Martimort, David; Stole, Lars A.

Article

Participation constraints in discontinuous adverse selection models

Theoretical Economics

Suggested Citation: Martimort, David; Stole, Lars A. (2022) : Participation constraints in discontinuous adverse selection models, Theoretical Economics, ISSN 1555-7561, The Econometric Society, New Haven, CT, Vol. 17, Iss. 3, pp. 1145-1181, https://doi.org/10.3982/TE3030

Theoretical Economics 17 (2022), 1145–1181 1555-7561/20221145

Participation constraints in discontinuous adverse selection models

David Martimort
Paris School of Economics and EHESS

Lars A. Stole
Booth School of Business, University of Chicago

We present a set of necessary and sufficient conditions for a class of optimal control problems with pure state constraints for which the objective function is linear in the state variable but the objective function is only required to be upper semicontinuous in the control variable. We apply those conditions to economic environments in contract theory where discontinuities in objectives prevail. Examples of applications include nonlinear pricing of digital goods and nonlinear pricing under competitive threat.

Keywords. Optimal control, nonsmooth optimization, convex analysis, type-dependent participation constraints, principal–agent models.

JEL classification. D82, D86.

1. Introduction

The textbook treatment of optimal screening contracts typically takes the agent's outside option as a fixed constant, independent of type.[1] More complex settings, which allow for competition by rival principals, nontrivial ownership rights on productive assets, and type-dependent fixed costs require a departure from this restrictive assumption. Lewis and Sappington (1989) initiated the seminal study of screening contracts in this more general setting by constructing the solution to a class of optimal control problems with type-dependent participation constraints. This class of problems was further enriched by Maggi and Rodriguez-Clare (1995) with the most general statement of the problem and its solution culminating in the analysis offered by Jullien (2000).

David Martimort: martimort.david@gmail.com
Lars A. Stole: lars.stole@chicagobooth.edu
We are especially thankful to John Birge for many helpful discussions. We also thank Simon Board, three referees for very useful comments and suggestions, and seminar participants at the 2019 Summer Meetings of the North American Econometric Society (Seattle) and the European Econometric Society (Manchester). A less general version of the main theorem in this paper appeared in an earlier, unpublished note, *"Necessary and sufficient conditions for nonsmooth linear-state optimal control problems"* (2009). The present paper provides the more general result along with a geometric intuition for its proof and several relevant applications. The usual disclaimer applies.

[1] See Laffont and Martimort (2002, Chapter 3) for instance.

These techniques have allowed modelers to apply the optimal contracting paradigm to more general economic contexts, unveiling new features of optimal contracts. Applications have spread through many fields including the design of nonlinear prices under the threat of bypass (Curien, Jullien, and Rey (1998), Biglaiser and Mezzetti (1993)), competition in nonlinear prices (Martimort and Stole (2009), Stole (1995, 2003), Calzolari and Denicolo (2013)), trade policy in open economies Brainard and Martimort (1997), regulation of privately-owned monopolies Caillaud (1990), and optimal contracting under liability constraints Ollier and Thomas (2013) and in dynamic contracting environments Deb and Said (2015), to name a few examples.

Unfortunately, the existing techniques also have their own limits. In particular, the need for tractability has led authors to restrict their analysis to economic environments which are sufficiently *smooth*. In many circumstances, such as when firms face nontrivial fixed costs or sunk investments, the environment is inherently discontinuous. In other settings, such as when principals compete against one another using payment schedules, equilibria may emerge, which exhibit discontinuities in each player's payoff function Martimort and Stole (2015). More broadly, discontinuities in principal-agent problems may directly come from the surplus function in the principal's objective (Section 3.1) or, in a more subtle manner, from nondifferentiability in how the agent's information rent depends on some control variables (Section 3.2). In such cases, we are left uncertain about the consequences of discontinuities for optimal contracts and the generality of results when environments (and equilibria) are not assumed to be smooth *a priori*. Important economic insights may go unnoticed because of such restricted attention. Developing the required techniques for *nonsmooth* environments and showing how they apply in practice are the purposes of this paper.

First, we present a set of necessary and sufficient conditions for a class of optimal control problems with pure state constraints and an objective function (linear in the state variable) that may exhibit kinks or discontinuities in the control variable. Second, we apply these techniques to quite natural contracting environments where existing techniques have previously restricted focus. Examples of applications include nonlinear pricing of digital goods and nonlinear pricing in the presence of competitive threat.

Section 2 presents our main result: A theorem that characterizes solutions to optimal control problems in which the objective function is only required to be upper semicontinuous. This theorem builds on earlier work by Vinter and Zheng (1998), but refines its application to the case of quasilinear objectives, which is prevalent in contract theory. While Vinter and Zheng (1998) focus on necessary conditions for optimality, we prove that these conditions are also sufficient in our context. We also discuss to what extent this theorem extends the existing literature and especially the work by Jullien (2000) under much weaker conditions. Section 3 develops applications of our framework, deriving economic insights that would not have been available without the use of new techniques. The contains the proofs of our theorem and of the various results related to our economic applications, but also provides a brief overview of optimization techniques in nonsmooth environments, illustrating the geometric intuition for our main theorem.

2. The theorem

We will consider control problems in which the state variable, u, is restricted to be an absolutely continuous function on the interval $\Theta = [\underline{\theta}, \overline{\theta}]$; $\mathrm{AC}(\Theta, \mathbb{R})$ denotes the set of such functions. As a motivation, in the context of principal-agent models the state variable is typically the agent's information rent as a function of his type. In such settings, incentive compatibility naturally implies absolute continuity.[2] We focus attention on problems in which that state variable must satisfy a nonnegativity constraint:

$$u(\theta) \geq 0 \quad \forall \theta \in \Theta \equiv [\underline{\theta}, \overline{\theta}]. \tag{1}$$

Using again our motivation of principal-agent problems, the nonnegativity constraint (1) corresponds to a participation constraint that ensures the agent will accept an offer rather than take an outside option, which is here normalized at zero. When the state variable u is both absolutely continuous and nonnegative, it is said *admissible*.

We are interested in the following pure-state control program:

$$(\mathcal{P}): \quad \underset{u \in \mathrm{AC}(\Theta, \mathbb{R})}{\text{Maximize}} \int_{\underline{\theta}}^{\overline{\theta}} \big(s\big(\theta, \dot{u}(\theta)\big) - u(\theta) f(\theta)\big)\, d\theta \quad \text{subject to (1)},$$

where we use the standard control-theoretic notation of $\dot{u}(\theta)$ to denote the derivative of $u(\theta)$ with respect to θ.

Although expressed in abstract terms at this stage, readers accustomed with the principal-agent literature will recognize the structure of such problem. The integrand features the familiar rent-efficiency trade-off. Below, we will push this analogy even further by demonstrating how principal-agent models expressed in more traditional terms can be transformed so as to apply the general methodology we now present. Sections 3.1 and 3.2 provide explicit examples of this transformation.

On the technical side, we only assume that the surplus function $s(\theta, v)$ (here expressed in terms of a control variable v) is an upper semicontinuous function of v for all θ, bounded from above, and that $f(\theta)$ is a positive, bounded function giving rise to an absolutely-continuous definite integral $F(\theta) \equiv \int_{\underline{\theta}}^{\theta} f(\theta)\, d\theta$. Without loss of generality, we normalize f such that $F(\overline{\theta}) = 1$, allowing us to interpret F as a continuous probability distribution and f as its associated density for our applications. We also make a minimal technical assumption that $s(\cdot, \cdot)$ is $\mathcal{L} \times \mathcal{B}$-measurable, where $\mathcal{L}$ denotes the set of Lebesgue measurable subsets of Θ and $\mathcal{B}$ is the set of Borel measurable subsets of $\mathbb{R}$.

We define the integrand for program $(\mathcal{P})$ as $L(\theta, u, \dot{u}) \equiv s(\theta, \dot{u}) - uf(\theta)$. We should make clear that the key restriction we have placed on $(\mathcal{P})$ is that, for any θ and $\dot{u}$, the maximand is a linear function of the state variable u. As we will see below in our applications, this linearity is found in a number of economic problems, especially in contract theory where agents are risk-neutral and payoffs are linear in money. The function $s(\theta, \dot{u})/f(\theta)$ can there be viewed as a surplus function while u is the share of this surplus that is captured by the agent—his information rent.

[2] See Milgrom and Segal (2002) and Carbajal and Ely (2013). As an example, if the agent's utility is continuously differentiable in type with a uniformly-bounded derivative, then the agent's indirect utility function is necessarily Lipschitz continuous (and therefore absolutely continuous).

The linearity restriction is the primary source of many sharp results in the analysis that follows, including the ability for us to relax the continuity of s, to characterize the solution by means of a simple generalized gradient condition, and to verify that necessary conditions for optimality are also sufficient. Indeed, nonsmooth techniques are particularly useful if one can find the solutions of such control problems as pointwise optima. This is where the assumption of linearity of the maximand in u provides purchase. Linearity allows for such a pointwise simplification—a well-known result in familiar quasilinear screening models (Myerson (1981, Lemma 3), Baron and Myerson (1982, Lemma 2), Laffont and Martimort (2002, Chapter 3)).[3] Complications arise if the agent's reservation utility is type-dependent, but the basic intuition remains. Linearity allows us to separate incentive and participation concerns[4] from any nonsmoothness of the surplus function, the latter of which is addressed by using the super-differential of the concave envelope of $s(\theta, v)$ in place of the gradient.

To better isolate the role of nonsmoothness, consider the case where the integrand is reduced to $L(\theta, u, v) \equiv s(\theta, v)$. In the contract theory applications below, this case would correspond to a scenario of complete information in which the principal can fully extract the agent's rent. The optimization problem so constructed can be solved pointwise by means of standard techniques for nonsmooth problems.[5] Any solution, $v^*(\theta)$, must satisfy the following pair of conditions:

$$\overline{\text{co}}(s)\big(\theta, v^*(\theta)\big) = s\big(\theta, v^*(\theta)\big) \quad \text{and} \quad 0 \in \partial\overline{\text{co}}(s)\big(\theta, v^*(\theta)\big), \tag{2}$$

where $\overline{\text{co}}(s)(\theta, v)$ is the concave majorization of the function s over v evaluated at (θ, v), and $\partial_v\overline{\text{co}}(s)(\theta, v)$ is the set of supported gradients of the majorized function evaluated at (θ, v). This is a nonsmooth generalization of the familiar first order necessary condition for optimality. Intuitively, the maximum of an upper semicontinuous function must also coincide with its concave majorization (the first part of (2)). If the maximand is differentiable at the maximizing point, then the derivative is necessarily zero; if it is not differentiable, it must nonetheless support a zero gradient (the second part of (2)). We briefly review this idea in an Appendix.[6] The solution to our original program $(\mathcal{P})$ will differ from v^* in (2) as a result of the addition of the linear term $-uf(\theta)$ in the Lagrangian. Condition (5) below indicates how the solution needs to be modified. As familiar from the principal-agent literature, $u(\theta)$ will have to maximize a virtual surplus function obtained by combining the impacts of participation and incentive constraints.

Heuristic Approach. Following Jullien (2000), one might address problem $(\mathcal{P})$ as follows. First, one could add a Lagrange multiplier $\mu(d\theta)$ to the participation constraint (1)

[3]When quasilinearity is not assumed, this familiar trick no longer works as, for instance, in the well-known model of the optimal taxation due to Mirrlees (1971).

[4]Incentive and participation concerns are captured by the term $F(\theta) - \overline{\gamma}(\theta)$ in the optimality condition (5) below.

[5]See Section A below for details.

[6]Appendix A provides a brief discussion and survey of nonsmooth, convex analysis. Because we are focused on maximization, our tools rely on concave majorizations (i.e., the minimal concave envelope of a function) rather than convex minorizations. Likewise, we are interested in the set of gradients of a concave function (superdifferentials) rather than the set of gradients of a convex function (subdifferentials).

with the complementarity slackness condition

$$\mu(d\theta) = 0 \quad \text{if } \theta \in \{\tilde{\theta} \mid u(\tilde{\theta}) > 0\}.$$

Second, one could then form a Lagrangian as

$$\int_{\underline{\theta}}^{\overline{\theta}} \big(s\big(\theta, \dot{u}(\theta)\big) - u(\theta) f(\theta)\big)\, d\theta + \int_{\underline{\theta}}^{\overline{\theta}} u(\theta)\mu(d\theta).$$

Third, with a simple integration by parts, the integrand could be written as

$$\int_{\underline{\theta}}^{\overline{\theta}} \big(s\big(\theta, \dot{u}(\theta)\big) + \big(F(\theta) - \overline{\gamma}(\theta)\big)\dot{u}(\theta)\big)\, d\theta$$

where the adjoint function $\overline{\gamma}(\theta) = \int_{[\underline{\theta},\theta)} \mu(d\tilde{\theta})$, is right continuous, strictly increasing at points where (1) is binding, and satisfies the boundary conditions $\overline{\gamma}(0) = 0$, $\overline{\gamma}(1) = 1$. These characteristics allow us to identify $\overline{\gamma}$ with a distribution function. Finally, pointwise optimization implies the optimality condition

$$\dot{\overline{u}}(\theta) \in \arg\max_{v}\, s(\theta, v) + \big(F(\theta) - \overline{\gamma}(\theta)\big)v.$$

Two difficulties arise with this simplistic approach. The first one is merely technical and puts conditions on the Lagrange multiplier. Indeed, integrating by parts requires that $\mu(d\theta)$ lies in the dual space of nonnegative functions in $\mathrm{AC}(\Theta, \mathbb{R})$; i.e., $\mu(d\theta)$ must be the "derivative" of a function of bounded variation. The second difficulty is that, once we proceed to pointwise optimization, we may have to deal with a nonsmooth objective since $s(\theta, v)$ is only required to be upper semicontinuous. The optimality conditions have to be expressed by means of tools imported from nonsmooth, convex analysis.

Main Result. We now present our main result for this class of problems.

Theorem 1. *$\overline{u}$ is a solution to program $(\mathcal{P})$ if and only if $\overline{u}$ is admissible and there exists a probability measure μ defined over the Borel subsets of Θ with an associated adjoint function, $\overline{\gamma} : \Theta \to [0, 1]$, defined by $\overline{\gamma}(\underline{\theta}) = 0$ and*

$$\overline{\gamma}(\theta) = \int_{[\underline{\theta},\theta)} \mu(d\tilde{\theta}), \quad \textit{for } \theta > \underline{\theta},$$

such that the following conditions are satisfied:

$$\int_{\underline{\theta}}^{\overline{\theta}} \overline{u}(\tilde{\theta})\mu(d\tilde{\theta}) = 0, \tag{3}$$

$$\overline{\mathrm{co}}(s)\big(\theta, \dot{\overline{u}}(\theta)\big) = s\big(\theta, \dot{\overline{u}}(\theta)\big) \quad \textit{for a.e. } \theta \in \Theta, \tag{4}$$

$$0 \in F(\theta) - \overline{\gamma}(\theta) + \partial_v \overline{\mathrm{co}}(s)\big(\theta, \dot{\overline{u}}(\theta)\big) \quad \textit{for a.e. } \theta \in \Theta. \tag{5}$$

The conditions in Theorem 1 are similar to those of Theorem 1 in Jullien (2000). In both theorems, necessary and sufficient conditions are stated in terms of a probability measure, which serves to express a "complementary slackness condition" (3) and a first-order optimality condition (5). Moreover, both theorems use a similar condition to establish the continuity of $\dot{\overline{u}}(\theta)$ in the solution to $(\mathcal{P})$. In contrast, the theorem in Jullien (2000) relies on results from Seierstad and Sydsaeter (1987, Theorems 2 and 3, Chapter 5), which use the stronger hypothesis that $s(\theta, v)$ is twice continuously differentiable in v. Our contribution is to demonstrate the force and the broader validity of these conditions for problems with integrands that are only upper semicontinuous through the use of nonsmooth analysis.

As in Jullien (2000), the measure $\mu(d\theta)$ stems for the shadow cost of the participation constraint (1) around θ. The adjoint function $\overline{\gamma}(\theta)$ can thus be interpreted as the sum of these shadow costs for all inframarginal types. It is thus nondecreasing and constant on any open interval where the participation constraint is slack. Replacing the right-hand side of (1) uniformly by $\epsilon < 0$ for all $\tilde{\theta} \leq \theta$ would relax the optimization problem and increase its value by $\overline{\gamma}(\theta)\epsilon$.

The adjoint function $\overline{\gamma}$ so constructed is right continuous. Since the probability measure μ may have mass points where the participation constraint begins to bind, $\overline{\gamma}(\theta)$ may have upward jumps at such points. This possibility may only arise at a countable number of points since any increasing function is almost everywhere differentiable.

We now investigate under which conditions the optimal solution remains continuously differentiable.

Proposition 1. *If*

$$\mathcal{V}(\theta, \sigma) \equiv \arg\max_{v \in \mathbb{R}} s(\theta, v) + \big(F(\theta) - \sigma\big)v \tag{6}$$

is single-valued and continuous over the domain $(\theta, \sigma) \in \Theta \times [0, 1]$, *then the solution* $\overline{u}$ *to* $(\mathcal{P})$ *is continuously differentiable.*

That $\mathcal{V}(\theta, \sigma)$ is single-valued and continuous is implied by strict concavity of $s(\theta, \cdot)$. It is also implied by the weaker condition in Jullien (2000, Assumption 2) that $s(\theta, v) - (\sigma - F(\theta))v$ is strictly quasiconcave in v for any $\sigma \in [0, 1]$. Together with the stronger hypothesis that $s(\theta, v)$ is twice continuously differentiable in v, Lemma 7 in Jullien (2000, p. 32) then provides a smooth version of (5) by means of a first-order condition, namely

$$F(\theta) - \overline{\gamma}(\theta) + \frac{\partial s}{\partial v}\big(\theta, \dot{\overline{u}}(\theta)\big) = 0 \quad \text{a.e.}$$ [7]

Proposition 1 is more general since it allows for the possibility that $s(\theta, v)$ fails to be continuous in v at its maximum.

That $\overline{u}$ is continuously differentiable captures the fact that often in applications the optimal control, say output in a principal-agent context, is itself continuous. Examples abound in the literature where continuity is not optimal if the virtual surplus is

[7] Galbraith and Vinter (2004) provide also alternative conditions ensuring Lipschitz-continuity of the optimal control.

not strictly concave. Models of bypass and regulation under the threat of entry due to Caillaud (1990), Laffont and Tirole (1990), and Curien, Jullien, and Rey (1998)), as an illustration, feature discontinuities in the optimal control because the bypass technology entails a fixed cost. In these papers, the authors generally deal with the discontinuities by using details specific to the setting to guess where the binding participation constraints lie (sometimes a complex task in itself), then constructing the agent's profile of information rent given the conjectured constraint set, and finally constructing the principal's nonconcave virtual surplus given the agent's conjectured rent profile—all before proceeding to optimization and confirming that the original conjecture was correct. This approach does not provide much guidance in settings with nonconcavities, especially when the discontinuities arise exactly where the marginal type's participation constraint binds.[8] In such a case, participation constraints truly interact with nonconvexities. In contrast to previous papers, our approach is more direct: (i) construct the concave envelope of s and (ii) compute the adjoint and indirect utility functions, which satisfy the first-order condition and complementary slackness. While this approach still entails jointly solving for two objects, $\overline{\gamma}$ and $\overline{u}$, which can be a complicated task, it is considerably more methodical.

A More Primitive Statement of the Problem. Principal-agent problems with type-dependent participation constraints as studied in the path-breaking works of Lewis and Sappington (1989), Maggi and Rodriguez-Clare (1995), and Jullien (2000) are often expressed under the form $(\mathcal{P}')$ below so as to make the nature of the agent's outside option, $\hat{U}(\theta)$, and its associated participation constraint more explicit:

$$(\mathcal{P}'): \quad \underset{U\in \mathrm{AC}(\Theta,\mathbb{R}),q}{\text{Maximize}} \int_{\underline{\theta}}^{\overline{\theta}} \big(\tilde{s}\big(\theta, q(\theta)\big) - U(\theta)\big) f(\theta)\, d\theta$$

$$\text{subject to } \dot{U}(\theta) = g\big(q(\theta), \theta\big) \text{ a.e., and } U(\theta) \geq \hat{U}(\theta) \text{ for all } \theta \in \Theta.$$

The control variable q, which is assumed to be measurable, is generally interpreted as a quantity vector that belongs to a feasible set $\mathcal{Q} \subseteq \mathbb{R}^k$.[9] The primitive surplus function $\tilde{s}(\theta, q)$ is defined over $\mathcal{Q}$.[10] The differential equation that defines $\dot{U}(\theta)$ immediately

[8]Another source of possible discontinuities comes when the types distribution has mass points; an assumption that we have ruled out for simplicity. See Lewis and Sappington (1993) and Cremer, Khalil, and Rochet (1998) for applications to information gathering where a mass point of agents remains uninformed; Hellwig (2010) provides a more general treatment.

[9]Incentive compatibility requires additional monotonicity conditions to hold. For instance, if $g_\theta(q, \theta) \geq 0$ for all (q, θ), then incentive compatibility requires $q(\theta)$ to be nondecreasing in θ. Often such monotonicity conditions are handled in the literature by adding one state variable and an associated law of motion (see Guesnerie and Laffont (1984)). To illustrate, provided that $q(\theta)$ is absolutely continuous, one can add a new control $p(\theta) = \dot{q}(\theta)$ a.e., and impose the constraint $p(\theta) \geq 0$. While this introduces multiple dimensions to the state space, we note that our analysis can readily be extended to multidimensional settings because Theorem 3 from Vinter and Zheng (1998), used in Appendix B in our one-dimensional setting, applies to multidimensional state variables. That said, this approach has limits. In a nonsmooth framework where discontinuities are pervasive, imposing that $q(\theta)$ is absolutely continuous may be excessive. Ironing techniques in the spirit of Myerson (1981) and Toikka (2011) could be powerful in this context but their development in the nonsmooth case lies outside the scope of this paper.

[10]Note that the domain of definition of q can be easily included into the objective to fit with the formalism of Theorem 1 if we set $\tilde{s}(\theta, q(\theta)) = -\infty$ for $q \notin \mathcal{Q}$.

follows from the well-known envelope condition for incentive compatibility. The participation constraint $U(\theta) \geq \hat{U}(\theta)$ allows the agent's outside option to vary by type.

The program $(\mathcal{P}')$ can easily be transformed into the canonical program $(\mathcal{P})$ we explore if $\hat{U}(\cdot)$ is differentiable almost everywhere. To illustrate, set $u(\theta) = U(\theta) - \hat{U}(\theta)$ and define

$$s(\theta, v) = \max_{q \in \mathcal{Q}} \big\{ \tilde{s}(\theta, q) f(\theta) \text{ s.t. } v = g(q, \theta) - \dot{\hat{U}}(\theta) \big\}.$$

This reduction is particularly easy when g is itself a bijection between q and $\dot{u}$ for all θ, which implies that the control q can be expressed as a function of (v, θ), namely $q(\theta) = g_q^{-1}(v + \dot{\hat{U}}(\theta), \theta)$. In that case, substitution yields

$$s(\theta, v) = \tilde{s}\big(\theta, g_q^{-1}\big(v + \dot{\hat{U}}(\theta), \theta\big)\big) f(\theta),$$

and $\mathcal{P}'$ reduces to $\mathcal{P}$. While it is possible that the function $s(\theta, v)$, obtained as a maximum over all controls that generate the same derivative $\dot{u}(\theta)$, may be smoother than $\tilde{s}(\theta, q)$, it will typically fall short of satisfying the twice continuous differentiability requirement in Jullien (2000).

3. Applications

This section shows the broad applicability of our approach by highlighting applications that require nonsmooth analysis. Our first application (Section 3.1) deals with nonlinear pricing of a digital good under the threat of competition by a low-quality fringe. Because the provider of a high-quality good needs to build capacity for extra services in discrete bundles, the cost function is discontinuous. This model illustrates an interesting intricacy. Avoiding the fixed cost requires leaving additional rent to loyal customers since, otherwise, they would switch to a competitive fringe. The discontinuity from the fixed cost of additional capacity thus determines the subset of types for which the participation constraint is binding.

Our second example, developed in Section 3.2, is another model of nonlinear pricing where a buyer may split his purchases between an incumbent firm and a competitive fringe. The fringe has limited capacity but sells a perfect substitute to the incumbent's product. We show that this possibility introduces a discontinuity in the surplus function simply because the buyer's rent has a different slope depending upon whether or not he purchases from the fringe. The discontinuity is endogenously derived from demand considerations.

3.1 *Nonlinear pricing by a digital firm*

Pricing for digital products, including internet services, online trading and advertising services, is complex. The first source of complexity comes from the specific cost structure of those goods. For an infrastructure of a given size, the marginal cost of service is zero while supramarginal blocks of infrastructure must be added discretely to satisfy higher demand. The second source of complexity comes from the fact that competing

firms might be quite different in their ability to expand infrastructure as required by such demand increase.[11]

The screening literature on this topic is sparse, probably because of the technical difficulty that comes when analyzing the jump discontinuities that arise with discrete capacity costs. Spulber (1993) and Thomas (2001) have studied nonlinear pricing when a monopolist faces a fixed capacity constraint in more general contexts. On top of the usual information distortions, the optimal consumption profile depends on the shadow cost of the capacity constraint—a result we quickly review below. For digital products, Huang and Sundararajan (2011) follow a similar path and determine the set of types who are bunched at the capacity level. Little, if anything, is known about optimal nonlinear pricing for a dominant firm with discrete costs that faces a competitive fringe. As a starting point, suppose that there is no cost to building extra services, so the logic of continuous screening models under the threat of competition applies (e.g., Champsaur and Rochet (1989), Stole (1995)). In this case, the ability of the dominant firm (the "monopolist") to screen consumers and extract their information rent by distorting consumption downwards is limited by the consumer's option to buy from the competitive fringe. When the monopolist must pay a fixed cost for additional production, however, an interesting tension appears between charging high prices to extract information rents and avoid additional capacity costs on the one hand, and inducing more consumers to switch from the fringe on the other.

Technology and Demand. A monopolist sells a digital product. We consider a simple model of cost: marginal cost is zero up to one unit, but to supply more, extra capacity must be built at cost k. The firm's cost function thus exhibits an upward jump discontinuity at $q = 1^+$, namely $C(q) = k\delta_{q>1}$. Let the set of feasible outputs be $\mathcal{Q} = [0, \overline{Q}]$ where $\overline{Q}$ is finite but sufficiently large to ensure interior solutions under all the circumstances we consider below.

On the demand side, there is one consumer with a valuation for q units of services equal to

$$(S - \theta)q - \frac{q^2}{2}. \tag{7}$$

The demand shock θ is a nonobservable heterogeneity parameter, uniformly distributed on $\Theta = [\underline{\theta}, \overline{\theta}]$ with $\overline{\theta} - \underline{\theta} = 1$. The parameter S reflects the quality of the monopolist's service.

Monopoly. Under complete information, the monopoly extracts all consumer's surplus. Because of the cost discontinuity, the bilateral surplus so obtained is

$$(S - \theta)q - \frac{q^2}{2} - k\delta_{q>1}.$$

Note that bilateral surplus is discontinuous, exhibiting a downward jump at $q = 1$, and its concave envelope consists of a linear segment for $q \in [1, 1 + \sqrt{2k}]$. Following the approach for solving nonsmooth problems that is outlined in Appendix A, we find that

[11] See Huang and Sundararajan (2011).

the complete information consumption level $q^{fb}(\theta)$ satisfies

$$q^{fb}(\theta) = \begin{cases} 1 & \text{if } \theta \in [S-1-\sqrt{2k}, S-1] \\ S-\theta & \text{otherwise.} \end{cases}$$

Throughout, we assume that $S-\overline{\theta} > \overline{\theta}-\underline{\theta}$ and $(S-1-\sqrt{2k}, S-1) \subsetneq [\underline{\theta}, \overline{\theta}]$. The former implies $S > \overline{\theta}$, which ensures that it is efficient for all types to consume. We make the stronger assumption that $S > 2\overline{\theta}-\underline{\theta}$ (used below) to ensure that a monopolist (absent fringe considerations) would find it optimal to sell to all types. The latter assumption implies it is efficient for a positive measure of consumers to choose $q=1$. Intermediate types in $[S-1-\sqrt{2k}, S-1]$ are bunched together at $q^{fb}(\theta)=1$ and consumption discontinuously jumps downwards at $\theta_0 \equiv (S-1-\sqrt{2k})$.

We now turn to the case where θ is private information for the consumer. Let $U(\theta)$ be the *informational rent* (indirect utility) when offered a nonlinear price $T(q)$[12]

$$U(\theta) \equiv \max_{q\in\mathcal{Q}} (S-\theta)q - \frac{q^2}{2} - T(q).$$

$U(\theta)$ is absolutely continuous and convex as the minimum of linear functions of θ. It is thus almost everywhere differentiable with a derivative (wherever it exists) given by the envelope condition

$$\dot{U}(\theta) = -q(\theta). \tag{8}$$

This envelope condition and the convexity of U are necessary and sufficient conditions for the implementability of $q(\cdot)$, which is achieved using the tariff $T(q(\theta)) = (S-\theta)q(\theta) - \frac{1}{2}q(\theta)^2 - U(\theta)$.[13] For the pure monopoly setting, standard techniques can be used to derive the optimal consumption levels under asymmetric information. Knowing that the participation constraint $U(\theta) \geq 0$ necessarily binds for the worst type $\overline{\theta}$, it is routine to reconstruct and maximize pointwise a virtual surplus that is expressed as

$$(S-2\theta+\underline{\theta})q - \frac{q^2}{2} - k\delta_{q>1}.$$

Taking again the concave envelope of this function, we replicate our earlier findings *mutatis mutandis* to get the following expression of the monopoly solution as

$$q^m(\theta) = q^{fb}(2\theta-\underline{\theta}), \tag{9}$$

which is positive given our maintained assumptions on S and θ. Because of the nonconcavity of the virtual surplus, the monopoly quantity is again discontinuous. As this example clearly shows, sometimes standard techniques may suffice. Unfortunately, it is

[12]From the taxation principle Rochet (1987), focusing on nonlinear tariffs is without loss of generality in the space of nonstochastic mechanisms.

[13]Because $U(\theta)$ is convex, the consumption level $q(\theta)$ is a nonincreasing selection within the best-response correspondence $\arg\max_{q\in\mathcal{Q}}(S-\theta)q - \frac{q^2}{2} - T(q)$. This monotonicity condition is first neglected and then verified ex post for the solution to the relaxed program.

no longer the case when participation constraints are more complex and the discontinuity of the surplus function precisely determines where these constraints are binding. This is the scenario we now investigate.

Competitive Fringe. We now assume that a competitive fringe, without capacity constraint, can supply a low-quality version of digital good at zero marginal cost. Specifically, we assume that the intercept of the inverse demand function for the low-quality good is $\hat{S} < S$ with $\Delta = S - \hat{S} > 0$, and the consumption of the two versions of the good is mutually exclusive. We also assume that $\Delta < 1$ so that the fringe poses a significant competitive threat to the dominant firm and $\hat{S} > \overline{\theta}$ to ensure positive consumption from the fringe. We denote the amount of low-quality good a consumer would buy from the competitive fringe as

$$\hat{q}(\theta) = \hat{S} - \theta,$$

which generates the net surplus of

$$\hat{U}(\theta) = \max_{q \in \mathcal{Q}} (\hat{S} - \theta) q - \frac{q^2}{2} \equiv \frac{(\hat{S} - \theta)^2}{2}.$$

Adopting our canonical form, we set $u(\theta) = U(\theta) - \hat{U}(\theta)$. The customer chooses to buy exclusively from the monopolist when the participation constraint (1) holds. With this notation, the envelope condition for incentive compatibility (8) becomes

$$\dot{u}(\theta) = \hat{q}(\theta) - q(\theta), \quad \text{a.e.} \tag{10}$$

Expressed in terms of $v = \dot{u}$, the surplus function becomes

$$s(\theta, v) = (S - \theta)\big(\hat{q}(\theta) - v\big) - \frac{1}{2}\big(\hat{q}(\theta) - v\big)^2 - k\delta_{\hat{q}(\theta) - v > 1}.$$

Note that the downward discontinuity implies that the concave envelope of $s(\theta, v)$ is linear over some range. We are now equipped to derive the optimal solution.

Smooth case: $k = 0$. To emphasize the difference between our techniques and those used in the smooth scenario studied by Jullien (2000), we start with the familiar case where there is no cost to additional capacity. The surplus function $s(\theta, v)$ is therefore smooth and strictly concave in v. Our Theorem 1 then takes the same form as Theorem 1 in Jullien (2000). The adjoint function $\overline{\gamma}_0(\theta) = \int_{\underline{\theta}}^{\theta} \mu_0(ds)$ and the optimal consumption level $\overline{q}_0(\theta) = \hat{q}(\theta) - \dot{\overline{u}}_0(\theta)$ must satisfy the following first-order condition for a smooth problem:

$$\overline{\gamma}_0(\theta) - \theta + \underline{\theta} = \frac{\partial s}{\partial v}\big(\theta, \dot{\overline{u}}_0(\theta)\big) \quad \Longleftrightarrow \quad \overline{q}_0(\theta) = S - 2\theta + \underline{\theta} + \overline{\gamma}_0(\theta). \tag{11}$$

Using (9), note that $\overline{q}_0(\theta) = q^m(\theta) + \overline{\gamma}_0(\theta)$, so the optimal solution reduces the distortion relative to the case in which the competitive fringe is absent. When there is competition with a fringe, strong downwards distortions of consumption for the lowest types are no longer so attractive for the monopolist since these types could switch to the fringe.

Proposition 2. *Suppose that $k = 0$ and let $\tilde{\theta} = \underline{\theta} + \Delta < \overline{\theta}$. The following optimal consumption and adjoint functions satisfy the necessary and sufficient conditions for optimality of Theorem 1:*

$$\overline{q}_0(\theta) = \max\{q^m(\theta), \hat{q}(\theta)\}, \tag{12}$$

$$\overline{\gamma}_0(\theta) = \max\{\theta - \tilde{\theta}; 0\}, \tag{13}$$

where $\tilde{\theta} \equiv \underline{\theta} + \Delta$. The participation constraint (1) is binding on $[\tilde{\theta}, \overline{\theta}]$, slack elsewhere.

The presence of the fringe limits the ability of the monopolist to price discriminate. Screening distortions are less pronounced than in a pure monopoly setting, a result that echoes findings in Champsaur and Rochet (1989), Stole (1995, 2003), and Calzolari and Denicolo (2015) among many others.

The intuition for the shape of the solution can be borrowed from the work of Maggi and Rodriguez-Clare (1995). In a smooth environment, those authors have proposed a typology of the various patterns of binding participation constraints that may arise at optimal contracts in structured environments. They argue that whether participation constraints bind on a whole interval or only at the extreme of the type set depends on the relative convexity of the rent profile that the monopolist would like to implement vis-à-vis the reservation payoff. Due to the lower quality of the alternative that it offers, buying from the fringe provides a type-dependent reservation payoff that is less convex than what the monopolist would like to implement without competition. It follows that the participation constraint is binding on a nondegenerate interval (here including the type with the lowest valuation).

Observe that the measure μ_0 is absolutely continuous with respect to the Lebesgue measure on $[\tilde{\theta}, \overline{\theta})$ but it has a mass point at $\overline{\theta}$, namely $\mu(\{\overline{\theta}\}) = \Delta < 1$. $\tilde{\theta}$ is determined by a "smooth-pasting" condition, $q^m(\tilde{\theta}) = \hat{q}(\tilde{\theta})$, which ensures that output is continuous at $\tilde{\theta}$ and $\overline{u}_0$ is differentiable at that point.

Discontinuity: $k > 0$. The surplus function $s(\theta, v)$ has a downward jump. Theorem 1 in Jullien (2000) now no longer applies while our Theorem 1 provides a valid solution. The optimality condition (5) must thus be expressed in terms of a subdifferential:

$$\overline{\gamma}(\theta) - (\theta - \underline{\theta}) \in \partial_v \overline{\mathrm{co}}(s)\big(\theta, \dot{\overline{u}}(\theta)\big). \tag{14}$$

To streamline exposition and limit the number of possible cases to study, we impose the following condition on $\hat{S}$:

$$\hat{S} > \underline{\theta} + \Delta + 1. \tag{15}$$

The impact of this condition (which is equivalent to $\hat{q}(\tilde{\theta}) > 1$) is that if the dominant firm wishes to implement the allocation $\overline{q}_0$ defined in Proposition 2 for $k = 0$ to types in a neighborhood above $\tilde{\theta}$, it must pay the additional capacity cost k because these types would buy more than one unit with their outside option. To avoid paying such cost, the monopolist would like to bunch the set of types just above $\tilde{\theta}$ and sell them each a single one unit of output. Yet, this objective conflicts with the possibility that those types would switch to the fringe. Hence, the monopolist must redistribute part

of these gains to customers under the form of extra information rent beyond what they receive from the fringe. These concerns give rise to a market segmentation with four different connected subsets of types. This pattern corresponds to an adjoint function $\overline{\gamma}$, which is obtained by slightly modifying $\overline{\gamma}_0$ (as explained in (18) below) according to our economic intuition.

Proposition 3. *Define θ_1, θ_2, θ_3 by*

$$\hat{q}(\theta_1) = 1 + (\sqrt{2} - 1)\sqrt{k}, \qquad \hat{q}(\theta_2) = 1 + \left(\frac{\sqrt{2}}{2} - 1\right)\sqrt{k}, \qquad \hat{q}(\theta_3) = 1 + (\sqrt{2} - 2)\sqrt{k}, \tag{16}$$

which satisfy $\tilde{\theta} < \theta_1 < \theta_2 < \theta_3 < \overline{\theta}$.

The following optimal consumption and adjoint functions satisfy the necessary and sufficient conditions for optimality of Theorem 1:

$$\overline{q}(\theta) = \begin{cases} q^m(\theta) & \text{if } \theta \in [\underline{\theta}, \tilde{\theta}), \\ \hat{q}(\theta) & \text{if } \theta \in [\tilde{\theta}, \theta_1) \text{ and } \theta \in [\theta_3, \overline{\theta}), \\ 1 & \text{if } \theta \in [\theta_1, \theta_2), \\ q^m(\theta) + \sqrt{k} + \theta_1 - \tilde{\theta} & \text{if } \theta \in [\theta_2, \theta_3); \end{cases} \tag{17}$$

$$\overline{\gamma}(\theta) = \begin{cases} 0 & \text{if } \theta \in [\underline{\theta}, \tilde{\theta}), \\ \theta - \tilde{\theta} & \text{if } \theta \in [\tilde{\theta}, \theta_1) \text{ and } \theta \in [\theta_3, \overline{\theta}), \\ \sqrt{k} + \theta_1 - \tilde{\theta} & \text{if } \theta \in [\theta_1, \theta_3); \end{cases} \tag{18}$$

The participation constraint (1) is binding on $[\tilde{\theta}, \theta_1] \cup [\theta_3, \overline{\theta}]$ and slack elsewhere.

The optimal consumption remains equal to $\overline{q}_0(\theta) = \max\{q^m(\theta), \hat{q}(\theta)\}$ for the bottom interval $[\underline{\theta}, \theta_1)$, i.e., for those types with the highest consumption levels for which the incumbent finds it worth to incur the extra capacity. There is, however, a downward jump in consumption at θ_1. Types in the right-neighborhood $[\theta_1, \theta_2)$ are bunched together and consume only one unit from the monopolist although they would consume more from the fringe. To save on the cost of additional capacity, the monopolist gives to those types a price discount to compensate for the constrained consumption and, as a result, those types get a payoff above what they would get purchasing from the fringe. Yet, types with a lower valuation (higher type) consume less than one unit of service up to the point where the lowest valuations in the interval $[\theta_3, \overline{\theta}]$ again consume the same amount as they would with the fringe.

Again, some intuition for this pattern of rents can be grasped from the typology offered by Maggi and Rodriguez-Clare (1995) although here the set of types with binding outside options is significantly more complex than in their study. Indeed, the fixed cost now forces a constant consumption at $q = 1$ for a bunch of intermediate types. The rent profile that the monopolist would like to implement becomes less convex than the outside option for such types. We know from Maggi and Rodriguez-Clare (1995) that, in

such contexts, the participation constraint may be slack for intermediate types. In contrast with Maggi and Rodriguez-Clare (1995) who focus on the rather pure cases where the reservation payoff is uniformly more (or less) convex than the monopoly's profile, the comparison is here reversed for the very highest types. In sum, the participation constraint is now slack for such intermediate types but it now binds on two intervals around that area.

Technically, the measure μ is again absolutely continuous with respect to the Lebesgue measure on the interior of the set of the types, namely $[\tilde{\theta}, \theta_1) \cup [\theta, \overline{\theta})$, where (1) is binding with a mass point at $\overline{\theta}$ given by $\mu(\{\overline{\theta}\}) = \Delta$. To cope with the consumption discontinuity at θ_1^- and the fact that a right neighborhood of that type receives rent beyond what it gets from the fringe, μ has also another charge at that discontinuity point θ_1, namely $\mu(\{\theta_1\}) = \sqrt{k}$.

3.2 *Nonlinear pricing with split purchases*

The previous model can be modified to cover different competitive scenarios. Instead of exclusively buying from one seller as in Section 3.1, here we assume that consumers may split their purchases across the monopolist and a competitive fringe. This scenario is relevant for a number of contracting environments where analyses have previously been limited by the assumption of exclusive contracting. As a first example, regulated monopolies can be subject to competition by entrants with alternative bypass technologies. Previous research (Caillaud (1990), Laffont and Tirole (1990), Curien, Jullien, and Rey (1998)) ruled out the possibility that consumers may also purchase from entrants. Recent technological advances in telecommunication services, broadcasting, and internet practices make split purchases more feasible, and thus motivates a theory of bypass that accounts for the possibility of nonexclusive purchases from incumbent operators and new entrants.

As a second example, exclusive contracts between wholesalers and retailers are restricted in some countries. This makes it important to study how vertical contracts are modified when retailers may sell products from rivals and exclusivity clauses cannot be enforced.[14] Lastly, competition in financial and insurance markets has often been modeled, following Rothschild and Stiglitz (1976), under the assumption that only exclusive contracts can be signed. But as noted by Attar, Mariotti, and Salanié (2011), these exclusivity clauses are difficult to enforce. A trader may buy a financial asset from one broker and then turn to another one to complete his portfolio. In each of these nonexclusive settings, standard screening distortions are limited by the buyer's ability to purchase from the competitive fringe. As we will see in the model below, such possibility introduces discontinuities in the virtual surplus. Instead of coming from the cost function

[14]In some environments, incumbent manufacturers may be able to offer exclusive contracts with their retailers to prevent their use of alternative suppliers. In Aghion and Bolton (1987)'s theory of contracts as a barrier to entry, retailers buy entirely from the entrant if they switch thanks to an assumption of unit demand. More recently, Calzolari and Denicolo (2015) have analyzed how an exclusivity clause endogenously arises in equilibrium. Choné and Linnemer (2015) also investigate the possibility of split purchases in related contexts.

as in Section 3.1, discontinuities come from the demand side and, more precisely, from how the buyer responds to changes in contractual terms with a dominant firm.

Model. A dominant firm produces a good at constant marginal cost $c \geq 0$ without any additional capacity costs. On the demand side, consumer preferences are again characterized by (7) where θ, uniformly distributed on $[\underline{\theta}, \overline{\theta}]$ (with $\overline{\theta} - \underline{\theta} = 1$), is private information. Importing *mutatis mutandis* our earlier findings of Section 3.1 with $k = 0$ and now a nonnegative marginal cost equal to $c \geq 0$, the monopoly solution (absent competition) consists in offering

$$q^m(\theta) = S - 2\theta + \underline{\theta} - c = q^{fb}(2\theta - \underline{\theta}).$$

A competitive fringe sells a perfect substitute to the incumbent's product at price $p > c$. We recycle our previous notation Δ, and define $\Delta = p - c > 0$. Taking p to be equal to the competitive fringe's unit cost, entry is inefficient in a first-best world. We denote by $\hat{q}$ the quantity bought from the fringe. The fringe has a capacity constraint $K > 0$ and thus $\hat{q} \in \hat{\mathcal{Q}} = [0, K]$. The key difference with Section 3.1 is that consumers can now always purchase from the fringe if the incumbent charges a price, which is too high. Formally, if a consumer with type θ chooses q units from the incumbent, his indirect utility function becomes

$$\hat{u}(\theta, q) = \max_{x \in \hat{\mathcal{Q}}} (S - \theta)(q + x) - \frac{(q + x)^2}{2} - px.$$

Define the highest level of consumption from the incumbent firm, which induces consumption of K from the competitive fringe as

$$\check{q}(\theta) \equiv S - \theta - p - k,$$

which we assume to be positive. For $q \leq \check{q}(\theta)$, the consumer purchases $\hat{q}(\theta) = K$ units from the competitive fringe; for $q \geq \check{q}(\theta) + K$, the consumer purchases $q_0(\theta) = 0$ from the fringe. Straightforward computations yield

$$\hat{u}(\theta, q) = \begin{cases} (S - \theta)(q + K) - \dfrac{(q + K)^2}{2} - pK & \text{if } q \leq \check{q}(\theta), \\ \dfrac{\left(\check{q}(\theta) + K\right)^2}{2} + pq & \text{if } q \in \left[\check{q}(\theta), \check{q}(\theta) + K\right], \\ (S - \theta)q - \dfrac{q^2}{2} & \text{if } q \geq \check{q}(\theta) + K. \end{cases}$$ [15]

[15] To better understand the shape of the indirect utility function $\hat{u}(\theta, q)$, we may think of the consumer as choosing between consuming 0 or K units from the fringe. His indirect utility function would be the (nonconcave) maximum of two concave functions because of a discontinuous jump in the corresponding choice. The possibility of consuming any arbitrary amount within the interval $\hat{\mathcal{Q}}$ concavifies this indirect utility function and introduces a linear segment for intermediate consumption levels from the dominant firm.

In particular, had the consumer not bought from the incumbent, i.e., $q = 0$, he would consume up to capacity from the fringe and get a payoff worth

$$\hat{U}(\theta) \equiv \hat{u}(\theta, 0) = \check{q}(\theta)K + \frac{K^2}{2}.^{16}$$

In what follows, we will assume that

$$q^m(\overline{\theta}) > K, \tag{19}$$

which ensures that the incumbent firm still wants to serve the type with the lowest possible demand $\overline{\theta}$ even when that type consumes up to the fringe's capacity.

A Discontinuous Surplus Function. When the consumer instead buys q units from the dominant firm at a nonlinear price $T(q)$, a consumer with type θ obtains

$$U(\theta) = \max_{q \in \mathcal{Q}} \hat{u}(\theta, q) - T(q).$$

To import our general formalism, we again introduce the state variable $u(\theta) = U(\theta) - \hat{U}(\theta)$. The participation constraint takes its usual form (1) while the standard envelope condition for incentive compatibility becomes

$$\dot{u}(\theta) = \begin{cases} -q(\theta) & \text{if } q(\theta) < \check{q}(\theta), \\ -\check{q}(\theta) & \text{if } q(\theta) \in \big[\check{q}(\theta), \check{q}(\theta) + K\big], \\ -q(\theta) + K & \text{if } q(\theta) > \check{q}(\theta) + K. \end{cases} \tag{20}$$

This relationship actually shows that $\dot{u}$, taken as a function of q, is not bijective.[17] This phenomenon captures the role of competition from the fringe. If, to facilitate rent extraction, the dominant firm is willing to slightly reduce the buyer's consumption $q(\theta)$ when it takes values in $(\check{q}(\theta), \check{q}(\theta) + K)$, the buyer can instead consume more from the fringe.

The dominant firm wants to maximize the bilateral surplus of contracting with the buyer,

$$\tilde{s}(\theta, q) = \hat{u}(\theta, q) - \hat{u}(\theta, 0) - cq,$$

net of the rent the buyer gets when purchasing exclusively from the fringe. That is, the dominant firm wishes to implement the allocation $q(\cdot)$ to maximize the expectation of

$$\tilde{s}\big(\theta, q(\theta)\big) - u(\theta) = v\big(\theta, q(\theta)\big) - v(\theta, 0) - cq(\theta) - u(\theta).$$

After replacing $q(\theta)$ by its expression in terms of $\dot{u}(\theta)$ wherever such inversion is possible (i.e., using (20), where $\dot{u}(\theta) \neq -\check{q}(\theta)$), and after straightforward computations (detailed in the proof of Proposition 4), we may express the gross surplus as

$$s(\theta, v) = -\big(\check{q}(\theta) + \Delta\big)v - \frac{v^2}{2} + K\Delta\delta_{v \leq -\check{q}(\theta)}. \tag{21}$$

[16] Since $v(\overline{\theta}) < 0$, this right-hand side remains positive.

[17] Carbajal and Ely (2016) present an interesting model of behavioral consumers having loss aversion that has similar features.

This surplus function is upper semicontinuous, has a downward jump discontinuity at $v_0(\theta) = -\check{q}(\theta)$, and is maximized at $v_1(\theta) = -\check{q}(\theta) - \Delta < -\check{q}(\theta)$. This downward jump captures the fact that, when consumption from the dominant firm is too low (which means $v(\theta) = -\dot{u}(\theta)$), the bilateral surplus between the dominant firm and the customer diminishes by $K\Delta$, reflecting the opportunity cost of purchasing K units from the fringe.

Proposition 4. *Let $\tilde{\theta} = \underline{\theta} + \Delta$ and $\tilde{\theta}_0 = \tilde{\theta} + \sqrt{2K\Delta}$. Suppose that $\tilde{\theta}_0 \leq \overline{\theta}$ and (19) holds. The following optimal consumption levels and adjoint functions satisfy the necessary and sufficient conditions for optimality of Theorem 1:*

$$\overline{q}(\theta) = \begin{cases} q^m(\theta) & \text{if } \theta \in [\underline{\theta}, \tilde{\theta}], \\ \text{arbitrary } \in \left[\check{q}(\theta), \check{q}(\theta) + K\right] & \text{if } \theta \in (\tilde{\theta}, \tilde{\theta}_0), \\ q^m(\theta) - K & \text{if } \theta \in [\tilde{\theta}_0, \overline{\theta}]. \end{cases} \tag{22}$$

$\overline{\gamma}(\theta)$ has a mass point at $\overline{\theta}$, $\mu(\{\overline{\theta}\}) = 1$. The participation constraint (1) is binding at $\overline{\theta}$ only.

Types who are the most eager to buy (i.e., $\theta \in [\underline{\theta}, \tilde{\theta})$) are not tempted to switch to the fringe. They consume the same downward distorted monopoly quantity $q^m(\theta)$ as what the incumbent firm would offer if the fringe was absent. This quantity nevertheless remains large enough to make it unattractive for the buyer to switch. For intermediate types in $[\tilde{\theta}, \tilde{\theta}_0)$, the monopolist loses some of his ability to screen. Any attempt to reduce consumption from the incumbent is entirely compensated by the buyer consuming more from the fringe. The incumbent's sales are indeterminate but screening distortions are reduced to avoid switching. Finally, types in $[\tilde{\theta}_0, \overline{\theta}]$ react to the downward screening distortion offered by the incumbent by consuming up to capacity from the fringe.

Although the discontinuity of $s(\theta, v)$ bears some similarity with that analyzed in Section 3.1, the characterization of the solution is rather deferent due to the different nature of the participation constraint. Over the range $(\tilde{\theta}, \tilde{\theta}_0)$, the solution satisfies $\overline{u}'(\theta) = -\check{q}(\theta)$ independently of the incumbent's sales, which explains the indeterminacy in those sales. Because of the downward jump discontinuity of $s(\theta, v)$ at $v = -\check{q}(\theta)$, $\overline{\text{co}}(s)(\theta, v)$ has a kink at that point and displays a flat segment for $v \in [-\check{q}(\tilde{\theta}_0), v_1(\tilde{\theta}_0)]$ where $v_1(\tilde{\theta}_0) = -\check{q}(\tilde{\theta}_0) + \sqrt{2K\Delta}$. At type $\tilde{\theta}_0$, the incumbent is thus indifferent between choosing $\dot{\overline{u}}(\tilde{\theta}_0) = -\check{q}(\tilde{\theta}_0)$ and moving up to $\dot{\overline{u}}(\tilde{\theta}_0) = v_1(\tilde{\theta}_0)$, reducing his sales to $q^m(\tilde{\theta}_0) - K$ and allowing the buyer to purchase from the fringe up to capacity.

Appendix A: NonSmooth optimization

This Appendix briefly reminds the reader about necessary and sufficient conditions for the general problem of maximizing an upper semicontinuous function, $h : \mathbb{R} \to \mathbb{R}$, over a compact set $X \subset \mathbb{R}$.

A generalization of the first-order condition for smooth, concave optimization programs can indeed be obtained for general upper semicontinuous programs by introducing a few concepts from nonsmooth convex analysis. The basic idea is that any solution

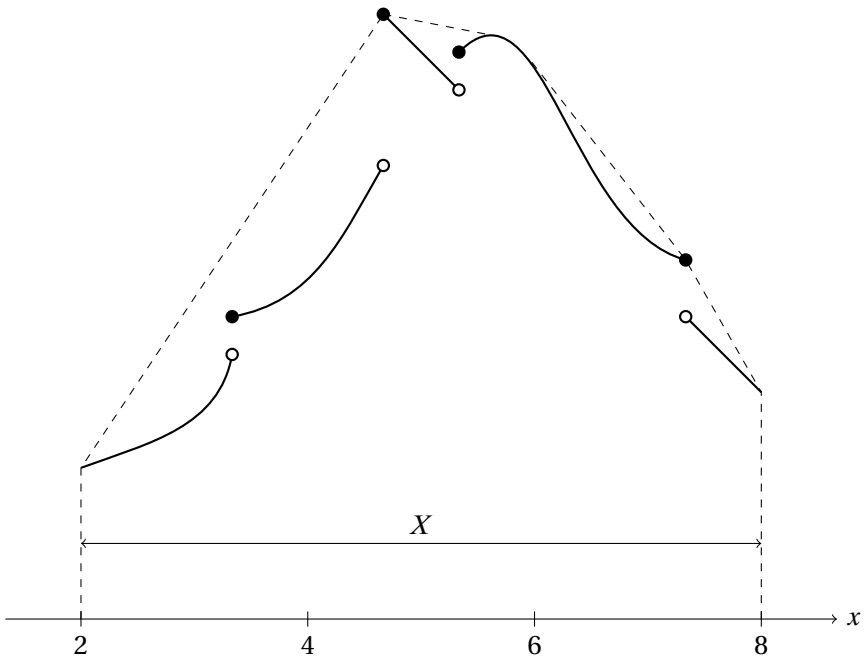

FIGURE 1. Concavification of discontinuous, but upper-semicontinuous function on X.

to the original upper semi-continuous program must lie on the minimal concave envelope or *concavification* of the objective. Consider, for example, the upper semicontinuous function graphed in Figure 1 in bold.

This function is defined over the real line, but the restricted domain of interest is $X = [\underline{x}, \overline{x}]$. The minimal-concave envelope over this domain is depicted by the dashed lines in the graph. Notice its value is negative infinity outside of $[\underline{x}, \overline{x}]$. Obviously, the maximum of *this* concave envelope is a solution to the original program. More generally, in the case in which there is a continuum of solutions (i.e., the maximum is achieved on a horizontal component of the majorization), there exist two solutions to the original program—the endpoints of the majorization. It is, in this sense, without loss to convert an upper semicontinuous program over a compact set into a concave (but possibly nondifferentiable) program over the same set. Formally, we will denote $\overline{\text{co}}_X(h)$ to refer to the concavification of an objective function, h, over a domain, X,[18] and $\overline{\text{co}}_X(h)(x)$ to refer to the value of this envelope evaluated at x.[19]

Having reached the conclusion that we may focus on the concave envelope of the program, we can now import the generalized notion of derivative from convex analysis. Formally, we will define a set of gradients at any point to be all those vectors which

[18]When $X = \mathbb{R}$, we simplify notation and omit the subscript.

[19]In the nonsmooth optimization literature, often one considers the minimal concave envelope of h over the real line instead of some domain X, but in this case with a penalty function, $\Psi_X(x)$, which equals 0 for $x \in X$ and $-\infty$ for $x \notin X$. Thus, in our notation, $\overline{\text{co}}_X(h) = \overline{\text{co}}_{\mathbb{R}}(h + \Psi_X)$.

"support" the graph at the given point, and we refer to this set-valued notion of derivative as the generalized gradient or the *superdifferential*, denoted $\partial h(x)$ when applied to a concave function h at point x.[20] Where h is differentiable, the superdifferential is single-valued and corresponds to the gradient. If h exhibits a kink and $X \subseteq \mathbb{R}$, the superdifferential is an interval of gradients with endpoints corresponding to the left- and right-side derivatives at the point. More generally, if $X \subseteq \mathbb{R}^n$, then

$$\partial h(x) = \big\{\tau \in \mathbb{R}^n \mid h(y) \leq h(x) + \langle \tau, y - x\rangle \ \forall\, y \in \mathbb{R}^n\big\}.$$

Using this generalization of gradient, we can now state the necessary and sufficient conditions for x^* to be a maximum of an upper-semicontinuous function, h, over some given domain $x \in X$: if x^* is a solution to the maximization program, then the following first-order condition must be satisfied:

$$0 \in \partial \overline{\mathrm{co}}_X(h)\big(x^*\big). \tag{23}$$

Furthermore, if x^* satisfies (23) and the envelope coincides with h at x^*, i.e.,

$$\overline{\mathrm{co}}_X(h)\big(x^*\big) = h\big(x^*\big) \tag{24}$$

then x^* solves the maximization program.

These conditions can be further tightened when a component of the objective function is affine. To this end, suppose that $h = g + f$ where g is affine (and slightly abusing notation, let us write $g(x) = gx$). Well-known identities from convex analysis give us:

$$\overline{\mathrm{co}}_X(h)(x) = gx + \overline{\mathrm{co}}_X(f)(x) \quad \text{and} \quad \partial\overline{\mathrm{co}}_X(h)(x) = g + \partial\overline{\mathrm{co}}_X(f)(x).$$

Thus, the linear linear part of the objective can be factored out and the "first order" necessary and sufficient condition for the optimality of x^* reduces to

$$-g \in \partial\overline{\mathrm{co}}_X(f)\big(x^*\big).$$

This property will be repeatedly used throughout our analysis, first, to derive generalized first-order conditions for our infinite-dimensional optimal control problem, and second, to tackle applications in contract theory where such a decomposition is frequently available.

[20]We use the term "support" from convex analysis given it is evocative and familiar. The term *subdifferential* is the parallel notion of superdifferential when applied to convex functions. When we refer to the generalized gradient of a function that is understood to be convex, we will abuse notation slightly by again using the notation $\partial h(x)$, where it is understood that when h is convex, then

$$\partial h(x) = \big\{\tau \in \mathbb{R}^n \mid h(y) \geq h(x) + \langle \tau, y - x\rangle \ \forall\, y \in \mathbb{R}^n\big\}.$$

See Ferrera (2014) for an introduction to nonsmooth analysis and an in-depth discussion of super and subdifferentials.

Appendix B: Proof of Theorem 1

Preliminaries for Nonsmooth Analysis. We draw heavily from Vinter and Zheng (1998) in the following presentation. A complete treatment can be found in the monograph of Vinter (2000). Theorem 3 from Vinter and Zheng (1998) appears as Theorem 10.2.1 in Vinter (2000).

Take a closed set $A \subseteq \mathbb{R}^n$ and a point $x \in A$. A vector $r \in \mathbb{R}^n$ is a *limiting normal* to A at x if there exists a sequence $(x_i, r_i) \to (x, r)$ with $x_i \in A$ and a constant $M \geq 0$ such that for each i in the sequence $r_i \cdot (x_i - x) \leq M\|x_i - x\|^2$, where $\|\cdot\|$ denotes Euclidean distance. The cone of limiting normal vectors to A at x is denoted $N_A(x)$. Given a lower semicontinuous function $g : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ and a point $x \in \mathbb{R}$ such that $g(x) < +\infty$, the *limiting subdifferential* of g at x is defined as

$$\partial g(x) \equiv \big\{\xi \mid (\xi, -1) \in N_{\mathrm{epi}\{g\}}\big(x, g(x)\big)\big\},$$

where $\mathrm{epi}\{g\}$ is the *epigraph* of the function g defined as

$$\mathrm{epi}\{g\} \equiv \big\{(x, \alpha) \in \mathbb{R} \times \mathbb{R} \mid \alpha \geq g(x)\big\}.$$

The *asymptotic limiting subdifferential* of g at x, written $\partial^\infty g(x)$, is defined as

$$\partial^\infty g(x) \equiv \big\{\xi \mid (\xi, 0) \in N_{\mathrm{epi}\{g\}}\big(x, g(x)\big)\big\}.$$

Finally, we define

$$\partial_x^> h(t, x) \equiv \overline{\mathrm{co}}\Big\{\lim_i \xi_i \;\Big|\; \exists t_i \to t,\, x_i \to x \text{ s.t. } h(t_i, x_i) > 0 \text{ and } \xi_i \in \partial_x h(t_i, x_i)\ \forall\, i\Big\}.$$

Two results from nonsmooth analysis (Vinter (2000, Propositions 4.3.3 and 4.3.4)) that we use are (1) $\partial^\infty g(x) = \{0\}$ if g is Lipschitz continuous and (2) for any x such that $g(x)$ is finite,

$$N_{\mathrm{epi}\{g\}}\big(x, g(x)\big) = \big\{(\xi d, -\xi) \mid \xi > 0,\, d \in \partial g(x)\big\} \cup \big\{\partial^\infty g(x) \times \{0\}\big\}.$$

A *local maximizer* of $\Lambda(x)$ is a feasible arc, $\overline{x}$, which maximizes $\Lambda(x)$ over all feasible arcs $x \in \mathrm{AC}(\Theta, \mathbb{R}_+)$ within an ε neighborhood of $\overline{x}$, $\|\overline{x} - x\|_{\mathrm{ac}} \leq \varepsilon$ where we denote the norm on the space of absolutely continuous functions by $\|x\|_{\mathrm{ac}} \equiv \|x(\underline{\theta})\| + \int_{\underline{\theta}}^{\overline{\theta}} \|\dot{x}(\theta)\|\, d\theta$. A *local minimizer* is defined analogously.

Necessity. First, and for completeness, we reproduce here Theorem 3 of Vinter and Zheng (1998), which provides necessary conditions for solutions to the following minimization program:

$$(\mathcal{P}'): \quad \text{Minimize } J(x) \equiv \int_{\underline{\theta}}^{\overline{\theta}} L\big(\theta, x(\theta), \dot{x}(\theta)\big)\, d\theta$$

$$\text{subject to } x \in \mathrm{AC}(\Theta, \mathbb{R}) \text{ and } h\big(\theta, x(\theta)\big) \leq 0 \text{ for all } \theta \in \Theta \equiv [\underline{\theta}, \overline{\theta}].$$[21]

[21] We specialize their theorem to our present problem in which the range of $x(\theta)$ is one-dimensional and there is no endpoint cost function.

We will prove necessity for Theorem 1 by specializing this theorem, exploiting the fact that our integrand in Λ is a linear function of x and $h(\theta, x) = -x$.

THEOREM 2 (Vinter and Zheng (1998, Theorem 3)). *Let $\overline{x}$ be local minimizer for $(\mathcal{P}')$ in $\mathrm{AC}(\Theta, \mathbb{R})$ such that $J(\overline{x}) < +\infty$. Assume that the following hypotheses are satisfied:*

H_1. *$L(\cdot, x, \cdot)$ is $\mathcal{L} \times \mathcal{B}$ measurable for each x and $L(\theta, \cdot, \cdot)$ is lower semicontinuous for a.e. $\theta \in \Theta$.*

H_2. *For every $N > 0$, there exists $\delta > 0$ and $k \in L^1$ such that*

$$\left\| L\big(\theta, x', v\big) - L(\theta, x, v) \right\| \le k(\theta) \left\| x' - x \right\|, \qquad L\big(\theta, \overline{x}(\theta), v\big) \ge -k(\theta)$$

for a.e. $\theta \in \Theta$, for all $x, x' \in \overline{x}(\theta) + \delta B$ and $v \in \dot{\overline{x}}(\theta) + NB$, where B is a unit Euclidean ball.

H_3. *h is upper semicontinuous near $(\theta, \overline{x}(\theta))$ for all $\theta \in \Theta$, and there exists a constant k_h such that*

$$\left\| h\big(\theta, x'\big) - h(\theta, x) \right\| \le k_h \left\| x' - x \right\|$$

for all $\theta \in \Theta$ and all $x', x \in \overline{x}(\theta) + \delta B$.

Then there exist an arc $p \in \mathrm{AC}(\Theta, \mathbb{R})$, a constant $\lambda \ge 0$, a nonnegative measure μ on the Borel subsets of Θ and a μ-integrable function $\zeta : \Theta \to \mathbb{R}$, such that

(i) $\lambda + \max_{\theta \in \Theta} |p(\theta)| + \int_{\underline{\theta}}^{\overline{\theta}} \mu(d\tilde{\theta}) = K > 0$ *(where K is an arbitrary normalization constant),*[22]

(ii)

$$\dot{p}(\theta) \in \overline{\mathrm{co}} \Bigg\{ \eta \;\Bigg|\; \Bigg(\eta,\, p(\theta) + \int_{[\underline{\theta}, \theta)} \zeta(\tilde{\theta}) \mu(d\tilde{\theta}), -\lambda \Bigg) \in N_{\mathrm{epi}\{L(\theta, \cdot, \cdot)\}} \big(\overline{x}(\theta), \dot{\overline{x}}(\theta), L\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big) \big) \Bigg\} \quad \textit{a.e.},$$

(iii)

$$p(\underline{\theta}) = p(\overline{\theta}) - \int_{\underline{\theta}}^{\overline{\theta}} \zeta(\tilde{\theta}) \mu(d\tilde{\theta}) = 0,$$

(iv)

$$\Bigg(p(\theta) + \int_{[\underline{\theta}, \theta)} \zeta(\tilde{\theta}) \mu(d\tilde{\theta}) \Bigg) \dot{\overline{x}}(\theta) - \lambda L\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big) \in \arg\max_{v \in \mathbb{R}} \Bigg(p(\theta) + \int_{[\underline{\theta}, \theta)} \zeta(\tilde{\theta}) \mu(d\tilde{\theta}) \Bigg) v - \lambda L\big(\theta, \overline{x}(\theta), v\big),$$

(v) $\zeta(\theta) \in \partial_x^{>} h(\theta, \overline{x}(\theta))$ *μ-a.e. and* $\mathrm{supp}\{\mu\} \subseteq \{\theta \mid h(\theta, \overline{x}(\theta)) = 0\}$.

[22]We choose to state the theorem using $K > 0$ as an arbitrary normalization rather than $K = 1$, which is the normalization chosen in Vinter and Zheng (1998). Later, by setting $K = 3$, we will succeed in normalizing μ to a probability measure, which is a more familiar object.

We apply this result to our setting by substituting $xf(\theta) - s(\theta, v)$ in program $(\mathcal{P})$ in place of $L(\theta, x, v)$ and thereby converting the maximization functional Λ in program $(\mathcal{P})$ to the minimization functional J in program $(\mathcal{P}')$. We complete the transformation by requiring that $h(\theta, x) = -x$, and that $L(\theta, x, v)$ is a linear function of x for any (θ, v).

First, we verify that hypotheses H_1–H_3 are satisfied for our program $(\mathcal{P})$. Because $s(\theta, \cdot)$ is upper semicontinuous and $\mathcal{B}$-measurable, and because $L(\theta, x, v)$ is linear in x, H_1 is satisfied. H_2 requires that $L(\theta, \cdot, v)$ is Lipschitz continuous, which is trivial given that L is linear in x with coefficient $f(\theta)$. Because the transformed program has $h(\theta, x) = -x$, h is a continuous linear function of x, and thus H_3 is also satisfied.

Next, we specialize the conclusions of Vinter and Zheng (1998) by making use of the additional restrictions on $L(\cdot)$ and $h(\cdot)$. We present this in the following lemma.

Lemma 1. *Suppose that $L(\theta, x, v)$ is a linear function of x and that $h(\theta, x) = -x$. Then the conclusions (i)–(v) of Theorem 2 imply*

(a) $\lambda + \max_{\theta\in\Theta} |p(\theta)| + \int_{\underline{\theta}}^{\overline{\theta}} \mu(d\tilde{\theta}) = K$,

(b) $\dot{p}(\theta) = \lambda f(\theta)$ *a.e.,*

(c) $p(\underline{\theta}) = p(\overline{\theta}) + \int_{\underline{\theta}}^{\overline{\theta}} \zeta(\tilde{\theta})\mu(d\tilde{\theta}) = 0$

(d) $\dot{\overline{x}}(\theta) \in \arg\max_{v\in\mathbb{R}} (p(\theta) + \int_{[\underline{\theta},\theta)} \zeta(\tilde{\theta})\mu(d\tilde{\theta}))v + \lambda s(\theta, v)$, *a.e.,*

(e) $\zeta(\theta) = -1$ *μ-a.e. and* $\operatorname{supp}\{\mu\} \subseteq \{\theta \mid \overline{u}(\theta) = 0\}$.

Proof of Lemma 1. Implications (i) and (a) are identical. Implication (ii) requires

$$\dot{p}(\theta) \in \overline{\operatorname{co}}\Big\{\eta \,\Big|\, \Big(\eta,\, p(\theta) + \int_{[\underline{\theta},\theta)} \zeta(\tilde{\theta})\mu(d\tilde{\theta}), -\lambda\Big) \in N_{\operatorname{epi}(L(\theta,\cdot,\cdot))}\big(\overline{x}(\theta), \dot{\overline{x}}(\theta), L\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big)\big)\Big\}, \quad \text{a.e.}$$

Because $L(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)) = f(\theta)\overline{x}(\theta) - s(\theta, \dot{\overline{x}}(\theta))$ is finite, the limiting normal cone in the above expression can be written as

$$\begin{aligned} &N_{\operatorname{epi}(L(\theta,\cdot,\cdot))}\big(\overline{x}(\theta), \dot{\overline{x}}(\theta), L\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big)\big) \\ &\quad = \big\{(\xi d_1, \xi d_2, -\xi) \mid \xi > 0, (d_1, d_2) \in \partial\big(f(\theta)\overline{x}(\theta) - s\big(\theta, \dot{\overline{x}}(\theta)\big)\big)\big\} \\ &\qquad \cup \big\{\partial^\infty\big(f(\theta)\overline{x}(\theta) - s\big(\theta, \dot{\overline{x}}(\theta)\big)\big) \times \{0\}\big\}. \end{aligned}$$

Using the fact that $L(\cdot)$ is additively separable in x and $\dot{x}$ yields (Rockafellar and Wets (2004, Proposition 10.5))

$$\begin{aligned} \partial\big(f(\theta)\overline{x}(\theta) - s\big(\theta, \dot{\overline{x}}(\theta)\big)\big) &= \partial\big(f(\theta)\overline{x}(\theta)\big) \times \partial\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big) \\ &= \big\{f(\theta)\big\} \times \partial\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big) \end{aligned}$$

and

$$\partial^\infty\big(f(\theta)\overline{x}(\theta) - s\big(\theta, \dot{\overline{x}}(\theta)\big)\big) \subseteq \partial^\infty\big(f(\theta)\overline{x}(\theta)\big) \times \partial^\infty\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big)$$
$$= \{0\} \times \partial^\infty\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big),$$

where the last equality uses the fact that a linear function is Lipschitz continuous, and hence $\partial^\infty(f(\theta)\overline{u}(\theta)) = \{0\}$. Substituting these subdifferentials into the expression for the limiting normal cone, we have a simple inclusion:

$$N_{\mathrm{epi}(L(\theta,\cdot,\cdot))}\big(\overline{x}(\theta), \dot{\overline{x}}(\theta), L\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big)\big)$$
$$\subseteq \big\{\big(\xi f(\theta), \xi d_2, -\xi\big) \mid \xi > 0, d_2 \in \partial\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big)\big\}$$
$$\cup \big\{\{0\} \times \partial^\infty\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big) \times \{0\}\big\}.$$

This simplifies again to the inclusion

$$N_{\mathrm{epi}(L(\theta,\cdot,\cdot))}\big(\overline{x}(\theta), \dot{\overline{x}}(\theta), \overline{L}\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big)\big)$$
$$\subseteq \big\{\big(\xi f(\theta), \xi d_2, -\xi\big) \mid \xi \geq 0, d_2 \in \partial\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big) \cup \partial^\infty\big(-s\big(\theta, \dot{\overline{x}}(\theta)\big)\big)\big\}.$$

The key point to note is that any vector in the limiting normal cone must point in the same direction in the $(\overline{x}, \overline{L})$ plane, regardless of d_2. Returning to implication (ii), we see that any point η in the given convex hull must satisfy $(\eta, \cdot, -\lambda) = (\xi f(\theta), \cdot, -\xi)$ for some $\xi \geq 0$, and hence the convex hull reduces to $\{\lambda f(\theta)\}$. We conclude that implication (ii) simplifies to implication (b) given that $L(\cdot)$ is both additively separable and linear in x.

Implication (iii) is identical to implication (c).

Using the transformation $L(\theta, x, v) = xf(\theta) - s(\theta, v)$, implication (iv) simplifies to implication (d). Lastly, the fact that $h(\theta, x) = -x$ yields $\partial_x h(\theta, \overline{u}(\theta)) = \partial_x^> h(\theta, \overline{x}(\theta)) = \{-1\}$. Thus, implication (v) simplifies to $\zeta(\theta) = -1$ μ-a.e. and

$$\mathrm{supp}\{\mu\} \subseteq \big\{\theta \mid \overline{u}(\theta) = 0\big\}. \tag{25}$$

This is implication (e) and completes the proof to the lemma. □

Returning to the proof of Theorem 1, an immediate inspection of conditions (a)–(e) suggest further simplifications by combining these conditions. Conditions (b) and (c) jointly yield

$$p(\theta) = \lambda F(\theta).$$

Because $p(\overline{\theta}) = \lambda$ and $\zeta(\theta) = -1$ a.e. with respect to μ, condition (c) also implies

$$\int_{\underline{\theta}}^{\overline{\theta}} \mu(d\tilde{\theta}) = \lambda.$$

Because we also have $\max_{\theta\in\Theta} |p(\theta)| = \lambda$, condition (a) implies $\lambda > 0$ and in particular $\lambda = \frac{K}{3}$. Because the choice of K is arbitrary, we choose $K = 3$ as a normalization, yielding

$\lambda = 1$ and $\int_{\underline{\theta}}^{\overline{\theta}} \mu(d\tilde{\theta}) = 1$. Thus, up to this normalization, μ is a probability measure on Θ. Defining now $\overline{\gamma}(\theta) = \int_{[\underline{\theta},\theta)} \mu(d\tilde{\theta})$, the implication in (d) is therefore

$$\dot{\overline{x}}(\theta) \in \arg\max_{v\in\mathbb{R}} s(\theta, v) + \big(F(\theta) - \overline{\gamma}(\theta)\big)v, \quad \text{a.e.} \tag{26}$$

This condition can finally be expressed as (4) and (5) of Theorem 1. Lastly, implication of (e) delivers the complementary slackness condition (3). We have therefore proven the necessity of the conditions in Theorem 1.

Sufficiency. Sufficiency is proven by generalizing Arrow's *sufficiency theorem* to non-smooth optimal control problems and specializing the theorem to the case in which the objective integrand is a linear function of x. We adapt the argument of Arrow's *sufficiency theorem* using the approach of Seierstad and Sydsaeter (1987) but relaxing their continuity and smoothness assumptions. The regularity of the optimal solution follows from arguments involving the necessary conditions.

Let x be any admissible arc satisfying thus $x \in \mathrm{AC}(\Theta, \mathbb{R})$ and $x(\theta) \geq 0$ for all $\theta \in \Theta$. Define

$$\Delta = \int_{\underline{\theta}}^{\overline{\theta}} \big\{\big(s\big(\theta, \dot{\overline{x}}(\theta)\big) - \overline{x}(\theta)f(\theta)\big) - \big(s\big(\theta, \dot{x}(\theta)\big) - x(\theta)f(\theta)\big)\big\}\, d\theta.$$

We will demonstrate that, under conditions (25) and (26) of Theorem 1, $\Delta \geq 0$.

To this end, it is useful to define the Hamiltonian for program (P) with $\overline{\gamma}(\theta)$ being the adjoint equation, which satisfies conditions (25) and (26):

$$H(\theta, x, v) \equiv s(\theta, v) - xf(\theta) - \big(\overline{\gamma}(\theta) - F(\theta)\big)v.$$

Note that $\overline{\gamma}(\theta)$ is defined for $\theta \in (\underline{\theta}, \overline{\theta}]$, and thus $H(\cdot)$ inherits the same domain. Nonetheless, because μ is not part of expression of Δ and F is absolutely continuous, we can ignore the point $\underline{\theta}$ in the integral and conclude that

$$\Delta = \int_{(\underline{\theta},\overline{\theta}]} \big(H\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big) - H\big(\theta, x(\theta), \dot{x}(\theta)\big)\big)\, d\theta + \int_{\underline{\theta}}^{\overline{\theta}} \big(F(\theta) - \overline{\gamma}(\theta)\big)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta.$$

Define the optimized Hamiltonian as

$$\hat{H}(\theta, x) \equiv \sup_{v\in\mathbb{R}} H(\theta, x, v).$$

Because $\overline{\gamma}(\theta) - F(\theta)$ is bounded on $(\underline{\theta}, \overline{\theta}]$ and $s(\theta, \cdot)$ is bounded from above by assumption, $\hat{H}(\cdot)$ must be finite. Condition (26) implies that

$$\hat{H}\big(\theta, \overline{x}(\theta)\big) = H\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big)$$

and for any admissible $x \in \mathrm{AC}(\Theta; \mathbb{R}_+)$,

$$\hat{H}\big(\theta, x(\theta)\big) \geq H\big(\theta, x(\theta), \dot{x}(\theta)\big).$$

Combining these facts, we obtain

$$H\big(\theta, \overline{x}(\theta), \dot{\overline{x}}(\theta)\big) - H\big(\theta, x(\theta), \dot{x}(\theta)\big) \geq \hat{H}\big(\theta, \overline{x}(\theta)\big) - \hat{H}\big(\theta, x(\theta)\big)$$
$$= f(\theta)\big(x(\theta) - \overline{x}(\theta)\big).$$

The last statement relies fundamentally on the linearity of $H(\cdot)$ in x. Substituting into the previous statement for Δ, we have

$$\begin{aligned}
\Delta &\geq \int_{(\underline{\theta}, \overline{\theta}]} f(\theta)\big(x(\theta) - \overline{x}(\theta)\big)\, d\theta + \int_{\underline{\theta}}^{\overline{\theta}} \big(F(\theta) - \overline{\gamma}(\theta)\big)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta \\
&= \int_{\underline{\theta}}^{\overline{\theta}} \big(f(\theta)\big(x(\theta) - \overline{x}(\theta)\big) + F(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\big)\, d\theta - \int_{(\underline{\theta}, \overline{\theta}]} \overline{\gamma}(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta \\
&= \int_{\underline{\theta}}^{\overline{\theta}} \frac{d}{d\theta}\big[F(\theta)\big(x(\theta) - \overline{x}(\theta)\big)\big]\, d\theta - \int_{(\underline{\theta}, \overline{\theta}]} \overline{\gamma}(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta \\
&= \big(x(1) - \overline{x}(1)\big) - \int_{(\underline{\theta}, \overline{\theta}]} \overline{\gamma}(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta.
\end{aligned}$$

It follows that $\Delta \geq 0$ if

$$\big(x(1) - \overline{x}(1)\big) - \int_{(\underline{\theta}, \overline{\theta}]} \overline{\gamma}(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta \geq 0.$$

If $\overline{\gamma}$ were absolutely continuous, we would be able to integrate the second term by parts and reach such a conclusion. Because $\overline{\gamma}$ has possibly countable upward discontinuities, we must proceed more carefully. Note that $\overline{\gamma}$ is nondecreasing on $(\underline{\theta}, \overline{\theta}]$ with at most a countable number of upward jump discontinuities. $\overline{\gamma}$ is thus the sum of a countable number of singular jump functions plus a measure $d\mu(\theta)$, which is absolute continuous with respect to the Lebesgue measure, and thus write as $d\mu(\theta) = \nu(\theta)\, d\theta$.[23] Denote the set of jump discontinuities by $\{\tau_1, \tau_2, \dots\}$, a possibly infinite but countable set. Let $\mathcal{I}$ be the index set of τ_i and let define the size of the jump discontinuity at any τ_i by $\Delta\mu(\tau_i) = \overline{\gamma}(\tau_i^+) - \overline{\gamma}(\tau_i) > 0$. We thus write

$$\overline{\gamma}(\theta) = \sum_{\tau_i \leq \theta, i \in \mathcal{I}} \Delta\mu(\tau_i) + \int_{\underline{\theta}}^{\theta} \nu(\theta)\, d\theta.$$

Since $\overline{\gamma}$ is absolutely continuous outside the discontinuities, we can integrate by parts between any pair of discontinuities. Also note that at any such upward jump point, τ, $\overline{\gamma}$ is left and right continuous with $\overline{\gamma}(\tau) < \overline{\gamma}(\tau^+)$ and (by condition (25)) we have $\overline{x}(\tau^+) = 0$.

Between any two points τ_i and τ_{i+1}, we know

$$\begin{aligned}
&\int_{(\tau_i, \tau_{i+1}]} \overline{\gamma}(\theta)\big(\dot{u}(\theta) - \dot{\overline{v}}(\theta)\big)\, d\theta \\
&\quad = \overline{\gamma}(\theta)\big(x(\theta) - \overline{x}(\theta)\big)\big|_{t=\tau_i^+}^{\tau_{i+1}} - \int_{(\tau_i, \tau_{i+1})} \big(x(\theta) - \overline{x}(\theta)\big)\nu(\theta)\, d\theta
\end{aligned}$$

[23] Royden (1988).

$$= \overline{\gamma}(\tau_{i+1})\big(x(\tau_{i+1}) - \overline{x}(\tau_{i+1})\big) - \overline{\gamma}\big(\tau_i^+\big)\big(x(\tau_i) - \overline{x}(\tau_i)\big)$$
$$- \int_{(\tau_i, \tau_{i+1})} \big(x(\theta) - \overline{x}(\theta)\big)\nu(\theta)\, d\theta.$$

The second equality above uses the fact that x and $\overline{x}$ are continuous on Θ.

Then we may write

$$\int_{(\underline{\theta}, \overline{\theta}]} \overline{\gamma}(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta$$
$$= \sum_{i \in \mathcal{I}} \overline{\gamma}(\tau_{i+1})\big(x(\tau_{i+1}) - \overline{x}(\tau_{i+1})\big) - \big(\Delta\mu(\tau_i) + \overline{\gamma}(\tau_i)\big)\big(x(\tau_i) - \overline{x}(\tau_i)\big)$$
$$- \int_{(\tau_i, \tau_{i+1})} \big(x(\theta) - \overline{x}(\theta)\big)\nu(\theta)\, d\theta$$
$$= \big(x(1) - \overline{x}(1)\big) - \sum_{i \in \mathcal{I}} \Delta\mu(\tau_i)\big(x(\tau_i) - \overline{x}(\tau_i)\big) - \int_{(\tau_i, \tau_{i+1})} \big(x(\theta) - \overline{x}(\theta)\big)\nu(\theta)\, d\theta.$$

By complementary slackness in condition (25), we know $\overline{x}(\theta)\nu(\theta) = 0$ and at any jump point τ_i we must have $\overline{x}(\tau_i) = 0$. Thus,

$$\int_{(\underline{\theta}, \overline{\theta}]} \overline{\gamma}(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta = \big(x(1) - \overline{x}(1)\big) - \sum_{i \in \mathcal{I}} \Delta\mu(\tau_i) x(\tau_i) - \int_{(\tau_i, \tau_{i+1})} x(\theta)\nu(\theta)\, d\theta.$$

We deduce

$$\Delta \geq \big(x(1) - \overline{x}(1)\big) - \int_{(\underline{\theta}, \overline{\theta}]} \overline{\gamma}(\theta)\big(\dot{x}(\theta) - \dot{\overline{x}}(\theta)\big)\, d\theta$$
$$= \sum_{i \in \mathcal{I}} \Delta\mu(\tau_i) x(\tau_i) + \int_{(\tau_i, \tau_{i+1})} x(\theta)\nu(\theta)\, d\theta.$$

Because $x(\theta) \geq 0$, μ is a nonnegative measure, and jump discontinuities $\Delta\mu(\tau_i)$ are positive, we conclude that $\Delta \geq 0$ as claimed. We have proven that conditions (25) and (26) are sufficient for a solution, thereby completing the proof of the theorem. □

Appendix C: Proofs of Propositions

Proof of Proposition 1: Smoothness of the Solution $\overline{x}$. We add the hypothesis that

$$\mathcal{V}(\theta, \sigma) \equiv \arg\max_{v \in \mathbb{R}} s(\theta, v) + \big(F(\theta) - \sigma\big)v$$

is single-valued and continuous for $(\theta, \sigma) \in \Theta \times [0, 1]$. It follows that $\mathcal{V}(\theta, \sigma)$ is nonincreasing in σ and from condition (26), that $\dot{\overline{x}}(\theta) = v(\theta, \overline{\gamma}(\theta))$ a.e.

Suppose to the contrary that $\dot{\overline{x}}$ is discontinuous at some point $\tau \in \Theta$. Initially, suppose that Condition (26) is extended to hold for all $\theta \in \Theta$ rather than for a.e. $\theta \in (\underline{\theta}, \overline{\theta}]$; call this Condition (26'). Condition (26') and the additional hypothesis that $\mathcal{V}(\theta, \sigma)$ is

continuous in (θ, σ) jointly imply that $\dot{\overline{x}}(\theta)$ is discontinuous at τ only if $\overline{\gamma}$ is also discontinuous at τ. Any discontinuity in $\overline{\gamma}$, however, must be an upward jump, $\overline{\gamma}(\tau^+) - \overline{\gamma}(\tau) > 0$, implying that $\dot{\overline{x}}(\theta)$ must jump downwards. Complementary slackness (Condition (25), however, imposes that $\overline{x}(\tau) = 0$, with the implication that a downward discontinuity at τ would violate the state constraint $u(\theta) \geq 0$ in the neighborhood to the immediate right of τ. Hence, continuity must hold for all points $\theta \in [\underline{\theta}, \theta)$ under Condition (26'). Furthermore, because $\overline{\gamma}$ is left continuous at $t = 1$, no jump in $\dot{\overline{x}}(\theta)$ is possible at this endpoint. We conclude that Condition (26') implies that $\dot{\overline{x}}(\theta)$ is continuous for all $\theta \in \Theta$. The weaker Condition (26) allows $\dot{\overline{x}}(\theta)$ to violate the maximization condition on sets of measure zero, including at $\theta = \underline{\theta}$. But such violations have no effect on the solution $\overline{x}$, which is absolutely continuous. Thus, $\overline{x}$ is smooth as posited. □

Proof of Proposition 2. We check that the consumption allocation (12) and the adjoint equation (13) satisfy the necessary and sufficient conditions of Theorem 1, (3)–(5).

The posited adjoint equation, $\overline{\gamma}_0$, is positive over the set $(\tilde{\theta}, \overline{\theta}]$ and zero elsewhere. Complementary slackness (3) is satisfied only if the participation constraint is binding on $[\tilde{\theta}, \overline{\theta}]$ and slack elsewhere. Because $k = 0$, the surplus function is smooth and concave, hence (4) is immediate to verify. Finally, the first-order condition (5) requires that

$$0 = (\theta - \underline{\theta}) - S + \theta + \overline{q}_0(\theta), \quad \text{for } \theta \in [\underline{\theta}, \tilde{\theta})$$

and

$$0 = (\theta - \underline{\theta}) - (\theta - \tilde{\theta}) - S + \theta + \overline{q}_0(\theta), \quad \text{for } \theta \in [\tilde{\theta}, \overline{\theta}].$$

In the first case, $\overline{q}_0(\theta) = S - 2\theta + \underline{\theta} = q^m(\theta)$; in the second case, $\overline{q}_0(\theta) = \hat{S} - \theta = \hat{q}(\theta)$. Because

$$q^m(\theta) > \hat{q}(\theta) \quad \Longleftrightarrow \quad S - 2\theta + \underline{\theta} > \hat{S} - \theta \quad \Longleftrightarrow \quad \theta < \tilde{\theta},$$

we conclude that (5) is satisfied if $\overline{q}_0(\theta) = \max\{q^m(\theta), \hat{q}(\theta)\}$ as in (12). □

Proof of Proposition 3. Using (16) and $\hat{q}(\theta) = \hat{S} - \theta$, we compute

$$\theta_1 = \hat{S} - 1 - (\sqrt{2} - 1)\sqrt{k}, \tag{27}$$

$$\theta_2 = \hat{S} - 1 - \left(\frac{\sqrt{2}}{2} - 1\right)\sqrt{k}, \tag{28}$$

$$\theta_3 = \hat{S} - 1 - (\sqrt{2} - 2)\sqrt{k}. \tag{29}$$

For $k > 0$, it follows that $\theta_1 < \theta_2 < \theta_3$. Observe that, for k sufficiently small, $\underline{\theta} < \tilde{\theta} < \theta_1$ and $\theta_3 < \overline{\theta}$ (using assumption (15)), thus allowing us to conclude

$$\underline{\theta} < \tilde{\theta} < \theta_1 < \theta_2 < \theta_3 < \overline{\theta}.$$

Starting from the expression of $s(\theta, v)$, we compute its majorization:

$$\overline{\mathrm{co}}(s)(\theta, v) = \begin{cases} (S-\theta)\big(\hat{q}(\theta)-v\big) - \frac{1}{2}\big(\hat{q}(\theta)-v\big)^2 \\ \quad \text{if } v \geq v_2(\theta), \\ (S-\theta)\big(\hat{q}(\theta)-v_1(\theta)\big) - \frac{1}{2}\big(\hat{q}(\theta)-v_1(\theta)\big)^2 - k - \big(\Delta + v_1(\theta)\big)\big(v - v_1(\theta)\big) \\ \quad \text{if } v \in \big[v_1(\theta), v_2(\theta)\big], \\ (S-\theta)\big(\hat{q}(\theta)-v\big) - \frac{1}{2}\big(\hat{q}(\theta)-v\big)^2 - k \\ \quad \text{if } v \leq v_1(\theta) \end{cases} \tag{30}$$

where $v_1(\theta) = \hat{q}(\theta) - 1 - \sqrt{2k}$ and $v_2(\theta) = \hat{q}(\theta) - 1$. Using (30), we obtain the expression of the subdifferential

$$\partial_v \overline{\mathrm{co}}(s)(\theta, v) = \begin{cases} -\Delta - v & \text{if } v \leq v_1(\theta) \text{ or } v > v_2(\theta), \\ -\Delta - v_1(\theta) & \text{if } v \in \big[v_1(\theta), v_2(\theta)\big), \\ \big[-\Delta - v_2(\theta), -\Delta - v_1(\theta)\big] & \text{if } v = v_2(\theta). \end{cases} \tag{31}$$

We next check that the pair $(\dot{\overline{u}} = \hat{q} - \overline{q}, \overline{\gamma})$, where $\overline{q}$ is defined in (17) and $\overline{\gamma}$ in (18), satisfies the necessary and sufficient conditions for optimality of Theorem 1. The corresponding conjecture is that (1) is binding on $[\tilde{\theta}, \theta_1] \cup [\theta_3, \overline{\theta}]$ and slack elsewhere. The new adjoint function $\overline{\gamma}$ is thus similar to $\overline{\gamma}_0$ on $[\underline{\theta}, \theta_1)$ but it has an upward jump of $\sqrt{k}$ at θ_1. This jump is followed by a plateau corresponding to the interval (θ_1, θ_3) where (1) is again slack.

Inserting this conjecture into (14) and using (31) yields the following conditions:

- If $\theta \in [\underline{\theta}, \tilde{\theta})$, (1) slack,

$$\overline{\gamma}(\theta) = 0, \tag{32}$$

and thus (14) implies

$$-\theta + \underline{\theta} = \partial_v \overline{\mathrm{co}}(s)\big(\theta, \dot{\overline{u}}(\theta)\big) = -\Delta - \dot{\overline{u}}(\theta) \quad \Longleftrightarrow \quad \overline{q}(\theta) = q^m(\theta). \tag{33}$$

For $\partial_v \overline{\mathrm{co}}(s)(\theta, \dot{\overline{u}}(\theta))$ to be single-valued and equal at $-\Delta - \dot{\overline{u}}(\theta)$ as conjectured, (31) implies that a sufficient condition is that $\dot{\overline{u}}(\theta) \leq v_1(\theta)$ or

$$-\theta + \underline{\theta} + \Delta \geq -v_1(\theta) = -\hat{S} + \theta + 1 + \sqrt{2k} \quad \Longleftrightarrow \quad q^m(\theta) \geq 1 + \sqrt{2k}. \tag{34}$$

Observe that $\theta \leq \tilde{\theta}$ means $q^m(\theta) \geq \hat{q}(\theta)$, and thus (34) holds when $\hat{q}(\tilde{\theta}) \geq 1 + \sqrt{2k}$, which is true when $\theta \leq \tilde{\theta} < \theta_1$.

- If $\theta \in [\tilde{\theta}, \theta_1]$, (1) is binding and, by differentiating, we get $\dot{\overline{u}}(\theta) = 0$ on $(\tilde{\theta}, \theta_1)$. Given (18),

$$\overline{\gamma}(\theta) = \theta - \tilde{\theta} \tag{35}$$

which is positive and nondecreasing. Inserting into (14), the optimality condition writes as

$$-\Delta \in \partial_v \overline{\mathrm{co}}(s)(\theta, 0).$$

When $\partial_v \overline{\mathrm{co}}(s)(\theta, 0) = -\Delta$, the optimal consumption level is such that

$$\overline{q}(\theta) = \hat{q}(\theta). \tag{36}$$

From (31), a sufficient condition for having having $\partial_v \overline{\mathrm{co}}(s)(\theta, 0) = -\Delta$ is thus that

$$\dot{\overline{u}}(\theta) = 0 \leq v_1(\theta) \quad \Longleftrightarrow \quad \hat{q}(\theta) \geq 1 + \sqrt{2k}$$

which is implied by $\theta \leq \theta_1$. Because (1) is binding at θ_1, we allow for some positive charge at θ_1, say γ_1, and we will show below that it is given by

$$\gamma_1 = \sqrt{k}. \tag{37}$$

We can thus write

$$\overline{\gamma}(\theta_1) = \theta_1 - \tilde{\theta} + \gamma_1.$$

- If $\theta \in [\theta_1, \theta_3)$, we conjecture that (1) is slack, and thus $\overline{\gamma}(\theta)$ is constant on this interval, namely

$$\overline{\gamma}(\theta) = \theta_1 - \tilde{\theta} + \gamma_1. \tag{38}$$

Inserting into (14), the optimality condition can now be written as

$$\theta_1 - \tilde{\theta} + \gamma_1 - \theta + \underline{\theta} \in \partial_v \overline{\mathrm{co}}(s)\big(\theta, \dot{\overline{u}}(\theta)\big). \tag{39}$$

We now consider two subcases.

– For $\theta \in (\theta_1, \theta_2]$, we have $\partial_v \overline{\mathrm{co}}(s)(\theta, \dot{\overline{u}}(\theta)) = [-\Delta - v_2(\theta), -\Delta - v_1(\theta)]$, which means

$$\dot{\overline{u}}(\theta) = v_2(\theta) \quad \Longleftrightarrow \quad \overline{q}(\theta) = 1. \tag{40}$$

Condition (39) becomes

$$\theta_1 - \tilde{\theta} + \gamma_1 - \theta + \underline{\theta} \in \big[-\Delta - v_2(\theta), -\Delta - v_1(\theta)\big] \quad \Longleftrightarrow \quad \theta - \theta_1 - \gamma_1 \in \big[v_1(\theta), v_2(\theta)\big].$$

We rewrite this condition as

$$\theta_1 - \gamma_1 - \theta \in \big[-\hat{q}(\theta) + 1 + \sqrt{2k}, -\hat{q}(\theta) + 1\big]$$

or $\theta \in [\theta_1, \theta_2]$ with the definitions (27) and (28).

– For $\theta \in (\theta_2, \theta_3]$, we have $\partial_v \overline{\mathrm{co}}(s)(\theta, \dot{\overline{u}}(\theta)) = -\Delta - \dot{\overline{u}}(\theta)$. Inserting into (39) yields

$$\theta_1 - \tilde{\theta} + \gamma_1 - \theta + \underline{\theta} = -\Delta - \dot{\overline{u}}(\theta) \quad \Longleftrightarrow \quad \overline{q}(\theta) = q^m(\theta) + \theta_1 - \tilde{\theta} + \gamma_1. \tag{41}$$

A sufficient condition for writing $\partial_v \overline{\mathrm{co}}(s)(\theta, \dot{\overline{u}}(\theta)) = -\Delta - \dot{\overline{u}}(\theta)$ as above is that $\dot{\overline{u}}(\theta) \geq v_2(\theta)$, which amounts to

$$\theta_1 - \theta \leq -v_2(\theta)$$

or $\theta \geq \theta_2$, which holds for this subcase.

- If $\theta \in [\theta_3, \overline{\theta}]$, (1) is binding and, by differentiating, we get $\dot{\overline{u}}(\theta) = 0$ on $(\theta_3, \overline{\theta})$. From (14), we obtain

$$\overline{\gamma}(\theta) - \theta + \underline{\theta} = \partial_v \overline{\mathrm{co}}(s)(\theta, 0) = -\Delta.$$

Thus, on $[\theta_3, \overline{\theta}]$, we have

$$\overline{\gamma}(\theta) = \theta - \underline{\theta} \tag{42}$$

and

$$\overline{q}(\theta) = \hat{q}(\theta). \tag{43}$$

The parameter γ_1 is chosen so that $\overline{u}(\theta_1) = \overline{u}(\theta_3) = 0$. This condition implies

$$0 = \int_{\theta_1}^{\theta_3} \dot{\overline{u}}(\theta)\, d\theta = \int_{\theta_1}^{\theta_3} \big(\hat{q}(\theta) - \overline{q}(\theta)\big)\, d\theta.$$

Because $\overline{q}$ is absolutely continuous on (θ_1, θ_3), we may integrate by parts using $\overline{q}(\theta_3) = \hat{q}(\theta_3)$ to obtain

$$0 = \int_{\theta_1}^{\theta_3} \big(\dot{\hat{q}}(\theta) - \dot{\overline{q}}(\theta)\big)(\theta - \theta_1)\, d\theta.$$

Rewriting,

$$0 = \int_{\theta_1}^{\theta_2} -(\theta - \theta_1)\, d\theta + \int_{\theta_2}^{\theta_3} (\theta - \theta_1)\, d\theta$$

or

$$2(\theta_2 - \theta_1)^2 = (\theta_3 - \theta_1)^2.$$

Geometrically, this condition just says that the algebraic area between the curves $\hat{q}$ and $\overline{q}$ is zero over $[\theta_1, \theta_3]$. Using the definitions (27), (28), and (29) yield

$$\gamma_1^2 = k$$

or (37).

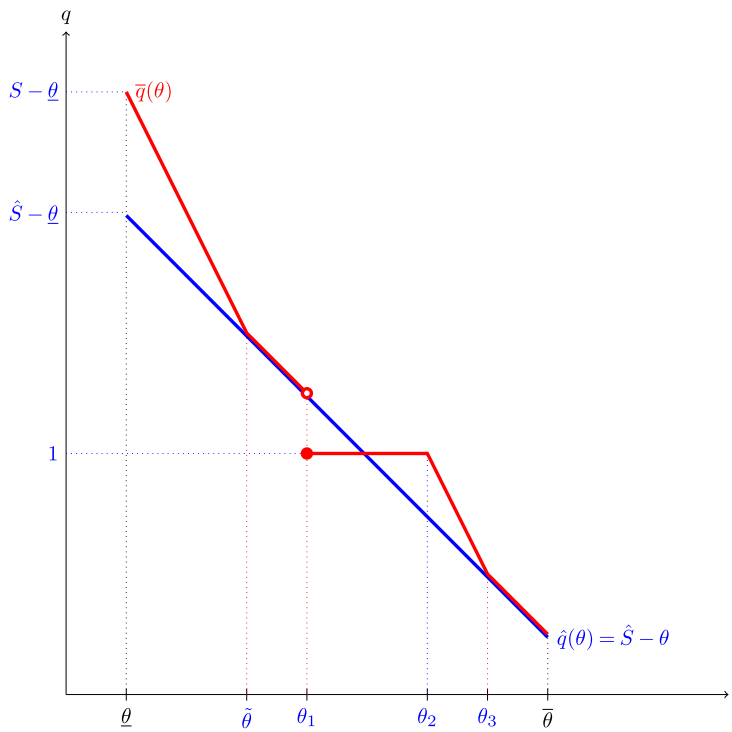

FIGURE 2. Consumption levels: Digital goods.

Summarizing our previous findings in (33), (36), (40), and (41) yield the expression of $\overline{q}$ in (17). (See Figure 2.) On the other hand, (32), (35), (38) and (42) yield the expression of $\overline{\gamma}$ in (18). □

PROOF OF PROPOSITION 4. Standard arguments (see footnote 2) establish that $U(\theta)$ so defined is absolutely continuous, and thus a.e. differentiable with

$$\dot{U}(\theta)=\begin{cases}-q(\theta)-K & \text{if } q(\theta)<\check{q}(\theta),\\ -\check{q}(\theta)-K & \text{if } q(\theta)\in\big(\check{q}(\theta),\check{q}(\theta)+K\big),\\ -q(\theta) & \text{if } (\theta)>\check{q}(\theta)+K.\end{cases} \tag{44}$$

From this and the fact that $\dot{\hat{U}}(\theta)=-K$, we get (20).

We can express $\tilde{s}(\theta,q)$ as

$$\tilde{s}(\theta,q)=\begin{cases}(S-\theta-K)q-\dfrac{q^2}{2}-cq & \text{if } q\le\check{q}(\theta),\\ \dfrac{\big(\check{q}(\theta)+K\big)^2}{2}+\Delta q-\left((S-\theta)K-\dfrac{K^2}{2}-pK\right) & \text{if } q\in\big[\check{q}(\theta),\check{q}(\theta)+K\big],\\ (S-\theta-c)q-\dfrac{q^2}{2}-\left((S-\theta)K-\dfrac{K^2}{2}-pK\right) & \text{if } q\ge\check{q}(\theta)+K.\end{cases}$$

Observing that $\check{q}(\theta)+K=S-\theta-p$ and simplifying yields

$$\tilde{s}(\theta,q)=\begin{cases}\big(\check{q}(\theta)+\Delta\big)q-\dfrac{q^2}{2} & \text{if } q\le\check{q}(\theta),\\ \dfrac{\check{q}(\theta)^2}{2}+\Delta q & \text{if } q\in\big[\check{q}(\theta),\check{q}(\theta)+K\big],\\ (S-\theta-c)q-\dfrac{q^2}{2}-\left((S-\theta-p)K-\dfrac{K^2}{2}\right) & \text{if } q\ge\check{q}(\theta)+K.\end{cases}$$

Expressing $q(\theta)$ in terms of $\dot{u}(\theta)$ over the different intervals yields

$$q(\theta)\begin{cases}=-\dot{u}(\theta) & \text{if } \dot{u}(\theta)>-\check{q}(\theta),\\ \in\big[\check{q}(\theta),\check{q}(\theta)+K\big] & \text{if } \dot{u}(\theta)=-\check{q}(\theta),\\ =-\dot{u}(\theta)+K & \text{if } \dot{u}(\theta)<-\check{q}(\theta).\end{cases}\tag{45}$$

Inserting these expressions of $q(\theta)$ into the definition of $\tilde{s}(\theta,q)$ above yield $s(\theta,v)$ as (21).

From there, we now compute

$$\begin{aligned}&\overline{\mathrm{co}}(s)(\theta,v)\\ &=\begin{cases}-\big(\check{q}(\theta)+\Delta\big)v-\dfrac{v^2}{2} & \text{if } v\ge v_2(\theta),\\ -(\sqrt{2\Delta K}+\Delta)\big(v-v_2(\theta)\big)-\big(\check{q}(\theta)+\Delta\big)v_2(\theta)-\dfrac{v_2^2(\theta)}{2} & \text{if } v\in\big[-\check{q}(\theta),v_2(\theta)\big],\\ -\big(\check{q}(\theta)+\Delta\big)v-\dfrac{v^2}{2}+\Delta K & \text{if } v<-\check{q}(\theta)\end{cases}\end{aligned}$$

where $v_2(\theta)=\sqrt{2\Delta K}-\check{q}(\theta)$.

This yields the following expression of the subdifferential for $\overline{\mathrm{co}}(s)(\theta,v)$:

$$\partial_v\overline{\mathrm{co}}(s)(\theta,v)=\begin{cases}-\check{q}(\theta)-\Delta-v & \text{if } v\ge v_2(\theta) \text{ and if } v<-\check{q}(\theta),\\ -\sqrt{2\Delta K}-\Delta & \text{if } v\in\big(-\check{q}(\theta),v_2(\theta)\big],\\ [-\sqrt{2\Delta K}-\Delta,-\Delta] & \text{if } v=-\check{q}(\theta).\end{cases}$$

With a uniform distribution, the optimality condition (5) becomes

$$\overline{\gamma}(\theta)-\theta+\underline{\theta}\in\partial_v\overline{\mathrm{co}}(s)\big(\theta,\dot{\overline{u}}(\theta)\big)$$

or

$$\overline{\gamma}(\theta)-\theta+\underline{\theta}\begin{cases}=-\check{q}(\theta)-\Delta-\dot{u}(\theta) & \text{if } \dot{u}(\theta)\ge v_2(\theta) \text{ and if } \dot{u}(\theta)<-\check{q}(\theta),\\ =-\sqrt{2\Delta K}-\Delta & \text{if } \dot{u}(\theta)\in\big(-\check{q}(\theta),v_2(\theta)\big],\\ \in[-\sqrt{2\Delta K}-\Delta,-\Delta] & \text{if } \dot{u}(\theta)=-\check{q}(\theta).\end{cases}\tag{46}$$

We conjecture a solution $(\overline{u}(\theta),\overline{\gamma}(\theta))$ such that (1) binds at $\overline{\theta}$ only, and thus $\mu(\{\overline{\theta}\})>0$ with $\overline{\gamma}(\theta)=0$ on $[\underline{\theta},\overline{\theta})$. Thanks to the sufficiency part of our theorem, we only check that this solution satisfies the necessary conditions for optimality.

- On the interval $[\underline{\theta}, \tilde{\theta})$, this conjecture implies $\overline{\gamma}(\theta) = 0$. Inserting into (46) yields

$$-\theta + \underline{\theta} = \partial_v \overline{\mathrm{co}}(s)\big(\theta, \dot{\overline{u}}(\theta)\big) = -\check{q}(\theta) - \Delta - \dot{\overline{u}}(\theta).$$

  Because $\theta \le \tilde{\theta} = \underline{\theta} + \Delta$, we thus have

$$\dot{\overline{u}}(\theta) = -\check{q}(\theta) + \theta - \tilde{\theta} < -\check{q}(\theta).$$

  From (45), we deduce that

$$-\overline{q}(\theta) + K = \dot{\overline{u}}(\theta) = -(S - \theta - p - K) + \theta - \tilde{\theta},$$

  and thus

$$\overline{q}(\theta) = S - 2\theta + \underline{\theta} - c = q^m(\theta).$$

- On the interval $[\tilde{\theta}, \tilde{\theta}_0]$, we have

$$\dot{\overline{u}}(\theta) = -\check{q}(\theta).$$

  Indeed, imposing our conjecture $\overline{\gamma}(\theta) = 0$ on (46) yields

$$-\theta + \underline{\theta} \in \partial_v \overline{\mathrm{co}}(s)\big(\theta, -\check{q}(\theta)\big) = [-\sqrt{2\Delta K} - \Delta, -\Delta] \quad \Longleftrightarrow \quad \theta \in [\tilde{\theta}, \tilde{\theta}_0].$$

  From (45), we deduce that

$$\overline{q}(\theta) \in [S - \theta - p - K, S - \theta - p] = \big[\check{q}(\theta), \check{q}(\theta) + K\big].$$

- On the interval $[\tilde{\theta}_0, \overline{\theta})$, our conjecture is $\overline{\gamma}(\theta) = 0$. Inserting into (46) yields

$$-\theta + \underline{\theta} = \partial_v \overline{\mathrm{co}}(s)\big(\theta, \dot{\overline{u}}(\theta)\big) = -\check{q}(\theta) - \dot{\overline{u}}(\theta) - \Delta.$$

  Because $\theta \ge \tilde{\theta}_0 > \tilde{\theta}$, we have

$$\dot{\overline{u}}(\theta) = -\check{q}(\theta) + \theta - \tilde{\theta} > -\check{q}(\theta) + \theta - \tilde{\theta}_0 \ge -\check{q}(\theta).$$

  From (45), we deduce that

$$-\overline{q}(\theta) = \dot{\overline{u}}(\theta) = -(S - \theta - p - K) + \theta - \tilde{\theta},$$

  and thus

$$\overline{q}(\theta) = S - 2\theta + \underline{\theta} - c - K = q^m(\theta) - K.$$

Gathering all the above findings yields (22). (See Figure 3 below.) The condition (19) ensures that the incumbent firm still wants to serve the type with the lowest possible demand $\overline{\theta}$ even when that type consumes up to the fringe's capacity. It implies that $\overline{u}(\theta)$ is everywhere decreasing, consistently with (1) being binding at $\overline{\theta}$ only. Hence, $\mu(\{\overline{\theta}\}) = 1$ as required. □

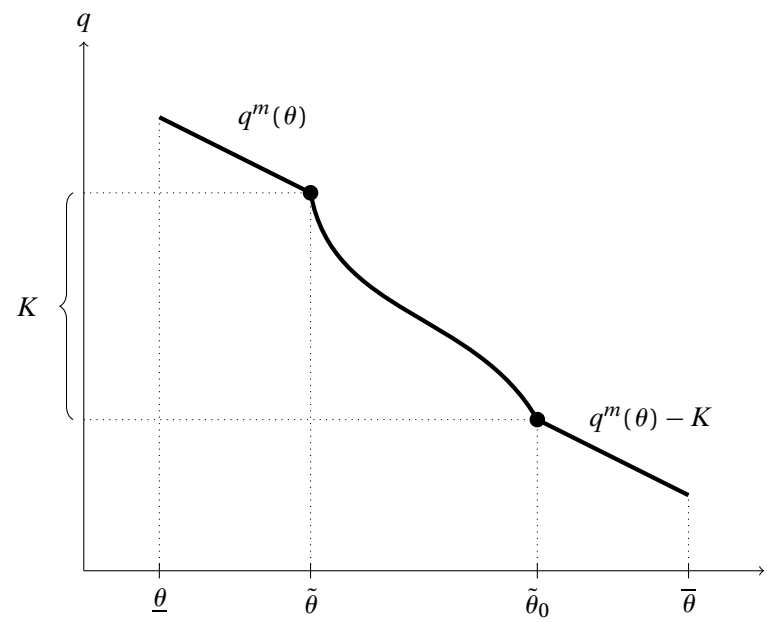

FIGURE 3. Consumption levels. Split purchases.

REFERENCES

Aghion, Philippe and Patrick Bolton (1987), "Contracts as a barrier to entry." *The American Economic Review*, 77, 388–401. [1158]

Attar, Andrea, Thomas Mariotti, and Francois Salanié (2011), "Nonexclusive competition in the market for lemons." *Econometrica*, 79, 1869–1891. [1158]

Baron, David and Roger Myerson (1982), "Regulating a monopolist with unknown costs." *Econometrica*, 50, 911–930. [1148]

Biglaiser, Gary and Claudio Mezzetti (1993), "Principals competing for an agent in the presence of adverse selection and moral hazard." *Journal of Economic Theory*, 61, 302–330. [1146]

Brainard, S. Lael and David Martimort (1997), "Strategic trade policy with incompletely informed policymakers." *Journal of International Economics*, 42, 33–65. [1146]

Caillaud, Bernard (1990), "Regulation, competition, and asymmetric information." *Journal of Economic Theory*, 52, 87–110. [1146, 1151, 1158]

Calzolari, Giacomo and Vincenzo Denicolo (2013), "Competition with exclusive contracts and market-share discounts." *The American Economic Review*, 103, 2384–2411. [1146]

Calzolari, Giacomo and Vincenzo Denicolo (2015), "Exclusive contracts and market dominance." *American Economic Review*, 105, 3321–3351. [1156, 1158]

Carbajal, Juan Carlos and Jeffrey Ely (2013), "Mechanism design without revenue equivalence." *Journal of Economic Theory*, 148, 104–133. [1147]

Carbajal, Juan Carlos and Jeffrey Ely (2016), "A model of price discrimination under loss aversion and state-contingent reference points." *Theoretical Economics*, 11, 455–485. [1160]

Champsaur, Paul and Jean-Charles Rochet (1989), "Multiproduct duopolists." *Econometrica*, 57, 533–557. [1153, 1156]

Choné, Philippe and Laurent Linnemer (2015), "Nonlinear pricing and exclusion: I. Buyer opportunism." *The RAND Journal of Economics*, 46, 217–240. [1158]

Cremer, Jacques, Fahad Khalil, and Jean-Charles Rochet (1998), "Contracts and productive information gathering." *Games and Economic Behavior*, 25, 174–193. [1151]

Curien, Nicolas, Bruno Jullien, and Patrick Rey (1998), "Pricing regulation under bypass competition." *The RAND Journal of Economics*, 29, 259–279. [1146, 1151, 1158]

Deb, Rahul and Maher Said (2015), "Dynamic screening with limited commitment." *Journal of Economic Theory*, 159, 891–928. [1146]

Ferrera, Juan (2014), *An Introduction to Nonsmooth Analysis*. Academic Press. [1163]

Galbraith, Grant and Richard Vinter (2004), "Regularity of optimal controls for state-constrained problems." *Journal of Global Optimization*, 28, 305–317. [1150]

Guesnerie, Roger and Jean-Jacques Laffont (1984), "A complete solution to a class of principal–agent problems with an application to the control of a self-managed firm." *Journal of Public Economics*, 25, 329–369. [1151]

Hellwig, Martin (2010), "Incentive problems with unidimensional hidden characteristics: A unified approach." *Econometrica*, 78, 1201–1237. [1151]

Huang, Ke-Wei and Arun Sundararajan (2011), "Pricing digital goods: Discontinuous costs and shared infrastructure." *Information Systems Research*, 22, 721–738. [1153]

Jullien, Bruno (2000), "Participation constraints in adverse selection models." *Journal of Economic Theory*, 93, 1–47. [1145, 1146, 1148, 1150, 1151, 1152, 1155, 1156]

Laffont, Jean-Jacques and David Martimort (2002), *The Theory of Incentives: The Principal–Agent Model*. Princeton University Press. [1145, 1148]

Laffont, Jean-Jacques and Jean Tirole (1990), "Optimal bypass and cream skimming." *The American Economic Review*, 80, 1042–1061. [1151, 1158]

Lewis, Tracy and David Sappington (1989), "Countervailing incentives in agency problems." *Journal of Economic Theory*, 49, 294–313. [1145, 1151]

Lewis, Tracy and David Sappington (1993), "Ignorance in agency problems." *Journal of Economic Theory*, 61, 169–183. [1151]

Maggi, Giovanni and Andrés Rodriguez-Clare (1995), "On countervailing incentives." *Journal of Economic Theory*, 66, 238–263. [1145, 1151, 1156, 1157, 1158]

Martimort, David and Lars Stole (2009), "Market participation in delegated and intrinsic common agency games." *The RAND Journal of Economics*, 40, 78–102. [1146]

Martimort, David and Lars Stole (2015), "Menu auctions and influence games with private information." Report, Chicago Booth School of Business. [1146]

Milgrom, Paul and Ilya Segal (2002), "Envelope theorems for arbitrary choice sets." *Econometrica*, 70, 583–601. [1147]

Mirrlees, James (1971), "An exploration in the theory of optimum income taxation." *The Review of Economic Studies*, 38, 175–208. [1148]

Myerson, Roger (1981), "Optimal auction design." *Mathematics of Operations Research*, 6, 58–73. [1148, 1151]

Ollier, Sandrine and Lionel Thomas (2013), "Ex post participation constraint in a principal–agent model with adverse selection and moral hazard." *Journal of Economic Theory*, 148, 2383–2403. [1146]

Rochet, Jean-Charles (1987), "A necessary and sufficient condition for rationalizability in a quasi-linear context." *Journal of Mathematical Economics*, 16, 191–200. [1154]

Rockafellar, R. Tyrell and Roger Wets (2004), *Variational Analysis*. Springer, Berlin. [1166]

Rothschild, Michael and Joseph Stiglitz (1976), "Equilibrium in competitive insurance markets: An essay on the economics of imperfect information." *The Quarterly Journal of Economics*, 90, 629–649. [1158]

Royden, Halsey (1988), *Real Analysis*. Prentice Hall, New York. [1169]

Seierstad, Atle and Knut Sydsaeter (1987), *Optimal Control Theory With Economic Applications*. North Holland, Amsterdam. [1150, 1168]

Spulber, David (1993), "Monopoly pricing of capacity usage under asymmetric information." *Journal of Industrial Economics*, 41, 241–257. [1153]

Stole, Lars (1995), "Nonlinear pricing and oligopoly." *Journal of Economics and Management Strategy*, 4, 529–562. [1146, 1153, 1156]

Stole, Lars (2003), "Price discrimination and imperfect competition." In *Handbook of Industrial Organization*, volume 3 (M. Armstrong and R. Porter, eds.), 34–47. [1146, 1156]

Thomas, Lionel (2001), "Cost functions in non-linear pricing." *Economics Letters*, 72, 53–59. [1153]

Toikka, Juuso (2011), "Ironing without control." *Journal of Economic Theory*, 146, 2510–2526. [1151]

Vinter, Richard (2000), *Optimal Control*. Birkhauser, Boston. [1164]

Vinter, Richard and Harry Zheng (1998), "Necessary conditions for optimal control problem with state constraints." *Transactions of the American Mathematical Society*, 350, 1181–1204. [1146, 1151, 1164, 1165, 1166]

Co-editor Simon Board handled this manuscript.

Manuscript received 14 October, 2017; final version accepted 31 May, 2021; available online 26 July, 2021.