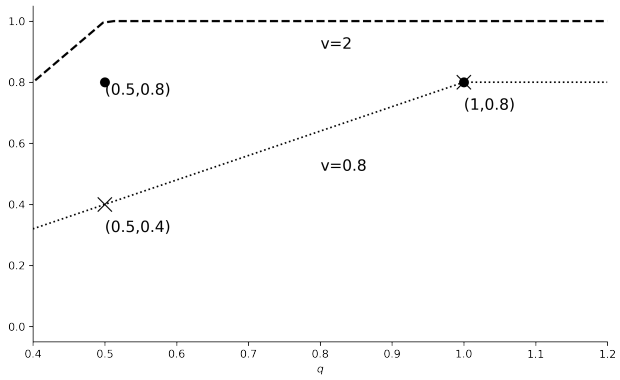


Figure 10: Utility curves for the two consumers (dotted line for consumer 1 and dashed line for consumer 2). The optimal menu is a fixed fee of 0.8. Each consumer consumes their maximum demand.

If $vd(v)$ is nowhere decreasing, then by Proposition 2, the optimal mechanism has only the all-you-can-eat and non-serving region. Corollary 3 states this result.

Corollary 3. *If* $d(\cdot)$ *is decreasing and* $vd(v)$ *is nowhere decreasing, the optimal mechanism is all-you-can-eat for a fixed price* $\tilde{p}$ *where* $\tilde{p} \in \arg\max p\mathbb{P}(vd(v) \geq p)$*.*

The optimality of an all-you-can-eat mechanism is not limited to consumer preferences with decreasing maximum demand. More generally, let $u(\theta, q)$ be the utility function of a consumer with type θ at quantity q and let $\bar{q}$ be the exogenous upper bound on

quantity. If the ranking of types by total utility and by marginal utility is consistent across quantities but reversed (i.e., higher-θ consumers have higher total utility at every q but lower marginal utility), then all-you-can-eat is optimal. Figure 11 illustrates this with an example when consumer preferences satisfy these conditions. Proposition 3 formalizes this result. See Appendix A.3 for its proof.

Proposition 3. *If $u_\theta(\theta, q) \geq 0$ and $u_{q\theta}(\theta, q) \leq 0$ for all q and θ, then optimal mechanism is all-you-can-eat for a fixed price $\tilde{p}$ where $\tilde{p} \in \arg\max p\mathbb{P}(u_\theta(\theta, \overline{q}) \geq p)$.*

Buffet restaurants, cruise and seasonal ski resort passes are common examples of all-you-can-eat pricing. The prevalent subscription models used in AI chat-bots, streaming services, and fitness gyms are also all-you-can-eat pricing. Consumers' preferences in these markets may not exactly satisfy the conditions in Corollary 3 or Proposition 3. However, Corollary 3 and Proposition 3 provide a rationale for the optimality and prevalence of these subscription and all-you-can-eat models: if the customers who value the product most also experience sharply diminishing return from quantity, pooling all the types and offering a single-item menu dominates menus that try to separate types.

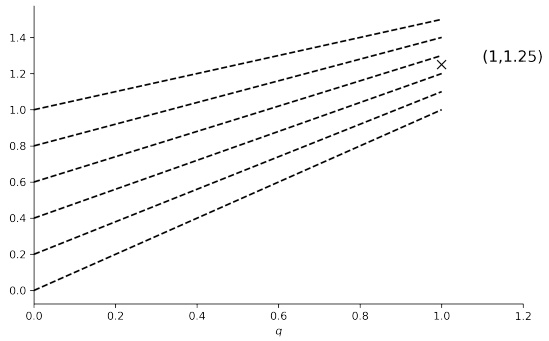


Figure 11: Let $\theta \sim U[0,1]$, $\overline{q} = 1$ and $u(\theta, q) = \theta + (1 - 1/2\theta)q$. Consumers with higher types θ have higher utilities for each q but lower marginal utility for each q. The optimal mechanism here is a fixed fee of 1.25 that allows each consumer to consume up to 1 unit.

Non-monotonic allocation With decreasing maximum demand, consumers with lower per-unit value have higher maximum demand while consumers with lower maximum demand have higher per-unit value. There are instances when consumers in the middle (with v falling in the medium range) have the lowest maximum willingness to pay, $vd(v)$, and thus are the actual 'low type' consumers that are not served in the optimal mechanism. Example 5 illustrates this.

Example 5. *Consider* $v \sim U[0,1]$, $d(v) = 1$ *for all* $v < 0.2$ *and* $d(v) = 0.3$ *for all* $v \geq 0.2$. *The maximum willingness to pay* $vd(v)$ *(dotted line in Figure 12) is high for* $v \in [0.2, 0.4]$ *and for* $v > 0.5$ *and low for* $v < 0.2$ *and* $v \in [0.4, 0.5]$. *The optimal mechanism is to charge* 0.2 *for 1 unit and* 0.15 *for* 0.3 *units. Consumers with* $v \in [0.4, 0.5]$ *do not consume.*

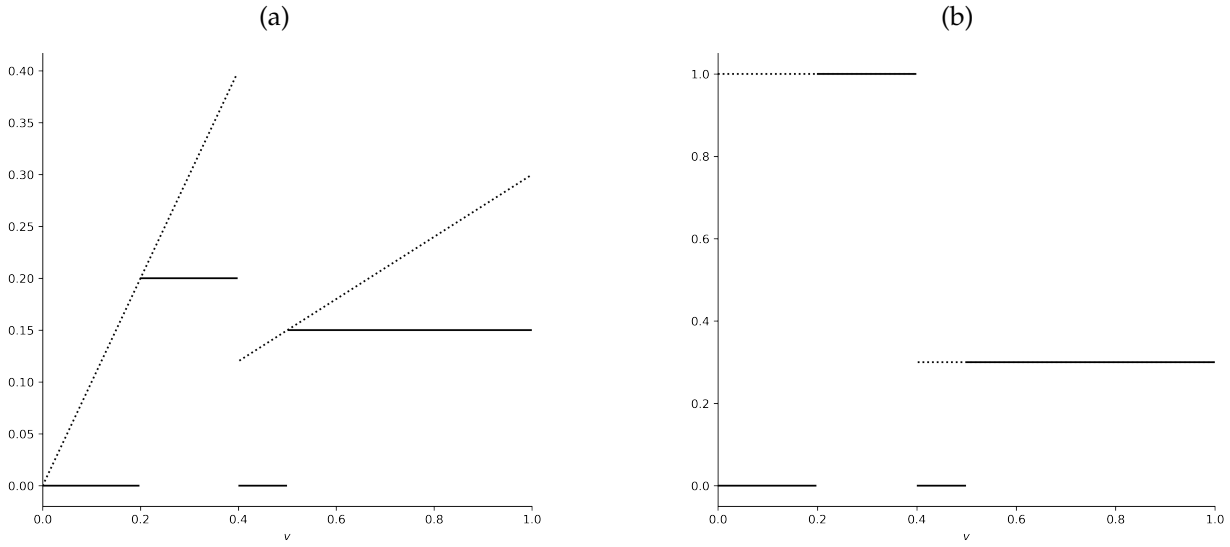


Figure 12: The optimal payment and allocation for Example 5. Panel (a): the optimal payment (solid line) and the maximum willingness to pay $vd(v)$ (dotted line). Panel (b): the optimal allocation (solid line) and the maximum demand $d(v)$ (dotted line). The optimal allocation first increases at $v = 0.2$, then decreases to 0 at $v = 0.4$, and then increases again at $v = 0.5$.

In the special case where the maximum demand $d(v)$ is a step function, as in Example 5, the optimal mechanism can be derived in closed form, avoiding the need to solve (1). The explicit solution is presented in the appendix. [BL: need to write it in the appendix.]

Figure 13 illustrates an example where the optimal pricing has all of the above features.

3.3 Quantity discount or quantity premium?

When consumer preferences satisfy the single-crossing condition, types with higher marginal utility at lower quantities also have higher marginal utility at higher quantities. In this case, the optimal mechanism allocates larger quantities to these higher types. Furthermore, if these types also have greater marginal utility for higher quantities compared to lower types at lower quantities, the optimal marginal price rises with quantity, making

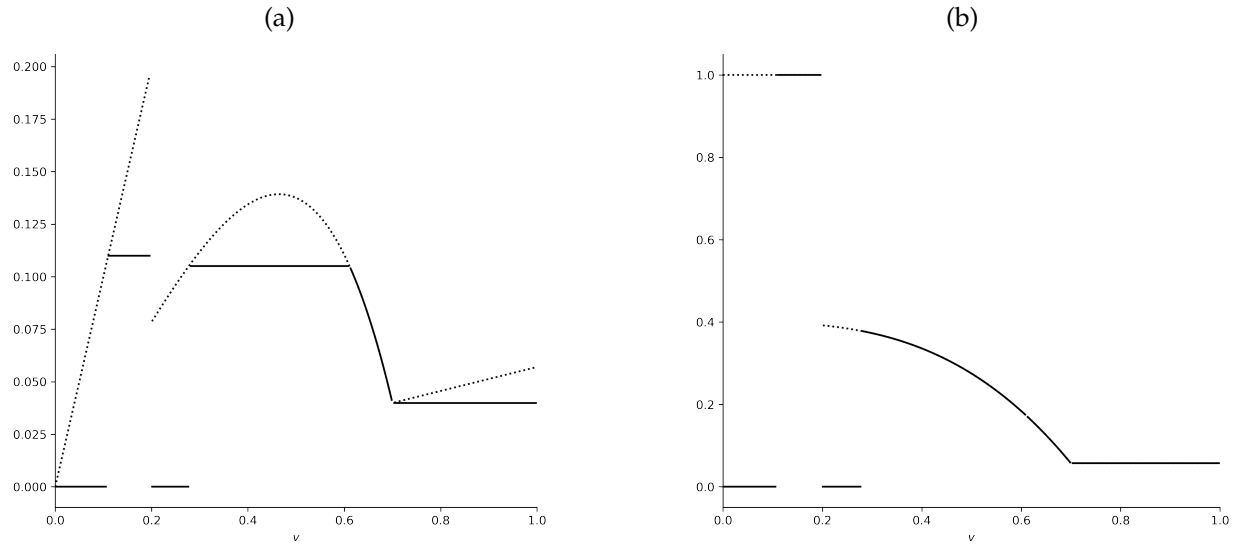


Figure 13: The optimal payment and allocation for $v \sim U[0,1]$ and $d(v) = 1$ for $v < 0.2$, $d(v) = 0.4 - v^3$ for $v \in [0.2, 0.7]$ and $d(v) = 0.057$ for $v > 0.7$. Panel (a): the optimal payment (solid line) and the maximum willingness to pay $vd(v)$ (dotted line). Panel (b): the optimal allocation (solid line) and the maximum demand $d(v)$ (dotted line).

a quantity premium optimal. This is the result in Corollary 1. Conversely, if these types have lower marginal utility at higher quantities compared to lower types at lower quantities, the optimal marginal price declines with quantity, leading to a quantity discount being optimal. This is the result in Maskin and Riley (1984).

When consumer preferences do not meet the single-crossing condition, types with lower marginal utilities at lower quantities may have higher marginal utilities at higher quantities. As a result, the optimal mechanism could allocate larger quantities to these types. However, due to the concavity of the consumer utility function, the marginal utility of these types at higher quantities is still lower than their marginal utility at higher quantities. Thus, the optimal marginal price decreases with quantity, making a quantity discount optimal in this scenario. This is what happens in Section 3.2.

We summarize this observation in Theorem 1 and leave the proof to Appendix A.4.

Theorem 1. *The optimal pricing exhibits quantity premium (respectively, discount) if $d(\cdot)$ is increasing (respectively, decreasing).*

The result that whether a quantity premium or a quantity discount is optimal depends on whether the maximum demand and the per-unit value is positively or negative affiliated extends to the setting when the maximum demand d and per-unit value v are stochastically correlated. We discuss this in Section 4.

4 Stochastic correlation

In this section, we allow the maximum demand d and the per-unit value v to be stochastically dependent. We identify conditions on their stochastic dependence such that the optimal mechanism exhibits a quantity discount or a quantity premium, and we further characterize the optimal pricing schedule in each case.

To this end, we derive the structures of the optimal per-unit price function $t(\cdot)$ or the marginal price function $\tau(\cdot)$, and relate them to the optimal cut-off type $\hat{v}(d)$ for each $d \in [\underline{d}, \overline{d}]$.

4.1 Positive affiliation and quantity premium

Intuitively, when v and d are positively affiliated, the probability that a consumer having a value greater than v conditional on having maximum demand greater than q, is higher than the probability that a consumer having a value greater than v, conditional on having maximum demand greater than $q' < q$. This is the driving force for the increasing marginal price.

To show the quantity premium formally, we begin by considering the relaxed problem of the seller selling each of the q-th unit with the relaxed IC constraint that the consumer of type v will purchase the q-th unit as long as their marginal utility is weakly higher than the marginal price at q. Solving this relaxed problem is equivalent to solving an optimal auction problem for each $q \in [\underline{d}, \overline{d}]$ where a consumer's value v follows the distribution $F(v; d \geq q)$. When v and d are positively affiliated, for any $q > q'$, the distribution $F(v; d \geq q)$ stochastically dominates $F(v; d \geq q')$ in monotone likelihood ratio. Let $\overline{\tau}(\cdot)$ be the optimal marginal price to the relaxed problem, then we have Lemma 2

Lemma 2. *If v and d are positively affiliated, $\overline{t}(\cdot)$ exhibits quantity premium.*

Lemma 3 is straightforward. One can find the formal proofs in Appendix A.6.

Lemma 3. *If $\overline{t}(\cdot)$ exhibits quantity premium, then $\overline{t}(\cdot)$ is the solution to the original problem.*

Proposition 4 follows directly from Lemmas 2 and 3.

Proposition 4. *If v and d are positively affiliated, the optimal pricing exhibits quantity premium.*

Not only we know that the optimal pricing has quantity premium, Lemmas 2 and 3 together also give further characterization of optimal pricing in terms of the marginal prices. Proposition 5 summarizes that.

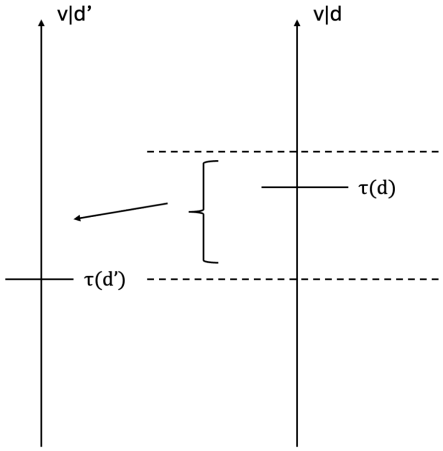


Figure 14: Illustration of the below maximum demand -consumption: consumers with higher maximum demand will optimally consume a lower quantity. The two vertically lines represent the per-unit value v of the consumers of two different maximum demand $d' < d$. Some consumers of values around the marginal price $\tau(d)$, shown as between the dashed lines, will consume a quantity below their maximum demand d.

Proposition 5. *For each $q \in [\underline{d}, \overline{d}]$, $\tau(q)$ satisfies that $1 - F(\tau(q); d \geq q) = \tau(q) f(\tau(q); d \geq q)$.*

Recall that when v and d are perfectly correlated, consumers either get their full maximum demand d or nothing. When v and d are stochastically correlated, this is not the case anymore. When v and d are positively affiliated, the increase in marginal price not only reflects that conditional on having a higher maximum demand, there is higher probability that a consumer has a higher value; but also that some consumers of higher maximum demand optimally choose to consume below their maximum demand. Figure 14 illustrates this.

Optimal substitution If the consumers commit to only consume their maximum demand conditional on consumption, then optimal cut-off type for each quantity q is $\hat{v}(q)$. The optimal pricing in Proposition 4 shows that the cut-off types in the optimal selling mechanism are higher than $\hat{v}(q)$: the seller optimally excludes consumers lower maximum demand d' and lower per-unit value $v \in [\hat{v}(d'), \tau(d')]$, and substitute them with consumers of higher maximum demand $d > d'$ and higher per-unit value $v > \tau(d')$. Similar to the standard models, this under-provision effect does not happen to the highest type– there is no distortion at the top: $\tau(\overline{d}) = \hat{v}(d)$. Corollary 4 summarizes this result.

Corollary 4. *$\tau(q) \geq \hat{v}(q)$ for all $q < \overline{d}$ and $\tau(\overline{d}) = \hat{v}(d)$.*

4.2 Negative affiliation

4.2.1 When quantity discount is not optimal

When v and d are negatively affiliated, some quantity premium may still be revenue-enhancing. Similar to the optimal substitution in Section 4.1, a quantity premium excludes consumers of lower maximum demand d' and lower per-unit value v, and substitutes them with consumers of higher maximum demand $d > d'$ and higher per-unit value v. In a model where all consumers have the same maximum demand, optimal pricing trades off between charging a higher price and attracting more consumers. In this model when consumers have different maximum demands, non-linear pricing, especially a quantity premium, eases this trade-off: the seller can simultaneously charge a high price, but for high quantity bundles, *while* attracting more consumers by setting lower per-unit prices for low quantity bundles. Example 6 demonstrates a case in which a quantity premium increases revenue even when v and d are negatively affiliated.

Example 6. *Consider two possible maximum demands* $d \in \{1, 1.01\}$ *and the conditional distributions of per-unit value* $v|1.01$ $U[0,1]$ *and* $v|1$ $U[0.001, 1.001]$*, so that* v *and* d *are negatively affiliated. Restricting to mechanisms that result in a quantity discount, the optimal menu*[5] *is to offer* 1 *unit at* 0.5005 *and* 1.01 *unit at* 0.505*. The total revenue from this is* 0.50275*. Now consider an alternative menu that offers* 1 *unit at* 0.499 *and* 1.01 *units at* 0.505*. This menu features a quantity premium since the average price for the* 1.01*-unit bundle is higher. Given this menu, consumers of maximum demand of* 1 *unit and per-unit higher than* 0.499 *will purchase the* 1*-unit bundle. Moreover, consumers of maximum demand of* 1.01 *units and per-unit higher than* 0.499 *but lower than* 0.6 *will* also *purchase the* 1*-unit bundle. The consumers of maximum demand of* 1.01 *units and per-unit higher than* 0.6 *will purchase the* 1.01*-unit bundle. The resulting revenue is* $0.502903 > 0.50275$*.*

We summarize this observation in Corollary 5.

Corollary 5. *Even when* v *and* d *are negatively affiliated, there are instances where quantity premium is optimal.*

More interestingly, when d and v are independent, the optimal pricing does not involve any quantity discount. On the other hand, some quantity premium might be optimal.

[5]Conditional on a quantity discount, consumers either purchase their maximum demand or nothing. The problem is equivalent to selling 1 unit to consumers with maximum demand equal to 1 and selling 1.01 unit to consumers with maximum demand equal to 1.01.

Proposition 6. *When d and v are independent, the optimal pricing is either linear with $t(d) = \hat{v}$ for all d, or contains some quantity premium.*

4.2.2 When quantity discount is optimal

When consumers have different maximum demand, lowering the average price for the d-unit bundle has two-opposing effects.[6] On the one hand, it lowers revenue from consumers with maximum demand equal to d and with high per-unit values. Second, it increases the revenue by recapturing back some low per-unit value consumers who would otherwise purchase a lower quantity bundle or make no purchase. Let the marginal change in $t(d)$ be $\Delta t(d)$. Let the boundary consumer be of type v: consumers with type higher than v will purchase the d-unit bundle and with type lower than v will purchase a lower-quantity bundle. For the first effect, the change in revenue is

$$d \cdot \left(\int F(y, d)dy - F(v, d) \right) \cdot \Delta t(d) \tag{2}$$

For the second effect, Let the lower-quantity bundle be a d'-unit bundle. Then v satisfies that

$$v = \frac{t(d)d - t(d')d'}{d - d'}.$$

For a small change $\Delta t(d)$, the change in the boundary type is $\frac{d}{d-d'} \cdot \Delta t(d)$. Also note that the boundary type is indifferent between the d-unit bundle and the d'-unit bundle. The increase in the boundary consumer welfare, $(d - d')v$, goes to the seller's revenue. Hence, the second effect is

$$\underbrace{(d - d')v}_{\text{increase in revenue}} \underbrace{f(v, d)\frac{d}{d - d'} \cdot \Delta t(d)}_{\text{mass of the boundary consumers}} = vf(v, d) \cdot \Delta t(d) \tag{3}$$

Comparing (2) and (3) gives us the first part of the intuition: for a d-unit bundle, it is revenue-enhancing to lower the average price such that the boundary types is $\hat{v}(d)$. So quantity discount follows from that $\hat{v}(d)$ decreases in d. Figure 15 provides an illustration of this intuition for the case when there are only two maximum demands d and d'. We summarize this result in Proposition 7 while the proof is in Appendix A.7. Assump-

[6] There is a potential third effect that it might lower revenue by 'attracting' some low per-unit value consumers with higher maximum demand who would otherwise purchase a higher quantity bundle. It turns out this is not the case under Assumption 3. We show this in the proof in Appendix A.7.

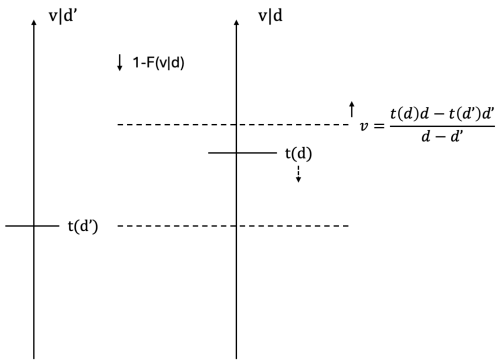


Figure 15: The effects of lowering the average price of the d-unit bundle, $t(d)$, in the presence of another bundle $d' < d$ with $t(d) < t(d')$. Here we assume that the mass of consumers with maximum demand equal to d and d' are both 1 each for neater presentation. v is the 'boundary' type that purchases the d-unit bundle. Lowering in $t(d)$ results in a lower revenue from a mass of $1 - F(v|d)$ consumers, but a higher revenue from consumers with type $v = \frac{t(d)d - t(d')d'}{d - d'}$ 'switching' back from d'-unit bundle to the d-unit bundle.

tion 3 rules out the scenario in Example 6 while Assumption 2 is a technical and convenient assumption to make sure the constraint that aggregate payment is non-decreasing in quantity is not binding.

Assumption 2. *$\hat{v}(d)d$ is non-increasing in d.*

Assumption 3. *For all $d \in [\underline{d}, \overline{d}]$ and $v < \hat{v}(d)$, $(\hat{v}(d) - v)\left(\int F(y,d)dy - F(\hat{v}(d),d)\right) > v\int_{x \geq d, \hat{v}(x) > v} \left(F(\hat{v}(x),x) - F(v,x)\right) dx$*

Proposition 7. *Under Assumption 2 and Assumption 3, if v and d are negatively affiliated, the optimal pricing exhibits quantity discount.*

Given a quantity discount, again, each consumer either consumes their maximum demand or nothing. This structure allows us to write the exact pricing. Proposition 8 states the optimal pricing and we include the formal proof in Appendix **??**.

Proposition 8. *Under Assumption 2 and Assumption 3, if v and d are negatively affiliated, then the optimal pricing is $t(d) = \hat{v}(d)$. All consumers with (v,d) such that $v \geq \hat{v}(d)$ consume d units.*

References

ARAUJO, A. AND H. MOREIRA (2010): "Adverse selection problems without the Spence–Mirrlees condition," *Journal of Economic Theory*, 145, 1113–1141.

ARAUJO, A., H. MOREIRA, AND S. VIEIRA (2015): "The marginal tariff approach without single-crossing," *Journal of Mathematical Economics*, 61, 166–184.

ARMSTRONG, M. (2016): "Nonlinear pricing," *Annual Review of Economics*, 8, 583–614.

CHEN, C.-H., J. ISHIDA, AND W. SUEN (2022): "Signaling under Double-Crossing Preferences," *Econometrica*, 90, 1225–1260.

GOLDMAN, M. B., H. E. LELAND, AND D. S. SIBLEY (1984): "Optimal nonuniform prices," *The Review of Economic Studies*, 51, 305–319.

KWAK, C. (2022): "Screening without Single Crossing," *Available at SSRN 4491074*.

LIU, B. (2025): "Countervailing Incentive in Mechanism Design," Tech. rep., Working Paper. Stanford University, Standford, CA.

MASKIN, E. AND J. RILEY (1984): "Monopoly with incomplete information," *The RAND Journal of Economics*, 15, 171–196.

MUSSA, M. AND S. ROSEN (1978): "Monopoly and product quality," *Journal of Economic theory*, 18, 301–317.

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

SCHOTTMÜLLER, C. (2015): "Adverse selection without single crossing: Monotone solutions," *Journal of Economic Theory*, 158, 127–164.

WILSON, R. B. (1993): *Nonlinear pricing*, Oxford University Press.

A Proofs

A.1 Proof of Proposition 1

In any optimal mechanism $q(v) \leq d(v)$ for all v since setting any $q(v) > d(v)$ worsens the downward IC constraint, does not relax the upward IC constraint and does not improve revenue. Then the problem is the same as the optimal auction problem in Myerson (1981) with the additional constraint that $q(v) \leq d(v)$ for all v. So the seller's problem can be re-written as

$$\max_q \int q(v)\phi(v)dF(v)$$

subject to $q(v)$ is non-decreasing in v and $q(v) \leq d(v)$ for all v. The mechanism in Proposition 1 yield the maximum revenue without the monotonicity constraint, an upper bound to the original problem. Since $d(v)$ is non-decreasing in v, the mechanism in Proposition 1 is also feasible to the original problem and hence optimal.

A.2 Proof of Proposition 2

We begin the proof by showing that the optimal mechanism exhibits price discount. Towards that end, we first construct a sequence of step functions $d^n(\cdot)$ that converges to $d(\cdot)$. Second let $\mathcal{C}$ be the correspondence that maps $d^n(\cdot)$ to the set of IR and IC mechanisms given the maximum demand is $d^n(\cdot)$. We show that $\mathcal{C}$ is continuous. Then by Berge's maximum theorem, the optimal solution to the problem with maximum demand $d^n(\cdot)$ converges to the optimal solution to the problem with maximum demand $d(\cdot)$ and thus the optimal solution to the problem with maximum demand $d(\cdot)$ has price discount.

For every $n \geq 1$, we divide the per-unit value space into 2^n equal intervals: $[v_0, v_1]$, $(v_1, v_2], \ldots, (v_{2^n-1}, v_{2^n}]$, where $v_0 = \underline{v}$, $v_{2^n} = \overline{v}$, $v_k = \underline{v} + \frac{k}{n}(\overline{v} - \underline{v})$. For each $v \in (v_{k-1}, v_k]$, let $d^n(v) = d(v_k)$. So $d^n \to d$ uniformly.

To show that $\mathcal{C}$ is upper hemicontinuous, let $(q^n, T^n)_n$ be a convergent sequence and let $(\tilde{q}, \tilde{T}) = \lim_{n\to\infty}(q^n, T^n)$, we show that $\in \mathcal{C}(d)$. Assume not. Then there exists types $v, v' \in [\underline{v}, \overline{v}]$ such that

$$v \min\{d(v), q(v)\} - T(v) < v \min\{d(v), q(v')\} - T(v').$$

Let N be such that $d^N(v) = d(v)$. Since $(\tilde{q}, \tilde{T}) = \lim_{n\to\infty}(q^n, T^n)$, there exists $N' > N$ large enough such that $(q^{N'}(v), T^{N'}(v))$ and $(q^{N'}(v'), T^{N'}(v'))$ is close enough to $(q(v), T(v))$

and $(q(v'), T(v'))$. Then the mechanism $q^{N'}, T^{N'}$ is also not IC. A constradiction.

To show that $\mathcal{C}$ is lower hemicontinuous, take any $(q, T) \in \mathcal{C}(d)$, we construct a convergent sequence of (q^n, T^n) and show that $(q^n, T^n) \in \mathcal{C}(d^n)$. Given (q, T), for any n and $k \leq 2^n$, let $M = \{(q(v_k), T(v_k))\}$ be the set of menus that consist only bundles offers to $(v_k)_{k \leq 2^n}$. For each $v \in [\underline{v}, \overline{v}]$, let $(q^n(v), T^n(v)) \in \arg\max_{(q,T) \in M} v \min(d^n(v), q) - T$. Clearly, $(q^n, T^n) \to (q, T)$ as $n \to \infty$ and satisfies IR and IC.

Given the price discounts, consumers of per-unit value v either purchase their maximum demand $d(v)$ or purchase nothing. So we search among the price schedules that result in allocations $q(v) \in \{0, d(v)\}$ for all $v \in [\underline{v}, \overline{v}]$. Thus $T(v)$ is non-increasing in v for all $v \in [\underline{v}, \overline{v}]$ and $q(v) > 0$. Now we show that $T(v)$ is either a constant for some interval or $T(v) = vd(v)$ for an interval where $vd(v)$ is non-increasing. Consider the solution to the following relaxed problem:

$$\max_{T(\cdot) \text{ non-increasing}} \int_{T(v) \leq vd(v)} T(v) - cd(v)F(v). \tag{4}$$

For any non-increasing $T(\cdot)$, if there exists an interval $[v_1, v_2]$ such that $T(v) < vd(v)$ for all $v \in [v_1, v_2]$ and $T(V)$ is not a constant, consider the following modification such that $T'(v) = \min\{T(v_1), vd(v)\}$ for all $v \in [v_1, v_2]$. Then $T'(v) > T(V)$ and is still non-increasing. So the solution to (4) satisfies that $T(v)$ is either a constant for some interval or $T(v) = vd(v)$ for an interval where $vd(v)$ is non-increasing.

We now show that solution satisfies IC and hence is the solution to the original problem. First, within each interval where $T(\cdot)$ is a constant, consumers have no incentive to deviate to another type within the interval. Second, within each interval where $T(v) = vd(v)$, consumers have no incentive to deviate to another type within the interval. To see this, for $v' < v$, $v'd(v) - T(v) = v'd(v) - vd(v) < 0$; for $v' > v$, $v' \min\{d(v), d(v')\} - T(v) = v'd(v') - vd(v) < 0$. Across the intervals, for v to deviate to a higher type v', $v \min\{d(v), d(v')\} - T(v') = vd(v') - v'd(v') < 0$; for v to deviate to a lower type v', $v \min\{d(v), d(v')\} - T(v') = vd(v) - T(v') < vd(v) - T(v)$;

A.3 Proof of Proposition 3

Let $M = (q, T(q))_q$ be any menu that consists of more than one quantity bundle. Let θ^M be type of the consumer that has the lowest type and purchases a bundle in M. We show that the menu with a single bundle $(\overline{q}, u(\theta^M, \overline{q}))$ generates more revenue than M. If $u(\theta^M, \overline{q}) \geq T(\overline{q})$, we are done. If $u(\theta^M, \overline{q}) < T(\overline{q})$, there must exists a bundle $(q', T(q')) \in$

M such that

$$u(\theta^M, \overline{q}) - u(\theta^M, q') < T(\overline{q}) - T(q')$$

Since $u_\theta(\theta, q) \geq 0$ and $u_{q\theta}(\theta, q) \leq 0$, for all $\theta' > \theta$,

$$u(\theta', \overline{q}) - u(\theta', q') < T(\overline{q}) - T(q').$$

So all consumers would purchase a bundle other than $\overline{q}$ in M. Assume there exists a bundle q'' with $T(q'') > u(\theta^M, \overline{q})$. Then $T(q'') - T(q') > u(\theta^M, \overline{q}) - u(\theta^M, q') > u(\theta', q'') - u(\theta', q')$. q'' cannot be the preferred bundle than q' to any type $\theta' > \theta^M$. Therefore any menu consists of more than a single bundle of quantity $\overline{q}$ cannot be optimal. The second part of the statement follows from this.

A.4 Proof of Theorem 1

For the quantity premium, using Proposition 1,

$$t(v) = \frac{T(v)}{d(v)} = \underline{v} + \int_{\underline{v}}^{v} 1 - \frac{d(x)}{d(v)} dx$$

For $v' > v \geq \hat{v}$, $q(v') > q(v)$ and

$$\begin{aligned} t(v') = \frac{T(v')}{d(v')} &= \underline{v} + \int_{\underline{v}}^{v} \underbrace{1 - \frac{d(x)}{d(v')}}_{> 1 - \frac{d(x)}{d(v)}} dx + \underbrace{\int_{v}^{\underline{v}} 1 - \frac{d(x)}{d(v')}}_{\geq 0} dx \\ &\geq t(v). \end{aligned}$$

For the quantity discount, using Proposition 2, the quantity discount in the all-you-can-eat region is obvious. For the quantity discount in the all-you-can-pay region, note that the per-unit price is $t(v) = \frac{T(v)}{q(v)} = v$ that increases in v while the quantity allocated $q(v)$ decreases in v.

A.5 Proof of Proposition 6

Proof. We begin by solving the relaxed problem where we replace the IC constraint with (ICr). Then the seller's problem is to optimally sell each q-th unit while the solution is to have $t(d) = \hat{v}$ for all d. Now we just need to show that $t(d) = \hat{v}$ for all d also satisfies the

full IC constraint: for all consumers with type (v, d) and $v > \hat{v}$, they purchase d units at the per unit price $\hat{v}$ and all the other consumers do not purchase anything.

Consider any (v, d) with $v < \hat{v}$. This consumer has no incentive to purchase any $d' > d$. If the consumer purchase $d' < d$, the consumer's payoff is $vd' - \hat{v}d' < 0$. Consider any (v, d) with $v > \hat{v}$. This consumer has no incentive to purchase any $d' > d$. If the consumer purchase $d' < d$, the consumer's payoff is $vd' - \hat{v}d' < vd - \hat{v}d$. □

A.6 Proofs of Lemma 3

If $\bar{t}(\cdot)$ exhibits quantity premium, the marginal price increases in quantity. If a consumer's marginal utility (v) is higher than the marginal price for the qth unit, their marginal utility (v) is higher than the marginal price for all $q' < q$th unit. Hence the solution to the problem with the relaxed IC constraint also satisfies the full IC constraint and the IR constraint.

A.7 Proof of Proposition 7

Let $t(q)$ be the average price for a quantity q. Given $t(\cdot)$, for each quantity q, there exists a cut-off type v_q such that any consumer with type (v, q) where $v > v_q$ purchases the q-unit bundle. We begin by determining the optimal v_q.

For $q = \overline{d}$, lowering the $t(\overline{d})$ for the $\overline{d}$-unit bundle has two-opposing effects. On the one hand, it lowers revenue from consumers with maximum demand equal to $\overline{d}$ and with high per-unit values. Second, it increases the revenue by recapturing back some low per-unit value consumers who would otherwise purchase a lower quantity bundle or make no purchase. Let the marginal change in $t(\overline{d})$ be $\Delta t(\overline{d})$. For the first effect, the change in revenue is

$$\overline{d} \cdot \left(\int F(y, \overline{d}) dy - F(v_{\overline{d}}, \overline{d}) \right) \cdot \Delta t(\overline{d}) \tag{5}$$

Assume that $v_{\overline{d}}$ purchases d'-unit bundle. Then $v_{\overline{d}}$ satisfies that

$$v_{\overline{d}} = \frac{t(\overline{d})\overline{d} - t(d')d'}{\overline{d} - d'}.$$

For a small change $\Delta t(\overline{d})$, the change in the boundary type is $\frac{\overline{d}}{\overline{d} - d'} \cdot \Delta t(\overline{d})$. Also note that the boundary type is indifferent between the $\overline{d}$-unit bundle and the d'-unit bundle.

The increase in the boundary consumer welfare, $(\overline{d} - d')v_{\overline{d}}$, goes to the seller's revenue. Hence, the second effect is

$$\underbrace{(\overline{d} - d')v_{\overline{d}}}_{\text{increase in revenue}} \underbrace{f(v_{\overline{d}}, \overline{d})\frac{\overline{d}}{\overline{d} - d'} \cdot \Delta t(\overline{d})}_{\text{mass of the boundary consumers}} = v_{\overline{d}} f(v_{\overline{d}}, \overline{d}) \cdot \Delta t(\overline{d}) \tag{6}$$

Comparing (5) and (6), $v_{\overline{d}} = \hat{v}(\overline{d})$.

For $d < \overline{d}$, Assumption 3 ensures that any $v_d < \hat{v}(d)$ is not optimal. For $v_d \geq \hat{v}(d)$, note that $\hat{v}(d') < \hat{v}(d)$ for any $d' > d$. Thus $v_d > \hat{v}(d')$ for any $d' > d$. Thus lowering the $t(d)$ for the d-unit bundle while $v_d \geq \hat{v}(d)$ again has only two-opposing effects as described in (5) and (6). The potential third effect of lowering revenue by 'attracting' some consumers with maximum demand $d' > d$ would never happen as $v_d > \hat{v}(d')$.