Theoretical Economics 17 (2022), 1719–1760 1555-7561/20221719

# Regulating a monopolist with uncertain costs without transfers

Manuel Amador
Federal Reserve Bank of Minneapolis and Department of Economics, University of Minnesota

Kyle Bagwell
Department of Economics, Stanford University

We analyze the Baron and Myerson (1982) model of regulation under the restriction that transfers are infeasible. Extending techniques from the delegation literature to incorporate an ex post participation constraint, we report sufficient conditions under which optimal regulation takes the form of price-cap regulation. We establish conditions under which the optimal price cap is set at a level such that no types are excluded and show that exclusion of higher cost types can be optimal when these conditions fail. We also provide conditions for the optimality of price-cap regulation when an ex post participation constraint is present and exclusion is infeasible.

Keywords. Regulation, price caps, delegation theory.

JEL classification. L43, L5, D82.

## 1. Introduction

The optimal regulatory policy for a monopolist is influenced by many considerations, including the possibility of private information, the objective of the regulator, and the feasibility and efficiency of transfers. Armstrong and Sappington (2007) survey the nature of optimal regulation in different settings and discuss as well the design of practical policies, such as price-cap regulation, that are frequently observed in practice. As they emphasize, an important question is whether practical policies perform well in realistic settings where private information may be present and transfer instruments may be limited.

In a seminal paper, Baron and Myerson (1982) consider the optimal regulation of a single-product monopolist with private information about its costs of production. In their model, a regulatory policy indicates, for every possible cost type, whether the

Manuel Amador: amador.manuel@gmail.com
Kyle Bagwell: kbagwell@stanford.edu
The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. We thank Mark Armstrong, Yeon-Koo Che, Marina Halac, Joe Harrington, Roger Noll, Alessandro Pavan, Michael Riordan, Peter Troyan, Robert Wilson, Frank Wolak, three anonymous referees, and seminar participants at Columbia, Hokkaido, Nottingham, Penn, Ryerson, SMU, Vanderbilt, and USC for helpful discussions. We thank Guillaume Sublet for excellent research assistance. Manuel Amador acknowledges NSF support under award 0952816. Kyle Bagwell thanks the Center for Advanced Studies in the Behavioral Sciences at Stanford for hospitality and support as a Fellow during 2014–2015.

monopoly is allowed to produce at all and, if so, the output and corresponding price that it selects and the transfer from consumers that it receives (where a negative transfer is a tax). A regulatory policy is feasible if it is incentive compatible and satisfies an ex post participation constraint. The regulator chooses over feasible regulatory policies to maximize a weighted social welfare function that weighs consumer surplus no less heavily than producer surplus.[1]

In a standard version of the Baron–Myerson model, the monopolist faces a commonly known and nonnegative fixed cost and is privately informed as to the level of its constant marginal cost, where the marginal cost has a continuum of possible types and is drawn from a commonly known distribution function. If the regulator gives greater welfare weight to consumer surplus, then the optimal regulatory policy defines a nondecreasing price schedule for active types with a positive mark up for all but the lowest cost type. Production is permitted only for types such that consumer surplus under the optimal pricing rule weakly exceeds the fixed cost of production.

In this paper, we characterize optimal regulatory policy in the Baron–Myerson model with constant marginal costs when transfers are infeasible. Our no-transfers assumption contrasts sharply with Baron and Myerson's assumption that all (positive and negative) transfers are available. We motivate our no-transfers assumption in three ways. First, regulators often do not have the authority to explicitly tax or pay subsidies.[2] Second, while transfers from consumers to firms may also be achieved via access fees in two-part tariff schemes, the scope for such transfers may be limited in practice, particularly when universal service is sought for heterogeneous consumers.[3] Finally, in other settings, the scope for a positive access fee may be limited by the possibility of consumer arbitrage, while the scope for a negative access fee may be limited by the prospect of strategic consumer behavior designed to capture "sign-up" bonuses. In view of these considerations, we remove the traditional assumption that all transfers are available and consider the opposite case in which all transfers are infeasible. Specifically, we assume that the regulated firm is restricted to a uniform price (i.e., linear pricing).[4] As our main finding, we report sufficient conditions under which price-cap regulation emerges as the optimal regulatory policy.

As mentioned above, price-cap regulation is a common form of regulation. The appeal of price-cap regulation is often associated with the incentive that it gives to the regulated firm to invest in endogenous cost reduction.[5] By contrast, we establish conditions for the optimality of price-cap regulation in a model in which costs are private and

[1]An alternative approach is developed by Laffont and Tirole (1993, 1986). They assume that the regulator maximizes aggregate social surplus and that transfers entail a social cost of funds.

[2]For further discussion, see, for example, Armstrong and Sappington (2007, p. 1607), Baron (1989, p. 1351), Church and Ware (2000, p. 840), Joskow and Schmalensee (1986, p. 5), Laffont and Tirole (1993, p. 130), and Schmalensee (1989, p. 418).

[3]As Laffont and Tirole (1993, p. 151) explain, "optimal linear pricing is a good approximation to optimal two-part pricing when there is concern that a nonnegligible fixed premium would exclude either too many customers or customers with low incomes whose welfare is given substantial weight in the social welfare function."

[4]In this respect, we follow the lead of Schmalensee (1989). Schmalensee (1989, p. 418) provides additional motivation for the practical relevance of linear pricing schemes in regulatory settings.

[5]See, for example, Armstrong and Sappington (2007, p. 1608) and the references cited therein.

exogenous. Our no-transfers assumption is critical: price-cap regulation is not optimal in the standard Baron–Myerson model with transfers. Our finding thus indicates that this practical regulatory policy may perform not just well but optimally when a regulator faces a privately informed monopolist and transfers are infeasible.

To develop this finding, we consider a "regulator's problem" in which the regulator chooses a menu of permissible outputs, with the understanding that the output choice intended for a monopolist with a given cost type must be the best choice for the monopolist relative to all other permitted output choices. In addition to this incentive compatibility constraint, the regulator faces an ex post participation, or individual rationality (IR), constraint: if the regulator seeks a positive output from a monopolist with a given cost type, then the monopolist must earn more by producing this output than by shutting down and avoiding the nonnegative fixed cost of production. Importantly, the regulator may choose a menu of permissible outputs such that, for some cost types, the monopolist elects to produce zero output and thus earn a profit of zero. As in the original Baron–Myerson model, the regulator may thus design the regulatory policy so as to "exclude" some cost types from production.

The IR constraint plays an important role in our analysis. If we were to ignore this constraint, then the regulator's problem would take the form of a traditional delegation problem and fit into the framework of Amador and Bagwell (2013). We could then use the sufficiency theorems developed in that paper and provide conditions under which a simple price cap (i.e., a quantity floor) is optimal. We show, however, that the IR constraint in fact would be violated for higher cost types when this simple price cap is used.

We consider instead a price-cap allocation where the cap is placed at a price level such that a threshold cost type earns zero profit and is thus indifferent to shut down. No exclusion occurs if the threshold cost type corresponds to the highest cost type in the full support, while exclusion occurs when the threshold cost type falls below the highest possible cost type. Within the set of nonexcluded cost types, higher cost types pool at the price cap, whereas lower cost types may select their monopoly prices. It is also possible that the price cap falls below the monopoly price for the lowest possible cost type, in which case all nonexcluded cost types pool at the price cap. The central task of our analysis is to identify sufficient conditions under which the described price cap with possible exclusion is optimal. We also seek to determine sufficient conditions that indicate when actual exclusion does or does not occur.

To establish our results, we proceed in three main steps. First, we consider the "regulator's truncated problem," wherein the regulator allocates production for cost types up to an exogenous upper-bound cost type and is not allowed to exclude any types in this truncated set. The upper-bound cost type can be fixed at any value that is above the lowest possible cost type and at or below the highest possible cost type in the full support. We then obtain sufficient conditions under which the optimal allocation for the regulator's truncated problem is a price cap set at a level such that the upper-bound cost type earns zero profit and is thus indifferent between producing or not. Second, we argue that this truncated allocation remains feasible when extended to the full support of possible costs if cost types above the upper-bound cost type are excluded (assigned zero output). Finally, we characterize the optimal level of exclusion. This exercise amounts to a single variable optimization problem defined over the upper bound, or threshold, cost type.

Our first proposition establishes a general set of sufficient conditions under which the described price-cap allocation solves the regulator's truncated problem. We then provide a second proposition, which establishes that, if the sufficient conditions for our first proposition hold for any upper-bound cost type, then a price-cap allocation with potential exclusion is optimal within the set of all feasible allocations for the regulator's problem. A key ingredient in making this argument is that the optimal price-cap allocation is such that the threshold cost type is indifferent to shut down.

We also provide several results that facilitate the application of our propositions. Three approaches are developed. First, we show that our sufficient conditions hold if the density is nondecreasing over the full support and if a "relative concavity" condition holds that concerns the relative curvature of the consumer surplus and profit functions, with each expressed as a function of quantity. The relative concavity condition is more likely to hold when the ratio of the concavity of the consumer surplus function to that of the profit function is higher. Second, we identify a family of demand functions under which the sufficient conditions for our propositions hold if a simple inequality is satisfied. The inequality condition holds when the density is nondecreasing over the full support, but it can hold as well when the density is decreasing over part or all of the full support. To illustrate the power of this approach, we show that the family includes linear demand, constant elasticity demand, and log demand functions, and we derive and interpret the corresponding inequality condition for each of these examples. The third approach is to check the sufficient conditions for our propositions directly. We illustrate this approach for an example with an exponential demand function.

Finally, we identify conditions under which actual exclusion does or does not occur, respectively. Our third proposition establishes that no exclusion is optimal under a general set of conditions; specifically, if the density is nondecreasing over the full support and the consumer surplus function is weakly concave, and if the sufficient conditions for our first proposition hold for any upper-bound cost type, then the optimal regulatory policy entails no exclusion and a price cap set at a price level such that the IR constraint for the highest cost type is binding. Thus, optimal regulation then takes the form of a standard second-best price cap that delivers zero profit for the highest cost type. We note that the consumer surplus function is weakly concave in quantity for the log demand and constant elasticity demand examples.

We also analyze the linear demand example. The consumer surplus function associated with this demand function is strictly convex, and so our third proposition cannot be applied. In our fourth and final proposition, we show that, if the distribution of cost types is uniform, the social planner maximizes aggregate social welfare, and the fixed cost of entry is strictly positive, then (a) the price cap is below the monopoly price of the lowest cost type, and thus induces pooling among all non-excluded types, and (b) some higher cost types must be excluded, provided that not all types would pool at the cap were no exclusion to occur (i.e., provided that the sub-monopoly price that generates zero profit for the highest possible cost type is above the monopoly price of the lowest possible cost type). This proposition demonstrates that exclusion of higher cost types can be optimal in some settings.[6,7]

The described results characterize optimal regulatory policy for market settings in which exclusion for some cost types is feasible. Our results thus directly apply when the monopolist provides an inessential service for a given market or region. Since Baron and Myerson (1982) also focus on settings where exclusion is feasible, our findings characterize how their analysis extends when transfers are infeasible.

We are also interested in the "no-exclusion" scenario, wherein the regulator must ensure that the monopolist earns nonnegative profit while providing positive output under all cost realizations. This scenario may be relevant for a monopolist that provides essential services with poor substitution alternatives. To characterize the optimal regulatory policy for this scenario, we refer to our first proposition for the special case in which the upper-bound cost type equals the highest cost type in the full support. Our first proposition then provides conditions under which optimal regulation for the no-exclusion scenario takes the form of a price-cap policy, where the price cap is set at the second-best level that generates zero profit for a monopolist with the highest possible cost type. Likewise, we can facilitate the application of our results to this scenario by using the three approaches described above.

Our work is related to research on optimal delegation. The delegation literature begins with Holmstrom (1977), who considers a setting in which a principal faces a privately informed and biased agent and in which contingent transfers are infeasible. The principal then selects a set of permissible actions from the real line, and the agent selects his preferred action from that set after privately observing the state of nature.[8] A key goal in this literature has been to identify general conditions under which the principal optimally defines the permissible set as an interval. Alonso and Matouschek (2008) consider a setting with quadratic utility functions and provide necessary and sufficient conditions for interval delegation to be optimal. Extending the Lagrangian techniques of Amador, Werning, and Angeletos (2006), Amador and Bagwell (2013) consider a general representation of the delegation problem and establish necessary and sufficient conditions for the optimality of interval delegation.[9]

Our analysis of the regulator's truncated problem builds on the Lagrangian methods used by Amador and Bagwell (2013), but a novel feature of the current paper is that the

[6]See Armstrong (1996) for an analysis of optimal exclusion in the different context of a model of multiproduct nonlinear pricing when the type space is multidimensional.

[7]The setting of linear demand and a uniform distribution is often treated in the literature. Alonso and Matouschek (2008) and Baron and Myerson (1982) illustrate their findings using this example.

[8]A large literature follows Holmstrom's work. See, for example, Amador and Bagwell (2012), Amador and Bagwell (2020), Amador, Bagwell, and Frankel (2018), Amador, Werning, and Angeletos (2006), Ambrus and Egorov (2017), Armstrong and Vickers (2010), Burkett (2016), Frankel (2014), Frankel (2016), Guo (2016), Koessler and Martimort (2012), Martimort and Semenov (2006), Melumad and Shibano (1991), and Mylovanov (2008). Related themes also arise in repeated games with private information; see Athey, Bagwell, and Sanchirico (2004), Athey, Atkeson, and Kehoe (2005), and Halac and Yared (2019).

[9]We note that a cap can be understood as a form of interval delegation, in which the maximum (minimum) action is defined by the cap (the lowest "flexible" choice for any agent type).

analysis is extended to include an ex post participation constraint.[10] A further distinction of the current paper is that, in our analysis of the regulator's problem, we allow for the possibility of excluded types and show further that actual exclusion can be optimal. In that case, the regulation contract can be understood as providing a disconnected set of quantities, namely, a quantity of zero for excluded types combined with an interval of positive quantities for nonexcluded types. The optimal regulation contract is then clearly distinct from an interval allocation.

Alonso and Matouschek (2008) were the first to argue that the monopoly regulation problem can be understood as a delegation problem. As an application of their analysis, they study optimal regulation when costs are privately observed by the regulated firm and transfers are infeasible, and they report conditions under which price-cap regulation is optimal. Our analysis differs in two ways. First, Alonso and Matouschek assume that the monopolist produces regardless of its cost type and do not include a participation constraint. Indeed, their price-cap solution would violate an ex post participation constraint. We include an ex post participation constraint, allow for exclusion, and consider as well the setting in which the ex post participation constraint holds but exclusion is infeasible. When exclusion is not optimal or is infeasible, the optimal price cap in our model is placed at a higher level than in their analysis. Second, Alonso and Matouschek assume that demand is linear and the regulator maximizes aggregate social surplus. We consider a more general family of demand functions and regulator objectives, and we provide conditions under which exclusion is optimal when demand is linear.

Recent work by Kolotilin and Zapechelnyuk (2019) is also related. They examine optimal delegation in a "linear delegation" framework and, as an application, provide conditions under which a price cap is the optimal regulatory policy in a delegation setting with a participation constraint. The two papers are complementary. We highlight three distinct features of our analysis. First, following Baron and Myerson (1982), we assume that the monopolist has a nonnegative fixed cost; by contrast, Kolotilin and Zapechelnyuk (2019) build from the assumption that the monopolist has no fixed costs. Second, the linear delegation framework corresponds in the regulation setting to the family of demand functions that we identify under which the sufficient conditions for our propositions hold if a simple inequality is satisfied; however, as noted above, we can go beyond this family and check the sufficient conditions for our propositions directly, as we do for the exponential demand function. Third, the two papers employ different proof methods: we analyze the delegation problem directly using a Lagrangian approach, whereas Kolotilin and Zapechelnyuk (2019) analyze the delegation problem by drawing a novel link to the literature on Bayesian persuasion.

Additional work in this area has explored alternative delegation environments where similar ex post participation constraints naturally arise. See, for example, Kartik, Kleiner,

[10] Amador and Bagwell (2020) also build on the Lagrangian methods used by Amador and Bagwell (2013). Amador and Bagwell (2020) provide sufficient conditions under which money burning expenditures are used in an optimal delegation contract. Building on work by Ambrus and Egorov (2017), they also consider an application with an ex ante participation constraint under the assumption that ex ante (noncontingent) transfers are feasible. The participation constraint can then be addressed using standard methods. In the present paper, by contrast, the participation constraint must hold ex post and cannot be addressed using standard methods since transfers are infeasible.

and Van Weelden (2021), Saran (2021), and Zapechelnyuk (2020), who consider applications to veto bargaining, monopolistic screening, and quality certification, respectively. Methodologically, one key difference is that the main results in these papers are obtained for a payoff specification for the principal that is independent of the private information parameter. This restriction is not appropriate for our regulation application, where the regulator's payoff function directly depends on the regulated firm's cost level, and as a consequence, we do not impose this restriction in our analysis.

The paper is organized as follows. Section 2 sets up the regulator's problem, and Section 3 examines cap allocations. Section 4 then focuses on the regulator's truncated problem and develops general sufficient conditions for the optimality of a cap allocation. Section 5 considers the global optimality of the cap allocation and develops further results and approaches that facilitate the application of our findings. Section 6 identifies conditions under which actual exclusion does or does not occur. Section 7 concludes. The Appendix contains the remaining proofs.

## 2. The regulator's problem

In this section, we present our basic model and formally define the problem that confronts the regulator. We also identify the bias in the monopolist's unrestricted output choice.

We consider a monopolist facing an inverse demand function given by $P(q)$ where $q$ is the quantity produced. The production quantity $q$ resides in the set $Q \equiv [0, q_{\max}]$, which is an interval of the real line with nonempty interior. The function $P(q)$ is well-defined and finite for all $q \in (0, q_{\max}]$.

We assume the monopolist faces a constant marginal cost of production $\gamma$ as well as a fixed cost $\sigma \geq 0$. The marginal cost $\gamma$ is private information to the monopolist and is distributed over the support $\Gamma = [\underline{\gamma}, \overline{\gamma}]$ where $\overline{\gamma} > \underline{\gamma} > 0$ with a differentiable cumulative distribution function $F(\gamma)$. The associated density, $f(\gamma) \equiv F'(\gamma)$, is strictly positive and differentiable.

We assume that the regulator has no access to transfers or taxes, and can only impose restrictions on the quantity produced by the monopolist. As discussed in the Introduction, our no-transfers assumption means that the regulator cannot impose taxes or subsidies, and it implicitly implies as well that the monopolist cannot use an access fee. We thus assume that the monopolist selects a uniform price, with the regulator determining the feasible menu of such prices through the selection of a feasible menu of quantities. We allow that the regulator's objective is to maximize a weighted social welfare function in which profits receive weight $\alpha \in (0, 1]$. The regulator maximizes aggregate social surplus when $\alpha = 1$ and gives greater weight to consumer interests when $\alpha < 1$.

We impose the following assumptions on primitives.

Assumption 1. *We impose the following assumptions:*

*(a)* $P(q)$ *is twice-continuously differentiable for* $q \in (0, q_{\max}]$ *with* $P'(q) < 0 < P(q)$.

*(b)* $\lim_{q\downarrow 0} P(q) > \overline{\gamma}$ *and* $P(q_{\max}) < \underline{\gamma}$.

*(c) There exist functions* $b(q)$, $v(q)$, *and* $w(\gamma, q)$, *which are twice-continuously differentiable for* $q \in (0, q_{\max}]$ *and that satisfy*

$$b(q) \equiv P(q)q,$$
$$v(q) \equiv \int_0^q P(z)\,dz - P(q)q,$$
$$w(\gamma, q) \equiv -\gamma q + b(q) + \frac{1}{\alpha}v(q),$$

*with* $\lim_{q\downarrow 0} b(q) = 0$ *and* $\lim_{q\downarrow 0} v(q) = 0$. *We define* $b(0) = v(0) = w(\gamma, 0) = 0$.

*(d)* $b''(q) < 0$ *and* $w_{qq}(\gamma, q) = b''(q) + \frac{1}{\alpha}v''(q) \leq 0$ *for all* $q \in (0, q_{\max}]$.

*(e)* $w_q(\underline{\gamma}, q_{\max}) < 0$.

In this assumption, $b(q)$ defines the total revenue for the monopolist, $v(q)$ represents consumer surplus, and $w(\gamma, q)$ represents the welfare to the regulator (gross of the fixed cost).[11] Using Assumption 1, we obtain that $v'(q) = -qP'(q) > 0$ for all $q > 0$. Similarly, using Assumption 1, we have that

$$\begin{aligned} w_{qq}(\gamma, q) &= b''(q) + \frac{1}{\alpha}v''(q) \\ &= P''(q)q + 2P'(q) - \frac{1}{\alpha}\big[P''(q)q + P'(q)\big] \leq 0 \quad \text{for } q > 0. \end{aligned}$$

Notice that $P'(q) < 0$ implies that $w$ is strictly concave when $\alpha = 1$. We make no assumption as regards the sign of $v''(q)$. If marginal revenue is steeper than demand (i.e., $b''(q) < P'(q)$), then $v''(q) > 0$.[12] For example, as we discuss below, $v''(q) > 0$ when demand is linear, and $v''(q) < 0$ when demand exhibits constant elasticity.

Assumption 1 also includes various regularity conditions. According to part (b), the inverse demand function exceeds the highest marginal cost for quantities that are sufficiently close to zero and falls below the lowest marginal cost for quantities that are sufficiently close to $q_{\max}$. Part (e) ensures that the welfare-maximizing quantity is below $q_{\max}$, even when marginal cost is at its lowest possible value.

We envision the regulator as choosing a menu of permissible outputs, with the understanding that a monopolist with cost type $\gamma$ selects its preferred output from this menu. Thus, if the regulator seeks to assign an output $q(\gamma)$ to a monopolist with type $\gamma$, then an incentive compatibility constraint must be satisfied. As well, if the regulator seeks a positive output from a monopolist with type $\gamma$, then type $\gamma$ must earn more by producing $q(\gamma) > 0$ than by shutting down and avoiding the fixed cost of production, $\sigma \geq 0$.[13] We allow that the regulator may choose a menu of permissible outputs such

[11] Our notation here is designed to facilitate easy comparison with Amador and Bagwell (2013).

[12] This condition holds if the demand function is log concave but fails otherwise.

[13] We have assumed that the fixed cost $\sigma$ is independent of $\gamma$. We have done so mostly for simplicity, as it is possible to generalize our main results (i.e., Propositions 1 and 2) to the case where $\sigma(\gamma)$ is a nondecreasing and nonnegative function of $\gamma$.

that some types produce zero output, incur no fixed cost, and thus earn a profit of zero. That is, the regulator may "exclude" some types from production.

The *regulator's problem* can then be written as follows:

$$\text{(P1)} \qquad \max_{q:\Gamma\to Q} \int_\Gamma \big(w\big(\gamma, q(\gamma)\big) - \mathbf{1}\big(q(\gamma)\big)\sigma\big)\, dF(\gamma) \quad \text{subject to:}$$

$$\gamma \in \arg\max_{\tilde{\gamma}\in\Gamma} -\gamma q(\tilde{\gamma}) + b\big(q(\tilde{\gamma})\big) - \mathbf{1}\big(q(\tilde{\gamma})\big)\sigma \quad \text{for all } \gamma \in \Gamma$$

$$0 \le -\gamma q(\gamma) + b\big(q(\gamma)\big) - \mathbf{1}\big(q(\gamma)\big)\sigma, \quad \text{for all } \gamma \in \Gamma$$

where $\mathbf{1}(\cdot)$ is an indicator function such that $\mathbf{1}(q) = 1$ if $q > 0$ and $\mathbf{1}(q) = 0$ if $q = 0$.

The first constraint in this problem is the incentive compatibility constraint, while the second constraint is the ex post participation or individual rationality (IR) constraint. The IR constraint requires that if a type produces, it needs to earn enough profit to cover its fixed cost, $\sigma$. The constraints also allow for the possibility of types for which $q(\gamma) = 0$, since the IR constraint holds when $q(\gamma) = 0$. We say that an allocation is *feasible* if it satisfies both of these constraints.

*The flexible allocation* Before characterizing the solution to the regulator's problem, it is convenient to define $q_f(\gamma)$ as the allocation that a monopolist would choose if it were forced to produce but were otherwise unrestricted. To this end, we let

$$\pi(\gamma, q) \equiv -\gamma q + b(q)$$

be the monopolist's profit function (gross of the fixed cost), and we then define the monopolist's flexible allocation as

$$q_f(\gamma) \equiv \arg\max_{q\in Q} \pi(\gamma, q).$$

The flexible allocation is simply the monopoly output as a function of the monopolist's cost type. The associated first-order condition is given by $b'(q) - \gamma = 0$.

We note that the $\lim_{q\to 0} P(q) > \overline{\gamma}$ and $b(0) = 0$ imply that $q_f(\overline{\gamma}) > 0$. Since $P(q_{\max}) < \underline{\gamma}$, we know that $q_f(\underline{\gamma}) < q_{\max}$. With these boundary results in place, we have that $q_f(\gamma)$ is differentiable, with $q_f'(\gamma) = 1/b''(q_f(\gamma)) < 0$ and $q_f(\gamma) \in (0, q_{\max})$ for all $\gamma \in \Gamma$. Note as well that $P(q_f(\gamma)) > \gamma$, and thus $\pi(\gamma, q_f(\gamma)) = -\gamma q_f(\gamma) + b(q_f(\gamma)) > 0$ for all $\gamma \in \Gamma$.

We further assume that it is optimal for all types to produce if given the ability to set their monopolist quantity.

Assumption 2. *For all types $\gamma \in \Gamma$, $\pi(\gamma, q_f(\gamma)) > \sigma$.*

An implication of Assumption 2 is that, for any given cost type, the regulator's welfare is higher when a monopolist with that cost type sets its monopoly output than when it shuts down and produces zero output. Thus, if the solution to the regulator's problem excludes a given cost type from production, then it must be that the regulator is able to improve the allocation for other cost types through this means.

Given the interiority of $q_f(\gamma)$, we may use the associated first-order condition and establish the following relationship: for all $\gamma \in \Gamma$,

$$w_q(\gamma, q_f(\gamma)) = \frac{1}{\alpha} v'(q_f(\gamma)) = -\frac{1}{\alpha} P'(q_f(\gamma)) q_f(\gamma) = \frac{1}{\alpha}[P(q_f(\gamma)) - \gamma] > 0$$

Thus, the model embodies downward or *negative bias*: the agent's (i.e., the monopolist's) preferred $q$ is too low from the principal's (i.e., the regulator's) perspective.

The presence of negative bias suggests the possibility of a solution that imposes a lower bound on $q$ for higher types (or equivalently a cap on the price for higher types). But note also that the unrestricted monopolist profits are decreasing in $\gamma$; thus, it is also possible that such a regulatory restriction could exclude higher-cost types from producing, if they are then unable to cover their fixed cost of production.

We show now that, if any exclusion occurs, then the excluded types are always defined by a threshold type, $\gamma_t \in \Gamma$.

Lemma 1. *In any feasible allocation* $q(\cdot)$, *there exists a cut-off* $\gamma_t \in [\underline{\gamma}, \overline{\gamma}]$ *such that* $q(\gamma) = 0$ *for* $\gamma > \gamma_t$ *and* $q(\gamma) > 0$ *for* $\gamma < \gamma_t$. *In addition, if* $\gamma_t \in (\underline{\gamma}, \overline{\gamma})$, *then* $-\gamma_t q(\gamma_t) + b(q(\gamma_t)) = \sigma$.

All proofs not in the text are in the Appendix.

The proof uses the property that an incentive compatible allocation for such a model must be monotonic, which in turn ensures the existence of the cut-off value $\gamma_t \in [\underline{\gamma}, \overline{\gamma}]$.

If we were to ignore the IR constraint, the regulator's problem would fit into the framework of Amador and Bagwell (2013), and we could use the sufficiency theorems in that paper to derive conditions under which a simple cap allocation is optimal. However, as we show below, the IR constraint will always be violated if ignored.

## 3. Optimality within the set of cap allocations

In this section, we study cap allocations when the IR constraint is ignored and also when exclusion is possible. Our analysis clarifies the role of the IR constraint and identifies a candidate allocation for the solution of the regulator's problem.

### 3.1 *The case without an IR constraint*

It is instructive to solve the regulator's problem under the restriction that the regulator can choose only among cap allocations, while ignoring the IR constraint. Let us define a cap allocation as follows.

Definition 1. A *cap allocation* indexed by $x$ is an allocation $q_c(\gamma; x)$ such that

$$q_c(\gamma; x) = \begin{cases} q_f(\gamma); & \text{if } q_f(\gamma) \geq x \\ x; & \text{otherwise} \end{cases}$$

for all $\gamma \in \Gamma$.

It is straightforward to confirm that a cap allocation is always incentive compatible. For a given cap allocation, there also exists a critical type $\gamma_c$, defined as follows.[14]

DEFINITION 2. Given $x \in Q$, let $\gamma_c(x)$ be the unique value in $\Gamma$ such that $q_f(\gamma) > x$ for all $\gamma \in [\underline{\gamma}, \gamma_c(x))$ and $q_f(\gamma) < x$ for all $\gamma \in (\gamma_c(x), \overline{\gamma}]$.

We allow in the definition of $\gamma_c(x)$ that $\gamma_c(x) = \underline{\gamma}$, in which case $x \geq q_f(\underline{\gamma})$, so that the flexible output for all types above $\underline{\gamma}$ is below $x$. Notice that the allocation $q_c(\gamma; x)$ actually defines a quantity floor rather than a cap. We still refer to this allocation as a cap allocation, since it corresponds to a cap on permissible prices and links thereby with the literature on price-cap regulation. Note also that the cap allocation only has bite in restricting the monopolist's choice if $x > q_f(\overline{\gamma})$.

We define an *optimal simple cap allocation* to be an optimal cap allocation when the IR constraint is ignored and all types produce. That is, the optimal simple cap allocation solves

$$\max_{x \geq q_f(\overline{\gamma})} W^c(x)$$

where $W^c(x)$ represents the regulator's welfare:

$$W^c(x) \equiv \int_{\underline{\gamma}}^{\gamma_c(x)} w\big(\gamma, q_f(\gamma)\big)\, dF(\gamma) + \int_{\gamma_c(x)}^{\overline{\gamma}} w(\gamma, x)\, dF(\gamma) - \sigma$$

We now present a necessary condition for an optimal simple cap allocation.[15]

LEMMA 2. *The cap allocation indexed by $x$ is an optimal simple cap allocation only if $x > q_f(\overline{\gamma})$ and*

$$\int_{\gamma_c(x)}^{\overline{\gamma}} w_q(\gamma, x)\, dF(\gamma) = 0$$

In the absence of a participation constraint, we could use results from Amador and Bagwell (2013) and establish a general set of environments under which the optimal simple cap allocation is optimal over the full class of incentive compatible allocations. As we now argue, however, the presence of an IR constraint implies that the optimal simple cap allocation is not feasible.

The basic point can be understood using Figure 1. The graph on the right in Figure 1 illustrates the optimal simple cap allocation in bold (for the case where $\gamma_c$ is in the interior of $\Gamma$). This allocation is illustrated relative to the flexible allocation, $q_f(\gamma)$, and the regulator's ideal (i.e., efficient) allocation, $q_e(\gamma)$, which we define as the allocation that maximizes $w(\gamma, q)$.[16] Notice that $q_e(\gamma)$ is downward sloping and that $q_e(\gamma) > q_f(\gamma)$,

[14] Here and in the rest of the paper, we use the convention that the intervals $[x, x)$ and $(x, x)$ correspond to the empty set.

[15] The existence of an optimal simple cap allocation follows from standard arguments, given Assumption 1.

[16] We assume for this graphical analysis that $q_e(\gamma)$ is uniquely determined.

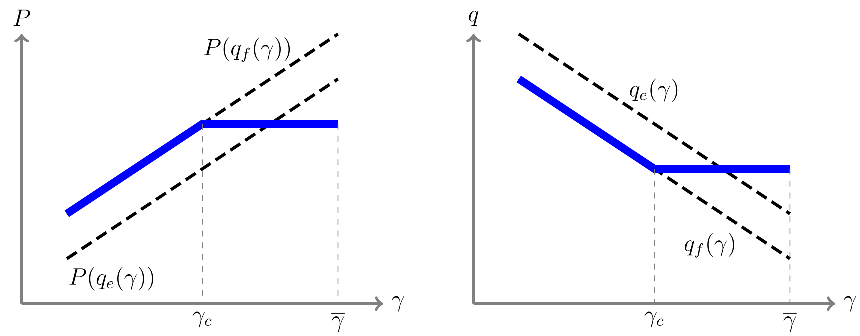


Figure 1. Optimal simple cap allocation fails IR.

where the inequality reflects the aforementioned downward bias. For given $\gamma$, $q_e(\gamma)$ induces a price equal to marginal cost (i.e., $P(q_e(\gamma)) = \gamma$) when $\alpha = 1$. When $\alpha < 1$, the regulator's ideal allocation entails even higher quantities, and thus drives price below marginal cost. The optimal simple cap allocation is such that the cap is ideal for the regulator *on average* for affected types (i.e., for $\gamma \geq \gamma_c$). The graph on the left in Figure 1 illustrates the same information in terms of the induced prices, which are also depicted in bold. As this graph illustrates, the optimal simple cap allocation places the price cap at a level that is ideal for the principal on average for affected types. This graph also suggests that the participation constraint is violated for the highest types when the optimal simple cap allocation is used. For type $\overline{\gamma}$, the optimal price cap lies strictly below the regulator's ideal price, $P(q_e(\overline{\gamma}))$, which equals $\overline{\gamma}$ when $\alpha = 1$ and is less than $\overline{\gamma}$ when $\alpha < 1$. The optimal price cap is thus strictly below $\overline{\gamma}$; hence, since the fixed cost $\sigma$ is nonnegative, the IR constraint must fail for the highest-cost type when the optimal simple cap allocation is used.

The following lemma offers a formal confirmation of this point.

Lemma 3. *The optimal simple cap allocation violates the IR constraint for the highest types.*

It is also straightforward to confirm that the IR constraint holds for a cap allocation if and only if it holds for the highest-cost type.

There are two ways a regulator could in principle deal with the problem that the optimal simple cap allocation violates the IR constraint. First, it could decide not to be so tough, and choose a cap that gives sufficient flexibility so that all types choose to produce. Alternatively, it could choose a cap that is sufficiently tight that some types choose not to produce. This leads us to consider the "best" cap allocation that satisfies the IR constraint while allowing types to be excluded from production. We thus proceed to characterize the class of allocations with caps and exclusion.

### 3.2 *IR constraint and exclusion*

Consider a situation where the regulator chooses a cap on the price, and as a result, some high-cost types may choose not to produce. This is a *cap allocation with potential* *exclusion*, and it is defined by a quantity $x$ such that any type is free to choose between producing a quantity higher or equal to $x$, or not producing at all.

Definition 3. A *cap allocation with potential exclusion* indexed by $x$ is an allocation $q(\gamma; x)$ such that

$$q(\gamma; x) = \begin{cases} q_f(\gamma); & \text{if } q_f(\gamma) \geq x, \\ x; & \text{if } q_f(\gamma) < x \text{ and } -\gamma x + b(x) - \sigma \geq 0, \\ 0; & \text{otherwise,} \end{cases}$$

for all $\gamma \in \Gamma$.

A cap allocation with potential exclusion is clearly incentive compatible. Without loss of generality, we can restrict attention to cap allocations such that $x \geq \underline{q} \equiv q_f(\overline{\gamma})$, as no type will ever choose to produce below $q_f(\overline{\gamma})$ if given the choice to produce more. Similarly, we can restrict attention to cap allocations such that $x \leq \overline{q}$ where $\overline{q} > q_f(\underline{\gamma})$ is the value that satisfies $-\underline{\gamma}\overline{q} + b(\overline{q}) = \sigma$. Imposing a bound $x$ above $\overline{q}$ is equivalent to assigning no production for all types (as not even the lowest cost type is willing to produce that much), and hence considering restrictions above that is unnecessary. Note that our assumptions guarantee $\overline{q} \in Q$.

Figure 2 presents a graphical representation of a cap allocation with exclusion where a nonzero measure of types are excluded, some types are constrained at the cap, and some other types are choosing their monopoly allocation. To describe such an allocation, recall that, from Lemma 1, we know that any allocation with exclusion satisfies a threshold property: types above some type $\gamma_t$ are excluded from production, while types below $\gamma_t$ produce. Thus, given a bound $x$, let $\gamma_t(x) \in [\underline{\gamma}, \overline{\gamma}]$ be the associated exclusion threshold. That is, $\gamma_t(x)$ is such that $\max_{q \geq x}\{-\gamma q + b(q) - \sigma\} < 0$ for all $\gamma \in (\gamma_t(x), \overline{\gamma}]$ and $\max_{q \geq x}\{-\gamma q + b(q) - \sigma\} > 0$ for all $\gamma \in [\underline{\gamma}, \gamma_t(x))$.

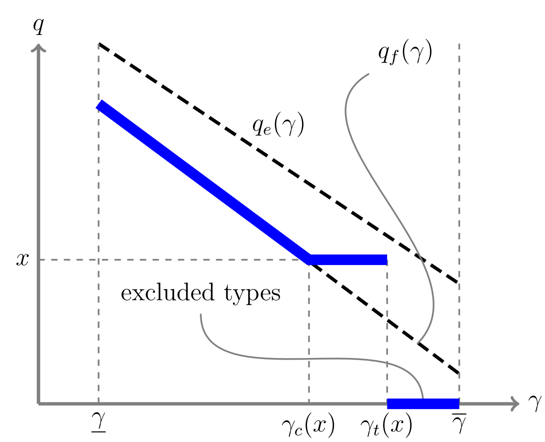


Figure 2. A cap allocation with exclusion. The solid thick line represents a cap allocation with exclusion.

However, not all the types that produce are able to do so at their monopoly level. Types with a cost smaller than $\gamma_c(x)$ would choose their monopoly level if forced to produce, while types above $\gamma_c(x)$ would choose the cap if forced to produce. Note that $\gamma_c(x) \leq \gamma_t(x)$ with strict inequality if $\underline{q} < x < \overline{q}$.

The welfare generated by a cap allocation with potential exclusion is thus

$$W(x) \equiv \int_{\underline{\gamma}}^{\gamma_c(x)} \big[w\big(\gamma, q_f(\gamma)\big) - \sigma\big] \, dF(\gamma) + \int_{\gamma_c(x)}^{\gamma_t(x)} \big[w(\gamma, x) - \sigma\big] \, dF(\gamma) \tag{1}$$

where the first term represents the regulator's payoff from giving flexibility to types below $\gamma_c(x)$, the second term represents the payoffs generated from types that produce at the cap, $x$, and where the payoff of the excluded types is zero.

Let $x^\star$ be such that $x^\star \in \operatorname{argmax}_{x \in [\underline{q}, \overline{q}]} W(x)$; that is, $x^\star$ represents the optimal cap that could be imposed.[17] Given this cap $x^\star$, the associated cap allocation $q^\star$ can be written as

$$q^\star(\gamma) = \begin{cases} q_f(\gamma); & \gamma \in \big[\underline{\gamma}, \gamma_c\big(x^\star\big)\big) \\ x^\star; & \gamma \in \big[\gamma_c\big(x^\star\big), \gamma_t\big(x^\star\big)\big] \\ 0; & \gamma \in \big(\gamma_t\big(x^\star\big), \overline{\gamma}\big] \end{cases} \tag{2}$$

This cap allocation with potential exclusion $q^\star$ is our candidate allocation for the solution to the regulator's problem. Our goal is thus to determine sufficient conditions under which $q^\star$ is also optimal within the set of all feasible allocations.

Having identified our candidate solution $q^\star$, we hasten to add that it is not obvious that the solution to the regulator's problem indeed takes this form. For example, and as we discuss in Section 4.2, we can also imagine that the optimal allocation might have jumps, and thus not take the form of a cap allocation. Further, a property of the allocation $q^\star$ is that, if exclusion is not used, then the highest type earns zero profit and satisfies the IR constraint with equality. This property, too, is not obvious in our no-transfer setting, since the allocation for this type affects as well incentive compatible allocations for lower types.

## 4. Toward sufficient conditions

We return now to consider the solution to the regulator's problem, Problem P1. As a general matter, we do not know whether a cap allocation with or without exclusion is optimal. Indeed, solving the regulator's problem directly seems difficult, since the possibility of excluding some types must be considered. We pursue an alternative approach, one that divides the problems into several subproblems.

The main idea is as follows:

1. Rather than working with the lower bound on production, we work with the excluded types directly. Based on Lemma 1, we fix a threshold for excluded types, $\gamma_t$,

[17] The existence of $x^\star$ follows from standard arguments, given Assumption 1.

and consider the problem of allocating production for types below $\gamma_t$ while ignoring the allocation for types higher than $\gamma_t$. That is, we restrict attention to the set of types $[\underline{\gamma}, \gamma_t]$ and study the problem for a regulator that only considers types in that truncated set and is not allowed to exclude any types in that set from production.[18] We obtain conditions under which an optimal allocation in this truncated problem is a cap allocation such that the threshold or upper-bound type, $\gamma_t$, is indifferent between producing or not.

2. Next, we argue that such a truncated allocation is incentive compatible when extended to the entire set $[\underline{\gamma}, \overline{\gamma}]$ by giving types above $\gamma_t$ zero output. The optimal allocation that results from considering only the truncated set is thus also optimal when considering the entire set of types for a given level of exclusion.

3. We then look for the best allocation by varying the level of exclusion, which in our case is indexed by $\gamma_t$. This is a single variable optimization problem.

Toward this goal, let us first consider the regulator's truncated problem.

### 4.1 *The regulator's truncated problem*

For this problem, we fix $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$ and define $\Gamma_t(\gamma_t) \equiv [\underline{\gamma}, \gamma_t]$.[19] The regulator's truncated problem is to find an allocation, $q_t : \Gamma_t(\gamma_t) \to Q$, that maximizes its payoff subject to the feasibility constraints and that no type in set $\Gamma_t(\gamma_t)$ is excluded.[20] Formally, the *regulator's truncated problem* may be written as

$$(\mathrm{P}_t) \qquad \max_{q_t:\Gamma_t(\gamma_t)\to Q} \int_{\Gamma_t(\gamma_t)} \big(w\big(\gamma, q_t(\gamma)\big) - \sigma\big)\, dF(\gamma) \quad \text{subject to:}$$

$$\gamma \in \arg\max_{\tilde{\gamma}\in\Gamma_t(\gamma_t)} \{-\gamma q_t(\tilde{\gamma}) + b\big(q_t(\tilde{\gamma})\big) - \sigma\} \quad \text{for all } \gamma \in \Gamma_t(\gamma_t)$$

$$0 \le -\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma, \quad \text{for all } \gamma \in \Gamma_t(\gamma_t)$$

Differently from Problem P1, in the regulator's truncated problem no type in $\Gamma_t(\gamma_t)$ is excluded, explaining why the indicator functions do not appear in Problem $\mathrm{P}_t$.[21] Similar to Section 3.1, if we were to look for a simple cap allocation in this truncated problem, the optimal one would violate the IR constraint for the highest cost type, which in this case is the threshold or upper-bound type, $\gamma_t$.

We conjecture that a cap allocation where type $\gamma_t$ is indifferent between producing or not is optimal. Let $q_i(\gamma_t)$ be the unique value such that

$$-\gamma_t q_i(\gamma_t) + b\big(q_i(\gamma_t)\big) = \sigma, \quad \text{and} \quad q_i(\gamma_t) > q_f(\gamma_t). \tag{3}$$

[18] Thus, the regulator assigns production levels and incurs the fixed cost $\sigma$ for all types in the truncated set. For an excluded type, by contrast, output is zero and the fixed cost $\sigma$ is not incurred.

[19] We ignore the case where $\gamma_t = \underline{\gamma}$, as this implies that almost all types are excluded, a situation that cannot be optimal under our assumptions.

[20] Within the set of cap allocations, it is sufficient to look for a quantity floor in $[\underline{q}, \overline{q}]$. When checking for optimality more generally, we do not impose that restriction, and hence $q_t : \Gamma(\gamma_t) \to Q$.

[21] A type in $\Gamma_t(\gamma_t)$ could still be assigned zero output, but the fixed cost is incurred. Note that if $\sigma > 0$, such assignment violates the type's IR constraint.

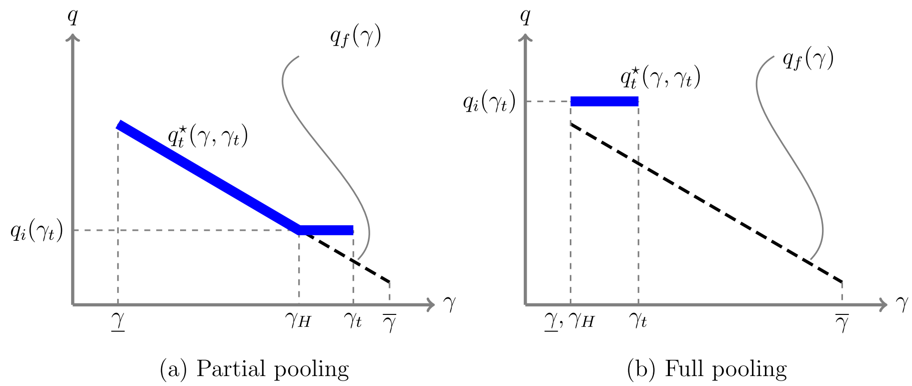


FIGURE 3. The truncated cap allocation, $q_t^\star(\gamma|\gamma_t)$.

Thus, $q_i(\gamma_t)$ is the output level that exceeds $\gamma_t$'s monopoly level and ensures that this type is indifferent between producing at that level or not. In other words, it corresponds to a price that equals the average cost for type $\gamma_t$. Note that under our assumptions, such $q_i(\gamma_t) \in Q$ exists.

We define $\gamma_H(\gamma_t) \in [\underline{\gamma}, \gamma_t]$ to be the value such that $q_i(\gamma_t) \leq q_f(\gamma)$ for $\gamma < \gamma_H(\gamma_t)$, and $q_i(\gamma_t) \geq q_f(\gamma)$ for $\gamma > \gamma_H(\gamma_t)$. Note that $\gamma_H(\gamma_t) = \gamma_c(q_i(\gamma_t))$ and that $\gamma_H(\gamma_t) < \gamma_t$ given $\gamma_t > \underline{\gamma}$.

With these objects, we can define the *truncated cap allocation*, $q_t^\star(\gamma|\gamma_t)$:

$$q_t^\star(\gamma|\gamma_t) = \begin{cases} q_f(\gamma); & \gamma \in [\underline{\gamma}, \gamma_H(\gamma_t)) \\ q_i(\gamma_t); & \gamma \in [\gamma_H(\gamma_t), \gamma_t] \end{cases} \tag{4}$$

The allocation $q_t^\star(\gamma|\gamma_t)$ is continuous in $\gamma$ and features full pooling if $\gamma_H(\gamma_t) = \underline{\gamma}$. If $\gamma_H(\gamma_t)$ is interior to the interval $\Gamma_t(\gamma_t)$, then $q_i(\gamma_t)$ coincides with the flexible quantity chosen by type $\gamma_H(\gamma_t)$. Figure 3 displays the two possible cases for $q_t^\star$ for two different values of $\gamma_t$. Panel (a) shows the case with partial pooling. Panel (b) shows the case where $\gamma_t$ is sufficiently small that full pooling of all types at the cap results.

We seek conditions under which $q_t^\star(\gamma|\gamma_t)$ is the optimal solution to the regulator's truncated problem. To present our next result, we require a couple of definitions. Let

$$\begin{aligned} G(\gamma|\gamma_t) \equiv {} & -\kappa F(\gamma_t) + \kappa\left[\frac{\gamma - b'\big(q_i(\gamma_t)\big)}{\gamma - \gamma_H(\gamma_t)}\right]F(\gamma) \\ & + \frac{1}{\gamma - \gamma_H(\gamma_t)} \int_{\gamma_H(\gamma_t)}^{\gamma} w_q\big(\tilde{\gamma}, q_i(\gamma_t)\big) f(\tilde{\gamma})\, d\tilde{\gamma}, \end{aligned} \tag{5}$$

for $\gamma > \gamma_H(\gamma_t)$ and where, following Amador and Bagwell (2013), $\kappa$ is a relative concavity parameter defined as

$$\kappa \equiv \min_{q \in Q} \left\{1 + \frac{v''(q)}{\alpha b''(q)}\right\}.$$

We let $G(\gamma_H(\gamma_t)|\gamma_t) \equiv \lim_{\gamma \downarrow \gamma_H(\gamma_t)} G(\gamma|\gamma_t)$, which exists and is a finite number.

We may now state our general sufficiency result as follows.

Proposition 1 (Sufficient conditions). *If:*

*(i)* $G(\gamma|\gamma_t) \le G(\gamma_t|\gamma_t)$ *for all* $\gamma \in [\gamma_H(\gamma_t), \gamma_t]$, *where G is given by (5); and*

*(ii)* $M_1(\gamma) \equiv \kappa F(\gamma) + w_q(\gamma, q_f(\gamma)) f(\gamma)$ *is nondecreasing in* $\gamma$ *for* $\gamma \in [\underline{\gamma}, \gamma_H(\gamma_t))$,

*then the cap allocation* $q_t^\star(\gamma|\gamma_t)$ *solves the regulator's truncated problem, Problem* $P_t$.

Our proof approach follows a guess-and-verify structure. To begin, we follow standard methods and rewrite the incentive constraint in the regulator's truncated problem as an integral equation and a monotonicity requirement (namely, that $q_t(\gamma)$ must be nonincreasing).[22] Next, we embed the monotonicity requirement into the choice set, and we express the integral equation equivalently in terms of two inequality conditions. The regulator's truncated problem is thereby represented as a maximization problem over functions belonging to a choice set of nondecreasing functions that satisfy three inequality constraints, where one of the constraints is the IR constraint. With the problem set up in this fashion, we conjecture that the cap allocation $q_t^\star(\gamma|\gamma_t)$ is the solution. To confirm this conjecture, we construct multiplier functions for each of the three inequality constraints. Under the conditions stated in Proposition 1 and for the constructed multiplier functions, we find that the multiplier functions are nondecreasing, the corresponding Lagrangian is concave, and the cap allocation satisfies first-order conditions and a complementary slackness condition. Building on work by Amador and Bagwell (2013), we show that these findings are sufficient to conclude that $q_t^\star(\gamma|\gamma_t)$ solves the regulator's truncated problem.[23]

### 4.2 *Intuition*

We now develop some intuition for the interpretation of Proposition 1. We begin with part (ii). Observe that part (ii) is more easily satisfied when $\kappa$ is big. Referring to the definition of $\kappa$, we thus conclude that part (ii) is more easily satisfied when the minimum value for $1 + \frac{v''(q)}{\alpha b''(q)}$ is big. Since $w_q(\gamma, q_f(\gamma)) > 0$, we see that part (ii) is also more easily satisfied when the density is nondecreasing for $\gamma \in [\underline{\gamma}, \gamma_H(\gamma_t))$.

To see why the relative sizes of $\frac{1}{\alpha} v''(q)$ and $b''(q)$ and the density slope matter, we consider alternatives to the truncated cap allocation. If the truncated cap allocation

[22] We emphasize that feasible allocations may be discontinuous. As illustrated in the intuition developed just below, our proof approach thus must establish that the cap allocation $q_t^\star(\gamma|\gamma_t)$ is optimal among a set of monotone and possibly discontinuous functions.

[23] It is instructive here to compare our regulator's truncated problem, in which transfers are unavailable, with the standard (Baron–Myerson) framework in which transfers are available. In the solution approach for the standard framework, the integral equation is substituted into the objective, the IR constraint is shown to bind for the highest type, the IR constraint for the highest type is substituted into the objective, and the resulting objective is then maximized pointwise. If the solution satisfies the monotonicity constraint, then the problem is solved. By contrast, in our no-transfers setting, we cannot substitute the integral equation into the objective, since we do not have a remaining transfer instrument with which to ensure that the solution of the resulting optimization problem satisfies the integral equation. For the same reason, we cannot substitute the IR constraint for the highest type into the objective. Indeed, as a general matter, when transfers are unavailable it is no longer obvious that the IR constraint for the highest type must bind.

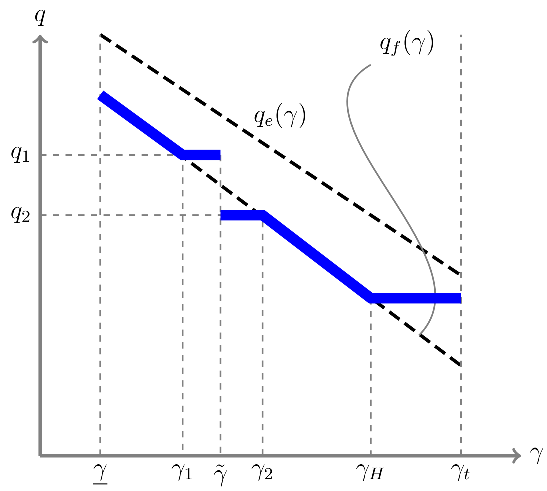


FIGURE 4. Drilling a hole (with $\sigma > 0$).

is optimal among all feasible allocations for the regulator's truncated problem, then it must be preferred by the regulator to alternative feasible allocations that are generated by "drilling holes" in the flexible part of the allocation. Figure 4 illustrates one such alternative allocation, in which output levels between $q_1 \equiv q_f(\gamma_1)$ and $q_2 \equiv q_f(\gamma_2)$ are prohibited and where $\underline{\gamma} < \gamma_1 < \gamma_2 < \gamma_H$. There then exists a unique type $\tilde{\gamma} \in (\gamma_1, \gamma_2)$ that is indifferent between $q_1$ and $q_2$. The alternative allocation thus induces a "step" at $\tilde{\gamma}$, with the allocation $q_1$ selected by $\gamma \in [\gamma_1, \tilde{\gamma})$ and the allocation $q_2$ selected by $\gamma \in [\tilde{\gamma}, \gamma_2]$, where for simplicity we place type $\tilde{\gamma}$ with the higher types.

In comparison to the truncated cap allocation, the alternative allocation has advantages and disadvantages. First, the alternative allocation generates output choices for $\gamma \in [\gamma_1, \tilde{\gamma})$ that are closer to the the regulator's ideal choices for such types; however, the alternative allocation also results in output choices for $\gamma \in [\tilde{\gamma}, \gamma_2]$ that are further from the regulator's ideal choices for such types. In line with our discussion above, these observations suggest that a nondecreasing density should work in favor of the truncated cap allocation. Second, the alternative allocation increases the variance of the induced allocation around $q_f(\gamma)$ over $[\gamma_1, \gamma_2]$. Consistent with our preceding discussion, this effect brings into consideration the relative magnitudes of $\frac{1}{\alpha} v''(q)$ and $b''(q)$, where the latter determines the slope of $q_f(\gamma)$. If $v(q)$ is concave, then the variance effect should work in favor of the truncated cap allocation, since the regulator would then not welcome an increase in variance. If instead $v(q)$ is convex, then the regulator would benefit from the greater variance afforded by the alternative allocation, with the benefit being larger when $\alpha$ is smaller. Based on this perspective, we may understand that the truncated allocation could remain optimal when $v(q)$ is convex, if the density rises fast enough, $\alpha$ is sufficiently large and/or $b''(q)$ is large in absolute value (so that $q_f(\gamma)$ is flat, in which case steps add little variation).

The intuitive discussion presented here considers only a subset of feasible alternative allocations that introduce variations in the flexible region. In our no-transfer setting, the incentive compatibility constraint implies that an allocation must be given by the flexible allocation over any interval for which the allocation is continuous and strictly decreasing; however, an incentive compatible allocation may include many points of discontinuity (steps), where any such point hurdles the flexible allocation as illustrated in Figure 4.[24] Our discussion above considers only an alternative allocation with a single step, but this discussion provides an intuitive foundation for understanding more generally the key forces at play.

We turn now to consider the intuition associated with part (i) of Proposition 1. For type $\gamma_t$, the IR constraint holds with equality at the output choices $q_i(\gamma_t)$ and $q'$, where $q' < q_f(\gamma_t)$ is defined so that type $\gamma_t$ is indifferent between $q_i(\gamma_t)$ and $q'$; thus, the IR constraint for type $\gamma_t$ is satisfied provided that the allocation for this type resides in the interval $[q', q_i(\gamma_t)]$. As noted above, it is also not obvious that the IR constraint must bind for type $\gamma_t$. Part (i) of Proposition 1 provides conditions under which the solution to the regulator's truncated problem is such that type $\gamma_t$ selects $q_i(\gamma_t)$ and has a binding IR constraint.

More formally, we show in the proof that the value of the multiplier function for the IR constraint of type $\gamma_t$ equals $G(\gamma_t|\gamma_t)$. The proposed allocation implies that $G(\gamma_t|\gamma_t) \geq 0$ (as shown in the proof of Proposition 1), confirming that the multiplier on the IR constraint is nonnegative; that is, the shadow price of relaxing the IR constraint for type $\gamma_t$ is nonnegative. Part (i) of Proposition 1 goes further and requires that $G(\gamma|\gamma_t) \leq G(\gamma_t|\gamma_t)$ for all $\gamma \in [\gamma_H(\gamma_t), \gamma_t]$. This condition ensures that the regulator cannot improve on the cap allocation $q_t^\star(\gamma|\gamma_t)$ by altering the allocation for types $\gamma \in [\gamma_H(\gamma_t), \gamma_t]$ while respecting the monotonicity requirement. In particular, it rules out alternative allocations that introduce steps within the $[\gamma_H, \gamma_t]$ region.[25]

For additional insight, we may compare condition (i) with the case where the IR constraint is not present (or not binding). Consider, for example, the analysis of Amador and Bagwell (2013), a case without IR constraints and with the additional requirement that $\gamma_H(\gamma_t)$ be interior. In that case, $G(\gamma_t|\gamma_t) = 0$, and their condition (c2) for the optimality of a cap is equivalent to our condition that $G(\gamma|\gamma_t) \leq G(\gamma_t|\gamma_t)$.[26]

[24] For further discussion, see Melumad and Shibano (1991).

[25] For the case of the special family of preferences introduced in (6), one can confirm by direct calculations that condition (i) is sufficient to guarantee that introducing a step in this region is not an improvement. Furthermore, for this family of preferences, condition (i) and ruling out the optimality of introducing a step in this pooling region are equivalent conditions when the IR constraint is not present, as shown in Amador and Bagwell (2013).

[26] For case where $\gamma_H = \underline{\gamma}$, Amador, Bagwell, and Frankel (2018) recover conditions for optimality of a cap without the presence of the IR constraint. Our condition is related but a bit more complex. For example, our condition requires that

$$\int_{\underline{\gamma}}^{\overline{\gamma}} w_q\big(\gamma, q_i(\overline{\gamma})\big)\, dF(\gamma) \geq 0,$$

which holds with equality in Amador, Bagwell, and Frankel (2018). In our case, the regulator would value an increase in $q$ for all types, but such an increase is infeasible because of the binding IR constraint.

## 5. Global optimality

In this section, we present a proposition that provides sufficient conditions for the global optimality of the cap allocation with potential exclusion $q^\star$. We then provide several results that facilitate the application of the sufficient conditions. Finally, we discuss how our results can be used to characterize optimal regulation in the "no-exclusion" scenario mentioned in the Introduction.

The results of the previous section offer a characterization of the optimal solution given an exogenous amount of exclusion as defined by the fixed threshold or upper-bound type, $\gamma_t$. For every exclusion threshold $\gamma_t$, we found sufficient conditions for the associated truncated cap allocation $q_t^\star$, defined in (4), to be optimal when restricting attention only to those types not excluded from production. However, it is direct to argue now that, given an amount of exclusion, the truncated cap allocation is optimal when attention is widened to include all types. Note that the only potential issue is that the $q_t^\star$ allocation when extended for all types must remain incentive compatible. But this is straightforward: since type $\gamma_t$ is indifferent between producing or not, all types above $\gamma_t$ strictly prefer not to produce, as they face a higher marginal cost.

We have the following result.

Proposition 2. *Assume that parts (i) and (ii) of Proposition 1 hold for all $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$. Then the cap allocation with potential exclusion $q^\star$ defined in (2) solves the regulator's problem, Problem P1.*

Proof. We know from Lemma 1 that any level of exclusion is given by a threshold $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$. Given any level of exclusion $\gamma_t$, the allocation $q_t^\star(\gamma|\gamma_t)$ defined in equation (4) remains a feasible allocation when the allocation is extended to entire type space by assigning no production to types strictly above $\gamma_t$. This follows because type $\gamma_t$ is indifferent between producing or not in the $q_t^\star(\gamma|\gamma_t)$ allocation, and thus all types higher than $\gamma_t$ strictly prefer not to produce, as prescribed by the allocation.

Thus, for a given level of exclusion, $\gamma_t$, Proposition 1 guarantees that the allocation $q_t^\star(\gamma|\gamma_t)$ extended over the entire type space is optimal within all feasible allocations that deliver the same level of exclusion.

Note that the allocation $q^\star(\gamma)$ is optimal among all $q_t^\star(\gamma|\gamma_t)$ allocations for all $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$. We can ignore any allocation where $\gamma_t = \underline{\gamma}$ (i.e., full exclusion) as such an allocation is dominated by the fully flexible allocation. As a result, $q^\star(\gamma)$ is optimal among the set of all feasible allocations. $\square$

We now provide several results that facilitate the application of our propositions. We develop three approaches. We first provide a corollary that states simple conditions under which our sufficient conditions for our propositions are sure to hold. We next provide a second corollary, which identifies for a family of demand functions a simple inequality condition that guarantees the satisfaction of the sufficient conditions for our propositions. We show that the demand family includes linear demand, constant elasticity demand, and log demand functions, and we derive and interpret the corresponding inequality condition for each of these demand specifications. Finally, a third approach is to check the sufficient conditions for our propositions directly. To illustrate this approach, we consider an example with exponential demand.

We begin with the following corollary, which provides simple and easy-to-check conditions for Proposition 1 and 2.

Corollary 1. *Suppose that $\kappa \geq 1/2$. For given $\gamma_t$, if $f(\gamma)$ is nondecreasing for all $\gamma \in [\underline{\gamma}, \gamma_t]$, then conditions (i) and (ii) of Proposition 1 hold. If $f(\gamma)$ is nondecreasing for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$, then the cap allocation with potential exclusion $q^\star$ is optimal within the set of all feasible allocations.*

The relative concavity and monotone density sufficient conditions in Corollary 1 are directly consistent with the intuition developed in Section 4.2. In particular, we note that $\kappa \geq 1/2$ is sure to hold if $v(q)$ is weakly concave; further, this inequality can hold as well when $v(q)$ is convex, if $\alpha$ is sufficiently large, and/or $b''(q)$ is sufficiently large in absolute value (so that $q_f(\gamma)$ is relatively flat).[27]

Let us also point out that the conditions in Corollary 1 hold independently of the value of the fixed cost, $\sigma$. That is, if $\kappa \geq 1/2$ and $f$ is nondecreasing for its entire support, then the optimal cap allocation with potential exclusion, $q^\star$, is optimal for any $\sigma$. Note that $q^\star$ is itself affected by the value of $\sigma$, but Corollary 1 guarantees that its global optimality is not. We note in particular that a higher $\sigma$ generates a lower value for $q_i(\gamma_t)$, corresponding to a higher price cap. Also, a different value of $\sigma$ may change the optimal value of $\gamma_t$ embedded in $q^\star$.

We proceed now to our second approach for facilitating the application of our propositions. In the Appendix proof of Corollary 1, we show that if the following $M_2(\gamma)$ function,

$$M_2(\gamma) \equiv \kappa F(\gamma) + \frac{1}{\alpha} v'\big(q_i(\gamma_t)\big) f(\gamma) + (\kappa - 1)\big(\gamma - b'\big(q_i(\gamma_t)\big)\big) f(\gamma),$$

is nondecreasing in $[\gamma_H(\gamma_t), \gamma_t]$, then part (i) of Proposition 1 holds. We now show that for a demand family (that includes several commonly used examples as we show below), $M_1(\gamma) = M_2(\gamma)$; thus, for this family, if part (ii) of Proposition 1 holds globally for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$, then part (i) holds as well.

Toward this end, we consider a family of demand functions such that[28]

$$\frac{P'(q)}{P(q)} q = a_0 + \frac{b_0}{P(q)} \quad \text{for all } q \in (0, q_{\max}] \text{ with } a_0 \neq -1. \tag{6}$$

We have the following result.

[27]The result that higher $\alpha$ makes the condition $\kappa \geq 1/2$ in Corollary 1 easier to hold is more general. Indeed, it is possible to show that a higher value of $\alpha$ makes conditions (i) and (ii) in Proposition 1 easier to satisfy.

[28]As we discuss later, this family of demand functions generates payoff functions $w$ and $b$ that belong to the preference family identified by Amador and Bagwell (2013) in their Proposition 2. This family is also the one studied by the linear delegation approach developed by Kolotilin and Zapechelnyuk (2019).

LEMMA 4. *Suppose that (6) holds. Then*

(a) $v(q) = -\frac{a_0}{1+a_0} b(q) - \frac{b_0}{1+a_0} q$ *for all* $q \in Q$,

(b) $\kappa = 1 + \frac{1}{\alpha}\frac{v''(q)}{b''(q)} = 1 - \frac{1}{\alpha}\frac{a_0}{1+a_0}$,

(c) $M_1(\gamma) = M_2(\gamma)$ *for all* $\gamma \in [\underline{\gamma}, \overline{\gamma}]$.

For the demand family stated in equation (6), we can obtain a general sufficient condition for the results in Proposition 1 (and 2) to hold.

COROLLARY 2. *Suppose that P satisfies (6). If*

$$(2\kappa - 1) f(\gamma) + \frac{1}{\alpha} v'\big(q_f(\gamma)\big) f'(\gamma) \geq 0 \tag{7}$$

*holds for all* $\gamma \in [\underline{\gamma}, \overline{\gamma}]$, *then conditions (i) and (ii) of Proposition 1 hold for all* $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$.

It is instructive to compare Corollaries 1 and 2. Inequality (7) clearly holds if $\kappa \geq 1/2$ and $f(\gamma)$ is nondecreasing for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$. For the demand family identified in equation (6), inequality (7) further indicates exactly how a relaxation of either of these conditions can be accommodated by an offsetting strengthening of the other. Note that similar to Corollary 1, the condition in Corollary 2 is independent of the value of $\sigma$, and thus implies that the optimal cap allocation with potential exclusion $q^\star$ as defined in (2) is optimal for any $\sigma$.

As we now illustrate, the demand family defined in equation (6) includes several common examples as special cases. For each of these examples, we also represent the form that inequality (7) takes and thereby derive sufficient conditions for the optimality of the cap allocation with potential exclusion $q^\star$.

*Linear demand* Consider $P(q) = \mu - \beta q$ with $\mu > \overline{\gamma}$, $\beta > 0$, and $Q = [0, \mu/\beta - \epsilon]$ for $\epsilon > 0$ small. For this example, $q_f(\gamma) = (\mu - \gamma)/(2\beta)$, $v(q) = \beta q^2/2$, and $\kappa = 1 - \frac{1}{2\alpha}$. Assumption 1 is satisfied for $\epsilon > 0$ sufficiently small if $\alpha \in [\mu/(\mu + \underline{\gamma}), 1]$ where $1 > \mu/(\mu + \underline{\gamma}) > 1/2$ follows from $\mu > \underline{\gamma} > 0$. Assumption 2 is satisfied if $q_f(\overline{\gamma}) > \sqrt{\sigma/\beta}$. This demand satisfies condition (6) with $a_0 = 1$ and $b_0 = -\mu$. Condition (7) is satisfied in this example if

$$\frac{f'(\gamma)}{f(\gamma)} \geq \frac{2(1-\alpha)}{\mu - \overline{\gamma}}$$

for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$.

*Constant elasticity demand* Consider $P(q) = q^{-\frac{1}{\eta}}$ with $\eta > 1$, and let $Q = [0, q_{\max}]$ where $q_{\max} > 0$. For this example, $q_f(\gamma) = (\frac{\gamma\eta}{\eta-1})^{-\eta}$, $v(q) = \frac{1}{\eta-1} q^{\frac{\eta-1}{\eta}}$, and $\kappa = 1 + \frac{1}{\alpha}\frac{1}{\eta-1}$. Assumption 1 is satisfied if ${q_{\max}}^{-\frac{1}{\eta}} < \frac{\gamma}{1-\frac{1}{\eta}(1-\frac{1}{\alpha})}$ where $0 < \frac{\gamma}{1-\frac{1}{\eta}(1-\frac{1}{\alpha})} \leq \underline{\gamma}$ follows from $\alpha \in (0, 1]$ and $\underline{\gamma} > 0$. Assumption 2 is satisfied if $(\frac{\overline{\gamma}\eta}{\eta-1})^{1-\eta}\frac{1}{\eta} > \sigma$. This demand satisfies condition (6) with $a_0 = -\frac{1}{\eta}$ and $b_0 = 0$. Condition (7) is satisfied in this example if

$$\frac{f'(\gamma)}{f(\gamma)} \geq -\frac{\alpha(\eta-1)+2}{\gamma}$$

for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$.

*Logarithmic demand* Consider $P(q) = \mu - \beta \ln q$ with $\beta > 0$ and $Q = [0, e^{\mu/\beta} - \epsilon]$ for $\epsilon > 0$ small. For this example, $q_f(\gamma) = e^{\frac{\mu-\beta-\gamma}{\beta}}$, $v(q) = \beta q$, and $\kappa = 1$. Assumption 1 is satisfied for $\epsilon > 0$ sufficiently small if $\beta(1-\alpha)/\alpha < \underline{\gamma}$. Assumption 2 is satisfied if $\beta e^{\frac{\mu-\beta-\overline{\gamma}}{\beta}} > \sigma$. This demand satisfies condition (6) with $a_0 = 0$ and $b_0 = -\beta$. Condition (7) is satisfied in this example if

$$\frac{f'(\gamma)}{f(\gamma)} \geq -\frac{\alpha}{\beta}$$

for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$.

Of course, the demand family defined by (6) includes examples beyond the three examples highlighted here.[29] The three examples, however, are commonly used in the literature and illustrate the breadth of the demand family defined by (6).

The sufficient conditions derived for the three examples admit an interpretation that is in line with the intuition developed previously whereby a rising density $f(\gamma)$ and a concave $v(q)$ work in favor of the optimality of the cap allocation. For the constant elasticity and log demand examples, $v(q)$ is concave and linear, respectively, and the sufficient conditions hold when $f(\gamma)$ is nondecreasing; indeed, for these examples, the sufficient conditions are satisfied even when $f(\gamma)$ is decreasing, provided that it does not fall too quickly. By contrast, for the linear demand example, $v(q)$ is convex, which works against the optimality of the cap allocation. The sufficient condition for this example thus places a more demanding restriction on the density: the condition fails if $f(\gamma)$ is anywhere decreasing, and it requires that $f(\gamma)$ is increasing (nondecreasing) when $\alpha < 1$ ($\alpha = 1$). We note as well that in all of these examples a higher value of $\alpha$ also supports the sufficient conditions (as expected from our previous discussion).

Interestingly, the demand family we have identified corresponds to the "linear delegation" case studied in Kolotilin and Zapechelnyuk (2019) for the regulation problem when $\sigma = 0$.[30] However, we are not restricted to demand functions within the family that satisfies condition (6). For other demand functions, we could use parts (i) and (ii)

[29] For example, the demand function $P(q) = \mu - \beta q^{\eta}$ satisfies (6) with $a_0 = \eta$ and $b_0 = -\mu\eta$.

[30] Kolotilin and Zapechelnyuk (2019) consider linear delegation problems where the principal's objective, $V(\gamma, q)$, satisfies $V_q(\gamma, q) = -\gamma - c(q)$ and where the agent's objective, $U(\gamma, q)$, satisfies $U_q(\gamma, q) = d(\gamma) - c(q)$ where $c$ and $d$ are continuous functions and $c$ is strictly increasing. This implies that $V_q(\gamma, q) - U_q(\gamma, q) = \gamma - d(\gamma)$. Thus, $V_q(\gamma, q) - U_q(\gamma, q)$ is independent of $q$. In our case, for $\sigma = 0$, $V_q(\gamma, q) = w_q(\gamma, q) = -\gamma + b'(q) + \frac{1}{\alpha} v'(q)$. Given that the objective of the agent can be modified by any strictly increasing affine transformation, we have that in our case, $U_q(\gamma, q) = A(-\gamma + b'(q))$ for any $A > 0$. Hence, linear delegation requires that there exists $A > 0$ such that $V_q(\gamma, q) - U_q(\gamma, q)$ is independent of $q$, or alternatively, that there exists $A > 0$ and $B$ such that

$$b'(q) + \frac{1}{\alpha} v'(q) - Ab'(q) = B.$$

of Propositions 1 and 2 directly. Alternatively, Corollary 1 also allows us to find simple conditions. Our results also apply when there is a fixed cost of production, $\sigma > 0$.

To illustrate the approach in which parts (i) and (ii) of Propositions 1 and 2 are directly used, we consider next an example with exponential demand. We note that this example does not fit in the family specified by condition (6).

*Exponential demand* Consider $P(q) = \beta e^{-q}$ with $\beta > \max\{\overline{\gamma}, \underline{\gamma} e^2\}$ with $Q = [0, 2 - \epsilon]$ for $\epsilon > 0$ small. For this example, the sign of $v''(q)$ varies over $Q$. We find that $v''(q) = \beta e^{-q}(1 - q)$ and $\kappa = 1 - \frac{1}{2\alpha}$. Assumption 1 is satisfied for $\epsilon > 0$ sufficiently small if $\alpha > 2/(1 + \underline{\gamma})$ where this inequality when combined with $\alpha \in (0, 1]$ implies that $\underline{\gamma} > 1$. Assumption 2 is satisfied when $\max_{q \in Q}(\beta e^{-q} - \overline{\gamma})q > \sigma$. Corollary 1 holds if $\alpha = 1$ and $f$ if nondecreasing for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$.

At this point, it is convenient to pause and consider the "no-exclusion" scenario mentioned in the Introduction, wherein the regulator must ensure that all types choose to produce so that exclusion never occurs. To characterize the optimal regulatory policy for this scenario, we may refer to the truncated cap allocation $q_t^\star(\gamma|\gamma_t)$, defined in (4), for the special case where $\gamma_t = \overline{\gamma}$. This allocation corresponds to a price-cap regulatory policy, where the price cap is set at the second-best level that leaves a monopolist with the highest possible cost, $\overline{\gamma}$, with zero profit (inclusive of the fixed cost, $\sigma$). To establish conditions for the optimality of this policy for the no-exclusion scenario, we simply set $\gamma_t = \overline{\gamma}$ and refer to Proposition 1, Corollaries 1 and 2, and the demand examples above. Thus, for example, this price-cap allocation is optimal for the no-exclusion scenario if the demand function takes a linear, constant elasticity or log form and if a simple inequality condition holds, respectively, where the inequality condition is sure to hold if the density is nondecreasing over the full support.

By contrast, a characterization of optimal regulation for the general scenario in which exclusion is allowed must also determine the optimal value for $\gamma_t$. In other words, the optimal regulatory policy for the general scenario must determine as well the degree (if any) of exclusion. We develop our results for the optimal degree of exclusion in the next section.

## 6. When to exclude

In the previous section, we obtain conditions that guarantee that the cap allocation with potential exclusion, $q^\star$ defined in (2), is optimal within the set of all feasible allocations. In this section, we study the properties of this optimal allocation, $q^\star$, and in particular, whether or not some types are excluded from production.

Given a level of exclusion $\gamma_t$, we can write the welfare function as

$$W(\gamma_t) = \int_{\underline{\gamma}}^{\gamma_H(\gamma_t)} \big(w\big(\gamma, q_f(\gamma)\big) - \sigma\big)\, dF(\gamma) + \int_{\gamma_H(\gamma_t)}^{\gamma_t} \big(w\big(\gamma, q_i(\gamma_t)\big) - \sigma\big)\, dF(\gamma)$$

Using $b'(q) = P'(q)q + P(q)$ and $v'(q) = -qP'(q)$, the above delivers condition (6). Note that demand functions within this family deliver payoff functions $w(\gamma, q)$ and $b(q)$ that belong to the restricted preference family previously identified by Amador and Bagwell (2013) in their Proposition 2.

where as before $q_i(\gamma_t)$ represents the quantity strictly above $q_f(\gamma_t)$ that makes type $\gamma_t$ indifferent between producing or not, as in (3).

Taking the derivative of the welfare function with respect to $\gamma_t$, we obtain that[31]

$$W'(\gamma_t) = \big(w\big(\gamma_t, q_i(\gamma_t)\big) - \sigma\big) f(\gamma_t) + \int_{\gamma_H(\gamma_t)}^{\gamma_t} w_q\big(\gamma, q_i(\gamma_t)\big) q_i'(\gamma_t)\, dF(\gamma)$$

This first-order condition has a simple intuition. Increasing the exclusion threshold $\gamma_t$ generates two effects. As indicated in the first term, it brings the marginal type that was previously excluded back into production. But in addition it changes the quantity at which all pooled types produce, generating an inframarginal effect that is captured by the second term. The following result shows that increasing the exclusion threshold reduces the quantity at which pooled types produce.

Lemma 5. *The quantity of the indifferent type, $q_i(\gamma_t)$, is such that $q_i'(\gamma_t) < 0$. In addition, $\gamma_t > \gamma_H(\gamma_t)$ for all $\gamma_t > \underline{\gamma}$.*

We now develop conditions under which no exclusion ($\gamma_t = \overline{\gamma}$) is optimal. Using the definition of $w$, together with the definitions of $\gamma_t$ and $\gamma_H(\gamma_t)$, we obtain

$$\begin{aligned} W'(\gamma_t) = {} & \frac{1}{\alpha} v\big(q_i(\gamma_t)\big) f(\gamma_t) - q_i(\gamma_t)\left(\frac{1}{\alpha} v'\big(q_i(\gamma_t)\big)\right) \frac{F(\gamma_t) - F\big(\gamma_H(\gamma_t)\big)}{\gamma_t - b'\big(q_i(\gamma_t)\big)} \\ & + \frac{q_i(\gamma_t)}{\gamma_t - b'\big(q_i(\gamma_t)\big)} \int_{\gamma_H(\gamma_t)}^{\gamma_t} \big(\gamma - b'\big(q_i(\gamma_t)\big)\big)\, dF(\gamma) \end{aligned}$$

for all $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$.[32]

We know that $\gamma > b'(q_i(\gamma_t))$ for all $\gamma > \gamma_H(\gamma_t)$. So, the last term in the above is strictly positive. Thus,

$$\begin{aligned} W'(\gamma_t) & > \frac{1}{\alpha} v\big(q_i(\gamma_t)\big) f(\gamma_t) - q_i(\gamma_t)\left(\frac{1}{\alpha} v'\big(q_i(\gamma_t)\big)\right) \frac{F(\gamma_t) - F\big(\gamma_H(\gamma_t)\big)}{\gamma_t - b'\big(q_i(\gamma_t)\big)} \\ & = \frac{1}{\alpha} v\big(q_i(\gamma_t)\big)\left[f(\gamma_t) - \frac{F(\gamma_t) - F\big(\gamma_H(\gamma_t)\big)}{\gamma_t - b'\big(q_i(\gamma_t)\big)}\right] \\ & \quad + \frac{1}{\alpha}\big[v\big(q_i(\gamma_t)\big) - v'\big(q_i(\gamma_t)\big) q_i(\gamma_t)\big] \frac{F(\gamma_t) - F\big(\gamma_H(\gamma_t)\big)}{\gamma_t - b'\big(q_i(\gamma_t)\big)}. \end{aligned}$$

Using $b'(q_i(\gamma_t)) \leq \gamma_H(\gamma_t) < \gamma_t$, we have that

$$f(\gamma_t) - \frac{F(\gamma_t) - F\big(\gamma_H(\gamma_t)\big)}{\gamma_t - b'\big(q_i(\gamma_t)\big)} \geq f(\gamma_t) - \frac{F(\gamma_t) - F\big(\gamma_H(\gamma_t)\big)}{\gamma_t - \gamma_H(\gamma_t)} = \frac{\int_{\gamma_H(\gamma_t)}^{\gamma_t} \big[f(\gamma_t) - f(\gamma)\big]\, d\gamma}{\gamma_t - \gamma_H(\gamma_t)} \geq 0,$$

[31] The function $\gamma_H(\gamma_t)$ may fail to be differentiable at the highest value for $\gamma_t$ at which $\gamma_H(\gamma_t) = \underline{\gamma}$; however, the differentiability of $\gamma_H$ does not affect the differentiability of the objective. An argument similar to the one used in the proof of Lemma 2 can be used to show differentiability of the objective.

[32] The value of $\alpha$ does not affect $q_i(\gamma_t)$. So, using the above equation one can show that $\alpha W'(\gamma_t)$ is strictly increasing in $\alpha$. Thus, the optimal exclusion threshold $\gamma_t$ is weakly increasing in $\alpha$.

where the last inequality follows if $f(\gamma)$ is nondecreasing. If $v(q)$ is weakly concave, then $v(q) - v'(q)q \geq 0$ as $v(0) = 0$. Hence, we have the following.

Proposition 3 (No exclusion). *If $f(\gamma)$ is nondecreasing for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$ and $v(q)$ is weakly concave for all $q \in Q$, then $W'(\gamma_t) > 0$ for all $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$. Under these conditions, Corollary 1 holds, and thus a cap with no exclusion ($\gamma_t = \overline{\gamma}$) solves the regulator's problem.*

Proof. The proof is given in the text. □

Proposition 3 delivers a general set of conditions under which there is no exclusion. The log demand and constant elasticity demand examples satisfy the requirement that $v$ is weakly concave. In addition, if $f$ is nondecreasing, then the optimal allocation in these examples is a cap allocation without exclusion.

It may be helpful to discuss the role of the fixed cost, $\sigma$, in these results. If $\sigma = 0$, then the possibility of exclusion is included in the truncated problem, as the regulator could assign zero output to some types and still satisfy the IR constraint for that problem. Thus, if Proposition 1 holds at $\gamma_t = \overline{\gamma}$, then a cap allocation *without exclusion* is optimal. When $\sigma > 0$, exclusion is no longer included in the truncated problem. Exclusion remains feasible in the regulator's problem, however, and indeed the assignment of zero output carries the extra benefit that the fixed cost is saved. Proposition 3 covers this case as well and shows that exclusion is still not used.

For the linear demand example, however, $v$ is strictly convex, and Proposition 3 thus does not apply. For this example, in the case of a uniform distribution with $\alpha = 1$ (so that the regulator maximizes aggregate social surplus), and $\sigma > 0$, we have a very different result.

Proposition 4 (Exclusion). *Consider the linear demand example and suppose that $F$ is uniform and $\alpha = 1$. If $\sigma > 0$, then*

*(a) In any optimal allocation, $\gamma_t$ is such that $\gamma_H(\gamma_t) = \underline{\gamma}$.*

*(b) If $q_i(\overline{\gamma}) < q_f(\underline{\gamma})$, then in any optimal allocation $\gamma_t < \overline{\gamma}$ and $q_i(\gamma_t) = q^\star$ where $q^\star$ is a solution of*

$$\frac{(\mu - \underline{\gamma})(\mu - \underline{\gamma} - 2\beta q^\star)(q^\star)^2}{\sigma - \beta (q^\star)^2} = \sigma.$$

This proposition contains two results. The first is that in the linear demand example with a uniform distribution and $\alpha = 1$, it is always optimal to pool all types at the cap (part (a)). Part (b) argues that if not all types pool at the cap when an allocation features no exclusion, that is, when $\gamma_t = \overline{\gamma}$, then some higher-cost types will necessarily be excluded in any optimal allocation.[33]

[33]The proposition only characterizes the solution for $\sigma > 0$. When $\sigma = 0$, if $q_i(\overline{\gamma}) < q_f(\underline{\gamma})$, we can show that any $\gamma_t$ such that $q_i(\gamma_t) \leq q_f(\underline{\gamma})$ is optimal. Thus, the regulator is indifferent between some exclusion or none. Kolotilin and Zapechelnyuk (2019) characterize the solution for this case ($\alpha = 1$, $\sigma = 0$, and linear demand) when the distribution $F$ is unimodal and show that optimal regulation involves a cap with exclusion. This case goes beyond our Corollary 2, which requires a nondecreasing density.

The above result demonstrates that no-exclusion result of Proposition 3 is not a general property. Because of its tractability, the linear demand example with a uniform distribution and $\alpha = 1$ is often used in the literature. For this case, we have shown that a cap allocation is optimal but that such an allocation also features the exclusion of higher-cost types.

In this example, the optimal regulation contract provides for a disconnected set of quantities (a quantity of zero for excluded types and an interval of quantities bounded away from zero for nonexcluded types). This contract is clearly distinct from the interval allocations that are typically featured in the delegation literature.

## 7. Conclusion

We analyze the Baron and Myerson (1982) model of regulation under the restriction that transfers are infeasible. To do this, we extend the Lagrangian approach to delegation problems of Amador and Bagwell (2013) to include an ex post participation constraint that allows for the possible exclusion of some types. We report sufficient conditions under which optimal regulation takes the simple and common form of price-cap regulation. We identify families of demand and distribution functions and welfare weights that satisfy our sufficient conditions. We also report conditions under which the optimal price cap is set at a level such that no types are excluded. Using a linear demand example, we show that exclusion of higher-cost types can be optimal when these conditions fail to hold. Our analysis also can be used to provide conditions for the optimality of price-cap regulation when an ex post participation constraint is present and exclusion is infeasible.

Our work points to several directions for future research. First, we provide general sufficient conditions so that a cap allocation with potential exclusion is optimal. These sufficient conditions guarantee that the Lagrangian approach can be used to show that a price cap is optimal *for any given level of exclusion*. Thus, the sufficient conditions may be stronger than necessary since the price-cap structure is required to be optimal even for exclusion levels that are suboptimal. It should be possible to relax these conditions by using the Lagrangian approach *only* at the optimal level of exclusion.

Second, if our sufficient conditions fail, it may be that the optimal allocation is not a price cap with potential exclusion. In that case, the Lagrangian approach requires us to identify the alternative solution candidate. It should be possible as well to construct Lagrange multipliers and generate sufficient conditions for such a case.

Third, we focus on a single-product monopolist and leave for future research the multiproduct expression of our findings. More generally, the characterization of optimal delegation contracts in multidimensional settings is a challenging and important avenue for future work.[34]

[34] For related work, see Ambrus and Egorov (2017), Amador and Bagwell (2020), Armstrong and Vickers (2010), Frankel (2014), Frankel (2016), and Koessler and Martimort (2012). The paper by Frankel (2016) is perhaps of special relevance here. He considers a model with multiple actions and establishes the exact optimality of a generalized cap rule, but under the assumptions that the loss function is quadratic, the agent has a constant bias, the ex ante distribution of states is normal i.i.d., and the participation constraint is absent.

Fourth, our analysis extends the optimal delegation literature to include an ex-post participation constraint that allows for potential exclusion within a regulation framework. We expect that other applications likewise may be naturally captured in versions of the delegation model developed here.

## Appendix A: Proof of Lemma 1

Proof. Suppose to the contrary that for some $\gamma_1$ and $\gamma_2$ with $\underline{\gamma} \leq \gamma_1 < \gamma_2 \leq \overline{\gamma}$, we have a feasible allocation such that $q(\gamma_1) = 0 < q(\gamma_2)$. A monopolist with type $\gamma_1$ then receives a profit of zero and would gain by violating its incentive compatibility constraint and selecting instead the positive output intended for type $\gamma_2$:

$$-\gamma_1 q(\gamma_2) + b\big(q(\gamma_2)\big) - \sigma > -\gamma_2 q(\gamma_2) + b\big(q(\gamma_2)\big) - \sigma \geq 0$$

where the first inequality follows since $q(\gamma_2) > 0$ and $\gamma_1 < \gamma_2$, and the second inequality follows since under incentive compatibility a monopolist with type $\gamma_2$ cannot gain from selecting $q(\gamma_1) = 0$ rather than $q(\gamma_2)$.

For the second part, suppose that for $\gamma_t \in (\underline{\gamma}, \overline{\gamma})$, $-\gamma_t q(\gamma_t) + b(q(\gamma_t)) > \sigma$. Then, for all sufficiently small $\epsilon > 0$, we have that $-(\gamma_t + \epsilon) q(\gamma_t) + b(q(\gamma_t)) > \sigma$. As a result, type $\gamma_1 = \gamma_t + \epsilon$ will prefer to produce rather than not, a contradiction of the cut-off property. Suppose instead that $-\gamma_t q(\gamma_t) + b(q(\gamma_t)) < \sigma$, so type $\gamma_t$ strictly prefers not to produce. For all sufficiently small $\epsilon > 0$, we have that $-(\gamma_t - \epsilon) q(\gamma_t - \epsilon) + b(q(\gamma_t - \epsilon)) > \sigma$, by the cut-off property and strict monotonicity of the profit function in $\gamma$ when $q > 0$. But this implies that $-\gamma_t q(\gamma_t - \epsilon) + b(q(\gamma_t - \epsilon)) > \sigma - \epsilon q(\gamma_t - \epsilon)$, and thus, for sufficiently small $\epsilon$, type $\gamma_t$ would strictly prefer to produce given Assumption 2 and choose type $\gamma_t - \epsilon$'s choice, a violation of feasibility. □

## Appendix B: Proof of Lemma 2

Proof. We start by observing that for any $\Delta \neq 0$,

$$\frac{W^c(x+\Delta) - W^c(x)}{\Delta} = \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \frac{w(\gamma, x+\Delta) - w\big(\gamma, q_f(\gamma)\big)}{\Delta} \, dF(\gamma) + \int_{\gamma_c(x)}^{\overline{\gamma}} \frac{w(\gamma, x+\Delta) - w(\gamma, x)}{\Delta} \, dF(\gamma) \tag{8}$$

Then we consider two different cases.

**Case 1.** $x > q_f(\underline{\gamma})$. Then for all $|\Delta| > 0$ small enough, we have that $x + \Delta > q_f(\underline{\gamma})$, and as a result $\gamma_c(x+\Delta) = \gamma_c(x) = \underline{\gamma}$, and thus

$$\frac{W^c(x+\Delta) - W^c(x)}{\Delta} = \int_{\gamma_c(x)}^{\overline{\gamma}} \frac{w(\gamma, x+\Delta) - w(\gamma, x)}{\Delta} \, dF(\gamma)$$

Taking the limit as $\Delta \to 0$, we obtain

$$\frac{dW^c(x)}{dx} = \int_{\gamma_c(x)}^{\overline{\gamma}} w_q(\gamma, x) \, dF(\gamma)$$

**Case 2.** $0 < q_f(\overline{\gamma}) < x \leq q_f(\underline{\gamma})$. Consider a neighborhood $U_x$ around $x$ such that that $0 \notin \text{cl}(U_x)$. Let $K_x = \max_{y \in \text{cl}(U_x)} |b'(y) + v'(y)/\alpha|$. Assumption 1 guarantees that such $K_x$ exists and is finite. The mean value theorem guarantees that $|\frac{(b(y)+v(y)/\alpha)-(b(x)-v(x)/\alpha)}{y-x}| \leq K_x$.

Note that $q_f(\gamma_c(x)) = x$, and that for $|\Delta| > 0$ small enough, $q_f(\gamma) \in U_x$ for $\gamma \in [\gamma_c(x + |\Delta|), \gamma_c(x - |\Delta|)]$, given that $q_f$ and $\gamma_c$ are continuous. Then

$$
\begin{aligned}
&\int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \left| \frac{w(\gamma, x+\Delta) - w(\gamma, q_f(\gamma))}{\Delta} \right| dF(\gamma) \\
&= \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \left| \frac{-\gamma(x + \Delta - q_f(\gamma))}{\Delta} \right. \\
&\quad + \left. \frac{(b(x+\Delta) + v(x+\Delta)/\alpha) - (b(q_f(\gamma)) - v(q_f(\gamma))/\alpha)}{\Delta} \right| dF(\gamma) \\
&\leq \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \gamma \left| \frac{x + \Delta - q_f(\gamma))}{\Delta} \right| dF(\gamma) \\
&\quad + \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \left| \frac{(b(x+\Delta) + v(x+\Delta)/\alpha) - (b(q_f(\gamma)) - v(q_f(\gamma))/\alpha)}{x + \Delta - q_f(\gamma)} \right. \\
&\quad \times \left. \frac{x + \Delta - q_f(\gamma)}{\Delta} \right| dF(\gamma) \\
&\leq \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \gamma \left| \frac{x + \Delta - q_f(\gamma))}{\Delta} \right| dF(\gamma) + \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} K_x \left| \frac{x + \Delta - q_f(\gamma)}{\Delta} \right| dF(\gamma) \\
&= (\gamma + K_x) \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \left| \frac{x + \Delta - q_f(\gamma)}{\Delta} \right| dF(\gamma) \\
&\leq (\gamma + K_x) \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} \left| \frac{x + \Delta - x}{\Delta} \right| dF(\gamma) \\
&= (\gamma + K_x) \int_{\gamma_c(x+\Delta)}^{\gamma_c(x)} dF(\gamma)
\end{aligned}
$$

The steps above are immediate except for the last inequality. For this, we use that if $x < q_f(\underline{\gamma})$, then for all sufficiently small $\Delta$, $q_f(\gamma_c(x + \Delta)) = x + \Delta$. If $x = q_f(\underline{\gamma})$, then for $\Delta > 0$, the integral range is empty (and thus the integral equals zero). For $\Delta < 0$, we still have that $q_f(\gamma_c(x + \Delta)) = x + \Delta$.

Now note that the last integral above tends to zero as $\Delta$ goes to zero, and thus, taking the limit of (8) as $\Delta \to 0$, we obtain that for $x > q_f(\overline{\gamma})$:

$$
\frac{dW^c(x)}{dx} = \int_{\gamma_c(x)}^{\overline{\gamma}} w_q(\gamma, x)\, dF(\gamma)
$$

Note that

$$\frac{dW^c(s)}{dx}\bigg|_{x=q_{\max}} = \int_{\gamma_c(q_{\max})}^{\overline{\gamma}} w_q(\gamma, q_{\max})\, dF(\gamma) < \int_{\underline{\gamma}}^{\overline{\gamma}} w_q(\underline{\gamma}, q_{\max})\, dF(\gamma) = w_q(\underline{\gamma}, q_{\max}) < 0$$

where we use that $\gamma_c(q_{\max}) = \underline{\gamma}$ as $q_{\max} > q_f(\underline{\gamma})$ and that $w_q(\underline{\gamma}, q_{\max}) > w_q(\gamma, q_{\max})$ for $\gamma > \underline{\gamma}$ to show the first inequality. For the last inequality, we use Assumption 1.

Note that $w_q(\overline{\gamma}, q_f(\overline{\gamma})) > 0$, as $v'(q) > 0$. Consider $x_0 > q_f(\overline{\gamma})$ such that $w_q(\overline{\gamma}, x_0) > 0$. Such an $x_0$ exists by continuity of $w_q$. Note that for all $q_0 \in (q_f(\overline{\gamma}), x_0]$, $\gamma_c(q_0) < \overline{\gamma}$ and $w_q(\overline{\gamma}, q_0) \geq w_q(\overline{\gamma}, x_0) > 0$, by weak concavity of $w$. It follows that $0 < w_q(\overline{\gamma}, q_0) \leq w_q(\gamma, q_0)$ for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$. Hence, for all $q_0 \in (q_f(\overline{\gamma}), x_0]$ we have that

$$\frac{dW^c(s)}{dx}\bigg|_{x=q_0} = \int_{\gamma_c(q_0)}^{\overline{\gamma}} w_q(\gamma, q_0)\, dF(\gamma) > 0$$

It follows then that the optimal value of $x$ is interior to $(q_f(\overline{\gamma}), q_{\max}]$ and must solve the first-order condition in the lemma. □

## Appendix C: Proof of Lemma 3

Proof. Recall from Lemma 2 that $x > q_f(\overline{\gamma})$ and $\int_{\gamma_c(x)}^{\overline{\gamma}} w_q(\gamma, x)\, dF(\gamma) = 0$. Using $w_{q\gamma}(\gamma, q) = -1 < 0$, $w_q(\overline{\gamma}, x) < 0$ follows. Next, observe $w_q(\overline{\gamma}, x) = -\overline{\gamma} + P(x) + (\frac{1-\alpha}{\alpha})(-P'(x)x) < 0$, and thus $P(x) - \overline{\gamma} < (\frac{1-\alpha}{\alpha})(P'(x)x) \leq 0$ given $P'(x) < 0$ and $\alpha \in (0, 1]$. Hence, $\pi(\overline{\gamma}, x) = (P(x) - \overline{\gamma})x < 0 \leq \sigma$, and so the IR constraint fails for the highest type. □

## Appendix D: Proof of Proposition 1

Proof. We proceed as follows. First, we restate the regulator's truncated problem by expressing the incentive compatibility constraints in their standard form as an integral equation and a monotonicity requirement:[35]

$$\max_{q_t:\Gamma_t(\gamma_t)\to Q} \int_{\Gamma_t(\gamma_t)} \big(w\big(\gamma, q_t(\gamma)\big) - \sigma\big)\, dF(\gamma) \quad \text{subject to:}$$

$$-\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma - \int_{\gamma}^{\gamma_t} q_t(\tilde{\gamma})\, d\tilde{\gamma} = \overline{U}, \quad \text{for all } \gamma \in \Gamma_t(\gamma_t)$$

$$q_t(\gamma) \quad \text{nonincreasing, for all } \gamma \in \Gamma_t(\gamma_t)$$

$$0 \leq -\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma, \quad \text{for all } \gamma \in \Gamma_t(\gamma_t)$$

where $\overline{U} \equiv -\gamma_t q_t(\gamma_t) + b(q_t(\gamma_t)) - \sigma$ is the profit enjoyed by the monopolist with the highest possible cost type in $\Gamma_t(\gamma_t)$.

Next, we follow Amador and Bagwell (2013) and rewrite the incentive constraints as a set of two inequalities and embed the monotonicity constraint in the choice set of $q_t(\gamma)$.

[35] See, for example, Milgrom and Segal (2002).

With the choice set for $q_t(\gamma)$ defined as $\Phi \equiv \{q_t | q_t : \Gamma_t(\gamma_t) \to Q$, and $q_t$ nonincreasing$\}$, the regulator's truncated problem may now be stated in final form as follows:

$$(\mathrm{P}'_t) \qquad \max_{q_t \in \Phi} \int_{\Gamma_t(\gamma_t)} \big(w\big(\gamma, q_t(\gamma)\big) - \sigma\big)\, dF(\gamma) \quad \text{subject to:}$$

$$\gamma q_t(\gamma) - b\big(q_t(\gamma)\big) + \sigma + \int_{\gamma}^{\gamma_t} q_t(\tilde{\gamma})\, d\tilde{\gamma} + \overline{U} \le 0, \quad \text{for all } \gamma \in \Gamma_t(\gamma_t) \tag{9}$$

$$-\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma - \int_{\gamma}^{\gamma_t} q_t(\tilde{\gamma})\, d\tilde{\gamma} - \overline{U} \le 0, \quad \text{for all } \gamma \in \Gamma_t(\gamma_t) \tag{10}$$

$$\gamma q_t(\gamma) - b\big(q_t(\gamma)\big) + \sigma \le 0, \quad \text{for all } \gamma \in \Gamma_t(\gamma_t) \tag{11}$$

Let $\Lambda_1(\gamma)$ and $\Lambda_2(\gamma)$ denote the (cumulative) multiplier functions associated with the two inequalities that define the incentive compatibility constraints in the final form of the regulator's truncated problem. The multiplier functions $\Lambda_1(\gamma)$ and $\Lambda_2(\gamma)$ are restricted to be nondecreasing in $\Gamma_t(\gamma_t)$. Letting $\Lambda(\gamma) \equiv \Lambda_1(\gamma) - \Lambda_2(\gamma)$, we can write the Lagrangian of the regulator's truncated problem as stated in Problem $\mathrm{P}'_t$ as follows:

$$\mathcal{L} = \int_{\Gamma_t} w\big(\gamma, q_t(\gamma)\big)\, dF(\gamma) - \int_{\Gamma_t} \left( \int_{\gamma}^{\gamma_t} q_t(\tilde{\gamma})\, d\tilde{\gamma} + \overline{U} + \gamma q_t(\gamma) - b\big(q_t(\gamma)\big) + \sigma \right) d\Lambda(\gamma)$$
$$+ \int_{\Gamma_t} \big(-\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma\big)\, d\Psi(\gamma),$$

where without loss of generality we have removed the constant $\sigma$ in the first integral and where to save notation we have removed the dependence of $\Gamma_t$ on $\gamma_t$. Notice that $\Psi(\gamma)$ is the multiplier for the ex post participation constraints. $\Psi(\gamma)$ is also restricted to be nondecreasing.

We propose the following multipliers:

$$\Lambda(\gamma) = \begin{cases} 0; & \gamma = \underline{\gamma} \\ w_q\big(\gamma, q_f(\gamma)\big) f(\gamma); & \gamma \in \big(\underline{\gamma}, \gamma_H(\gamma_t)\big) \\ A + \kappa\big(F(\gamma_t) - F(\gamma)\big); & \gamma \in \big[\gamma_H(\gamma_t), \gamma_t\big] \end{cases}$$

and

$$\Psi(\gamma) = \begin{cases} 0; & \gamma \in [\underline{\gamma}, \gamma_t) \\ A; & \gamma = \gamma_t \end{cases}$$

where

$$A = \frac{1}{\gamma_t - \gamma_H(\gamma_t)} \left[ \int_{\gamma_H(\gamma_t)}^{\gamma_t} w_q\big(\gamma, q_i(\gamma_t)\big) f(\gamma)\, d\gamma + \kappa\big(\gamma_H(\gamma_t) - b'\big(q_i(\gamma_t)\big)\big) F(\gamma_t) \right]. \tag{12}$$

Note that while defining $\Lambda(\gamma)$, we allow for the possibility that $\gamma_H(\gamma_t) = \underline{\gamma}$, and the intermediate case in the definition then does not apply. This is the case where there is full pooling of all types.

We show below that the hypothesis of Proposition 1 guarantees that $R(\gamma) \equiv \kappa F(\gamma) + \Lambda(\gamma)$ is nondecreasing; thus, we may write $\Lambda(\gamma)$ as the difference between two nondecreasing functions, $\Lambda_1(\gamma) = R(\gamma)$ and $\Lambda_2(\gamma) = \kappa F(\gamma)$.[36] We also require that $A \geq 0$ as $\Psi(\gamma)$ must be nondecreasing. We establish this inequality below.

We note that the cap allocation $q_t^\star(\gamma|\gamma_t)$ together with the proposed multipliers satisfy complementary slackness. The incentive compatibility constraints bind under the cap allocation, and $\Psi(\gamma)$ is constructed to be zero whenever the participation constraint holds with slack.

When these multipliers are used, the Lagrangian becomes

$$\begin{aligned}\mathcal{L} = &\int_{\Gamma_t} w\big(\gamma, q_t(\gamma)\big)\, dF(\gamma) - \int_{\Gamma_t}\left(\int_{\gamma}^{\gamma_t} q_t(\tilde{\gamma})\, d\tilde{\gamma} + \overline{U} + \gamma q_t(\gamma) - b\big(q_t(\gamma)\big) + \sigma\right) d\Lambda(\gamma)\\ &+ \big(-\gamma_t q_t(\gamma_t) + b\big(q_t(\gamma_t)\big) - \sigma\big) A.\end{aligned}$$

Recalling the definition of $\overline{U}$ and using $\Lambda(\underline{\gamma}) = 0$ and $\Lambda(\gamma_t) = A$, we can then write the Lagrangian as

$$\mathcal{L} = \int_{\Gamma_t} w\big(\gamma, q_t(\gamma)\big)\, dF(\gamma) - \int_{\Gamma_t}\left(\int_{\gamma}^{\gamma_t} q_t(\tilde{\gamma})\, d\tilde{\gamma} + \gamma q_t(\gamma) - b\big(q_t(\gamma)\big) + \sigma\right) d\Lambda(\gamma)$$

Integrating the Lagrangian by parts, we get[37]

$$\mathcal{L} = \int_{\Gamma_t} \big(w\big(\gamma, q_t(\gamma)\big) f(\gamma) - \Lambda(\gamma) q_t(\gamma)\big)\, d\gamma + \int_{\Gamma_t} \big(-\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma\big)\, d\Lambda(\gamma) \tag{13}$$

Let us now consider the concavity of the Lagrangian. Using (13), we may rewrite the Lagrangian as

$$\begin{aligned}\mathcal{L} = &\int_{\Gamma_t} \big(w\big(\gamma, q_t(\gamma)\big) - \kappa\big(-\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma\big)\big) f(\gamma)\, d\gamma - \int_{\Gamma_t} \Lambda(\gamma) q_t(\gamma)\, d\gamma\\ &+ \int_{\Gamma_t} \big(-\gamma q_t(\gamma) + b\big(q_t(\gamma)\big) - \sigma\big)\, d\big(\kappa F(\gamma) + \Lambda(\gamma)\big)\end{aligned}$$

From the definition of $\kappa$, $w(\gamma, q_t(\gamma)) - \kappa b(q_t(\gamma))$ is concave in $q_t(\gamma)$. We may thus conclude that the Lagrangian is concave in $q_t(\gamma)$ if

$$\kappa F(\gamma) + \Lambda(\gamma)$$

is nondecreasing for all $\gamma \in [\underline{\gamma}, \gamma_t]$. Using the constructed $\Lambda(\gamma)$ and referring to part (ii) of Proposition 1, we see that $\kappa F(\gamma) + \Lambda(\gamma)$ is nondecreasing for all $\gamma \in [\underline{\gamma}, \gamma_t]$ if the jumps

[36]For our analysis, only the difference between $\Lambda_1(\gamma)$ and $\Lambda_2(\gamma)$ matters, and so we need only show that there exists two nondecreasing functions, $\Lambda_1(\gamma)$ and $\Lambda_2(\gamma)$, whose difference delivers $\Lambda(\gamma)$.

[37]Observe that $h(\gamma) \equiv \int_{\gamma}^{\gamma_t} q_t(\widetilde{\gamma})\, d\widetilde{\gamma}$ exists (as $q_t$ is bounded and measurable by monotonicity) and is absolutely continuous. Observe as well that $\Lambda(\gamma) \equiv \Lambda_1(\gamma) - \Lambda_2(\gamma)$ is a function of bounded variation, as it is the difference between two nondecreasing and bounded functions. We may thus conclude that $\int_{\underline{\gamma}}^{\gamma_t} h(\gamma)\, d\Lambda(\gamma)$ exists (it is the Riemman–Stieltjes integral), and integration by parts can be done as follows: $\int_{\underline{\gamma}}^{\gamma_t} h(\gamma)\, d\Lambda(\gamma) = h(\overline{\gamma})\Lambda(\overline{\gamma}) - h(\underline{\gamma})\Lambda(\underline{\gamma}) - \int_{\underline{\gamma}}^{\gamma_t} \Lambda(\gamma)\, dh(\gamma)$. Given that $h(\gamma)$ is absolutely continuous, we can replace $dh(\gamma)$ with $-q_t(\gamma)\, d\gamma$.

in $\Lambda(\gamma)$ at $\underline{\gamma}$ and $\gamma_H(\gamma_t)$ are nonnegative. We verify these jumps are indeed nonnegative below.

We now show that the cap allocation $q_t^\star$ maximizes the Lagrangian. To this end, we use the sufficiency part of Lemma A.2 in Amador, Werning, and Angeletos (2006), which concerns the maximization of concave functionals on a convex cone. In our case, we need to extend the set $Q$ to be $[0, \infty)$, making our choice set $\Phi$ a convex cone. To do this, we follow Amador and Bagwell (2013) and extend $b$ and $w$ to the entire nonnegative ray of the real line. We can then apply Lemma A.2 to the extended Lagrangian with the choice set $\widehat{\Phi} \equiv \{q | q : \Gamma_t \to \Re_+, \text{ and } q \text{ nonincreasing}\}$.

Following the arguments in Amador and Bagwell (2013), we can then establish that the cap allocation $q_t^\star$ maximizes the Lagrangian if the Lagrangian is concave and the following first-order conditions hold:

$$\begin{aligned} \partial\mathcal{L}(q_t^\star; q_t^\star) &= 0 \\ \partial\mathcal{L}(q_t^\star; x) &\leq 0 \quad \text{for all } x \in \Phi, \end{aligned}$$

where $\partial\mathcal{L}(q_t^\star; x)$ is the Gateaux differential of the Lagrangian in (13) in the direction $x$.[38] Importantly, the Lagrangian in (13) is evaluated using our constructed multiplier functions.

Taking the Gateaux differential of the Lagrangian in (13) in direction $x \in \Phi$, we get[39]

$$\begin{aligned} \partial\mathcal{L}(q_t^\star; x) = &\int_{\Gamma_t} \big(w_q\big(\gamma, q_t^\star(\gamma)\big) f(\gamma) - \Lambda(\gamma)\big) x(\gamma)\, d\gamma \\ &+ \int_{\Gamma_t} \big(-\gamma + b'\big(q_t^\star(\gamma)\big)\big) x(\gamma)\, d\Lambda(\gamma). \end{aligned}$$

Using $b'(q_f(\gamma)) = \gamma$ and our knowledge of $\Lambda$ and $\Psi$, we get that

$$\begin{aligned} \partial\mathcal{L}(q_t^\star; x) = &\int_{\gamma_H(\gamma_t)}^{\gamma_t} \big(w_q\big(\gamma, q_i(\gamma_t)\big) f(\gamma) - A - \kappa\big(F(\gamma_t) - F(\gamma)\big) \\ &- \kappa\big(b'\big(q_i(\gamma_t)\big) - \gamma\big) f(\gamma)\big) x(\gamma)\, d\gamma \end{aligned}$$

Hence, integrating by parts, we get

$$\begin{aligned} \partial\mathcal{L}(q_t^\star; x) = \bigg[&\int_{\gamma_H(\gamma_t)}^{\gamma_t} \big(w_q\big(\gamma, q_i(\gamma_t)\big) f(\gamma) - A - \kappa\big(F(\gamma_t) - F(\gamma)\big) \\ &- \kappa\big(b'\big(q_i(\gamma_t)\big) - \gamma\big) f(\gamma)\big)\, d\gamma\bigg] x(\gamma_t) \end{aligned}$$

[38] Given a function $T : \Omega \to Y$, where $\Omega \subset X$ and $X$ and $Y$ are normed spaces, if for $x \in \Omega$ and $h \in X$ the limit

$$\lim_{\alpha \downarrow 0} \frac{1}{\alpha}\big[T(x + \alpha h) - T(x)\big]$$

exists, then it is called the Gateaux differential at $x$ with direction $h$ and is denoted by $\partial T(x; h)$.

[39] Existence of the Gateaux differential follows from Lemma A.1 in Amador, Werning, and Angeletos (2006). See Amador and Bagwell (2013) for further details concerning the application of this lemma.

$$
\begin{aligned}
&- \int_{\gamma_H(\gamma_t)}^{\gamma_t} \biggl[ \int_{\gamma_H(\gamma_t)}^{\gamma} \bigl( w_q\bigl(\tilde{\gamma}, q_i(\gamma_t)\bigr) f(\tilde{\gamma}) - A - \kappa\bigl(F(\gamma_t) - F(\tilde{\gamma})\bigr) \\
&- \kappa\bigl(b'\bigl(q_i(\gamma_t)\bigr) - \tilde{\gamma}\bigr) f(\tilde{\gamma})\bigr)\, d\tilde{\gamma} \biggr] dx(\gamma)
\end{aligned}
$$

Now, we use $\int_b^a \{(F(c) - F(x)) + (d - x) f(x)\}\, dx = (a - b)(F(c) - F(a)) + (d - b)(F(a) - F(b))$ to get that

$$
\begin{aligned}
\partial \mathcal{L}\bigl(q_t^{\star}; x\bigr) = {}& \biggl[ \int_{\gamma_H(\gamma_t)}^{\gamma_t} w_q\bigl(\gamma, q_i(\gamma_t)\bigr) f(\gamma)\, d\gamma - \bigl(\gamma_t - \gamma_H(\gamma_t)\bigr) A \\
&- \kappa\bigl(b'\bigl(q_i(\gamma_t)\bigr) - \gamma_H(\gamma_t)\bigr)\bigl(F(\gamma_t) - F\bigl(\gamma_H(\gamma_t)\bigr)\bigr) \biggr] x(\gamma_t) \\
&- \int_{\gamma_H(\gamma_t)}^{\gamma_t} \biggl[ \int_{\gamma_H(\gamma_t)}^{\gamma} w_q\bigl(\tilde{\gamma}, q_i(\gamma_t)\bigr) f(\tilde{\gamma})\, d\tilde{\gamma} - \bigl(\gamma - \gamma_H(\gamma_t)\bigr) A \\
&- \kappa\bigl(\bigl(\gamma - \gamma_H(\gamma_t)\bigr)\bigl(F(\gamma_t) - F(\gamma)\bigr) \\
&+ \bigl(b'\bigl(q_i(\gamma_t)\bigr) - \gamma_H(\gamma_t)\bigr)\bigl(F(\gamma) - F\bigl(\gamma_H(\gamma_t)\bigr)\bigr)\bigr) \biggr] dx(\gamma).
\end{aligned}
$$

Given that $(b'(q_i(\gamma_t)) - \gamma_H(\gamma_t))F(\gamma_H(\gamma_t)) = 0$, as $\gamma_H(\gamma_t) < \gamma_t$ and $b'(q_i(\gamma_t)) = \gamma_H(\gamma_t)$ if $\gamma_H(\gamma_t) \in (\underline{\gamma}, \overline{\gamma})$, the above becomes

$$
\begin{aligned}
\partial \mathcal{L}\bigl(q_t^{\star}; x\bigr) = {}& \biggl[ \int_{\gamma_H(\gamma_t)}^{\gamma_t} w_q\bigl(\gamma, q_i(\gamma_t)\bigr) f(\gamma)\, d\gamma - \bigl(\gamma_t - \gamma_H(\gamma_t)\bigr) A \\
&+ \kappa\bigl(\gamma_H(\gamma_t) - b'\bigl(q_i(\gamma_t)\bigr)\bigr) F(\gamma_t) \biggr] x(\gamma_t) \\
&- \int_{\gamma_H(\gamma_t)}^{\gamma_t} \biggl[ \int_{\gamma_H(\gamma_t)}^{\gamma} w_q\bigl(\tilde{\gamma}, q_i(\gamma_t)\bigr) f(\tilde{\gamma})\, d\tilde{\gamma} - \bigl(\gamma - \gamma_H(\gamma_t)\bigr) A \\
&- \kappa\bigl(\gamma - \gamma_H(\gamma_t)\bigr) F(\gamma_t) + \kappa\bigl(\gamma - b'\bigl(q_i(\gamma_t)\bigr)\bigr) F(\gamma) \biggr] dx(\gamma).
\end{aligned}
$$

Using the definition of $G$ in equation (5), we can rewrite the above as

$$
\begin{aligned}
\partial \mathcal{L}\bigl(q_t^{\star}; x\bigr) = {}& \bigl(G(\gamma_t|\gamma_t) - A\bigr)\bigl(\gamma_t - \gamma_H(\gamma_t)\bigr) x(\gamma_t) \\
&- \int_{\gamma_H(\gamma_t)}^{\gamma_t} \bigl(G(\gamma|\gamma_t) - A\bigr)\bigl(\gamma - \gamma_H(\gamma_t)\bigr)\, dx(\gamma).
\end{aligned}
$$

Using (5) and (12), we also observe that

$$
G(\gamma_t|\gamma_t) = A, \tag{14}
$$

and thus

$$
\partial \mathcal{L}\bigl(q_t^{\star}; x\bigr) = -\int_{\gamma_H(\gamma_t)}^{\gamma_t} \bigl(G(\gamma|\gamma_t) - A\bigr)\bigl(\gamma - \gamma_H(\gamma_t)\bigr)\, dx(\gamma). \tag{15}
$$

We are now ready to evaluate the first-order conditions.

Note that it follows immediately that $\partial\mathcal{L}(q_t^\star; q_t^\star) = 0$ as $q_t^\star$ is constant for $\gamma \in [\gamma_H(\gamma_t), \gamma_t]$.

If $G(\gamma|\gamma_t) \leq A = G(\gamma_t|\gamma_t)$ for all $\gamma \in [\gamma_H(\gamma_t), \gamma_t]$, then for any nonincreasing $x \in \Phi$, it follows that $\partial\mathcal{L}(q_t^\star; x) \leq 0$, which is provided by part (i) of Proposition 1.

Recall also that we require $A \geq 0$, since $\Psi(\gamma)$ must be nondecreasing. To see that this inequality holds, note that

$$A = \kappa\left[\frac{\gamma_H(\gamma_t) - b'\big(q_i(\gamma_t)\big)}{\gamma_t - \gamma_H(\gamma_t)}\right]F(\gamma_t) + \frac{1}{\gamma_t - \gamma_H(\gamma_t)}\int_{\gamma_H(\gamma_t)}^{\gamma_t} w_q\big(\tilde{\gamma}, q_i(\gamma_t)\big)f(\tilde{\gamma})\,d\tilde{\gamma}.$$

By the definition of $\gamma_H$, we have that $q_i(\gamma_t) \geq q_f(\gamma_H(\gamma_t))$. Note also that $b'(q_f(\gamma_H(\gamma_t))) = \gamma_H(\gamma_t)$, and concavity of $b$ implies that $b'(q_i(\gamma_t)) \leq b'(q_f(\gamma_H(\gamma_t))) = \gamma_H(\gamma_t)$. So, the first term in the previous equation is nonnegative. Finally, note that $w_q(\gamma, q) = P(q) - \gamma + P'(q)q + \frac{1}{\alpha}v'(q) = P(q) - \gamma - \frac{1-\alpha}{\alpha}qP'(q)$. Thus,

$$\begin{aligned} w_q\big(\gamma_t, q_i(\gamma_t)\big) &= P\big(q_i(\gamma_t)\big) - \gamma_t - \frac{1-\alpha}{\alpha}q_i(\gamma_t)P'\big(q_i(\gamma_t)\big) \\ &= \frac{\sigma}{q_i(\gamma_t)} - \frac{1-\alpha}{\alpha}q_i(\gamma_t)P'\big(q_i(\gamma_t)\big) \geq 0 \end{aligned}$$

where the last equality follows from $(P(q_i(\gamma_t)) - \gamma_t)q_i(\gamma_t) = \sigma$, by the definition of $q_i$. But we also have that

$$w_q(\gamma, q) > w_q\big(\gamma', q\big)$$

for all $\gamma < \gamma'$, and thus

$$w_q\big(\gamma, q_i(\gamma_t)\big) \geq w_q\big(\gamma_t, q_i(\gamma_t)\big) \geq 0$$

for $\gamma \leq \gamma_t$. Hence, we can also sign the integral term: $\int_{\gamma_H(\gamma_t)}^{\gamma_t} w_q(\tilde{\gamma}, q_i(\gamma_t))f(\tilde{\gamma})\,d\tilde{\gamma} \geq 0$. Taken together, the above implies that $A \geq 0$.

As discussed above, we now finish the argument that $\kappa F(\gamma) + \Lambda(\gamma)$ is nondecreasing for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$ by showing that the potential jumps in $\Lambda(\gamma)$ are nonnegative. There are two cases to consider. The first case is where $\gamma_H(\gamma_t) > \underline{\gamma}$. In this case, there are two jumps, one at $\underline{\gamma}$ and one at $\gamma_H(\gamma_t)$. For the jump at $\gamma_H(\gamma_t)$, we get

$$A + \kappa\big(F(\gamma_t) - F\big(\gamma_H(\gamma_t)\big)\big) - w_q\big(\gamma_H(\gamma_t), q_f\big(\gamma_H(\gamma_t)\big)f\big(\gamma_H(\gamma_t)\big) = G(\gamma_t|\gamma_t) - G\big(\gamma_H(\gamma_t)|\gamma_t\big)$$

where $G(\gamma_H(\gamma_t)|\gamma_t) = -\kappa[F(\gamma_t) - F(\gamma_H(\gamma_t))] + w_q(\gamma_H(\gamma_t), q_f(\gamma_H(\gamma_t))f(\gamma_H(\gamma_t))$. Part (i) of Proposition 1 guarantees that $G(\gamma_t|\gamma_t) \geq G(\gamma_H(\gamma_t)|\gamma_t)$, and thus the jump at $\gamma_H(\gamma_t)$ is nonnegative.

The jump in $\Lambda(\gamma)$ at $\underline{\gamma}$ is nonnegative, since $w_q(\underline{\gamma}, q_f(\underline{\gamma}))f(\underline{\gamma}) > 0$.

Finally, for the case where $\gamma_H(\gamma_t) = \underline{\gamma}$, there is only one jump, at $\underline{\gamma}$. The jump is

$$A + \kappa F(\gamma_t)$$

which is positive, given that we have shown that $A \geq 0$.

To complete the proof, we use Theorem 1 in Amador and Bagwell (2013). To apply this theorem, we set (i) $x_0 \equiv q_t^\star$; (ii) $X \equiv \{q_t | q_t : \Gamma_t \to Q\}$; (iii) $f$ to be given by the negative of the objective function, $\int_{\Gamma_t} w(\gamma, q_t(\gamma))\, dF(\gamma)$, as a function of $q_t \in X$; (iv) $Z \equiv \{(z_1, z_2, z_3) | z_1 : \Gamma_t \to \mathbb{R}, z_2 : \Gamma_t \to \mathbb{R}$ and $z_3 : \Gamma_t \to \mathbb{R}$ with $z_1, z_2, z_3$ integrable $\}$; (v) $\Omega \equiv \Phi$; (vi) $P \equiv \{(z_1, z_2, z_3) | (z_1, z_2, z_3) \in Z$ such that $z_1(\gamma) \geq 0$, $z_2(\gamma) \geq 0$ and $z_3(\gamma) \geq 0$ for all $\gamma \in \Gamma_t\}$; (vii) $\hat{G}$ (which is referred to as $G$ in Theorem 1) to be the mapping from $\Phi$ to $Z$ given by the left-hand sides of inequalities (9), (10), and (11); (viii) $T$ to be the linear mapping:

$$T\big((z_1, z_2, z_3)\big) \equiv \int_{\Gamma_t} z_1(\gamma)\, d\Lambda_1(\gamma) + \int_{\Gamma_t} z_2(\gamma)\, d\Lambda_2(\gamma) + \int_{\Gamma_t} z_3(\gamma)\, d\Psi(\gamma)$$

where $\Lambda_1$, $\Lambda_2$, and $\Psi$ being nondecreasing functions implies that $T(z) \geq 0$ for $z \in P$. We have that

$$\begin{aligned} T\big(\hat{G}(x_0)\big) \equiv & \int_{\Gamma_t} \left( \int_{\gamma}^{\gamma_t} q_t^\star(\tilde{\gamma})\, d\tilde{\gamma} + \overline{U} + \gamma q_t^\star(\gamma) - b\big(q_t^\star(\gamma)\big) + \sigma \right) d\big(\Lambda_1(\gamma) - \Lambda_2(\gamma)\big) \\ & - \int_{\Gamma_t} \big(-\gamma q_t^\star(\gamma) + b\big(q_t^\star(\gamma)\big) - \sigma\big)\, d\Psi(\gamma) = 0 \end{aligned}$$

where $\overline{U}$ is evaluated at the $q_t^\star$ allocation, and where the last equality follows from the $q_t^\star$ allocation and the proposed multipliers. We have found conditions under which the proposed allocation, $q_t^\star$, minimizes $f(x) + T(\hat{G}(x))$ for $x \in \Omega$. Given that $T(\hat{G}(x_0)) = 0$, then the conditions of Theorem 1 hold and it follows that $q_t^\star$ solves $\min_{x \in \Omega} f(x)$ subject to $-\hat{G}(x) \in P$, which is Problem $\mathrm{P}_t'$. □

### D.1 *Proof of Corollary 1*

Proof. Letting $q_i$ and $\gamma_H$ represent $q_i(\gamma_t)$ and $\gamma_H(\gamma_t)$, respectively, we start with the following manipulations:

$$\begin{aligned} G(\gamma|\gamma_t) &= -\kappa F(\gamma_t) + \kappa \frac{\gamma - b'(q_i)}{\gamma - \gamma_H} F(\gamma) + \frac{1}{\gamma - \gamma_H} \int_{\gamma_H}^{\gamma} \left( -\tilde{\gamma} + b'(q_i) + \frac{1}{\alpha} v'(q_i) \right) f(\tilde{\gamma})\, d\tilde{\gamma} \\ &= -\kappa F(\gamma_t) + \kappa \frac{\gamma}{\gamma - \gamma_H} F(\gamma) - \kappa \frac{b'(q_i)}{\gamma - \gamma_H} F(\gamma) \\ &\quad + \kappa \frac{b'(q_i)}{\gamma - \gamma_H} F(\gamma_H) - \kappa \frac{b'(q_i)}{\gamma - \gamma_H} F(\gamma_H) \\ &\quad + \frac{1}{\gamma - \gamma_H} \int_{\gamma_H}^{\gamma} \left( -\tilde{\gamma} + b'(q_i) + \frac{1}{\alpha} v'(q_i) \right) f(\tilde{\gamma})\, d\tilde{\gamma} \\ &= -\kappa F(\gamma_t) + \frac{\kappa}{\gamma - \gamma_H} \int_{\gamma_H}^{\gamma} \big(\tilde{\gamma} f(\tilde{\gamma}) + F(\tilde{\gamma})\big)\, d\tilde{\gamma} - \kappa \frac{b'(q_i)}{\gamma - \gamma_H} \big(F(\gamma) - F(\gamma_H)\big) \\ &\quad + \frac{1}{\gamma - \gamma_H} \int_{\gamma_H}^{\gamma} \left( -\tilde{\gamma} + b'(q_i) + \frac{1}{\alpha} v'(q_i) \right) f(\tilde{\gamma})\, d\tilde{\gamma} \\ &= -\kappa F(\gamma_t) + \frac{\kappa}{\gamma - \gamma_H} \int_{\gamma_H}^{\gamma} \big(\tilde{\gamma} f(\tilde{\gamma}) + F(\tilde{\gamma}) - b'(q_i) f(\tilde{\gamma})\big)\, d\tilde{\gamma} \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\gamma-\gamma_H}\int_{\gamma_H}^{\gamma}\left(-\tilde{\gamma}+b'(q_i)+\frac{1}{\alpha}v'(q_i)\right)f(\tilde{\gamma})\,d\tilde{\gamma}\\
&=-\kappa F(\gamma_t)+\frac{1}{\gamma-\gamma_H}\int_{\gamma_H}^{\gamma}\left[\kappa F(\tilde{\gamma})+\frac{1}{\alpha}v'(q_i)f(\tilde{\gamma})+(\kappa-1)\big(\tilde{\gamma}-b'(q_i)\big)f(\tilde{\gamma})\right]d\tilde{\gamma}\\
&=-\kappa F(\gamma_t)+\frac{1}{\gamma-\gamma_H}\int_{\gamma_H}^{\gamma}M_2(\tilde{\gamma})\,d\tilde{\gamma},
\end{aligned}
$$

where we use in the third equality above that $\frac{\gamma_H-b'(q_i)}{\gamma-\gamma_H}F(\gamma_H)=0$ and where we define

$$
M_2(\tilde{\gamma})\equiv\kappa F(\tilde{\gamma})+\frac{1}{\alpha}v'(q_i)f(\tilde{\gamma})+(\kappa-1)\big(\tilde{\gamma}-b'(q_i)\big)f(\tilde{\gamma}).
$$

Thus,

$$
(\gamma-\gamma_H)G(\gamma|\gamma_t)=-\kappa(\gamma-\gamma_H)F(\gamma_t)+\int_{\gamma_H}^{\gamma}M_2(\tilde{\gamma})\,d\tilde{\gamma}.
$$

Taking a derivative with respect to $\gamma$, for $\gamma>\gamma_H$, we obtain

$$
(\gamma-\gamma_H)G'(\gamma|\gamma_t)+G(\gamma|\gamma_t)=-\kappa F(\gamma_t)+M_2(\gamma),
$$

and thus

$$
(\gamma-\gamma_H)G'(\gamma|\gamma_t)=M_2(\gamma)-\frac{1}{\gamma-\gamma_H}\int_{\gamma_H}^{\gamma}M_2(\tilde{\gamma})\,d\tilde{\gamma}.
$$

It follows that, if $M_2'(\gamma)\geq 0$, then $G'(\gamma|\gamma_t)\geq 0$. Now note that

$$
\begin{aligned}
M_2'(\gamma)&=\kappa f(\gamma)+\frac{1}{\alpha}v'(q_i)f'(\gamma)+(\kappa-1)\big(\gamma-b'(q_i)\big)f'(\gamma)+(\kappa-1)f(\gamma)\\
&=(2\kappa-1)f(\gamma)+\kappa\big(\gamma-b'(q_i)\big)f'(\gamma)+\big(-\gamma+b'(q_i)+v'(q_i)/\alpha\big)f'(\gamma).
\end{aligned}
$$

Recall that

$$
\gamma-b'(q_i)\geq 0
$$

for $\gamma\geq\gamma_H$. In addition,

$$
\begin{aligned}
-\gamma+b'(q_i)+v'(q_i)/\alpha&=-\gamma+b'(q_i)+v'(q_i)+\left(\frac{1}{\alpha}-1\right)v'(q_i)\\
&=\big(P(q_i)-\gamma\big)+\left(\frac{1}{\alpha}-1\right)v'(q_i)\geq 0\quad\text{for }\gamma\geq\gamma_H
\end{aligned}
$$

where we use that $b'(q_i)+v'(q_i)=P(q_i)$ and where the inequality follows from $v'(q_i)>0$, $\alpha\in(0,1]$, and that $P(q_i)\geq\gamma$ for all types in $[\gamma_H,\gamma_t]$ (so that they can make profits and cover the fixed cost $\sigma\geq 0$). Hence,

$$
M_2'(\gamma)=(2\kappa-1)f(\gamma)+\kappa(\text{nonnegative term})f'(\gamma)+(\text{nonnegative term})f'(\gamma).
$$

Thus, $\kappa \geq 1/2$ and $f'(\gamma) \geq 0$ together are sufficient to guarantee that $M_2'(\gamma) \geq 0$, and thus that $G(\gamma|\gamma_t)$ is nondecreasing for any $\gamma_t$. Hence, part (i) of Proposition 1 then holds for all $\gamma_t \in (\underline{\gamma}, \overline{\gamma}]$.

Finally, note that

$$M_1'(\gamma) = \kappa f(\gamma) + \frac{1}{\alpha} v''\big(q_f(\gamma)\big) q_f'(\gamma) f(\gamma) + \frac{1}{\alpha} v'\big(q_f(\gamma)\big) f'(\gamma).$$

Using $q_f'(\gamma) = 1/b''(q_f(\gamma))$ and the definition of $\kappa$, we obtain that

$$M_1'(\gamma) \geq (2\kappa - 1) f(\gamma) + \frac{1}{\alpha} v'\big(q_f(\gamma)\big) f'(\gamma) \geq 0$$

where the second inequality follows from $\kappa \geq 1/2$ and $f$ nondecreasing. Thus, part (ii) of Proposition 1 also holds for all $\gamma_t \in (\gamma_t, \overline{\gamma}]$. We can thus use Proposition 2 to obtain the desired result. □

### D.2 *Proof of Lemma 4*

Proof. For parts (a) and (b), recall that $v'(q) = -P'(q)q$ and that $b'(q) = P(q) + qP'(q)$. Using equation (6), it follows that, for all $q \in (0, q_{\max}]$,

$$\begin{aligned} -v'(q) &= a_0\big(b'(q) + v'(q)\big) + b_0 \\ v'(q) &= -\frac{a_0}{1+a_0} b'(q) - \frac{b_0}{1+a_0} \end{aligned} \tag{16}$$

Integrating the above in $[q_0, q]$ where $q > q_0 > 0$, we have that

$$v(q) + \frac{a_0}{1+a_0} b(q) + \frac{b_0}{1+a_0} q = v(q_0) + \frac{a_0}{1+a_0} b(q_0) + \frac{b_0}{1+a_0} q_0$$

From Assumption 1, using the limit condition $\lim_{q_0 \downarrow 0} v(q_0) = \lim_{q_0 \downarrow 0} b(q_0) = 0$, we get part (a) for all $q \in Q$.

Differentiating (16), we get that $\frac{1}{\alpha} \frac{v''(q)}{b''(q)} = -\frac{1}{\alpha} \frac{a_0}{1+a_0}$, and thus part (b) follows.

To show part (c), note that

$$\begin{aligned} w_q\big(\gamma, q_f(\gamma)\big) &= \frac{1}{\alpha} v'\big(q_f(\gamma)\big) = -\frac{1}{\alpha} \frac{a_0}{1+a_0} b'\big(q_f(\gamma)\big) - \frac{1}{\alpha} \frac{b_0}{1+a_0} \\ &= \underbrace{-\frac{1}{\alpha} \frac{a_0}{1+a_0}}_{\kappa - 1} \big(\underbrace{b'\big(q_f(\gamma)\big)}_{\gamma} - b'(q_i)\big) - \frac{1}{\alpha} \underbrace{\left[\frac{a_0}{1+a_0} b'(q_i) - \frac{b_0}{1+a_0}\right]}_{-v'(q_i)} \\ &= (\kappa - 1)\big(\gamma - b'(q_i)\big) + \frac{1}{\alpha} v'(q_i) \end{aligned}$$

It follows then that

$$\begin{aligned} M_1(\gamma) &= \kappa F(\gamma) + w_q\big(\gamma, q_f(\gamma)\big) f(\gamma) \\ &= \kappa F(\gamma) + (\kappa - 1)\big(\gamma - b'(q_i)\big) f(\gamma) + \frac{1}{\alpha} v'(q_i) f(\gamma) = M_2(\gamma) \end{aligned}$$

which with $q_i = q_i(\gamma_t)$ delivers part (c). □

### D.3 *Proof of Corollary 2*

Proof. For this family, we already know that if part (ii) of Proposition 1 holds globally, then so does part (i). Taking a derivative of $M_1(\gamma)$ with respect to $\gamma$, and using that $w_q(\gamma, q_f(\gamma)) = \frac{1}{\alpha}v'(q_f(\gamma))$ together with $\frac{dv'(q_f(\gamma))}{d\gamma} = \frac{v''(q_f(\gamma))}{b''(q_f(\gamma))}$ delivers the result. □

### D.4 *Proof of Lemma 5*

Proof. By (3), $q_i(\gamma_t)$ satisfies $\gamma_t = P(q_i(\gamma_t)) - \frac{\sigma}{q_i(\gamma_t)}$. It follows that

$$q_i'(\gamma_t) = \frac{1}{P'\big(q_i(\gamma_t)\big) + \sigma/\big(q_i(\gamma_t)\big)^2}$$

Given that $q_i(\gamma_t) > q_f(\gamma_t)$, it follows that $\pi_q(\gamma_t, q_i(\gamma_t)) < 0$. Hence, $\pi_q(\gamma_t, q_i(\gamma_t)) = P'(q_i(\gamma_t))q_i(\gamma_t) + \pi(\gamma_t, q_i(\gamma_t))/q_i(\gamma_t) < 0$. Using that $\pi(\gamma_t, q_i(\gamma_t)) = \sigma$, we obtain the first result of the lemma.

For the second result, there are two cases to consider, one where $\gamma_H(\gamma_t) > \underline{\gamma}$ and the other where $\gamma_H(\gamma_t) = \underline{\gamma}$. For the latter case, the result is immediate. For the former case, we have that $\gamma_H(\gamma_t) = b'(q_f(\gamma_H(\gamma_t))) = b'(q_i(\gamma_t)) = P'(q_i(\gamma_t))q_i(\gamma_t) + P(q_i(\gamma_t))$. Thus,

$$\gamma_t - \gamma_H(\gamma_t) = -\left(\frac{\sigma}{q_i(\gamma_t)} + P'\big(q_i(\gamma_t)\big)q_i(\gamma_t)\right)$$

The first result of the lemma establishes that the bracketed expression is negative; thus, it follows that $\gamma_t > \gamma_H(\gamma_t)$. □

## Appendix E: Proof of Proposition 4

Proof. First, note that $q_i'(\gamma_t) < 0$ implies that $\gamma_H(\gamma_t)$ is strictly increasing in $\gamma_t$, as long as $\gamma_H(\gamma_t) > \underline{\gamma}$. That is, there exists a $\hat{\gamma} \in (\underline{\gamma}, \overline{\gamma}]$ such that $\gamma_H(\gamma_t) = \underline{\gamma}$ for all $\gamma_t \leq \hat{\gamma}$ and $\gamma_H(\gamma_t) > \underline{\gamma}$ for all $\gamma_t > \hat{\gamma}$. It is possible that $\hat{\gamma} = \overline{\gamma}$, and thus for any level of exclusion, all types are pooled.

Consider a situation where $\hat{\gamma} < \overline{\gamma}$. Then, for $\gamma_t \in (\hat{\gamma}, \overline{\gamma}]$, using the functional forms, $\alpha = 1$, the uniform distribution assumption, and that $b'(q_i(\gamma_t)) = \gamma_H(\gamma_t)$, we have that

$$\begin{aligned} W'(\gamma_t)/f_U &= v\big(q_i(\gamma_t)\big) - v'\big(q_i(\gamma_t)\big)q_i(\gamma_t) + \frac{q_i(\gamma_t)}{\gamma_t - \gamma_H(\gamma_t)} \int_{\gamma_H(\gamma_t)}^{\gamma_t} \big(\gamma - \gamma_H(\gamma_t)\big)\, d\gamma \\ &= -\beta q_i(\gamma_t)^2/2 - q_i(\gamma_t)\gamma_H(\gamma_t) + \frac{q_i(\gamma_t)}{2\big(\gamma_t - \gamma_H(\gamma_t)\big)}\big(\gamma_t^2 - \gamma_H(\gamma_t)^2\big) \\ &= \frac{q_i(\gamma_t)}{2}\big[-\beta q_i(\gamma_t) + \gamma_t - \gamma_H(\gamma_t)\big] \end{aligned}$$

where $f_U = \frac{1}{\overline{\gamma} - \underline{\gamma}}$ denotes the uniform density.

Using that $\gamma_t - \gamma_H(\gamma_t) = -\frac{\sigma}{q_i(\gamma_t)} + \beta q_i(\gamma_t)$, we obtain that $W'(\gamma_t)/f_U = -\frac{1}{2}\sigma < 0$. Thus, for all $\gamma_t \in (\hat{\gamma}, \overline{\gamma}]$, $W'(\gamma_t) < 0$, and thus, in an optimal allocation $\gamma_t \leq \hat{\gamma}$, guaranteeing that all types are pooled. This completes the proof of part (a).

For part (b), note that the conditions imply that $\hat{\gamma} < \overline{\gamma}$. To see this, note that if $\gamma_t = \overline{\gamma}$, then $q_i(\gamma_t) = q_i(\overline{\gamma}) < q_f(\underline{\gamma})$ under the conditions in part (b). There then exists $\gamma_0 > \underline{\gamma}$ such that $q_f(\gamma_0) = q_i(\gamma_t)$, from which it follows that $\gamma_H(\gamma_t) = \gamma_0 > \underline{\gamma}$, a contradiction to part (a). Given that $\gamma_t \leq \hat{\gamma} < \overline{\gamma}$, part (a) implies that some types will be excluded.

For the case where $\gamma_t < \hat{\gamma}$, then $\gamma_H(\gamma_t) = \underline{\gamma}$, and

$$W'(\gamma_t)/f_U = -\frac{1}{2}\left(\sigma - \frac{(\mu - \underline{\gamma})\big(\mu - \underline{\gamma} - 2\beta q_i(\gamma_t)\big)\big(q_i(\gamma_t)\big)^2}{\sigma - \beta\big(q_i(\gamma_t)\big)^2}\right).$$

Note that

$$\lim_{\gamma_t \to \underline{\gamma}} W'(\gamma_t)/f_U = -\frac{1}{2}\left(\sigma - \frac{(\mu - \underline{\gamma})\big(\mu - \underline{\gamma} - 2\beta q_i(\underline{\gamma})\big)\big(q_i(\underline{\gamma})\big)^2}{\sigma - \beta\big(q_i(\underline{\gamma})\big)^2}\right).$$

By definition of $q_i$, we have that $(\mu - \beta q_i(\underline{\gamma}) - \underline{\gamma})q_i(\underline{\gamma}) = \sigma$. Using this, we get

$$\big(\mu - \underline{\gamma} - 2\beta q_i(\underline{\gamma})\big)q_i(\underline{\gamma}) = \sigma - \beta\big(q_i(\underline{\gamma})\big)^2,$$

and thus

$$\begin{aligned}\lim_{\gamma_t \to \underline{\gamma}} W'(\gamma_t)/f_U &= -\frac{1}{2}\left(\sigma - \frac{(\mu - \underline{\gamma})\big(\mu - \underline{\gamma} - 2\beta q_i(\underline{\gamma})\big)\big(q_i(\underline{\gamma})\big)^2}{\sigma - \beta\big(q_i(\underline{\gamma})\big)^2}\right) = -\frac{1}{2}\big(\sigma - (\mu - \underline{\gamma})q_i(\underline{\gamma})\big) \\ &= -\frac{1}{2}\big((\mu - \beta q_i(\underline{\gamma}) - \underline{\gamma})q_i(\underline{\gamma}) - (\mu - \underline{\gamma})q_i(\underline{\gamma})\big) = \frac{1}{2}\beta q_i(\underline{\gamma})^2 > 0.\end{aligned}$$

We have already shown that $W'(\gamma) < 0$ for $\gamma \in (\hat{\gamma}, \overline{\gamma}]$. Together with the above limiting result, and that $W'$ is continuous at $\hat{\gamma}$ (see footnote 31), it follows that the optimal $\gamma_t$ is interior in $[\underline{\gamma}, \hat{\gamma}]$, and thus satisfies $W'(\gamma_t) = 0$. □

## References

Alonso, Ricardo and Niko Matouschek (2008), "Optimal delegation." *The Review of Economic Studies*, 75, 259–293. http://www.jstor.org/stable/4626195. [1723, 1724]

Amador, Manuel and Kyle Bagwell (2012), "Tariff revenue and tariff caps." *American Economic Review, Papers and Proceedings*, 102, 459–465. [1723]

Amador, Manuel and Kyle Bagwell (2013), "The theory of optimal delegation with an application to tariff caps." *Econometrica*, 81, 1541–1599. [1721, 1723, 1724, 1726, 1728, 1729, 1734, 1735, 1737, 1739, 1742, 1745, 1748, 1751, 1753]

Amador, Manuel and Kyle Bagwell (2020), "Money burning in the theory of delegation." *Games and Economic Behavior*, 121, 382–412. https://www.sciencedirect.com/science/article/pii/S089982562030035X. [1723, 1724, 1745]

Amador, Manuel, Kyle Bagwell, and Alex Frankel (2018), "A note on interval delegation." *Economic Theory Bulletin*, 6, 239–249. [1723, 1737]

Amador, Manuel, Iván Werning, and George-Marios Angeletos (2006), "Commitment vs. flexibility." *Econometrica*, 74, 365–396. http://www.jstor.org/stable/3598805. [1723, 1751]

Ambrus, Attila and Georgy Egorov (2017), "Delegation and nonmonetary incentives." *Journal of Economic Theory*, 171, 101–135. [1723, 1724, 1745]

Armstrong, Mark (1996), "Multiproduct nonlinear pricing." *Econometrica*, 64, 51–75. [1723]

Armstrong, Mark and David E. M. Sappington (2007), "Recent developments in the theory of regulation." In *Handbook of Industrial Organization*, volume 3 (Mark Armstrong and Robert Porter, eds.), 1557–1700, North-Holland, Amsterdam. [1719, 1720]

Armstrong, Mark and John Vickers (2010), "A model of delegated project choice." *Econometrica*, 78, 213–244. http://ideas.repec.org/a/ecm/emetrp/v78y2010i1p213-244.html. [1723, 1745]

Athey, Susan, Andrew Atkeson, and Patrick J. Kehoe (2005), "The optimal degree of discretion in monetary policy." *Econometrica*, 73, 1431–1475. http://www.jstor.org/stable/3598880. [1723]

Athey, Susan, Kyle Bagwell, and Chris Sanchirico (2004), "Collusion and price rigidity." *The Review of Economic Studies*, 71, 317–349. http://www.jstor.org/stable/3700628. [1723]

Baron, David P. (1989), "Design of regulatory mechanisms and institutions." In *Handbook of Industrial Organization*, volume 2 (Richard Schmalensee and Robert D. Willig, eds.), 1347–1447, North-Holland, Amsterdam. [1720]

Baron, David P. and Roger B. Myerson (1982), "Regulating a monopolist with unknown costs." *Econometrica*, 50, 911–930. http://www.jstor.org/stable/1912769. [1719, 1723, 1724, 1745]

Burkett, Justin (2016), "Optimally constraining a bidder using a simple budget." *Theoretical Economics*, 11, 133–155. [1723]

Church, Jeffrey and Roger Ware (2000), *Industrial Organization: A Strategic Approach*. McGraw-Hill. [1720]

Frankel, Alexander (2014), "Aligned delegation." *American Economic Review*, 104, 66–83. [1723, 1745]

Frankel, Alexander (2016), "Delegating multiple decisions." *American Economic Journal: Microeconomics*, 8, 16–53. [1723, 1745]

Guo, Yingni (2016), "Dynamic delegation of experimentation." *American Economic Review*, 106, 1969–2008. [1723]

Halac, Marina and Pierre Yared (2019), "Fiscal rules and discretion under limited enforcement." Nber working paper 25463. [1723]

Holmstrom, Bengt (1977), *On Incentives and Control in Organizations*. Ph.D. thesis. Stanford University. [1723]

Joskow, Paul L. and Richard Schmalensee (1986), "Incentive regulation for electric utilities." *Yale Journal of Regulation*, 4, 1–49. [1720]

Kartik, Navin, Andreas Kleiner, and Richard Van Weelden (2021), *Delegation in Veto Bargaining*. Technical report, working paper. [1724, 1725]

Koessler, Frederic and David Martimort (2012), "Optimal delegation with multidimensional decisions." *Journal of Economic Theory*, 147, 1850–1881. [1723, 1745]

Kolotilin, Anton and Andriy Zapechelnyuk (2019), "Persuasion meets delegation." Working paper. [1724, 1739, 1741, 1744]

Laffont, Jean-Jacques and Jean Tirole (1986), "Using cost observation to regulate firms." *Journal of Political Economy*, 94, 614–641. [1720]

Laffont, Jean-Jacques and Jean Tirole (1993), *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA. [1720]

Martimort, David and Aggey Semenov (2006), "Continuity in mechanism design without transfers." *Economics Letters*, 93, 182–189. http://www.sciencedirect.com/science/article/B6V84-4KV2RBN-D/2/6204fb6e11cfd78415823100ad0aaf18. [1723]

Melumad, Nahum D. and Toshiyuki Shibano (1991), "Communication in settings with no transfers." *The RAND Journal of Economics*, 22, 173–198. http://www.jstor.org/stable/2601016. [1723, 1737]

Milgrom, Paul and Ilya Segal (2002), "Envelope theorems for arbitrary choice sets." *Econometrica*, 70, 583–601. http://ideas.repec.org/a/ecm/emetrp/v70y2002i2p583-601.html. [1748]

Mylovanov, Tymofiy (2008), "Veto-based delegation." *Journal of Economic Theory*, 138, 297–307. http://ideas.repec.org/a/eee/jetheo/v138y2008i1p297-307.html. [1723]

Saran, Rene (2021), *Monopolistic Screening With Single-Peaked Preferences*. Technical report, working paper. [1725]

Schmalensee, Richard (1989), "Good regulatory regimes." *The RAND Journal of Economics*, 20, 417–436. [1720]

Zapechelnyuk, Andriy (2020), "Optimal quality certification." *American Economic Review: Insights*, 2, 161–176. https://www.aeaweb.org/articles?id=10.1257/aeri.20190387. [1725]

---

Co-editor Marina Halac handled this manuscript.

Manuscript received 13 January, 2021; final version accepted 6 October, 2021; available online 19 October, 2021.