CAHIER DE RECHERCHE #2502E
Département de science économique
Faculté des sciences sociales
Université d'Ottawa

WORKING PAPER #2502E
Department of Economics
Faculty of Social Sciences
University of Ottawa

Optimal Contracts under General Mixed Constraints: Continuity, Structure, and Applications

Aggey Simons (Semenov)*

April 2025

* Department of Economics, University of Ottawa, 9039-120 University Private, Ottawa, Ontario, Canada, K1N 6N5; e-mail: aggey.simons@uottawa.ca.

Abstract

This paper characterizes optimal contract structures under adverse selection when the principal faces a general class of mixed (involving allocation and transfer) constraints. We establish conditions for the existence and the continuity of the optimal allocation. We show that under regularity conditions, the optimal continuous contract features at most three distinct regions: segments where the constraint is slack and the allocation follows a modified Baron-Myerson path, alternating with segments where the constraint binds. Assuming non-generic cases are excluded, the binding constraint forces a constant allocation (bunching) over a range of agent types. Our analysis demonstrates how bunching can arise endogenously from optimal design under smooth constraints, distinct from exogenously induced behavioural responses documented empirically.

Key words: *Adverse Selection, Optimal Contracts, Mixed Constraints, Endogenous Bunching, Continuity, Allocation Dynamics.*

JEL Classification: C61, D82, D86, H21, L51.

1 Introduction

Optimal contract design under asymmetric information is a cornerstone of modern economic theory (Laffont and Martimort, 2002). Seminal works like Mirrlees (1971) and Mussa and Rosen (1978) laid the foundation for understanding how principals can design mechanisms to elicit private information efficiently. While standard models often yield continuous and monotonic allocation rules, practical contracting and policy environments frequently impose additional constraints which may lead to discontinuities or pooling. These constraints go beyond the canonical incentive compatibility and individual rationality conditions, potentially introducing distortions, inefficiencies, or systematic deviations in allocations across different agent types. Understanding how optimal mechanisms behave under these restrictions is essential for advancing theory and for making policy useful in practice.

In this paper, using an optimal control approach, we analyse a canonical adverse selection model (Laffont and Martimort, 2002) subject to a general *mixed*, i.e., involving state (t or U) and control q variables, constraint:

$$a_1 W(q(\theta)) - a_2 t(\theta) + a_3 \geq 0, \tag{1}$$

where q is the allocation, t is the transfer, θ is the agent's private type, $W(\cdot)$ captures quality, standards, verifiable output, or other non-utility aspects of allocation (e.g., quality, environmental measure, or potentially the principal's benefit $S(q)$ as a special case), and a_1, a_2, a_3 parametrize the constraint. The principal's objective is to maximize expected $S(q) - t$. Our goal is to characterize when such constraints alter the optimal allocation and to derive structural properties of the resulting contract.

The main results in the paper are:

1. **Existence and Continuity:** We prove existence of a solution and continuity of the optimal allocation under a generalized monotonic hazard condition.
2. **Structure and Bunching:** We show that under genericity or curvature assumptions, the optimal allocation partitions the type space into at most three regions, with endogenous bunching on binding intervals.
3. **Application Scope:** Our framework nests regulatory, enforcement, and budget constraints as special cases, unifying existing models under a common structure.

A well-documented pattern in applied microeconomics is "bunching"—where individuals cluster their choices (like reported income or consumption levels) around specific points in policy schedules (see Kleven, 2016, for a comprehensive survey). This empirical literature often links this behavior to sharp policy discontinuities, such as kinks (Saez, 2010; Chetty et al., 2011) or

notches (Kleven and Waseem, 2013), in tax systems, welfare programs, or regulations. While these studies have been extremely useful for measuring behavioural responses, they mostly focus on exogenously imposed policy features. This leaves an open question: how might pooling or other systematic patterns emerge naturally from optimal contract design, especially when constraints are smoother or take different forms?

Many realistic constraints faced by principals do not create sharp kinks or notches in the agent's choice set but rather serve as limitations on the implementable contract itself. Examples include budget caps on total spending or transfers (Thomas, 2002), regulatory requirements setting minimum standards for quality, environmental impact, or service (Kühn and Siciliani (2009) and Laffont and Tirole (1993)), limited collateral or imperfect enforcement restricting the promises a principal can credibly make (Martimort et al., 2017), or participation constraints ensuring the principal (e.g., a government agency or firm) achieves a minimum net payoff (Laffont and Martimort, 2002). These scenarios are often captured by smooth constraints involving the allocation, transfers, and potentially functions related to the allocation (like quality or environmental impact) that might differ from the principal's direct surplus. This raises the following questions: how do such constraints influence optimal contracts, and can they generate specific allocation patterns like discontinuities or bunching?

Standard mechanism design often relies on regularity properties like continuity. Acemoglu (2009) states in the case without constraints that, "The important and stringent assumption here is that ... is a solution that never hits boundaries and does not involve discontinuities. Even though this feature is true of optimal controls in most economic applications, in purely mathematical terms it is a strong assumption." Continuity requires formal verification, especially when constraints might induce jumps or complex allocation patterns. Discontinuities, if they arise, can be problematic, potentially leading to fragility, renegotiation (Laffont and Tirole, 1993), or significant welfare losses (Stiglitz and Weiss, 1981). This paper demonstrates, however, that even when smooth constraints bind, the optimal contract often maintains continuity.[1]

The analysis highlights how smooth constraints interact with incentive provision. Deviations from the standard efficiency-rent extraction trade-off (based on $S(q)$) occur optimally precisely when that trade-off would lead to a violation of the constraint involving $W(q)$ and t. Our results offer a potential bridge between the theoretical design of optimal mechanisms under realistic constraints and the empirical observation of pooled or systematically structured allocations, providing conditions under which constant bunching arises.

The paper proceeds as follows: Section 2 lays out the model. Section 3 derives the main theoretical results using optimal control. Section 4 discusses applications and the connection to

[1]This result is consistent with the intuition as presented in (Acemoglu, 2009):"Fortunately, in most economic problems there will be enough structure to make optimal solutions continuous functions."

empirical observations. Section 5 concludes. Proofs are in the Appendix.

Literature Review

This paper builds on several strands of the mechanism design literature and connects to recent empirical findings.

First, it relates to the foundational analysis of adverse selection models, pioneered by Mussa and Rosen (1978) and Maskin and Riley (1984) for non-linear pricing, and Mirrlees (1971) for optimal taxation. These works established the standard techniques, often involving reducing the problem to finding an optimal allocation rule subject to implementability constraints. A common assumption, often implicit, is the regularity condition on the distribution of types (monotonic hazard rate) which ensures the monotonicity of the unconstrained solution (Laffont and Martimort, 2002). Our work explicitly incorporates an additional mixed constraint involving a general function $W(q)$ and focuses on establishing continuity and characterizing the structure when this constraint binds, requiring generalized regularity conditions (Assumption 1) and conditions ensuring bunching (Assumptions 2 or 3).

Second, our paper contributes to the literature specifically analyzing optimal contracts under various forms of constraints.

- *Mixed constraints and allocation patterns:* Thomas (2002) studies nonlinear pricing when buyers face budget constraints ($t \leq t_{max}$), corresponding to $a_1 = 0, a_2 > 0$ in our setup. He finds pooling (bunching) at the top. Our Proposition 1 (Case 2b) confirms this structure under conditions ensuring bunching. Martimort et al. (2017) investigate optimal stationary contracts in infinitely repeated principal-agent relationships with limited enforcement, leading to a constraint $\delta S(q) - t \geq$ const (a special case of (1), where $W = S$, $a_1 = \delta$, $a_2 = 1$). They find bunching among efficient types. Our Proposition 1 (Case 2a) aligns with this, and our Assumptions provide conditions for the continuity and bunching they observe.

  Our work also relates to the burgeoning empirical literature that uses bunching methods to estimate behavioural elasticities, as surveyed by Kleven (2016). This empirical literature typically identifies bunching as a behavioural response by agents massing at exogenously given discontinuities, such as kinks (Saez, 2010; Chetty et al., 2011) or notches (Kleven and Waseem, 2013). In contrast, the bunching analyzed in our paper arises endogenously as part of the optimally designed contract by the principal facing informational asymmetries and a smooth mixed constraint.[2]

- *Participation constraints:* Jullien (2000) examines models where the agent's reservation

[2] Kleven (2016) pointed out that: "The key assumptions that make notches suboptimal is that the underlying fundamentals are continuous, that agents are sophisticated optimizers, and that the available policy instruments are sufficiently flexible. If we break any of those assumptions, notches may become optimal".

utility depends on type ($U(\theta) \geq U_0(\theta)$). This also leads to potential bunching. Martimort and Stole (2022) provide a comprehensive analysis of principal-agent problems with general state constraints, including type-dependent participation constraints, using tools from non-smooth optimal control. Our analysis uses similar optimal control techniques but focuses on the constraint (1) and derives a sharp characterization of the optimal contract's structure (alternating regions, conditions for bunching).

- *Failures of standard conditions:* Challenges to monotonicity and continuity can also arise from failures of standard assumptions, like the Spence-Mirrlees single-crossing condition (SC). Araujo and Moreira (2010) show that violating SC can lead to discontinuities or bunching. However, Schottmüller (2015) demonstrates that continuity can persist even without SC under certain conditions. Our work maintains SC but focuses on the impact of the explicit constraint (1).

Third, methodologically, we employ optimal control theory with state and control variables subject to a mixed constraint. The use of optimal control in dynamic economic problems is standard (e.g., Acemoglu (2009) in growth theory), but its application to static mechanism design problems, particularly with constraints, is less common, though growing (e.g., (Martimort and Stole, 2022) and Jullien (2000)). Our problem features a Hamiltonian linear in the state variable $U(\theta)$, requiring techniques drawing on results like Seierstad and Sydsæter (1987) to handle the constraint and establish continuity.

2 Set up

We consider a contractual relationship between a buyer (the principal) and a seller (the agent), where the seller provides a good or service q in exchange for a payment t. The buyer's and seller's utilities are given by:

$$V(q,t) = S(q) - t \quad \text{and} \quad U(q,t,\theta) = t - \theta q,$$

respectively, where $S(\cdot)$ is the buyer's gross surplus function from the allocation q, and θ is the agent's marginal cost. The function S is twice differentiable, increasing, strictly concave, and satisfies $S(0) = 0$. We also assume standard Inada conditions: $\lim_{q\to 0} S'(q) = \infty$, and $\lim_{q\to\infty} S'(q) = 0$.

The seller has private information about the cost parameter $\theta \in \Theta = \left[\underline{\theta}, \overline{\theta}\right]$. The buyer knows that θ follows a cumulative distribution function $F(\cdot)$, with a non-atomic, differentiable density function $f(\cdot)$.

By the Revelation Principle, any mechanism can be represented as a direct, truthful mechanism $(q(\theta), t(\theta))_{\theta\in\Theta}$. The allocation $q(\theta)$ is implementable if there exists a payment function $t(\theta)$ such

that the contract $(q(\theta), t(\theta))$ is incentive compatible (IC):

$$\theta \in \arg\max_{\theta'} U\left(q(\theta'), t(\theta'), \theta\right), \quad \text{for all } \theta.$$

We normalize the agent's reservation utility to zero (Individual Rationality, IR):

$$t(\theta) - \theta q(\theta) \geq 0, \quad \text{for all } \theta. \tag{2}$$

The Constraint: We assume that the allocation vector $(q(\cdot), t(\cdot))$ satisfies the constraint:

$$a_1 W(q(\theta)) - a_2 t(\theta) + a_3 \geq 0, \quad \text{for all } \theta, \tag{3}$$

where $a_i \geq 0$ for $i = 1, 2$, and $a_3 \in \mathbb{R}$. The function $W(q)$ represents the aspect of the allocation relevant for the constraint. We assume $W(q)$ is twice differentiable, increasing, concave[3] and satisfies Inada conditions. This general formulation allows us to model quality standards, budget caps, or enforcement limits depending on the application.

Remark 1 (Examples of Constraints). The constraint (3) encompasses several economically relevant cases:

1. If $a_1 = a_2 = 0, a_3 \geq 0$, the constraint is trivial. The solution is the standard Baron-Myerson (BM) allocation, where $S'(q^{BM}(\theta)) = \theta + F(\theta)/f(\theta)$.
2. Cap on Transfers: If $a_1 = 0, a_2 > 0$, the constraint becomes $t(\theta) \leq a_3/a_2 = \widehat{t}$, as in Thomas (2002).
3. Quality/Standard Floor: If $a_1 > 0$ and $a_2 = 0$, the constraint imposes a threshold on $W(q)$, $W(q(\theta)) \geq -a_3/a_1$. If W is increasing, this implies $q(\theta) \geq q_{\min}$ for some $q_{\min}$. $W(q)$ represents measured quality or environmental standard compliance (Kühn and Siciliani, 2009).
4. Enforcement Constraint: If $a_1 = \delta > 0, a_2 = 1$, we have $\delta W(q(\theta)) - t(\theta) + a_3 \geq 0$. If $W(q)$ represents the verifiable part of the output or value used for enforcement, this generalizes the constraint in Martimort et al. (2017) (which used $W = S$).
5. Principal's Participation/Budget Constraint: If $a_1 = a_2 = 1$, the constraint is $S(q(\theta)) - t(\theta) \geq -a_3 = \underline{V}$ (setting $W = S$ as in Laffont and Martimort (2002)). If $W(q)$ represents some other measure relevant for the principal's overall budget or payoff (e.g., a different valuation), then the constraint takes the general form.
6. Regulatory Constraint on Externality: If $W(q)$ represents (negative of) an externality (e.g.,

[3]We primarily assume W is strictly concave ($W'' < 0$) for the applicability of certain analytical conditions (Assumptions 2, 3). We discuss the implications of relaxing this to just concavity in Section 4.5.

pollution), $a_1 > 0, a_2 = 0$, then the constraint imposes a cap on the externality. If $a_1 > 0, a_2 > 0$, it is related the allowed externality to the transfer paid with specific weights.

With the change of variables, $U(\theta) = t(\theta) - \theta q(\theta)$, we characterize feasible allocations $(q(\theta), U(\theta))$.[4]

Lemma 1 (IC/IR Characterization)**.** *An allocation rule $q(\theta)$ is implementable if and only if $q(\theta)$ is weakly decreasing. For a given implementable $q(\theta)$, the corresponding rent profile $U(\theta)$ satisfies:*

$$\dot{U}(\theta) = -q(\theta), \quad a.e. \tag{4}$$

The individual rationality constraint (2) is satisfied if and only if

$$U(\overline{\theta}) \geq 0. \tag{5}$$

Furthermore, $U(\theta)$ is absolutely continuous and convex.

We rewrite the constraint (3) using $t(\theta) = U(\theta) + \theta q(\theta)$:

$$h(\theta, q(\theta), U(\theta)) = a_1 W(q(\theta)) - a_2\big(\theta q(\theta) + U(\theta)\big) + a_3 \geq 0, \quad \forall \theta \in \Theta. \tag{6}$$

Define the set of admissible allocations $(U(\theta), q(\theta))$ for the optimization problem $(\mathscr{P})$ such that: $q(\theta) \geq 0$ is weakly decreasing; $U(\theta)$ is absolutely continuous; $(U(\theta), q(\theta))$ satisfies (4), (5), and (6).

The principal's problem $(\mathscr{P})$ is:

$$(\mathscr{P}): \quad \max_{(U,q)\text{is admissible}} \int_{\underline{\theta}}^{\overline{\theta}} \underbrace{\big(S(q(\theta)) - (\theta q(\theta) + U(\theta))\big)}_{=S(q(\theta))-t(\theta)} dF(\theta).$$

To solve $(\mathscr{P})$, we follow a standard relaxation approach. We drop the monotonicity requirement on $q(\theta)$ and solve the relaxed program $(\mathscr{P}^*)$, which maximizes the same objective subject to (4), (5), and the mixed constraint (6). If the solution $q^*(\theta)$ to $(\mathscr{P}^*)$ happens to be weakly decreasing, it is also a solution to the original problem $(\mathscr{P})$.[5]

For these purposes we define the set $\mathscr{F}$ of feasible allocations:

Definition 1 (Feasible Allocations)**.** An allocation $(U(\theta), q(\theta))$ is feasible if:

- $U(\theta)$ is absolutely continuous;

[4] See Rochet (1985) or Laffont and Martimort (2002).

[5] If the solution to $(\mathscr{P}^*)$ is not monotone (e.g., because Assumption 1 fails), one may apply the "ironing" procedure (Guesnerie and Laffont, 1984). Analysing ironing under the mixed constraint (6) is beyond the scope of this paper's primary results.

- $(U(\theta), q(\theta))$ satisfies (4), (5) (typically $U(\overline{\theta}) = 0$ at optimum), and (6).

We assume for simplicity that the set $\mathscr{F}$ is non-empty.[6]

The principal's relaxed problem becomes:

$$\max_{(U,q)\in\mathscr{F}} \int_{\underline{\theta}}^{\overline{\theta}} [S(q(\theta)) - \theta q(\theta) - U(\theta)]\, f(\theta) d\theta.$$

We first characterize existence and continuity in this relaxed setup before returning to the full incentive-compatible problem.

3 Results

We establish existence using Filippov–Cesari (Seierstad and Sydsæter (1987, Theorem 2, p. 285)).

Lemma 2 (Existence)**.** *Under the stated continuity and concavity assumptions on $S(\cdot)$ and $W(\cdot)$, and boundedness of Θ, a solution $(U^*(\theta), q^*(\theta))$ to the relaxed problem $(\mathscr{P}^*)$ exists within the space of absolutely continuous states U and piecewise continuous controls q.*

Proof. See Appendix A.1. □

The Hamiltonian for problem $(\mathscr{P}^*)$ is:

$$H(\theta, U, q, \lambda) = (S(q) - \theta q - U) f(\theta) - \lambda q,$$

where λ is the co-state variable associated with $\dot{U}(\theta) = -q(\theta)$. We treat θ as time in a static-to-dynamic transformation.

The Lagrangian associated with the Hamiltonian and the mixed constraint (6) is:

$$L(\theta, U, q, \lambda, \mu) = H(\theta, U, q, \lambda) + \mu(\theta)\left(a_1 W(q) - a_2(\theta q + U) + a_3\right),$$

where $\mu(\theta) \geq 0$ is the Lagrange multiplier for constraint (6).

Since the objective is concave in (U, q) (as S is concave and it's linear in U) and the constraint function $h(\theta, q, U)$ is concave in (U, q) (as W is concave and it's linear in U, q), the necessary conditions from Pontryagin's Maximum Principle are also sufficient for optimality (Mangasarian sufficiency, Seierstad and Sydsæter (1987, Theorem 5, p. 287)).[7]

[6] Non-emptiness of $\mathscr{F}$ requires that the constraint (6) is not overly restrictive. For example, if $a_3 \geq 0$ and $W(0) \geq 0$, the allocation $(q(\theta), U(\theta)) = (0, 0)$ is feasible. If $a_3 < 0$, feasibility requires the existence of an allocation satisfying all conditions. A sufficient condition is the existence of a constant allocation $\bar{q} \geq 0$ such that $a_1 W(\bar{q}) - a_2 \overline{\theta} \bar{q} + a_3 \geq 0$, which holds if the maximum value of the concave function $a_1 W(q) - a_2 \overline{\theta} q$ is at least $-a_3$.

[7] Since $H(U, q, \lambda(\theta), \theta)$ is concave in (U, q) and $h(U, q, \theta)$ is quasi-concave in (U, q), no constraint qualification is required.

An admissible pair $(U(\theta), q(\theta))$ for $(\mathscr{P}^*)$ is optimal if and only if there exist a piecewise continuous multiplier function $\mu(\theta) \geq 0$ and an absolutely continuous co-state function $\lambda(\theta)$ such that for almost all $\theta \in \Theta$:

1. Maximality condition: $q(\theta)$ maximizes $L(\theta, U(\theta), q, \lambda(\theta), \mu(\theta))$. The first-order condition (FOC) is:

$$\frac{\partial L}{\partial q} = f(\theta)\left(S'(q(\theta)) - \theta\right) - \lambda(\theta) + \mu(\theta)\left(a_1 W'(q(\theta)) - a_2\theta\right) = 0. \tag{7}$$

2. *Co-state equation:*

$$\dot{\lambda}(\theta) = -\frac{\partial L}{\partial U} = f(\theta) + a_2\mu(\theta). \tag{8}$$

3. *Boundary conditions:*

$$\lambda(\underline{\theta}) = 0, \qquad U(\overline{\theta}) = 0. \tag{9}$$

4. *Complementary slackness:*

$$\mu(\theta) \geq 0, \quad h(\theta, q(\theta), U(\theta)) \geq 0, \quad \mu(\theta)h(\theta, q(\theta), U(\theta)) = 0. \tag{10}$$

Integrating the co-state equation (8) using $\lambda(\underline{\theta}) = 0$, we get:

$$\lambda(\theta) = F(\theta) + a_2 \int_{\underline{\theta}}^{\theta} \mu(\xi)\, d\xi = F(\theta) + a_2\Psi(\theta), \tag{11}$$

where $\Psi(\theta) = \int_{\underline{\theta}}^{\theta} \mu(\xi)\, d\xi \geq 0$ is continuous and non-decreasing.

Substituting $\lambda(\theta)$ into the rearranged FOC (7), we get the main equation defining the optimal allocation $q(\theta)$:

$$f(\theta)S'(q(\theta)) + a_1\mu(\theta)W'(q(\theta)) = (f(\theta) + a_2\mu(\theta))\,\theta + F(\theta) + a_2\Psi(\theta). \tag{12}$$

Definition 2 (Partition based on constraint binding)**.** Let the optimal multiplier be $\mu(\theta)$. We partition Θ into regions where $\mu(\theta) > 0$ (constraint binds) and $\mu(\theta) = 0$ (constraint slack). An n-partition is the minimal set of adjacent intervals $\{\Theta_1, \ldots, \Theta_n\}$ such that $\Theta = \bigcup \Theta_i$, and $\mu(\theta)$ is either strictly positive or identically zero on the interior of each Θ_i. Let boundaries be $\theta_0 = \underline{\theta}, \ldots, \theta_n = \overline{\theta}$.

The requirement of minimality of the set $\{\Theta_1, \ldots, \Theta_n\}$ means that the multiplier μ alternates

between adjacent intervals. On any interval Θ_k where $\mu(\theta) = 0$, $\Psi(\theta)$ is constant, say Ψ_k. Equation (12) simplifies to the generalized Baron-Myerson condition:

$$S'(q(\theta)) = \underbrace{\theta + \frac{F(\theta) + a_2\Psi_k}{f(\theta)}}_{=:V(\theta, a_2\Psi_k)}. \tag{13}$$

The allocation follows a path based on the principal's objective function $S(q)$, adjusted by the accumulated shadow cost $a_2\Psi_k$. We define $V(\theta, a_2\Psi_k)$ as the (*modified*) virtual cost.

To ensure the solution $q(\theta)$ derived from (13) is non-increasing on slack intervals, we require:

Assumption 1 (Generalized Monotone Hazard Rate Condition). The virtual cost function $V(\theta, a) = \theta + \frac{F(\theta)+a}{f(\theta)}$ is non-decreasing in θ for any constant $a \geq 0$.

Assumption 1 necessitates $f'(\theta) \leq 0$.[8] This assumption ensures implementability on slack intervals. Under Assumption 1, $q(\theta)$ derived from (13) is weakly decreasing. If, the virtual cost function is not monotone, one may apply the "ironing" procedure proposed by Guesnerie and Laffont (1984).[9]

We now establish continuity of the optimal allocation $q(\theta)$.

Lemma 3 (Continuity at Boundaries)**.** *Under Assumption 1, the optimal output $q(\theta)$ is continuous at the boundary points θ_k between intervals Θ_k.*

Proof. See Appendix A.2. □

Lemma 3 ensures that the allocation rule doesn't suddenly jump just because the constraint starts or stops binding. Although the economic forces determining $q(\theta)$ change at these boundary points (specifically, the shadow price $\mu(\theta)$ becomes positive or zero), the continuity of the agent's utility $U(\theta)$ and the integrated shadow cost $\Psi(\theta)$ provides a smooth link. The proof leverages the optimality conditions (FOCs) holding on both sides of the boundary, combined with the concavity properties of S and W, to rule out jumps that would create inconsistencies.

Lemma 4 (Continuity within Binding Intervals)**.** *Under Assumption 1, the optimal output $q(\theta)$ is continuous within the interior of intervals Θ_k where the constraint binds ($\mu(\theta) > 0$).*

Proof. See Appendix A.3. □

Lemma 4 addresses continuity *within* an interval where the constraint is already actively shaping the allocation ($\mu(\theta) > 0$). It confirms that even as the type θ changes while $h(\theta, q(\theta), U(\theta)) =$

[8] Assumption 1 requires $V'(\theta, a) = 1 + \frac{f(\theta)^2 - (F(\theta)+a)f'(\theta)}{f(\theta)^2} \geq 0$ for all $\theta, a \geq 0$. If $f'(\theta) > 0$, the term $-(F(\theta) + a)f'(\theta)$ becomes arbitrarily negative for large a, violating the condition.

[9] Analysing ironing under mixed constraint (6) is beyond the scope of this paper's main results.

0, the optimal allocation adjusts smoothly rather than discretely. The proof here relies critically on the fact that both $q(\theta^-)$ and $q(\theta^+)$ would need to satisfy the binding constraint $h = 0$ at a hypothetical jump point θ, which, together with the FOC and strict concavity of S, forces the quantities to be equal.

These lemmas establish that under Assumption 1:

Corollary 1. *The optimal allocation $q(\theta)$ is continuous and weakly decreasing over the entire type space Θ.*

3.1 Allocation Dynamics on Binding Intervals and Bunching

Consider the behaviour of $q(\theta)$ within an interval $\Theta_k = (\theta_a, \theta_b)$ where the constraint binds: $h(\theta, q(\theta), U(\theta)) = 0$. Differentiating this condition with respect to θ yields (a.e.):[10]

$$\dot{q}(\theta)\left(a_1 W'(q(\theta)) - a_2\theta\right) = 0. \tag{14}$$

This implies that a.e. on a binding interval, either $\dot{q}(\theta) = 0$ (bunching) or , assuming $a_1 > 0$, $q(\theta) = q^W(\theta)$ defined by

$$W'(q(\theta)) = \frac{a_2}{a_1}\theta. \tag{15}$$

If the solution were to follow the path $q = q^W(\theta)$, it must also satisfy the FOC (12), which requires $S'(q^W(\theta)) = V(\theta, a_2\Psi(\theta))$. This means the path dictated by the constraint's derivative (W') must coincide exactly with the (modified) BM path dictated by the principal's objective (S') and the endogenous shadow price accumulation (Ψ). Such alignment between potentially distinct functions (S, W, F) represents a *non-generic coincidence*. Unless this specific functional relationship holds, case $q = q^W(\theta)$ is ruled out over any interval of positive measure.

Therefore, the generic outcome on a binding interval is $\dot{q}(\theta) = 0$, implying the allocation is constant.

3.2 Conditions Guaranteeing Bunching

Beyond the genericity argument we can derive explicit conditions ensuring this path is suboptimal. Specifically, these conditions guarantee that following the $q = q^W(\theta)$ path would necessitate a negative shadow price ($\mu_{req}(\theta) < 0$) to satisfy the FOC, which contradicts the requirement $\mu(\theta) \geq 0$. If the $q = q^W(\theta)$ path is thus ruled out on an interval where the constraint binds, the only remaining possibility from equation (14) is $\dot{q}(\theta) = 0$, implying bunching. These conditions generally involve the second derivatives S'' and W'', highlighting the relevance of the curvature of the objective and constraint functions (see Section 4.5 for discussion on relaxing strict concavity of W).

[10] We established that the optimal allocation $q(\theta)$ is continuous (Lemmas 3, 4) and weakly decreasing. A continuous, monotonic function is differentiable almost everywhere. Additionally, $U(\theta)$ is absolutely continuous, ensuring $\dot{U}(\theta)$ exists a.e. Therefore, applying the chain rule to differentiate $h(\theta, q(\theta), U(\theta)) = 0$ is justified a.e.

Assumption 2 (Exact Condition Ruling Out $q = q^W$ Path). Assume $W''(q^W(\theta))$ is strictly concave. Then, the following inequality is satisfied:

$$\frac{S''(q^W(\theta))}{W''(q^W(\theta))} < \frac{a_1}{a_2}\frac{\partial V(\theta, a_2\Psi(\theta))}{\partial \theta}, \tag{16}$$

where $\frac{\partial V}{\partial \theta}$ is the partial derivative of the modified virtual cost $V(\theta, a_2\Psi(\theta)) = \theta + \frac{F(\theta)+a_2\Psi(\theta)}{f(\theta)}$ with respect to θ.

The path $q = q^W(\theta)$ requires $\mu_{req}(\theta) < 0$ if and only if Assumption 2 is satisfied (see appendix). If this inequality holds over an interval (θ_a, θ_b) where the constraint might bind, then bunching ($\dot{q} = 0$) must occur instead of following the $q = q^W$ path.

Assumption 2 provides the exact condition under which the candidate path $q = q^W(\theta)$ is inconsistent with the optimality requirements. It compares the ratio of second derivatives, S''/W'', which measures the relative concavity of the principal's objective versus the constraint function, to a threshold. This threshold depends on the constraint parameters (a_1/a_2) and, crucially, on the rate of change of the modified virtual cost with respect to type, $\partial V/\partial\theta$. Bunching is guaranteed if the principal's objective function S is sufficiently concave relative to the constraint function W (i.e., S''/W'' is small enough) compared to this threshold. While exact, evaluating this condition can be complex because $\partial V/\partial\theta$ depends on the distribution (F, f, f') and potentially the accumulated shadow cost $\Psi(\theta)$.

Assumption 3 (Sufficient Condition Ruling Out $q = q^W$ Path). Assume Assumption 1 and $W''(q^W(\theta))$ is strictly concave. Bunching ($\dot{q} = 0$) must occur on a binding interval if:

$$\frac{S''(q^W(\theta))}{W''(q^W(\theta))} < \frac{2a_1}{a_2} \tag{17}$$

holds over the interval.

Assumption 3 offers a simpler, sufficient condition for bunching that is easier to verify, provided the density function $f(\theta)$ is non-increasing. Like Assumption 2, it compares the relative concavity S''/W'' to a threshold involving the constraint parameters a_1/a_2. However, the threshold here is simply the constant $2a_1/a_2$, independent of the type θ or the virtual cost's derivative. If S is sufficiently concave relative to W such that their second derivative ratio falls below this constant threshold, the path $q = q^W$ is guaranteed to be suboptimal, forcing the allocation to be bunched ($\dot{q} = 0$) on binding segments.

Comparison and Illustration: Assumption 2 is precise but involves the potentially complex term $\partial V/\partial\theta$. Assumption 3 provides a simpler check under the condition $f' \leq 0$, yielding a sufficient condition by effectively bounding the virtual cost derivative (specifically, $f' \leq 0$ contributes

to ensuring $\partial V/\partial\theta \leq 2$ under Assumption 1, making 3 potentially easier to satisfy than 2 if $\partial V/\partial\theta$ is large, or harder if $\partial V/\partial\theta < 2$). Both conditions formalize the intuition that bunching occurs if the principal's objective S is sufficiently more concave than the constraint function W, relative to the weights a_1, a_2 in the constraint.

To build further intuition, consider the important special case where the function in the constraint is the same as the principal's objective function, $W = S$. In this scenario, the path $q^W(\theta)$ becomes $q^S(\theta)$ defined by $a_1 S'(q^S(\theta)) = a_2\theta$. Assumption 3 simplifies significantly, as $S''/W'' = 1$. The condition requires:

$$1 < \frac{2a_1}{a_2} \implies a_2 < 2a_1.$$

This condition recovers results found in related literature. For instance, in the limited enforcement model analysed by Martimort et al. (2017), the relevant constraint corresponds to $a_1 = \delta$ (discount rate) and $a_2 = 1$. Assumption 3 then yields $\delta > 1/2$, matching a condition derived from comparing slopes. Indeed, the condition $a_2 < 2a_1$ (when $W = S$) connects directly to the slopes of relevant allocation paths. Bunching ($\dot{q} = 0$) occurs if the path preferred by the principal under incentive compatibility (the modified BM path q_Ψ solving $S'(q_\Psi) = V(\theta, a_2\Psi)$) is *steeper* (i.e., has a more negative slope $\dot{q}_\Psi$) than the path dictated purely by the constraint's derivative (the path q^S solving $a_1 S' = a_2\theta$, with slope $\dot{q}^S$). The condition $a_2 < 2a_1$ (derived from Assumption 3 under $W = S$ and $f' \leq 0$) ensures this relative steepness, making it suboptimal for the solution to follow the q^S path. When $W \neq S$, the comparison becomes more complex, involving the relative second derivatives S''/W'' as captured by Assumptions 2 and 3. However, the underlying idea remains: bunching occurs when the path implied by the constraint (q^W) conflicts sufficiently with the incentive-compatible path adjusted for the constraint's shadow cost (the path q_Ψ implicitly defined by the full FOC (12)).

3.3 Structure of the Optimal Contract

Combining the continuity results (Lemmas 3, 4) and the analysis of binding intervals (leading to bunching under Assumptions 2 or 3, or the genericity argument) leads to the main structural characterization of the optimal contract.

Proposition 1 (Structure of the Optimal Contract)**.** *Under Assumption 1, the optimal allocation $q(\theta)$ for problem $(\mathscr{P}^*)$ (and thus for $(\mathscr{P})$) is continuous and weakly decreasing. Furthermore, assuming Assumption 2 or 3 holds, the optimal solution partitions the type space $\Theta = [\underline{\theta}, \overline{\theta}]$ into at most three connected intervals ($n \leq 3$), characterized by the behavior of the constraint multiplier $\mu(\theta)$:*

1. ***Universally Slack (n=1, S):*** *$\mu(\theta) = 0$ for all $\theta \in \Theta$. The optimal allocation follows the standard Baron-Myerson path $q(\theta) = q^{BM}(\theta)$, where $S'(q^{BM}(\theta)) = V(\theta, 0)$.*

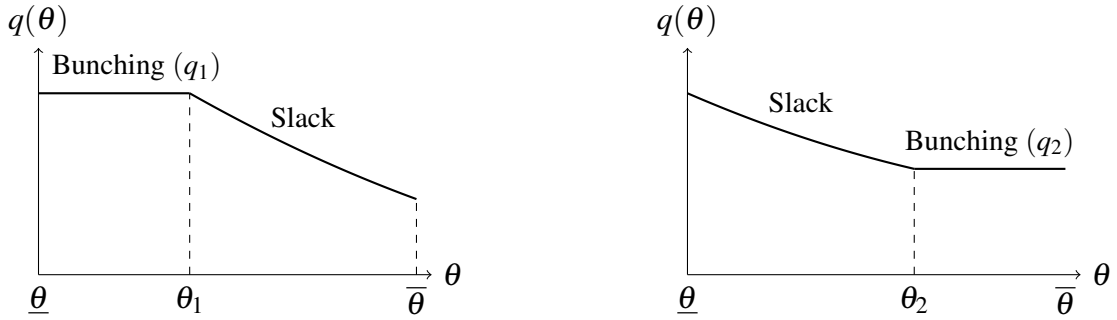
2. ***Single Binding Region (n=2):*** *$\mu(\theta) > 0$ on exactly one connected interval.*

   - ***Case 2a (Binding at Bottom, B-S):*** *The constraint binds for efficient types. $\mu(\theta) > 0$ on $[\underline{\theta}, \theta_1]$ and $\mu(\theta) = 0$ on $(\theta_1, \overline{\theta}]$. The allocation is constant $q(\theta) = q_1$ on $[\underline{\theta}, \theta_1]$ (Bunching) and follows a modified BM path $q(\theta) = q_{\Psi_1}(\theta)$ where $S'(q_{\Psi(1)}(\theta)) = V(\theta, a_2\Psi_1)$ on $(\theta_1, \overline{\theta}]$, with $\Psi_1 = \int_{\underline{\theta}}^{\theta_1} \mu(\xi)d\xi > 0$ (Figure 1a).*
   - ***Case 2b (Binding at Top, S-B):*** *The constraint binds for inefficient types. $\mu(\theta) = 0$ on $[\underline{\theta}, \theta_2)$ and $\mu(\theta) > 0$ on $[\theta_2, \overline{\theta}]$. The allocation follows the standard (unmodified) BM path $q(\theta) = q^{BM}(\theta)$ where $S'(q^{BM}(\theta)) = V(\theta, 0)$ on $[\underline{\theta}, \theta_2)$[11] and is constant $q(\theta) = q_2$ on $[\theta_2, \overline{\theta}]$ (Bunching) (Figure 1b).*

3. ***Binding at Both Ends (n=3, B-S-B):*** *The constraint binds for both efficient and inefficient types. $\mu(\theta) > 0$ on $[\underline{\theta}, \theta_1]$ and $[\theta_2, \overline{\theta}]$, separated by a slack interval (θ_1, θ_2) where $\mu(\theta) = 0$. The allocation is constant $q(\theta) = q_1$ on $[\underline{\theta}, \theta_1]$, follows a modified BM path $q(\theta) = q_{\Psi_1}(\theta)$ where $S'(q_{\Psi_1}(\theta)) = V(\theta, a_2\Psi_1)$ on (θ_1, θ_2) (with $\Psi_1 = \int_{\underline{\theta}}^{\theta_1} \mu(\xi)d\xi > 0$), and is constant $q(\theta) = q_3$ (with $q_3 \leq q_{\Psi_1}(\theta_2)$) on $[\theta_2, \overline{\theta}]$ (Figure 2).*

Proof. See Appendix A.5. □

Figures 1 and 2 illustrate the possible structures of the optimal allocation $q(\theta)$ described in Proposition 1. They depict how the continuous and weakly decreasing allocation combines constant segments (bunching), where the constraint $h \geq 0$ binds ($\mu(\theta) > 0$), with strictly decreasing segments, where the constraint is slack ($\mu(\theta) = 0$) and the allocation follows a standard or modified Baron-Myerson path.



(a) B-S Structure (Bunching-Slack)

(b) S-B Structure (Slack-Bunching)

Figure 1: Two-Region Allocation Structures (Cases 2a and 2b)

[11] In this initial slack interval starting from $\underline{\theta}$, the accumulated shadow cost $\Psi(\theta) = \int_{\underline{\theta}}^{\theta} \mu(\xi)d\xi$ is necessarily zero since $\mu(\xi) = 0$ for $\xi < \theta_2$.

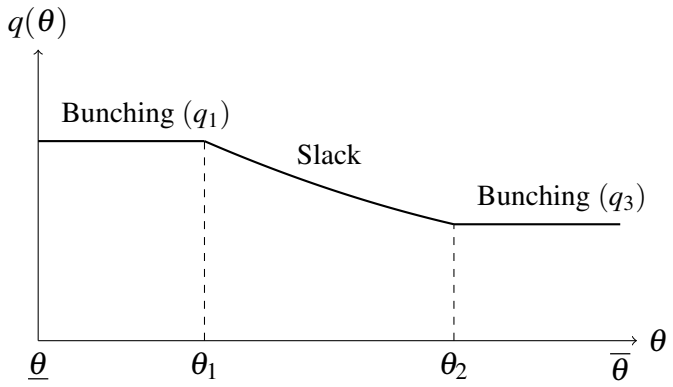


Figure 2: Three-Region Structure B–S–B (Case 3): Bunching at the extremes with an interior slack region.

Continuity is established by Lemmas 3 and 4. Constancy on binding intervals follows from equation (14) under the condition that the non-generic case $q = q^W(\theta)$ is ruled out (e.g., by Assumption 2 or 3, or the genericity argument). Weak monotonicity holds due to Assumption 1 on slack intervals and constancy on binding intervals. The proof regarding the maximum number of regions ($n \leq 3$) and the impossibility of the S-B-S structure relies on the properties of the virtual cost function $V(\theta, a_2\Psi)$ and the non-decreasing nature of $\Psi(\theta)$ at the boundaries (see Appendix A.5).

Remark 2 (Predicting the Optimal Structure: The Role of the BM Test). The structure (S, B-S, S-B, B-S-B) can often be anticipated by checking where the standard unconstrained Baron-Myerson solution $(q^{BM}(\theta), t^{BM}(\theta))$ would violate the constraint $a_1 W(q) - a_2 t + a_3 \geq 0$. Let $h^{BM}(\theta) = a_1 W(q^{BM}(\theta)) - a_2 t^{BM}(\theta) + a_3$, where $t^{BM}(\theta)$ is the transfer associated with $q^{BM}(\theta)$ (satisfying $U(\overline{\theta}) = 0$, so $t^{BM}(\theta) = \int_\theta^{\overline{\theta}} q^{BM}(x)dx + \theta q^{BM}(\theta)$). The sign pattern of $h^{BM}(\theta)$ suggests where the constraint needs to bind ($\mu(\theta) > 0$):

- **Structure S (Universally Slack):** Occurs if $h^{BM}(\theta) \geq 0$ for all θ. The unconstrained solution is feasible and thus optimal.
- **Structure B-S (Binding at Bottom):** Predicted if $h^{BM}(\theta) < 0$ only for low θ (near $\underline{\theta}$). This happens when the unconstrained solution for the most efficient types violates the constraint. Common reasons include:
  - Their high information rent leads to a large transfer t^{BM}, making $-a_2 t^{BM}$ strongly negative (especially if a_2 is large).
  - The corresponding $a_1 W(q^{BM})$ term is not large enough (especially if a_1 is small) to compensate for the negative transfer term or satisfy a baseline requirement set by a_3.
- **Structure S-B (Binding at Top):** Predicted if $h^{BM}(\theta) < 0$ only for high θ (near $\overline{\theta}$). This

happens when the unconstrained solution for the least efficient types violates the constraint. Common reasons include:

  - Their low quantity q^{BM} results in $a_1 W(q^{BM})$ being too small, failing to meet a minimum standard implied by the constraint (e.g., if $a_1 > 0$ and a_3 sets a floor).
  - Their transfer t^{BM} is also typically low. The primary driver of the constraint binding is often the insufficient $a_1 W(q^{BM})$ term, rather than a large negative impact from the transfer term.

- **Structure B-S-B (Binding at Both Ends):** Predicted if $h^{BM}(\theta) < 0$ near *both* $\underline{\theta}$ and $\overline{\theta}$. This requires a combination of the factors above: the constraint is violated by the high transfers (or low $a_1 W(q)$) of efficient types *and* by the low quantity (or low transfers) of inefficient types. Such a situation arises when the parameters (e.g., a relatively large a_2 combined with a significant a_1 and baseline a_3) and the functional forms (S, W, F) interact such that the unconstrained BM solution simultaneously puts pressure on the constraint at both the low-θ and high-θ ends of the type distribution.

This diagnostic check on the unconstrained solution provides valuable intuition about which types are most affected by the constraint and thus helps predict the qualitative pattern of slack versus binding regions in the optimal contract, such as those depicted in Figures 1 and 2.

4 Applications and Discussion

The general framework analysed, featuring the mixed constraint, provides a unified approach for a variety of constrained mechanism design problems. Proposition 1 offers a prediction: under reasonable regularity conditions (Assumptions 1, and 2 or 3), the optimal contract remains continuous and exhibits a structure with at most three alternating regions of bunching (where the constraint binds) and modified incentive-compatible allocations (where the constraint is slack). This section discusses specific applications, highlights the connection to observed economic phenomena like bunching, and clarifies the distinction between the endogenous pooling derived here and behavioural responses documented in the empirical literature.

4.1 Economic Applications of the Constraint

The constraint (3), $a_1 W(q) - a_2 t + a_3 \geq 0$, provides a versatile tool for modeling various real-world limitations faced by principals. The function $W(q)$ represents the specific aspect of the allocation (potentially quality, environmental impact, verifiable output, or the principal's own surplus $S(q)$ in some cases) that is subject to the constraint, interacting with the transfer $t(\theta)$ according to the parameters a_1, a_2, a_3.

- **Regulatory Constraints:** This framework covers various forms of regulation impacting contracts.
  - Ensuring a minimum principal payoff $S(q) - t \geq \underline{V}$ fits the framework by setting $W = S$ and $a_1 = a_2 = 1$.
  - More broadly, $W(q)$ can represent specific regulated aspects such as measured quality, environmental impact (e.g., $W(q) = -\text{Pollution}(q)$), or compliance with technical standards. Examples include quality floors (where $a_1 > 0, a_2 = 0$, imposing $W(q) \geq -a_3/a_1$), emission caps, or rules linking externalities to transfers (where $a_1 > 0, a_2 > 0$).
  - Price caps ($t \leq \bar{t}$) are captured by setting $a_1 = 0, a_2 > 0$.
- **Public Finance and Procurement:**
  - In optimal taxation or procurement, constraints might arise from budget limitations or distributional goals. While the full Mirrlees model has a different structure, analogous constraints appear in simpler settings. For instance, ex-post budget balance or minimum net benefit requirements ($S(q) - t \geq \underline{V}$) correspond to $W = S, a_1 = 1, a_2 = 1$. (cf. Laffont and Martimort, 2002).
  - Constraints might include limits on total payments ($t \leq \bar{B}$), corresponding to $a_1 = 0, a_2 > 0$.
  - The framework also accommodates situations where constraints apply to specific verifiable project milestones or outputs, represented by a function $W(q)$ that captures only that constrained dimension.
- **Enforcement Constraints:**
  - Limited enforcement, where only certain aspects $W(q)$ are pledgeable or verifiable, can be modeled as $\delta W(q) - t \geq \text{const}$ (corresponding to $a_1 = \delta, a_2 = 1$). This applies whether the verifiable component $W(q)$ is the full surplus $S(q)$ (as in Martimort et al. (2017)) or a different measure.
- **Finance and Credit Markets:**
  - Constraints could model borrower repayment limits ($t \leq t_{max}$, i.e., $a_1 = 0, a_2 > 0$) potentially arising from liability constraints or regulations.
- **Labor Contracts:**
  - While minimum wage constraints ($t \geq t_{min}$) typically fall under the $a_2 < 0$ case (see Section 4.4), other constraints related to maximum hours (if linked to q) or required

non-monetary investments or standards (represented by $W(q)$) might fit the $a_2 \geq 0$ framework.

This versatility allows the framework to address a wide range of regulatory and contractual environments where principals face constraints related to performance metrics, external standards, budget limitations, or enforcement possibilities, potentially involving different aspects $W(q)$ of the allocation.

4.2 Connection to Empirical Bunching Literature

A significant body of empirical work, surveyed by Kleven (2016), exploits bunching behavior at sharp discontinuities (kinks and notches) in policy schedules to estimate behavioural responses (e.g., Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013). This literature typically assumes the kink/notch is exogenously given and studies how optimizing agents react to it.

Our theoretical results offer a complementary perspective. We show that bunching – intervals where the optimal allocation $q(\theta)$ is constant – can arise as part of an optimal incentive scheme designed by a principal facing a smooth constraint (like (3)) and asymmetric information. Key distinctions include:

- **Source of Bunching:** Empirical literature focuses on behavioural responses to *exogenous discontinuities* (kinks/notches). Our model generates bunching from the principal's optimal design problem under *smooth constraints*.
- **Nature of Constraint:** Empirical bunching studies sharp changes in marginal incentives (kinks) or average payoffs/costs (notches). Our model uses a constraint that is typically smooth in the contract variables (q,t) via the smooth functions $S(q)$ or $W(q)$.
- **Continuity vs. Discontinuity:** Kinks maintain continuity of the choice set level, while notches introduce jumps. Our framework, under Assumption 1, predicts a *continuous* optimal allocation $q(\theta)$, even when bunching occurs (i.e., the function $q(\theta)$ is flat over the bunching interval, but doesn't jump at the boundaries).

Despite these differences, our findings are potentially relevant for applied economists:

1. **Alternative Source of Observed Bunching:** In settings where bunching is observed but clear, sharp kinks or notches are absent, our model suggests that underlying smooth constraints (e.g., participation, budget, enforcement, regulation) combined with optimal responses to asymmetric information could be a contributing factor. The resulting pooling might be observationally similar in some respects.
2. **Justification for Continuity:** Our continuity result provides a theoretical basis for assuming or modeling continuous allocation rules even when constraints are known to bind, which can simplify empirical analysis or calibration.

3. **Understanding Policy Design:** The model sheds light on *why* certain features (like pooling) might appear in optimally designed policies or contracts when facing realistic constraints, beyond simply being ad-hoc rules or approximations of unconstrained optima.

Further research could explore the observational equivalence between bunching generated by these different mechanisms and how to potentially distinguish them empirically. The generalization to $W(q)$ further broadens the applicability of this perspective to constraints not directly tied to the principal's objective $S(q)$.

4.3 On Uniqueness and Sufficiency of Optimality Conditions

The characterization of the optimal contract structure relies on the necessary conditions derived from the Pontryagin Maximum Principle. It is natural to inquire about the uniqueness of the solution and the sufficiency of these conditions.

Uniqueness: While the Maximum Principle provides necessary conditions, uniqueness of the optimal contract $(q^*(\theta), U^*(\theta))$ is not automatically guaranteed by standard theorems in this setting. The strict concavity of the principal's objective $S(q)$ ensures that, for a given state U, costate λ, and multiplier μ, the control q that maximizes the Lagrangian L is unique. However, uniqueness of the state and costate trajectories (U^*, λ^*) often relies on stricter curvature conditions on the Hamiltonian with respect to the state variable U. Since our Hamiltonian is linear in U, these standard uniqueness theorems do not directly apply. Nevertheless, in many economic applications of this type, the specific structure combined with the boundary conditions $(\lambda(\underline{\theta}) = 0, U(\overline{\theta}) = 0)$ often pins down a unique solution trajectory in practice. Establishing formal uniqueness might require exploiting more specific features of the functions S, W, F beyond the general assumptions made here.

Sufficiency and Construction of $\mu(\theta)$: The concavity of the objective function and the constraint function h in (q, U) ensures that the necessary conditions from the Maximum Principle are also *sufficient* for optimality (Mangasarian sufficiency). That is, if we find a trajectory $(q(\theta), U(\theta))$ and corresponding costate $\lambda(\theta)$ and multiplier $\mu(\theta)$ that satisfy all the conditions — the FOC (7), the state equation (4), the co-state equation (8), the boundary conditions (9), and the complementary slackness conditions (10) (importantly, requiring $h \geq 0$ and $\mu \geq 0$) — then $(q(\theta), U(\theta))$ is indeed the optimal solution.

Proposition 1 characterizes the necessary structure (alternating S and B regions, with bunching in B regions under genericity or Assumptions 2/3). The sufficiency theorem confirms that if we can construct a non-negative multiplier $\mu(\theta)$ consistent with one of these structures and satisfying all conditions, then that structure indeed corresponds to the optimum.

- **In Slack Regions (S):** By definition, $\mu(\theta) = 0$. This is consistent with $\mu \geq 0$.
- **In Binding Regions (B):** Here, $q(\theta) = q_k$ (constant) and $\mu(\theta) \geq 0$ is required. The FOC (7)

must hold. We can solve it for $\mu(\theta)$:

$$\mu(\theta)\left[a_1 W'(q_k) - a_2\theta\right] = f(\theta)(\theta - S'(q_k)) + \lambda(\theta).$$

Assuming $a_1 W'(q_k) - a_2\theta \neq 0$ (which holds if bunching occurs, i.e., if the non-generic case is excluded), we get:

$$\mu(\theta) = \frac{f(\theta)(\theta - S'(q_k)) + \lambda(\theta)}{a_1 W'(q_k) - a_2\theta}. \tag{18}$$

Here, $\lambda(\theta) = F(\theta) + a_2\Psi(\theta) = F(\theta) + a_2\int_{\underline{\theta}}^{\theta}\mu(\xi)d\xi$ depends on the integral of μ itself. For a candidate solution (e.g., a B-S structure with parameters q_1, θ^*, Ψ^*), one would need to:

1. Define $q(\theta)$ based on the structure (q_1 on $[\underline{\theta}, \theta^*]$, $q_{\Psi^*}(\theta)$ on $(\theta^*, \overline{\theta}]$).
2. Define $\mu(\theta) = 0$ on $(\theta^*, \overline{\theta}]$.
3. Use (18) to define $\mu(\theta)$ on $[\underline{\theta}, \theta^*]$, where $\lambda(\theta)$ inside the formula involves integrating this same $\mu(\xi)$ from $\underline{\theta}$ to θ.
4. Verify that the resulting $\mu(\theta)$ is indeed non-negative on $[\underline{\theta}, \theta^*]$.
5. Determine the structural parameters (q_1, θ^*, Ψ^*) using continuity of q, the binding constraint condition $h = 0$ (e.g., $h(\theta^*, q_1, U(\theta^*)) = 0$), and the relationship $\Psi^* = \int_{\underline{\theta}}^{\theta^*}\mu(\xi)d\xi$.

While the structure is determined by necessity, verifying a specific solution involves ensuring a valid non-negative multiplier $\mu(\theta)$ can be constructed consistently with all optimality conditions. The sufficiency theorem guarantees that once such a consistent solution is found, it is the optimal one. The analysis leading to Proposition 1 shows that solutions generally conform to one of the simple described structures. For instance, the condition $h^{BM}(\theta) < 0$ near $\underline{\theta}$ (Remark 2) typically ensures that the $\mu(\theta)$ constructed for the B-S structure using the FOC will indeed be positive in that region.

In essence, the necessary conditions dictate the possible forms of the solution (Proposition 1), and the sufficiency property ensures that if we can find parameters and a multiplier $\mu(\theta) \geq 0$ fitting one of these forms and satisfying all equations, then we have found the optimum.

4.4 The Case $a_2 < 0$, Limited Liability

If $a_2 = -|a_2| < 0$, the constraint is $a_1 W(q) + |a_2|t + a_3 \geq 0$. This scenario requires separate analysis as some key properties change:

- Co-state: $\dot{\lambda} = f - |a_2|\mu$.

- Slack intervals: $S'(q) = V(\theta, -|a_2|\Psi_k)$. Monotonicity of $V(\theta, a')$ for $a' \leq 0$ is not guaranteed by Assumption 1, potentially requiring ironing.
- Binding intervals: Differentiating $h = 0$ leads to $\dot{q} = -|a_2|q/(a_1 W' + |a_2|\theta) < 0$ (if $q > 0$). The allocation is strictly *decreasing* on binding intervals; bunching does *not* occur.
- Continuity: Continuity arguments (Lemmas 3, 4) still holds.
- Structure: The proof that $n \leq 3$ fails because $V(\theta, a_2\Psi)$ is now decreasing in Ψ. More complex structures ($n > 3$) or non-monotonicity requiring ironing are possible.

In summary, the $a_2 < 0$ case yields continuity and dynamically decreasing allocations on binding intervals, but monotonicity on slack intervals and the simple 3-region structure are not guaranteed.

4.5 On the Concavity of W

Our main analysis assumes $W(q)$ is strictly concave ($W'' < 0$), primarily for the applicability of the analytical conditions ruling out the $q = q^W$ path (Assumptions 2 and 3), which involve W'' in the denominator. Relaxing this to just concavity ($W'' \leq 0$) has the following implications:

- Key results: existence (Lemma 2), continuity (Lemmas 3, 4), and the derivation leading to the dichotomy $\dot{q}(a_1 W' - a_2\theta) = 0$ do not strictly require $W'' < 0$, relying only on concavity or differentiability. The strict concavity of S remains crucial for the continuity proofs.
- The $q = q^W$ path ($W' = (a_2/a_1)\theta$) cannot coincide with a linear segment of W (where $W'' = 0$) over an interval if $a_2 > 0$. If $a_2 = 0$, $\dot{q} = 0$ holds on binding intervals.
- The non-generic coincidence argument ($S'(q^W) = V(\theta, a_2\Psi)$) for ruling out the $q = q^W$ path is robust, as it relies on the distinct functional definitions of the two sides, not necessarily on W'' being strictly negative everywhere.
- The analytical conditions (Assumptions 2, 3) are technically ill-defined if $W''(q^W(\theta)) = 0$.

Therefore, the main qualitative conclusions – continuity and bunching as the generic outcome on binding intervals (when $a_2 \geq 0$) – are robust to relaxing the assumption to W being merely concave.

5 Conclusion

This paper provides a theoretical characterization of optimal contracts under adverse selection when subject to a general class of linear mixed constraints involving an arbitrary function $W(q)$ of the allocation, distinct from the principal's objective $S(q)$. Using optimal control, we establish conditions for existence and continuity, even when the constraint binds. Our main structural result (Proposition 1) shows that the optimal allocation partitions the type space into at most three regions, alternating between slack segments (following a modified BM rule based on $S(q)$) and

binding segments. Under explicit regularity conditions (or genericity), the allocation is constant (bunched) on binding segments.

The framework unifies applications where constraints relate to budgets, quality/standards (measured by $W(q)$), enforcement, or participation. Key findings include the persistence of continuity and the endogenous generation of bunching as part of the optimal constrained mechanism. This provides a theoretical counterpart to the empirical bunching literature, suggesting that observed pooling might sometimes arise from optimal responses to smooth constraints rather than solely from behavioural reactions to sharp discontinuities. This generalization allows the constraint to depend on features of the allocation ($W(q)$) distinct from the principal's objective ($S(q)$), thereby broadening the model's applicability. Future work could explore non-linear constraints or relax regularity conditions requiring ironing.

A Proofs

A.1 Proof of Lemma 2

We verify the conditions for the Filippov–Cesari existence theorem (Seierstad and Sydsæter, 1987, Theorem 2, p. 285).

(a) **Existence of an admissible pair:** We assumed the set ($\mathscr{F}$) of feasible allocations (as defined in Definition 1) is non-empty.

(b) **Convexity of the set** ($N(U,\theta)$)**:** Define the set $N(U,\theta) = \{((S(q)-\theta q-U)f(\theta)+\gamma, -q) : \gamma \leq 0, q \geq 0, h(U,q,\theta) \geq 0\}$. Let $Q(U,\theta) = \{q \geq 0 : h(U,q,\theta) \geq 0\}$ be the set of feasible controls for a given state U and type θ. Since $h(U,q,\theta) = a_1 W(q) - a_2\theta q - a_2 U + a_3$ is concave in q, the set $Q(U,\theta)$ is a convex subset of $\mathbb{R}_+$ (an interval, possibly empty or unbounded). Let $f_0(q,U,\theta) = (S(q)-\theta q-U)f(\theta)$ be the integrand of the objective. Since $S(q)$ is concave, f_0 is concave in q. The set $N(U,\theta)$ represents the pairs of achievable by feasible controls. Under Inada condition Standard arguments show that for concave objective integrands and convex control constraint sets, the set $N(U,\theta)$ is convex in $(\mathbb{R}\times\mathbb{R})$ (see e.g., Seierstad and Sydsæter, 1987, Chapter 4).

(c) **Uniform boundedness of the state variable** $U(\theta)$**:** For any admissible pair, $\dot{U}(\theta) = -q(\theta) \leq 0$ and $U(\overline{\theta}) \geq 0$, implying $U(\theta)$ is non-increasing and non-negative. Thus $U(\theta) \geq 0$. Since, the constraint (6) has to be satisfied, by Inada conditions the optimal $U(\theta)$ will be bounded above. Thus, for the relevant set of feasible pairs that could be optimal, there exists a constant b such that $0 \leq U(\theta) \leq b$.

(d) **Boundedness of the relevant control set:** We need the set of controls q that could potentially be optimal to be bounded. Let $V(U,\theta) = \{q \geq 0 : h(U,q,\theta) \geq 0\}$. We need this set,

intersected with the controls that could plausibly maximize the Hamiltonian, to be contained within a fixed compact set (e.g., $[0, q_{max}]$) for all $\theta \in \Theta$ and $U \in [0, b]$.

- $a_2 = 0$. Since $a_1 \geq 0$ and W is assumed increasing, the mixed constraint (6) requires $W(q(\theta)) \geq -a_3/a_1$. This imposes a lower bound on q.

  Under Inada condition, for sufficiently large q, $S'(q) = 0$, the direct payoff $S(q) - \theta q$ decreases. Maximizing the Hamiltonian $H = (S(q) - \theta q - U)f(\theta) - \lambda q$ or the Lagrangian $L = H + \mu h$ will therefore lead to a finite optimal $q(\theta)$.

- $a_2 > 0$. The constraint (6) rearranges to $a_1 W(q) - a_2 \theta q \geq a_2 U - a_3$. Under Inada conditions for $W(q)$ and using the fact that optimal U is bounded for large q we have $-a_2 \theta q + const \geq a_2 U - a_3$. Thus, q is bounded from above.[12]

  In both cases, the set of control values q that are relevant for the optimization problem is contained within a fixed compact interval.

Therefore, under standard assumptions on S' and the behavior of W, the set of controls relevant for the optimization problem is effectively bounded, i.e., contained within some interval.

Since the conditions of the Filippov–Cesari theorem are satisfied under standard assumptions, an optimal solution exists. ■

A.2 Proof of Lemma 3 (Continuity at Boundaries)

Consider boundary θ_k where μ switches from 0 to positive (S to B). Let $q^- = q(\theta_k^-)$, $q^+ = q(\theta_k^+)$. Continuity of U and Ψ ensures $h(\theta_k, q^+, U(\theta_k)) = 0$ and $h(\theta_k, q^-, U(\theta_k)) \geq 0$. This implies $a_1 W(q^-) - a_2(\theta_k q^- + U) \geq a_1 W(q^+) - a_2(\theta_k q^+ + U)$, and, therefore:

$$\int_{q^+}^{q^-} (a_1 W'(q) - a_2 \theta_k) dq \geq 0. \tag{A.1}$$

The FOC (12) at θ_k gives: $fS'(q^-) = f\theta_k + \lambda(\theta_k)$ (since $\mu(\theta_k^-) = 0$) and $fS'(q^+) + a_1 \mu^+ W'(q^+) = (f + a_2 \mu^+)\theta_k + \lambda(\theta_k)$ (where $\mu^+ = \mu(\theta_k^+) \geq 0$). Subtracting yields

$$f(S'(q^+) - S'(q^-)) = \mu^+(a_2 \theta_k - a_1 W'(q^+)). \tag{A.2}$$

Suppose that $q^+ < q^-$. Then $S'(q^+) > S'(q^-)$, so the LHS of (A.2) is positive. Since $q^+ \neq q^-$ the multiplier $\mu^+ > 0$, and we must have $a_2 \theta_k - a_1 W'(q^+) > 0$, or $a_1 W'(q^+) - a_2 \theta_k < 0$. Let $g(q) = a_1 W'(q) - a_2 \theta_k$. We know $g(q^+) < 0$. Since W is concave, W' is non-increasing, making $g(q)$ non-increasing. Thus, for $q \in [q^+, q^-]$ we have $g(q) \leq g(q^+) < 0$. The integral in (A.1) is

[12] A sufficient Inada condition is $\lim_{q \to \infty} W'(q) = c \geq 0$, where we can take $c = \frac{a_2}{a_1}\underline{\theta}$. Note, that we can also use a similar Inada condition for S.

$\int_{q^+}^{q^-} g(q)dq$. Since the integrand is negative over the interval $[q^+, q^-]$ (where $q^+ < q^-$), the integral must be negative. This contradicts (A.1). So $q^+ < q^-$ is impossible.

Suppose now that $q^+ > q^-$. Then $S'(q^+) < S'(q^-)$, so so the LHS of (A.2) is negative. This requires $a_2\theta_k - a_1W'(q^+) < 0$, or $a_1W'(q^+) - a_2\theta_k > 0$. Let $g(q) = a_1W'(q) - a_2\theta_k$. We have $g(q^+) > 0$. Since $g(q)$ is non-increasing, for $q \in [q^-, q^+]$ the following inequality holds $g(q) \geq g(q^+) > 0$. The integral in (A.1) is $\int_{q^+}^{q^-} g(q)dq = -\int_{q^-}^{q^+} g(q)dq$. Since the integrand is positive over $[q^-, q^+]$, the integral $\int_{q^-}^{q^+} g(q)dq > 0$. Thus, $-\int_{q^-}^{q^+} g(q)dq < 0$, which again contradicts (A.1). So $q^+ > q^-$ is impossible. Therefore, we must have $q^+ = q^-$.

The argument for continuity at a B to S boundary is analogous, reversing the roles of q^-, q^+ and the inequalities, leading to the same conclusion $q^- = q^+$. ■

A.3 Proof of Lemma 4 (Continuity within Binding Intervals)

Suppose there is a jump discontinuity at $\hat{\theta}$ within a binding interval Θ_k. Let $q^- = q(\hat{\theta}^-)$, $q^+ = q(\hat{\theta}^+)$, with $q^- \neq q^+$. Since $h = 0$ just before and just after $\hat{\theta}$, continuity of U implies $h(\hat{\theta}, q^-, U(\hat{\theta})) = h(\hat{\theta}, q^+, U(\hat{\theta})) = 0$. This yields $a_1W(q^-) - a_2(\hat{\theta}q^- + U(\hat{\theta})) = a_1W(q^+) - a_2(\hat{\theta}q^+ + U(\hat{\theta}))$, which leads to

$$\int_{q^-}^{q^+} (a_1W'(q) - a_2\hat{\theta})dq = 0. \tag{A.3}$$

Let $g(q) = a_1W'(q) - a_2\hat{\theta}$. Since W is concave, W' is non-increasing, and $g(q)$ is non-increasing. For the integral of a non-increasing function over an interval $[q^-, q^+]$ (or $[q^+, q^-]$) to be zero when $q^- \neq q^+$, the function must be identically zero, $g(q) \equiv 0$, over that range. Thus, we must have $g(q) = a_1W'(q) - a_2\hat{\theta} = 0$ for all q between q^- and q^+.

Now consider the FOC (12). Let $\lambda = \lambda(\hat{\theta}), \mu^- = \mu(\hat{\theta}^-) > 0, \mu^+ = \mu(\hat{\theta}^+) > 0$. The FOC can be written as $f(\theta)S'(q) + \mu(\theta)g(q) = f(\theta)\theta + \lambda(\theta)$. Since $g(q) = 0$ for q between q^- and q^+, the FOC simplifies to $f(\hat{\theta})S'(q) = f(\hat{\theta})\hat{\theta} + \lambda(\hat{\theta})$ for $q = q^-$ and $q = q^+$. So, $f(\hat{\theta})S'(q^-) = f(\hat{\theta})\hat{\theta} + \lambda(\hat{\theta})$ and $f(\hat{\theta})S'(q^+) = f(\hat{\theta})\hat{\theta} + \lambda(\hat{\theta})$. This implies $f(\hat{\theta})S'(q^-) = f(\hat{\theta})S'(q^+)$. Since $f(\hat{\theta}) > 0$, we have $S'(q^-) = S'(q^+)$. Because S is strictly concave, S' is strictly decreasing. Therefore, $S'(q^-) = S'(q^+)$ implies $q^- = q^+$. This contradicts the initial assumption of a jump ($q^- \neq q^+$). Thus, q must be continuous. ■

A.4 Derivations for Conditions (16) and (17)

If the optimal path were $q(\theta) = q^W(\theta)$, where $a_1W'(q^W(\theta)) = a_2\theta$ (assuming $a_1 > 0, a_2 > 0$), it must satisfy $S'(q^W(\theta)) = V(\theta, a_2\Psi(\theta))$. We derive conditions under which this path requires $\mu_{req}(\theta) < 0$.

Derivation 1: Differentiate $S'(q^W) = V(\theta, a_2\Psi(\theta))$ w.r.t. θ:

$$S''(q^W)\dot{q}^W = \frac{dV}{d\theta} = \frac{\partial V}{\partial \theta} + \frac{\partial V}{\partial (a_2\Psi)}\frac{d(a_2\Psi)}{d\theta}.$$

Using $\dot{q}^W = a_2/(a_1 W''(q^W))$, $\partial V/\partial(a_2\Psi) = 1/f$, $d(a_2\Psi)/d\theta = a_2\mu_{req}$ and substituting, we obtain for μ_{req}:

$$\mu_{req}(\theta) = \frac{f(\theta)}{a_1}\frac{S''(q^W)}{W''(q^W)} - \frac{f(\theta)}{a_2}\frac{\partial V}{\partial \theta}.$$

Requiring $\mu_{req}(\theta) < 0$ gives Condition (16).

Derivation 2: Differentiate $a_2\Psi(\theta) = (S'(q^W(\theta)) - \theta)f(\theta) - F(\theta)$ wrt θ:

$$a_2\mu_{req}(\theta) = [S''(q^W)\dot{q}^W - 1]f(\theta) + [S'(q^W) - \theta]\dot{f}(\theta) - f(\theta).$$

Substituting $\dot{q}^W = a_2/(a_1 W''(q^W))$ yields:

$$a_2\mu_{req}(\theta) = f(\theta)\frac{S''(q^W)}{W''(q^W)}\frac{a_2}{a_1} - 2f(\theta) + [S'(q^W) - \theta]\dot{f}(\theta).$$

Since $\dot{f}(\theta) \leq 0$ and $S'(q^W) - \theta = (F + a_2\Psi)/f \geq 0$, the last term is non-positive. A sufficient condition for $a_2\mu_{req}(\theta) < 0$ is $f(\theta)\frac{S''(q^W)}{W''(q^W)}\frac{a_2}{a_1} - 2f(\theta) < 0$, which yields Condition (17). ■

A.5 Proof of Proposition 1 (Structure)

The allocation $q(\theta)$ is continuous (Lemmas 3, 4) and weakly decreasing (Assumption 1 on slack intervals; $\dot{q} = 0$ on binding intervals, assuming Assumption 2 or 3 the non-generic $q = q^W$ case is excluded). Therefore, the path alternates between constant (Binding, B) and decreasing (Slack, S) segments.

Arguments ruling out $n \geq 4$ and the S-B-S structure rely on the properties of $V(\theta, a_2\Psi)$ at boundaries θ_i:

- $V(\theta, a)$ is non-decreasing in θ for $a \geq 0$ (Assumption 1).
- $V(\theta, a)$ is strictly increasing in a if $a_2 > 0$.
- $\Psi(\theta)$ is continuous, non-decreasing, and strictly increasing across B regions.

Suppose optimal partition contains S $\to$ B $\to$ S with boundaries $\theta_1 < \theta_2$. On bunching interval the constant is q_2. On slack intervals the output is equal to q_{Ψ_1} and q_{Ψ_2} with $\Psi_2 > \Psi_1$. Therefore, $q_2 = V(\theta_2, a_2\Psi_2) > V(\theta_1, a_1\Psi_1) = q_2$. Contradiction. Thus, S-B-S is impossible.

Since any partition with $n \geq 4$ necessarily contains S $\to$ B $\to$ S, we have that $n \leq 3$ and the only possible structures are S, B-S, S-B, B-S-B. ■

References

ACEMOGLU, D. (2009): *Introduction to Modern Economic Growth*, Princeton University Press.

ARAUJO, A. AND H. MOREIRA (2010): "Adverse Selection Problems Without the Spence–Mirrlees Condition," *Journal of Economic Theory*, 145, 1113–1141.

CHETTY, R., J. N. FRIEDMAN, T. OLSEN, AND L. PISTAFERRI (2011): "Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records," *The quarterly journal of economics*, 126, 749–804.

GUESNERIE, R. AND J.-J. LAFFONT (1984): "A Complete Solution to a Class of Principal-Agent Problems with an Application to the Control of a Self-Managed Firm," *Journal of Public Economics*, 25, 329–369.

JULLIEN, B. (2000): "Participation Constraints in Adverse Selection Models," *Journal of Economic Theory*, 93, 1–47.

KLEVEN, H. J. (2016): "Bunching," *Annual Review of Economics*, 8, 435–464.

KLEVEN, H. J. AND M. WASEEM (2013): "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan," *The Quarterly Journal of Economics*, 128, 669–723.

KÜHN, M. AND L. SICILIANI (2009): "Performance Indicators for Quality with Costly Falsification," *Journal of Economics & Management Strategy*, 18, 1137–1154.

LAFFONT, J.-J. AND D. MARTIMORT (2002): *The Theory of Incentives: The Principal-Agent Model*, Princeton University Press, reprint 2009 ed.

LAFFONT, J.-J. AND J. TIROLE (1993): *A Theory of Incentives in Procurement and Regulation*, MIT Press.

MARTIMORT, D., A. SEMENOV, AND L. STOLE (2017): "Optimal Stationary Contract with Two-Sided Imperfect Enforcement and Persistent Adverse Selection," *Economics Letters*, 159, 18–22.

MARTIMORT, D. AND L. A. STOLE (2022): "Participation Constraints in Discontinuous Adverse Selection Models," *Theoretical Economics*, 17, 1145–1181.

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

MIRRLEES, J. A. (1971): "An Exploration in the Theory of Optimum Income Taxation," *The Review of Economic Studies*, 38, 175–208.

MUSSA, M. AND S. ROSEN (1978): "Monopoly and Product Quality," *Journal of Economic Theory*, 18, 301–317.

ROCHET, J.-C. (1985): "Bilateral Monopoly with Imperfect Information," *Journal of Economic Theory*, 36, 214–236.

SAEZ, E. (2010): "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2, 180–212.

SCHOTTMÜLLER, C. (2015): "Adverse Selection Without Single Crossing: Monotone Solutions," *Journal of Economic Theory*, 158, 127–164.

SEIERSTAD, A. AND K. SYDSÆTER (1987): *Optimal Control Theory with Applications*, North Holland.

STIGLITZ, J. E. AND A. WEISS (1981): "Credit Rationing in Markets with Imperfect Information," *The American Economic Review*, 71, 393–410.

THOMAS, L. (2002): "Non-linear Pricing with Budget Constraint," *Economics Letters*, 75, 257–263.