Simultaneous Adverse Selection and Moral Hazard*

Daniel Gottlieb and Humberto Moreira†

First Version: August, 2011. This Version: March, 2013.

Abstract

We study a principal-agent model with moral hazard and adverse selection. Agents have private information about the distribution of outcomes conditional on each effort. We characterize the solution of the resulting multidimensional screening problem, and establish several general properties. A positive mass of types with low conditional probabilities of success gets a constant payment and zero rents. Exclusion is desirable if and only if it is first-best efficient. When agents are risk neutral, an intermediate mass of types is also pooled, although they are offered contracts with variable payments and get positive rents. In addition, the region of types who exert high effort is contained in the first-best high-effort region and, unlike in pure adverse selection models, the model may feature "distortion at the boundary." Under additional conditions, the optimal mechanism offers only finitely many contracts. We apply our framework to multidimensional generalizations of canonical models in insurance, regulation, and optimal taxation and show that it generates novel results.

1 Introduction

Most contracting situations feature both adverse selection and moral hazard. Potential insurees, for example, frequently have better knowledge of their risks. Meanwhile, they can often influence their risks by engaging in preventive effort. Managers often have better knowledge of their abilities and take actions that affect the firm's profits. Borrowers may have more precise information about their ability to repay a loan but may also be able to influence this probability. Doctors are better informed about the adequacy of each medical treatment to patients but also have some ability to substitute between treatments. Taxpayers are often better informed about their earning abilities but can also choose between activities with different distribution of outputs. Regulated firms have more precise information about their technologies but can also engage in cost-reducing actions. Still, the agency literature has focused on models in which only one of these features is present. Hence, the consequences of the interaction between adverse selection and moral hazard are poorly understood.

In this paper, we introduce adverse selection in a standard moral hazard model. Agents choose between two costly actions ("efforts"), and there are two potential outcomes. They have private

*Preliminary version; comments are especially welcome. We thank Alex Edmans, Faruk Gul, Lucas Maestri, Stephen Morris, Luca Rigotti, Yuliy Sannikov, Jean Tirole, Glen Weyl, and seminar audiences at Pittsburg/Carnegie Mellon, Princeton, and Johns Hopkins for comments and suggestions.

†Gottlieb: The Wharton School, The University of Pennsylvania, dgott@wharton.upenn.edu. Moreira: EPGE, Getulio Vargas Foundation, humberto@fgv.br.

information about the distribution of outcomes conditional on each action. Therefore, types are two-dimensional vectors.[1] The principal has a continuous prior over the set of conditional probability distributions. We characterize the optimal mechanism and establish several properties that arise under joint adverse selection and moral hazard.

If the agents' efforts were observable (no moral hazard), the principal would be able to implement the efficient allocation by compensating them for the full cost of their effort. This would keep agents indifferent between each effort and, therefore, ensure that they would have no incentives to deviate. The unobservability of effort requires the principal to leave rents to agents in order to induce a high effort choice, and to prevent each type from pretending to be another type with a slightly less favorable distribution. This generates a standard adverse selection trade-off between rent extraction and effort distortion through the local incentive-compatibility constraints. However, moral hazard also allows agents to pretend to be "distant" types by deviating in the effort dimension. For example, any type with some distribution of output conditional on low effort can always pretend to be someone who has the same distribution conditional on a high effort. Consequently, moral hazard generates *binding global incentive constraints*, which introduces new features in the model.

Because some types of agents can pretend to be less productive and shirk, they receive variable payments but still exert low effort. When reservation utilities are type independent, a positive mass of types with low conditional probabilities of success always gets a constant payment and zero rents; all other types get variable payments and positive rents. Moreover, exclusion of some types is desirable if and only if exclusion is first-best efficient.

We establish several additional properties in the special case of risk-neutral agents. A region with intermediate types (those immediately above the ones with zero rents) are also all pooled, although their contract offers variable payments. Moreover, the region of types who exert high effort under asymmetric information is contained in the first-best high-effort region. Unlike models of adverse selection with both one-dimensional and multi-dimensional types, the solution can involve *distortion at all points* (including at the top).

In some cases, the informational rents required to prevent an agent from deviating are so high that the optimal mechanism involves offering a very limited number of contracts to the agent. For example, when the distribution of types satisfies an increasing rents condition and the incremental output does not exceed twice the incremental cost of effort, the optimal mechanism involves offering *at most* three *contracts,* despite the presence of a two-dimensional continuum of types. When the probability of a high outcome is bounded away from zero and the incremental output is "not much larger" than the incremental cost of effort, the optimal mechanism involves offering *at most two contracts.* Thus, our model provides a rationale for the fact that large menus of contracts are rarely offered in practice: In the presence of simultaneous adverse selection and moral hazard, offering large menus of contracts gives too many incentives for gaming, thereby requiring the principal to leave significant informational rents. Whenever the value from effort is "not too large," it may be optimal to offer a small number of contracts instead.

Our framework builds on the principal-agent model of Grossman and Hart (1983), which has a natural interpretation in terms of the employment relationships. However, we illustrate its applicability beyond the canonical principal-agent model by considering models of insurance, procurement and regulation, and optimal taxation featuring both adverse selection and moral

[1]Grossman and Hart (1983) characterize the solution of the pure moral hazard model when there are two outcomes. However, apart from existence, they show that very little can be said about the optimal incentive scheme when there are more than two outcomes. Accordingly, this paper focuses on the two-outcome model but allows the agent to have general private information about the distribution of outcomes.

hazard.

Although the consequences of either adverse selection or moral hazard on insurance are well understood, only a few papers have studied the theoretical implications of their joint presence. This is a specially important issue since many empirical papers have found that simultaneous moral hazard and adverse selection is a key feature of several insurance markets.[2] The main difference between an insurance model and the standard principal-agent framework is the presence of type-dependent reservation utilities, since riskier types have a higher cost of remaining uninsured. We show that *exclusion is always optimal* in the insurance model. The optimality of exclusion is a consequence of the multidimensionality of types, and contrasts with one-dimensional type models where exclusion is not optimal if there are "enough low types" in the population (Stiglitz, 1977; Chade and Schlee, 2012). We also show that, because of moral hazard, the second-best high-effort region is strictly contained in the high-effort region in the absence of insurance. Therefore, policyholders *under-provide effort*.

We then consider an extension of the canonical regulation model of Laffont and Tirole (1986, 1993), with the main distinctive feature that effort in our model affects the regulated firm's costs stochastically. As a result, the regulator's incentive problem cannot be reduced to a pure adverse selection problem. We characterize the optimal regulatory mechanism and show that it has the following features. A non-degenerate region comprising firms with low conditional probabilities of reducing costs are pooled into a cost-plus contract. Types in an intermediate region are also pooled, although they receive a contract with positive power. The high-effort region is always weakly contained in the first-best high-effort region. Moreover, either this inclusion is strict (in the sense that it is contained in the interior of the first-best effort region), or the optimal mechanism features only two contracts: a cost-plus and a fixed-price contract.

Our last application consists of an optimal taxation model in the tradition of Mirrlees (1971), with the new feature that the mapping between effort and income is stochastic. Thus, the model can no longer be reduced to a pure adverse selection model. Individuals in our model differ in their *conditional probabilities* of generating high or low incomes given each effort level. We study the optimal nonlinear income tax for a Rawlsian planner. Tax rates are decreasing and there is always bunching at the bottom of the distribution, where all types face tax rates of 100%. Under the additional assumption of quasi-linear utilities, we show that there is also bunching in an intermediate region, although they face lower tax rates. Either the high-effort region is contained in the interior of the first-best high-effort region, or the optimal tax system features only two tax brackets: tax rates of 100% for lower types and 0% on higher types. We then establish conditions under which the optimal tax system generically features a distortion at the top. We also obtain conditions under which the optimal tax system can be implemented with a limited number of tax brackets.

Related Literature

Adding private information to conditional probability distributions naturally leads to a multidimensional screening environment. It is often challenging to characterize the solutions of such problems since one cannot determine from the outset the direction in which incentive constraints bind. While most of the multidimensional screening literature has focused on generalizations of the non-linear pricing model, we study a different class of models. The class of models covered by our framework includes, for example, generalizations of the principal-agent model common in

[2]See, for example, Karlan and Zinman (2009), Bajari, Hong, and Khwaja (2011), and Einav, Finkelstein, Ryan, Schrimpf, and Cullen (2013).

corporate finance and labor economics, the model of insurance provision by a monopolist, the Mirrleesian model of taxation, as well as procurement and regulation models.

There are some key differences between our framework and the non-linear pricing framework of multidimensional screening. First, *conditional on effort*, only one dimension of the type vector matters. Therefore, payoffs conditional on effort are not strictly monotone in all dimensions. However, since effort is not observable, the optimal mechanism has to provide incentives for the agent to pick the appropriate effort level. As a result, local incentive compatibility is no longer sufficient to ensure global incentive compatibility: types may also deviate in the effort dimension, thereby pooling with "distant" types. In fact, all types who exert high effort in our optimal mechanism have binding global incentive-compatibility constraints. The principal's program, therefore, has to take into account a continuum of binding global constraints. Intuitively, this program corresponds to a non-standard optimal control problem with a continuum of intermediate constraints (on top of the local first- and second-order conditions). Although there exists no general method for this class of problems, we are able to obtain optimality conditions using a calculus of variations approach.

Despite these differences, versions of classic results from the multi-dimensional screening literature also hold in our framework. For example, Armstrong (1996) established that it is generically optimal to exclude a positive mass of buyers with low valuations. Rochet and Choné (1998) showed that Armstrong's result can be generalized but, instead of exclusion, the principal would typically extract all the surplus from a positive mass of types. While it is not optimal to exclude types in our framework (as long as exclusion is not first-best optimal and participation constraints are type independent), it is also the case that the principal extracts the full surplus from a region of types with low conditional probabilities of success. In contrast, exclusion is always optimal in the insurance application of our model because of the type-dependent reservation utilities. Rochet and Choné also established that bunching was a generic property of multidimensional screening models. In our framework, the solution always entails "bunching at the bottom." In fact, bunching can be so extreme that, in some cases, the optimal mechanism features only a finite number of contracts.

We obtain several new results that are not present in the non-linear pricing model. For example, because all types who exert high effort have binding global constraints, the optimal allocation typically features a distortion at all points. This result contrasts with the "no distortion at the top" property from one-dimensional models, as well as Rochet and Choné's (1998, pp. 811) generalization of it ("no distortion at the boundary").[3]

Our paper relates to and extends several lines of work. The first one is that on screening in insurance markets with both adverse selection and moral hazard. Stewart (1994) argued that adverse selection and moral hazard may partially offset the welfare loss associated with each other. Since low risk types are offered incomplete coverage because of adverse selection, they may exert more effort than if they were fully insured. Chassagnon and Chiappori (1997) introduced precautionary effort in the seminal model of Rothschild and Stiglitz (1976) and characterized the set of separating equilibria. De Meza and Webb (2001) and Jullien, Salanie, and Salanie (2007) considered models where consumers have private information about their risk aversion and may

[3]Laffont, Maskin, and Rochet (1987) considered a natural departure from the nonlinear pricing models of Mussa and Rosen (1978) or Maskin and Riley (1984), where agents have quadratic utility functions (linear demands) and types are two-dimensional. Rochet and Stole (2002) introduced independently distributed reservation utilities in the standard nonlinear pricing model. In the monopolistic case, they show that there is no distortion at the top, and either no distortion or bunching at the bottom. For a survey of the multidimensional screening literature, see Rochet and Stole (2003).

engage in precautionary effort, and showed that the correlation between risk and coverage may be negative.[4] Similarly, Chiu and Karni (1998) presented an explanation for the lack of private unemployment insurance based on the interaction between employees' preferences for work and the unobservable effort exerted on the job.

While these papers studied models with two types of consumers, we allow for general continuous type distributions on the space of conditional probabilities. Therefore, our paper extends the literature by characterizing optimal insurance contracts when consumer's private information about riskiness is unrestricted. The continuous-type model allows us to determine which are the relevant binding constraints and provides a clearer representation of the richness of the incentive problem.[5]

The second line of related work concerns optimal taxation models with multidimensional taxpayer types. The seminal model of Mirrlees (1971), and most of the literature that followed, assumed that taxpayers differ only through a one-dimensional productivity parameter. In reality, however, taxpayer heterogeneity is multi-dimensional. Nonetheless, the theoretical difficulty of characterizing the solution of multidimensional screening programs has been a substantial barrier for the analysis of optimal taxes when taxpayer types are multidimensional. Accordingly, most of the literature has either assumed a discrete number of types, or used simulations in order to obtain properties of the optimal tax system.[6] A few recent notable exceptions are Kleven, Kreiner, and Saez (2009), Choné and Laroque (2010), Rothschild and Scheuer (2012), and Rothschild and Scheuer (2013), who study continuous-type two-dimensional screening problems resulting from the design of taxes for couples, heterogeneity in the opportunity cost of work, self-selection into different sectors, or rent seeking, respectively.

Our paper also contributes to the literature on optimal regulation and procurement. The classic model of Laffont and Tirole (1986, 1993) considers an environment that features both adverse selection (the regulated firm has private information about its technology) and moral hazard (the regulator cannot observe the firm's cost-reducing effort). However, because the link between effort, types, and output is deterministic, the model can be reduced to a pure adverse selection model.[7]

Caillaud, Guesnerie, and Rey (1992) and Picard (1987) allow for noise in the relationship between output and effort and show that, under certain conditions, the principal can achieve the same utility as in the absence of noise.[8] In our model, pure adverse selection does not

[4]In De Meza and Webb (2001), one of the two types is risk-neutral, whereas the other type is risk averse, and insurance firms have positive administrative costs. Jullien, Salanie, and Salanie (2007) studied consumers with CARA utilities and showed that the power of incentives always decreases with risk aversion.

[5]As in our model, most of the insurance literature – including all the papers above – focus on two states (loss and no loss). Furthermore, with the exception of Jullien, Salanie, and Salanie (2007), these papers also assume two effort levels. However, they study competitive equilibria whereas we study the monopolist case.

[6]Tarkiainen and Tuomala (1999) and Judd and Su (2006) discuss the theoretical difficulties of characterizing optimal taxes with multidimensional types and present simulations showing that optimal taxes in multidimensional- and one-dimensional-type models can be substantially different. Several papers have analyzed models with two types in each of two dimensions, which can be suitably mapped into a one-dimensional model consisting of four types. For example, Boadway, Marchand, Pestieau, and del Mar Racionero (2002) study optimal income taxes and Cremer, Pestieau, and Rochet (2001) show that the uniform commodity tax result fails to hold when types are multidimensional. Diamond (2005) and Diamond and Spinnewijn (2011) study the optimal taxation of individuals with heterogeneous skills and discount factors using a model with two types in each dimension, while Tenhunen and Tuomala (2010) consider three types in each dimension.

[7]This kind of environments, which also includes the Mirrleesian optimal taxation model, are often labeled 'false moral hazard' models (c.f. Laffont and Martimort, 2002).

[8]Caillaud, Guesnerie, and Rey (1992) describe these as 'noisy adverse selection models' rather than models of

entail any welfare losses compared to the first best, whereas pure moral hazard does. Moreover, welfare under joint moral hazard and adverse selection is lower than in the cases of both pure moral hazard and pure adverse selection. The reason for the contrasting welfare results is that agents in our model have private information about the conditional distribution of outcomes given efforts, whereas agents in their models have private information about the cost of effort. Another difference between our models is that we are able to characterize the solution under both risk neutrality and risk aversion, whereas they only consider risk-neutral agents.

We believe that the robustness of bunching indicates a non-trivial relationship between the complexity of the environment and the number of contracts offered to the agents. When the distribution of outcomes given efforts is observable (pure moral hazard), the principal is able to perfectly design the contract for each type. Consequently, each type who exerts high effort is offered a different contract. Moreover, all types who exert low effort obtain a constant payment. When the conditional distributions of outcomes are unobservable, offering a large number of contracts introduces too many possible deviations by the agents, which requires the principal to leave very large informational rents. Offering a limited number of contracts can be an efficient way to prevent gaming by the agents. In some cases, these informational rents are so large that the optimal mechanism features a finite number of contracts only.

The optimality of offering simple contracts in some "complex" environments is related to the robustness intuition of the seminal work of Holmstrom and Milgrom (1987). However, the notion of robustness in our model is different from the one in their paper. Here, offering a limited number of contracts is robust in that it reduces the agents' incentives to misrepresent their private information about the environment. In Holmstrom and Milgrom's model, linear contracts are robust in the sense that they prevent the agent from readjusting effort over time.[9]

The structure of the paper is as follows. Section 2 presents the basic framework and Section 3 derives some general properties of the solution. Section 4 then characterizes the solution and establishes several additional properties under the assumption of risk neutral agents, whereas Section 5 generalizes the characterization for situations where agents may be risk-averse. Section 6 applies our framework to multidimensional models of insurance (6.1), regulation (6.2), and optimal taxation (6.3). Then, Section 7 concludes.[10]

2 Model

2.1 Statement of the Problem

There is a risk-neutral principal and an agent who may be either risk neutral or risk averse. The agent exerts an effort $e \in \{0, 1\}$, which is unobservable by the principal. The principal does, however, observe the outcome from the partnership $x \in \{x_L, x_H\}$, which is stochastically affected

joint adverse selection and moral hazard since they "restrict attention to risk-neutral agents, which eliminates the insurance question that characterizes moral hazard problems."

[9]Edmans and Gabaix (2011) extend the linearity results to a model in which the realization of noise occurs before the action in each period and the principal desires to implement a fixed action in all states. Relatedly, Chassang (2011) introduces a class of calibrated contracts that are detail-free and approximate the performance of the best linear contract in dynamic environments when players are patient, while Carroll (2013) shows that the best contract for a principal who faces an agent with uncertain technology and evaluates contracts in terms of their worst-case performance is linear.

[10]Appendix A presents the benchmark cases of pure moral hazard and pure adverse selection. All proofs are available from the authors upon request.

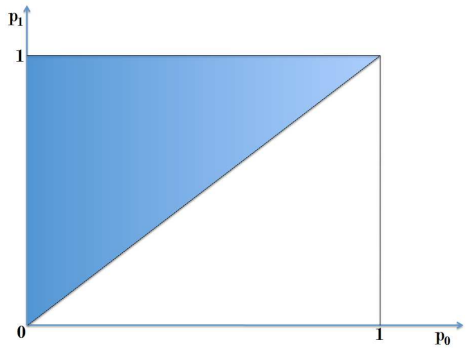

Figure 1: Type Space (shaded area).

by the agent's effort. We refer to $\Delta x \equiv x_H - x_L > 0$ as the incremental output. Let p_e denote the probability of outcome x_H given effort e. We refer to x_H and x_L as high and low outcomes and to $e = 1$ and $e = 0$ as high and low efforts.

The agent has better knowledge about the distribution of outcomes than the principal. Therefore, the conditional distribution of outcomes given efforts is private information. The agent's type is identified by the vector $\mathbf{p} \equiv (p_0, p_1)$. The principal has a continuous prior distribution over types, denoted by f.

Types satisfy the Monotone Likelihood Ratio Property (MLRP), which states that exerting higher effort increases the probability of the high outcome: $p_1 \geq p_0$. Under MLRP, the type space is contained in the area above the 45° line in Figure 1. Let $\bar{\Delta} = \{(p_0, p_1) : p_1 \geq p_0\}$ denote the space of types satisfying MLRP. We assume that the distribution of types f has full support on $\bar{\Delta}$.[11]

The agent's preferences over money W and effort e are represented by an additively separable von-Neumann Morgenstern utility function, $u(W) - c_e$, where $c_0 < c_1$, and u is continuously differentiable, increasing, and weakly concave, and the marginal utility function $\dot{u}$ is bounded.

There is no loss of generality in focusing on direct mechanisms in which the agent follows 'honest and obedient' strategies (Myerson, 1982). Accordingly, we can restrict mechanisms to be a fixed payment function $W : \bar{\Delta} \to \mathbb{R}$, a bonus function $B : \bar{\Delta} \to \mathbb{R}$, and an effort recommendation function $e : \bar{\Delta} \to \{0, 1\}$. We refer to the pair of payments $W(\mathbf{p})$ and $B(\mathbf{p})$ as a *contract*. An agent who reports type $\mathbf{p}$ agrees to exert (unobservable) effort $e(\mathbf{p})$ and receives $W(\mathbf{p})$ in case of low output and $W(\mathbf{p}) + B(\mathbf{p})$ in case of high output.

As in Grossman and Hart (1983), it is convenient to express these mechanisms in terms of the agent's utility. Let $w \equiv u(W)$ denote the agent's utility from the fixed payment W and let $b \equiv u(W + B) - u(W)$ denote the 'power' of the contract. The power of a contract is the utility gain from a high output relative to a low output. A mechanism is completely characterized by a pair of functions w and b and an effort recommendation function e. With a slight abuse of notation, we will therefore refer to a mechanism as a function $(w, b, e) : \bar{\Delta} \to \mathbb{R}^2 \times \{0, 1\}$, and we will refer to the pair $w(\mathbf{p})$ and $b(\mathbf{p})$ as a contract.

Given a mechanism (w, b, e), a type-$\mathbf{p}$ agent obtains expected utility

$$U(\mathbf{p}) \equiv w(\mathbf{p}) + p_{e(\mathbf{p})} b(\mathbf{p}) - c_{e(\mathbf{p})}. \tag{1}$$

[11] It is immediate to generalize our results for distributions that do not satisfy MLRP as long as their support contains $\bar{\Delta}$, by projecting types outside $\bar{\Delta}$ onto the 45° line.

The agent follows honest and obedient strategies if the following *incentive-compatibility* constraint is satisfied:

$$U(\mathbf{p}) \geq w(\hat{\mathbf{p}}) + p_e b(\hat{\mathbf{p}}) - c_e, \quad \forall \mathbf{p}, \hat{\mathbf{p}} \in \bar{\Delta}, \ \forall e \in \{0,1\}. \tag{IC}$$

The mechanism satisfies *individual-rationality* if the following participation constraint is satisfied:[12]

$$U(\mathbf{p}) \geq 0, \quad \forall \mathbf{p} \in \bar{\Delta}. \tag{IR}$$

We assume that the agent can costlessly reduce any amount of output (*free disposal*).[13] Therefore, payments have to be nondecreasing in the output:

$$b(\mathbf{p}) \geq 0, \quad \forall \mathbf{p} \in \bar{\Delta}. \tag{FD}$$

A mechanism is *feasible* if it satisfies incentive-compatibility, individual-rationality, and free disposal.

The principal's expected utility is:

$$\int_{\bar{\Delta}} \left\{ \ p_{e(\mathbf{p})} \left[x_H - u^{-1}(w(\mathbf{p}) + b(\mathbf{p})) \right] + \left(1 - p_{e(\mathbf{p})}\right) \left[x_L - u^{-1}(w(\mathbf{p})) \right] \ \right\} f(\mathbf{p}) d\mathbf{p}. \tag{2}$$

Two mechanisms are *equivalent* if they induce the same expected payoff to the principal and all agent types. A mechanism is *optimal* if it maximizes the principal's expected utility (1) within the class of feasible mechanisms.

2.2 Principal's Problem

In this subsection, we obtain necessary and sufficient conditions for the optimality of a mechanism. The first result establishes that there is no loss of generality in considering mechanisms for which there exists a continuous and non-decreasing function separating the sets of types who exert high and low efforts:[14]

Lemma 1. *For any feasible mechanism, there exists an equivalent mechanism* (w, b, e) *such that* $e(p_0, p_1) = 1$ *if and only if* $p_1 > \varphi(p_0)$ *for a continuous and non-decreasing function* $\varphi : [0,1] \to [0,1]$.

The intuition behind Lemma 1 is the following. Suppose a feasible mechanism recommends that type $\mathbf{p} = (p_0, p_1)$ exerts high effort, and consider a type $\hat{\mathbf{p}} = (p_0, p_1 + \varepsilon)$ for some $\varepsilon > 0$. Type $\hat{\mathbf{p}}$ has the same distribution of outcomes conditional on low effort as $\mathbf{p}$, but has a higher probability of high outcome conditional on high effort. Therefore, $\hat{\mathbf{p}}$ has an even higher incentive to exert high effort.

Next, suppose that the mechanism recommends that type $\mathbf{p} = (p_0, p_1)$ exerts low effort, and consider some type $\hat{\mathbf{p}} = (p_0 + \varepsilon, p_1)$ for some $\varepsilon > 0$. Incentive compatibility implies that $\hat{\mathbf{p}}$ will have a (weakly) higher incentive to exert low effort than type $\mathbf{p}$ has. If type $\hat{\mathbf{p}}$ is indifferent, the principal can improve by asking it to exert low effort.

[12] This formulation of the participation constraint with type-independent reservation utilities is standard in principal-agent models. In Section 6, we allow for type-dependent reservation utilities in order to study optimal insurance contracts.

[13] Free disposal is assumed in many principal-agent models, including Innes (1990), Acemoglu (1998), and Poblete and Spulber (2012).

[14] We will adopt the convention that indifferent types choose low effort. This will not affect our results since these types have measure zero.

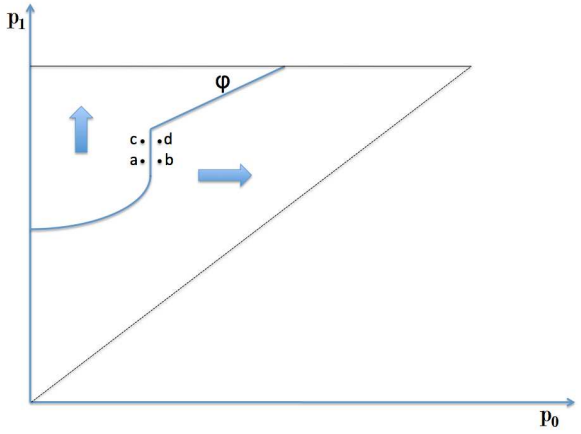

Figure 2: Intuition behind Lemma 1 (continuity of φ).

The continuity of φ follows from the indirect utility function U being continuous, strictly increasing in p_1 in the region of high effort, and constant in p_1 in the region of low effort. Figure 2 illustrates the argument. The arrows indicate the direction of growth of the informational rent function U. Consider points $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{d}$. Since U is continuous, if the distances between $\boldsymbol{a}$ and $\boldsymbol{b}$ and, $\boldsymbol{c}$ and $\boldsymbol{d}$ are small enough, we must have $U(\mathbf{a}) \approx U(\mathbf{b})$ and $U(\mathbf{c}) \approx U(\mathbf{d})$. Moreover, because the informational rent increases in p_1 in the region above φ, we must have $U(\mathbf{c}) > U(\mathbf{a})$, and because the informational rent is constant in p_1 in the region below φ, we must have $U(\mathbf{b}) = U(\mathbf{d})$. Therefore, we must have

$$U(\mathbf{c}) > U(\mathbf{a}) \approx U(\mathbf{b}) = U(\mathbf{d}) \approx U(\mathbf{c}),$$

which is a contradiction.

For a given feasible mechanism (w, b, e), we refer to the function φ as the *effort frontier* associated with it.[15] The effort frontier partitions the type space into types who exert low and high efforts:

$$e(p_0, p_1) = 1 \iff p_1 > \varphi(p_0). \tag{3}$$

The local first- and second-order conditions for incentive-compatibility yield the following necessary conditions:

Lemma 2. *Let (w, b, e) be a feasible mechanism and let φ and U be the effort frontier and informational rent functions associated with it. Then:*

a. $U(p_0, p_1)$ is convex, differentiable a. e., and has gradient

$$\nabla U(p_0, p_1) = \begin{cases} (b(p_0, p_1), 0), \text{ if } p_1 < \varphi(p_0) \\ (0, b(p_0, p_1)), \text{ if } p_1 > \varphi(p_0) \end{cases};$$

b. $b(p_0, p_1)$ is constant in p_1 for $p_1 < \varphi(p_0)$ and constant in p_0 for $p_1 > \varphi(p_0)$;

[15] Due to the equivalence result of Lemma 1, we focus on mechanisms for which an effort frontier function φ exists. Any other feasible mechanism will give the same payoff to the principal and all types of agents and will differ only in a set of zero measure (see the proof of the lemma). Moreover, such a mechanism exists whenever afeasible mechanism exists.

c. $U(0,0) \geq 0$ and $b(0,0) \geq 0$.

The two first properties follow from the adverse selection incentive-compatibility constraints, which requires reporting each type truthfully while following the principal's effort recommendation to maximize the agent's payoff. Properties (a) and (b) are the local first- and second-order conditions of this maximization program. Property (c) is a direct consequence of the participation and free disposal constraints.

The conditions from Lemma 2 are implied by adverse selection alone. Moral hazard introduces additional incentive-compatibility constraints. In particular, under moral hazard, satisfying the local constraints is not enough to prevent global deviations from being profitable since a type may now want to pretend to be another "distant" type by choosing a different effort level. The following lemma presents necessary conditions to avoid global deviations:

Lemma 3. *Let (w, b, e) be a feasible mechanism and let φ and U be the effort frontier and informational rent functions associated with it. Then:*

d. $U(p_1, p_1) = U(p_0, p_1) + \Delta c$ for $p_1 > \varphi(p_0)$;

e. $b(p_1, p_1) = b(p_0, p_1)$ for almost all (p_0, p_1) such that $p_1 > \varphi(p_0)$.

Because the conditional distribution over outputs under high and low efforts is the same for types on the 45° line and high effort is costly, these types will never exert high effort. Thus, type (p_1, p_1) exerts low effort and has the same distribution of outputs as any type (p_0, p_1) who exerts high effort (i.e., $p_1 > \varphi(p_0)$). Therefore, in any incentive-compatible mechanism, they must get the same utility net of their different effort costs (d). Property (e) is a consequence of the envelope theorem applied to deviations along the 45° line.

In models of pure adverse selection, the local necessary conditions (a)-(c) are also sufficient for the feasibility of a mechanism. We have seen that moral hazard introduces additional necessary conditions (d) and (e). We now establish that these necessary conditions are also sufficient. Thus, any optimal mechanism maximizes the principal's payoff subject to these conditions.[16]

Let (P) denote the following program:

$$\max_{(w,b,e,\varphi)} \int_{\bar{\Delta}} \left\{ x_L - u^{-1}(w(\mathbf{p})) + p_{e(\mathbf{p})} \left\{ \Delta x - \left[u^{-1}(w(\mathbf{p}) + b(\mathbf{p})) - u^{-1}(w(\mathbf{p})) \right] \right\} \right\} f(\mathbf{p}) d\mathbf{p} \quad \text{(P)}$$

subject to equations (1) and (3), and conditions (a)-(e).

Proposition 1. *A mechanism (w, b, e) is optimal if and only if there exists an effort frontier φ such that (w, b, e, φ) solves program (P).*

The direct usefulness of the characterization from Proposition 1 is limited by the fact that Program (P) is not very tractable as stated. In the next section, we will rewrite (P) as a one dimensional program, which will be key to our study of the properties of optimal mechanisms.

[16] Modulo the conventions from footnotes 14 and 15.

2.3 One-Dimensional Conditions

Let (w, b, e) be a feasible mechanism and let φ and U denote the effort frontier and informational rent functions associated with it. Let the *rent projection* associated with this mechanism be the function $\mathcal{U} : [0,1] \to \mathbb{R}^2_+$ defined as $\mathcal{U}(t) \equiv U(t,t)$. We say that a mechanism is trivial if it recommends low effort to almost all types.[17] The following lemma establishes that any nontrivial feasible mechanism is characterized by the one-dimensional functions $\mathcal{U}$ and φ:[18]

Lemma 4. *Let (w, b, e) be a nontrivial feasible mechanism and let φ and $\mathcal{U}$ denote the effort frontier and rent projection functions associated with it. Then:*

$$b(p_0, p_1) = \begin{cases} \dot{\mathcal{U}}(p_0) & \text{if } p_1 \leq \varphi(p_0) \\ \dot{\mathcal{U}}(p_1) & \text{if } p_1 > \varphi(p_0) \end{cases} \quad (a.e.), \tag{4}$$

$$w(p_0, p_1) = \begin{cases} \mathcal{U}(p_0) - p_0\dot{\mathcal{U}}(p_0) + c_0 & \text{if } p_1 \leq \varphi(p_0) \\ \mathcal{U}(p_1) - p_1\dot{\mathcal{U}}(p_1) + c_0 & \text{if } p_1 > \varphi(p_0) \end{cases} \quad (a.e.), \quad \text{and} \tag{5}$$

$$\mathcal{U}(\varphi(p_0)) = \min\{\mathcal{U}(p_0) + \Delta c; \mathcal{U}(1)\}. \tag{6}$$

Lemma 4 establishes that any nontrivial mechanism is (a.e.) characterized by the rent projection function $\mathcal{U}$, representing the information rent along the 45° line. Given such a rent projection function $\mathcal{U}$, the effort recommendation (characterized by the effort frontier function) is obtained by equation (6), which depicts types $(p_0, \varphi(p_0))$ who are indifferent between exerting high and low efforts.

Using the necessary and sufficient properties established previously, we can recover w, b, and φ from the rent projection function. First note that Properties (a) and (b) imply that the derivative of the rent projection function, $\dot{\mathcal{U}}$, equals the power of the contract along the 45° line, $b(p_0, p_0)$. Because the power of the contract is constant in the region of low effort for types with the same probability of success given low effort, the derivative of the rent projection function determines b in the low effort region. Moreover, since the power of the contract along the 45° line equals the one in the high effort region for each fixed probability of success given low effort (Property e), it also determines the power of the contract for types who are recommended high effort (see Figure 3). Finally, equation (6), which is the counterpart of Property (e), allows us to recover the effort frontier φ.[19]

Finally, note that Property (a) establishes that iso-rent functions have an inverse L shape with the kink at the effort frontier (Figure 4). Therefore, the informational rent U is determined by the rent along the 45° line $\mathcal{U}$ and the effort frontier φ. Then, using the definition of the informational rent (equation 1), we can recover the fixed component of the mechanism w.

It is more convenient to work with the one-dimensional functions $\mathcal{U}$ and φ rather than the original two-dimensional mechanism (w, b, e). The principal's cost of providing rent $\mathcal{U}$ and power $\dot{\mathcal{U}}$ to type t is:

$$G(\mathcal{U}, \dot{\mathcal{U}}, t) \equiv tu^{-1}(\mathcal{U} + (1-t)\dot{\mathcal{U}} + c_0) + (1-t)u^{-1}(\mathcal{U} - t\dot{\mathcal{U}} + c_0). \tag{7}$$

[17] The optimal trivial mechanism offers the same payments $w = c_0$ and $b = 0$ and recommends low effort $e = 0$ to (almost) all types. Similarly to Grossman and Hart (1983), it is convenient to solve for the optimal nontrivial mechanism and verify whether it generates an expected profit greater than the optimal trivial mechanism.

[18] Without loss of generality we can assume that $\dot{\mathcal{U}}(t)$ is a càdlàg function (i.e., right continuous with left limits at every point).

[19] This technique is not entirely new. For instance, Laffont, Maskin, and Rochet (1987) use a similar projection method to determine a boundary condition of the partial differential equation that characterizes incentive-compatible mechanisms in their model.

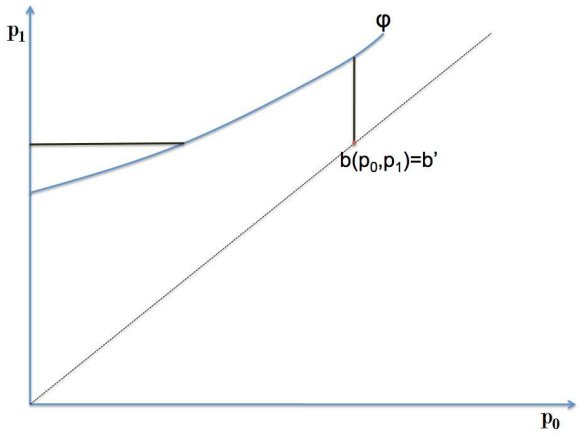

Figure 3: Iso-power functions: Types who are offered the same contract power b'.

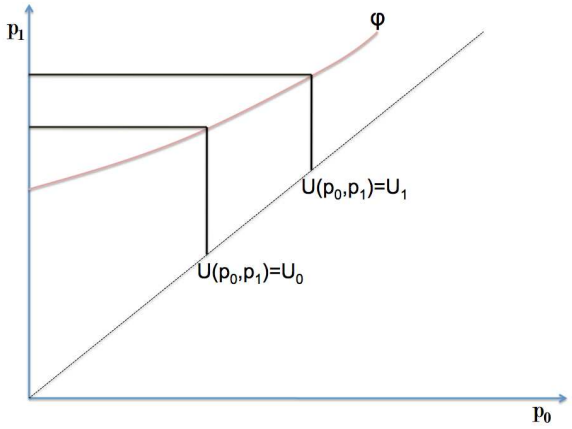

Figure 4: Iso-rent functions: Types with the same informational rents U_0 and U_1 (with $U_1 > U_0$ because informational rents are increasing).

Rewriting the principal's expected payoff in terms of $\mathcal{U}$ and φ yields

$$x_L + \int_0^1 \int_t^{\varphi(t)} (t\Delta x - G(\mathcal{U}(t), \dot{\mathcal{U}}(t), t)) f(t,s) ds dt + \int_0^1 \int_{\varphi(t)}^1 (s\Delta x - G(\mathcal{U}(s), \dot{\mathcal{U}}(s), s)) f(t,s) ds dt. \quad (8)$$

Applying Fubini's theorem, the principal's payoff (8) becomes:

$$x_L + \int_0^1 \int_t^{\varphi} \left(t\Delta x - G(\mathcal{U}, \dot{\mathcal{U}}, t)\right) f(t,s)\, ds dt + \int_{\varphi(0)}^1 \int_0^{\varphi^{-1}} \left(t\Delta x - G(\mathcal{U}, \dot{\mathcal{U}}, t)\right) f(s,t)\, ds dt$$

$$= x_L + \int_0^1 \left(t\Delta x - G(\mathcal{U}, \dot{\mathcal{U}}, t)\right) F_0(t, \varphi)\, dt + \int_{\varphi(0)}^1 \left(t\Delta x - G(\mathcal{U}, \dot{\mathcal{U}}, t)\right) F_1\left(\varphi^{-1}, t\right) dt,$$

where $F_0(t,s) \equiv \int_t^s f(t,z) dz$ and $F_1(s,t) \equiv \int_0^s f(z,t)\, dz$, and we are omitting the dependence of the functions $\mathcal{U}$, φ and φ^{-1} on t for notational simplicity.

Therefore, the principal's program (P) can be written in terms of of the one-dimensional functions $\mathcal{U}$ and φ as:

$$\max_{\mathcal{U},\varphi} x_L + \int_0^1 \left(t\Delta x - G(\mathcal{U},\dot{\mathcal{U}},t)\right) F_0\left(t,\varphi\right) dt + \int_{\varphi(0)}^1 \left(t\Delta x - G(\mathcal{U},\dot{\mathcal{U}},t)\right) F_1\left(\varphi^{-1},t\right) dt \qquad \text{(P')}$$

subject to (6), $\mathcal{U}$ nondecreasing and convex, and $\mathcal{U}\left(0\right) \geq 0$.

Proposition 2. ***a.*** *Let (w,b,e) be a feasible mechanism, let U and φ be the informational rent and effort frontier functions associated with it, and let $\mathcal{U}\left(t\right) \equiv U\left(t,t\right)$. Then, $\mathcal{U}\left(t\right)$ and $\varphi\left(t\right)$ are feasible for program (P').*
b. *Suppose $\mathcal{U}\left(t\right)$ and $\varphi\left(t\right)$ are feasible for program (P'), and define (w,b,e) according to equations (3), (4) and (5). Then, (w,b,e) is a feasible mechanism.*

Proposition 2 simplifies the search for optimal mechanisms by restating the original maximization over the set of two-dimensional functions (P) as a maximization over the set of one-dimensional rent projection and effort frontier functions (P'). Although this is a one-dimensional program, it is different from the standard programs from one-dimensional screening models in two important ways. First, there is no standard probability distribution or utility function that ensures the concavity of the objective function. Second, equation (6) corresponds to a non-standard constraint connecting a each type t to its projection along the effort frontier $\varphi\left(t\right)$. Mathematically, this corresponds to a continuum of intermediate value constraints. Economically, this means that, in addition to the local incentive compatibility constraints, there is also a continuum of binding global incentive-compatibility constraints. Using the equivalence between Programs (P) and (P'), we will say that a function $\mathcal{U} : [0,1] \to \mathbb{R}_+$ is *feasible* if it is nondecreasing and convex.[20]

3 General Properties

This section presents general properties of optimal mechanisms. Our first proposition establishes that in any optimal mechanism, there exists a positive mass of agents that do not receive any informational rents $(U\left(\mathbf{p}\right) = 0)$:

Proposition 3 (**Zero Rents at the Bottom**). *No mechanism that gives strictly positive informational rents for almost all types is optimal.*

Because the rent projection function is nondecreasing and convex, there exists a cutoff type $t^* > 0$ such that $\mathcal{U}\left(t\right) = 0$ if and only if $t \leq t^*$. Since $\dot{\mathcal{U}}\left(t\right) = b\left(t,t\right)$, types in the (interior of the) zero-rent region get the zero-power contract: $w = c_0$, $b = 0$. Then, equation (6) implies that the effort frontier $\varphi(t)$ is flat in the interval $[0,t^*]$. Figure 5 depicts these results graphically.

[20] The idea of working with a dual approach, which treats the informational rent as the instrument, is justified by Rochet (1987). In their classic analysis, Rochet and Choné (1998) follow this approach in a multidimensional-type model. Our approach is different from theirs in three aspects: (i) local constraints are necessary and sufficient in their model, whereas moral hazard introduces binding global constraints here; (ii) their instrument is the entire (multidimensional) informational rent function, whereas the domain of the instrument here is a one-dimensional subspace of the type space; and (iii) their number of instruments is equal to the dimension of the type space. In our model, the global moral hazard constraint reduces the dimensionality of the instrument from two (the dimension of the type space) to one through the one-dimensional projection method.

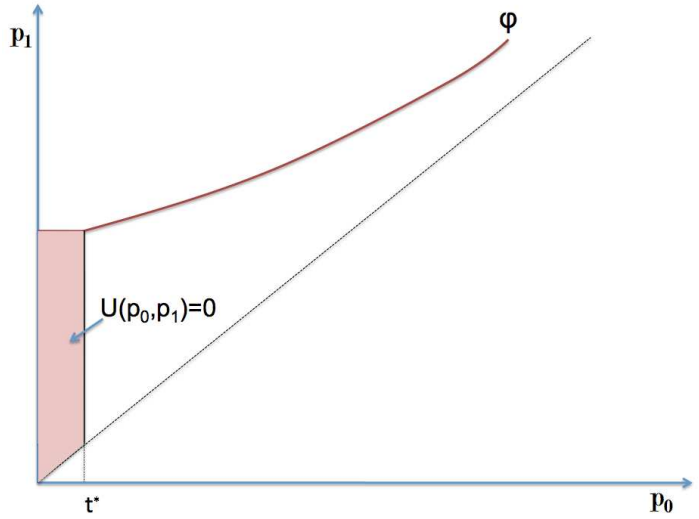

Figure 5: Types who are offered the zero-power contract and get zero rents.

Recall the original description of the mechanism in monetary units (W, B, e), where $W(\mathbf{p})$ is the fixed payment and $B(\mathbf{p})$ is the bonus. Our next result establishes that any mechanism involving a bonus greater than the incremental output Δx is not optimal:

Lemma 5 (**Bounded Bonus**). *Let (W, B, e) be a feasible mechanism, and suppose that $B(\boldsymbol{p}) > \Delta x$ in a set with positive measure. Then, (W, B, e) is not optimal.*

The principal gets $x_L - W(\mathbf{p})$ if type $\mathbf{p}$ produces a low output and $x_L + \Delta x - W(\mathbf{p}) - B(\mathbf{p})$ if he produces a high output. Therefore, Lemma 5 states that it cannot be optimal for the principal to offer a contract in which she loses money in case of a high output relative to a low output. The intuition is the following. If the principal were offering such a large bonus, reducing the bonus would have two effects. First, it would reduce the region of effort. In general, this would be detrimental to the principal. However, because she is losing money from a high output, this effect is actually positive in this case. Second, it would reduce the informational rents of all types above this one (according to the projection on the 45^o line), which would, again, raise the principal's payoff. Thus, she can unambiguously improve her expected payoff by reducing the bonus.

Using the non-optimality of mechanisms with unbounded bonus functions, we are able to establish existence. This is an important issue since we are focusing on pure strategy mechanisms and, therefore, cannot use available existence theorems.[21]

Theorem 1 (**Existence**). *There exists an optimal mechanism.*

Our next result concerns the slope of the effort frontier φ. As a benchmark, consider a situation where the principal can observe both the agent's type and his effort choice (*first best*). The principal would then offer an expected payment $u^{-1}(c_e)$ and would require the effort level that maximizes the total expected surplus.[22] Her payoff from requiring high effort is $x_L + p_1 \Delta x - u^{-1}(c_1)$, whereas her payoff from requiring low effort is $x_L + p_1 \Delta x - u^{-1}(c_0)$. Therefore, the

[21] See, for example, Kadan, Reny, and Swinkels (2011) and references therein.

[22] If the agent is risk neutral, the principal can offer any random payment with expected value equal to $u^{-1}(c_e)$.

principal would require high effort from types (p_0, p_1) such that

$$p_1 \geq p_0 + \frac{u^{-1}(c_1) - u^{-1}(c_0)}{\Delta x}. \tag{9}$$

This inequality determines the *first-best effort frontier*, which has slope 1.

Recall that, by equation (6), the effort frontier in any feasible mechanism satisfies

$$\mathcal{U}(\varphi(t)) = \mathcal{U}(t) + \Delta c$$

for all points in which $\varphi(t) < 1$. Thus, $\dot{\varphi}(t) = \frac{\dot{\mathcal{U}}(t)}{\dot{\mathcal{U}}(\varphi(t))}$ almost everywhere. Since $\varphi(t) > t$ and $\mathcal{U}$ is convex, it then follows that $\dot{\varphi} \leq 1$ (a.e.). Thus, the effort frontier function in any feasible mechanism is flatter than the first-best effort frontier. Moreover, by Proposition 3, in any optimal mechanism there exists $t^* > 0$ such that $\varphi = \varphi^*$ for all $t \in [0, t^*)$. We formally state this result in the following lemma:

Lemma 6 (**Slope of Effort Frontier**). *Let (w, b, e) be an optimal mechanism and let φ be the effort frontier function associated with it. Then, φ is continuous, differentiable (a.e.), and $\dot{\varphi} \leq 1$ at all points of differentiability. Moreover, there exists $t^* > 0$ such that $\varphi(t) = \varphi^*$ for all $t < t^*$.*

Next, we turn to the optimality of exclusion of types.

Exclusion

Our individual rationality constraint (IR) required all types to participate in the mechanism. In many situations, however, the principal can exclude some types by not offering any contracts that dominate their reservation utility (normalized to zero). This subsection considers mechanisms when the principal is allowed to exclude types.

Let $\pi(\mathbf{p}) \in \{0, 1\}$ denote the agent's participation. When $\pi(\mathbf{p}) = 0$, type $\mathbf{p}$ does not participate in the mechanism and gets utility zero. When $\pi(\mathbf{p}) = 1$, he participates and gets the utility specified in equation (1). As before, there is no loss of generality in focusing on direct mechanisms in which the agent follows 'honest and obedient' strategies. A *mechanism in the model with exclusion of types* consists of a pair of functions $w(\mathbf{p})$ and $b(\mathbf{p})$, and recommended effort and participation functions $e(\mathbf{p})$ and $\pi(\mathbf{p})$. Given a mechanism (w, b, e, π), a type-$\mathbf{p}$ agent obtains expected utility:

$$U(\mathbf{p}) \equiv \pi(\mathbf{p}) \left[w(\mathbf{p}) + p_{e(\mathbf{p})} b(\mathbf{p}) - c_{e(\mathbf{p})} \right], \tag{10}$$

and the principal gets expected utility:

$$\int_{\bar{\Delta}} \left\{ \begin{array}{c} x_L - u^{-1}(w(\mathbf{p})) + \\ p_{e(\mathbf{p})} \left\{ \Delta x - [u^{-1}(w(\mathbf{p}) + b(\mathbf{p})) - u^{-1}(w(\mathbf{p}))] \right\} \end{array} \right\} \pi(p) f(\mathbf{p}) \, d\mathbf{p}.$$

The individual-rationality and incentive-compatibility constraints are analogous to the ones in the no-exclusion model with the appropriate substitution of the utility function (1) by (10). All previous results can be adjusted to model with exclusion of types by restricting attention to the set of types who participate. In order to exclude a type of agent, the principal must ensure that this type gets at most zero expected utility from participating.

Before studying conditions for exclusion in the optimal mechanism, let us consider the exclusion rule when the principal can observe the both agent's type and his effort choice (*first-best*

exclusion rule). From condition (9), the principal's expected utility when contracting with type (p_0, p_1) is

$$\max\left\{x_L + p_0\Delta x - u^{-1}(c_0)\,;\; x_L + p_1\Delta x - u^{-1}(c_1)\right\}.$$

It is therefore optimal to exclude this type if the expected payoff is negative:

$$\max\left\{x_L + p_0\Delta x - u^{-1}(c_0)\,;\; x_L + p_1\Delta x - u^{-1}(c_1)\right\} < 0.$$

Because the expression on the left is increasing in the p_0 and p_1, it follows that exclusion is optimal if and only if it is optimal to exclude type $(0, 0)$. Substituting $p_0 = p_1 = 0$ in the previous expression, yields the condition for exclusion to be first-best optimal:

$$x_L < u^{-1}(c_0). \tag{11}$$

The following proposition establishes that the same condition for exclusion to be optimal holds when types and efforts are not observable. Therefore, exclusion is second-best optimal if and only if it is first-best optimal.

Proposition 4 (**Exclusion**). *It is optimal to exclude a strictly positive mass of types if and only if exclusion of types is first-best optimal (i.e., condition (11) holds).*

The result from Proposition 4 contrasts with the celebrated exclusion result from Armstrong (1998) for multidimensional screening in the context of a multi-product monopolist. We return to this issue on Subsection 6.1, when we consider an application to an insurance market, where the participation constraint is type-dependent. Note that Proposition 4 only refers to the "extensive margin," by showing that there is (almost) no exclusion if and only if the first-best features no exclusion. It does *not* imply that the exclusion regions in these two environments must coincide. In fact, it can be shown that when exclusion is optimal, the region of excluded types may either contain or be contained in the first-best exclusion region.

4 Risk Neutrality

This section characterizes optimal mechanisms when the agent is risk neutral: $u(X) = X$. In this case, the first-best region of effort (9) is determined by $(p_1 - p_0)\Delta x \geq \Delta c$. Therefore, a mechanism implements the first-best effort if its effort frontier solves $(\varphi(t) - t)\Delta x = \Delta c$ whenever $\varphi(t) < 1$. Let $\mathcal{U}$ be a feasible rent projection and let φ be the effort frontier associated with it. The effort distortion associated with projected type t under the effort frontier φ is

$$(\varphi(t) - t)\Delta x - \Delta c. \tag{12}$$

The effort distortion is zero if the mechanism implements the first-best effort frontier at t; it is positive if there is less effort than under first best and negative if there is more effort than under first best.

Let $\varphi^* \equiv \mathcal{U}^{-1}(\Delta c)$ denote the lowest projected type who exerts high effort, let $t^* \equiv inf\{t : \mathcal{U}(t) = 0\}$ denote the lowest projected type who gets positive rents, and let $\xi^* \equiv \sup\{t : \varphi(t) = 1\}$ denote the projected type where the effort frontier hits $p_1 = 1$. In the spirit of Myerson (1981), we define the *expected virtual surplus* as

$$\int_0^1 S_0(t,\mathcal{U})\mathcal{U}(t)\, f(t,\varphi)\, dt + \int_0^1 S_1(t,\mathcal{U})\mathcal{U}(t)\, f(\varphi^{-1},t)dt + S^*(\mathcal{U})\,\mathcal{U}(\varphi^*), \tag{13}$$

where

$$S_0(t,\mathcal{U}) := \begin{cases} -\frac{(\varphi-t)\Delta x-\Delta c}{\dot{\mathcal{U}}(\varphi)} - \frac{F_0(t,\varphi)}{f(t,\varphi)} \text{ if } t < \xi^* \\ -\frac{F_0(t,1)}{f(t,1)} \text{ if } t \geq \xi^* \end{cases}, \tag{14}$$

$$S_1(t,\mathcal{U}) := \begin{cases} 0 \text{ if } t \leq \varphi^* \\ \frac{(t-\varphi^{-1})\Delta x-\Delta c}{\dot{\mathcal{U}}(\varphi^{-1})} - \frac{F_1(\varphi^{-1},t)}{f(\varphi^{-1},t)} \text{ if } t > \varphi^* \end{cases}, \tag{15}$$

$$S^*(\mathcal{U}) := \frac{(\varphi^* - E[t|t \leq t^*, \varphi^*])\Delta x - \Delta c}{\dot{\mathcal{U}}(\varphi^*)} F_1(t^*, \varphi^*),$$

and $E[t|t \leq t^*, \varphi^*] := \frac{\int_0^{t^*} tf(t,\varphi^*)dt}{F_1(t^*,\varphi^*)}$. S_0 and S_1 are the *marginal virtual surpluses* in the regions of low and high effort, and S^* is the *inframarginal virtual surplus*. These marginal and inframarginal virtual surpluses will be key in the interpretation of the trade-offs that determine the optimal mechanism. Our expected virtual surplus (13) differs from Myerson's classic formula (and standard multidimensional generalizations of it) in one important way. Because global incentive constraints are now binding, the virtual surplus also takes into account the informational rents that are left to nonadjacent types with binding incentive constraints.

The marginal virtual surpluses S_0 and S_1 describe the costs and benefits from leaving informational rents. Recall that types above (t,t) in the low effort region and types to the left of (t,t) in the high effort region get the same contract (see Figure 3). Consider a small increase in the projected rents function $\mathcal{U}(t)$ for t in the region where some types exert high effort ($t > \varphi^*$). This perturbation affects all types who get the same contract as (t,t). Therefore, there is a "vertical effect" on types above (t,t) in the low effort region and a "horizontal effect" on types to the left of (t,t) in the high effort region.

Let us consider the vertical effect first. There is a marginal effect through the effort frontier and an inframarginal effect through the informational rents left to all those that keep exerting low effort but receive a higher rent. For the marginal effect, recall that type (t,φ) is indifferent between exerting high or low effort (for notational simplicity, we will omit the term t from $\varphi(t)$). Exerting high effort yields expected payoff

$$w(\varphi,\varphi) + \varphi b(\varphi,\varphi) - c_1 = \mathcal{U}(\varphi) - \Delta c.$$

Exerting low effort yields

$$w(t,t) + tb(t,t) - c_0 = \mathcal{U}(t).$$

If we increase $\mathcal{U}(t)$ while leaving $\mathcal{U}(\varphi)$ constant, type (t,φ) will strictly prefer to exert low effort (that is, condition (6) will no longer hold). The type who will now be indifferent between high and low efforts $(t,\hat{\varphi})$ will be above the original one: $\hat{\varphi} > \varphi$. Therefore, increasing the rent projection at t increases the effort frontier φ, thereby reducing the effort region. Recall that the distortion associated to the effort frontier is $(\varphi - t)\Delta x - \Delta c$, for $\varphi < 1$. The cost of increasing the effort frontier (and, thereby, increasing the distortion) is captured by the distortion per unit of bonus paid to the marginal type (t,φ):

$$\frac{(\varphi - t)\Delta x - \Delta c}{\dot{\mathcal{U}}(\varphi)}, \text{ for } t < \xi^*.$$

For the intramarginal effect, note that all types in the vertical line above (t,φ) are now receiving a higher rent. The total mass of those types is $F_0(t,\varphi)$. Since the marginal type (t,φ) has mass

$f(t,\varphi)$, the cost of leaving higher rents relative to the marginal type is captured by the hazard rate:

$$\frac{F_0(t,\varphi)}{f(t,\varphi)}, \text{ for } t < \xi^*.$$

Combining both terms yields the vertical effect $S_0(t)$ (with negative signs due to these effects being costs).

Next, consider the horizontal effect. Again, there is a marginal effect through the effort frontier and an inframarginal effect through the informational rents left to all those that keep exerting high effort but receive a higher rent. Recall that type (φ^{-1}, t) is indifferent between high and low efforts. His expected payoff from high effort is

$$w(t,t) + tb(t,t) - c_1 = \mathcal{U}(t) - \Delta c.$$

His expected payoff from exerting low effort is

$$w\left(\varphi^{-1}, \varphi^{-1}\right) + \varphi^{-1} b\left(\varphi^{-1}, \varphi^{-1}\right) - c_0 = \mathcal{U}\left(\varphi^{-1}\right).$$

Raising $\mathcal{U}(t)$ while keeping $\mathcal{U}(\varphi^{-1})$ unchanged makes type (φ^{-1}, t) strictly prefer to exert high effort. Thus, the effort frontier shift to the right (the type who will now be indifferent between both effort levels is $(\hat{\varphi}^{-1}, t)$ with $\hat{\varphi}^{-1} > \varphi^{-1}$), so that the region of high effort increases. The benefit from the higher effort region is captured by the distortion per unit of bonus paid to the marginal type (φ^{-1}, t):

$$\frac{(t - \varphi^{-1})\Delta x - \Delta c}{\dot{\mathcal{U}}(\varphi^{-1})}, \quad \text{for } t > \varphi^*.$$

The inframarginal effect arises from the rents left for all types to the left of (φ^{-1}, t) (who still exert high but now obtain higher informational rents). The cost of leaving total these rents is given by the mass of those types relative to the marginal type:

$$\frac{F_1(\varphi^{-1}, t)}{f(\varphi^{-1}, t)}, \quad \text{for } t > \varphi^*.$$

Adding the marginal and inframarginal effects yields the horizontal effect $S_1(t)$.

The inframarginal effect S^* is the discrete counterpart of S_0 at $\varphi^* = \varphi(0)$. In order to interpret it, consider an increase in the informational rent left in a small neighborhood of φ^*. Recall that types (t, φ^*) with $t \leq t^*$ get the same contract as (φ^*, φ^*) and are indifferent between exerting high and low efforts. Thus, the expected payoff from high effort is

$$w(\varphi^*, \varphi^*) + \varphi^* b(\varphi^*, \varphi^*) - c_1 = \mathcal{U}(\varphi^*) - \Delta c = 0.$$

The payoff from low effort is zero (since, by Proposition 3, types (t,t) with $t \leq t^*$ get zero rents). Therefore, an increase in $\mathcal{U}(\varphi^*)$ makes all those types strictly prefer to exert high effort so that the effort frontier moves down: there exists $\hat{\varphi}^* < \varphi^*$ that satisfies condition (6). Incorporating each of these types reduces the distortion $(\varphi^* - t)\Delta x - \Delta c$. Since there exists a mass $F_1(t^*, \varphi^*)$ of such types, the total gain from incorporating them equals:

$$\frac{\varphi^* - E\left[t | t \leq t^*, \varphi^*\right]\Delta x - \Delta c}{\dot{\mathcal{U}}(\varphi^*)} F_1(t^*, \varphi^*).$$

Moreover, because all of these types get zero payoffs, no informational rents have to be left. Hence, the hazard rate that appears in the expressions of S_0 and S_1 vanishes from S^*.

For notational clarity, let $\mathcal{S}(t,\mathcal{U}) \equiv S_0(t,\mathcal{U})\, f(t,\varphi) + S_1(t,\mathcal{U})\, f(\varphi^{-1},t)$ denote the marginal virtual surplus weighted by its probability density. The following lemma establishes that any optimal mechanism must maximize the expected virtual surplus among the class of feasible mechanisms.

Lemma 7. *Let $\mathcal{U}$ be the rent projection associated to an optimal mechanism. Then, for any feasible rent projection $\mathcal{V}:[0,1]\to\mathbb{R}$,*

$$\int_0^1 [\mathcal{U}(t) - \mathcal{V}(t)]\, \mathcal{S}(t,\mathcal{U})\, dt + [\mathcal{U}(\varphi^*) - \mathcal{V}(\varphi^*)]\, S^*(\mathcal{U}) \geq 0.$$

In our characterization result, we will use the following notions:

Definition 1. Let $f:[0,1]\to\mathbb{R}$ be a function with a càdlàg derivative $\dot{f}:[0,1]\to\mathbb{R}$.

- f is *strongly convex* in an interval $[a,b]\subset[0,1]$ if there exists $m>0$ such that $\dot{f}(y)-\dot{f}(x)\geq m(y-x)$ for all $x,y\in[a,b]$;
- We say f has a *kink* at $x_0\in(0,1]$ if $\lim\limits_{x\nearrow x_0}\dot{f}(x)\neq\dot{f}(x_0)$; and
- An interval $[a,b]\subset[0,1]$ is called a *maximal interval where f is affine* if there exists $m\in\mathbb{R}$ such that $\dot{f}(x)=m$, for all $x\in[a,b]$, and there is no open interval containing $[a,b]$ such that $\dot{f}(x)=m$ for all x in that interval.

The following theorem gives the necessary optimality conditions:

Theorem 2 (**Optimal Mechanisms under Risk Neutrality**)**.** *Suppose that $\mathcal{U}$ is an optimal rent projection. Then:*

1. *(pointwise condition) If $\mathcal{U}$ is strongly convex in a non-degenerated interval $[a,b]\subset[0,1]$, then $\mathcal{S}(t,\mathcal{U})=0$ for almost all $t\in[a,b]$.*
2. *(bunching conditions) Let $[a,b]\subset[0,1]$ be a maximal interval where $\mathcal{U}$ is affine.*

- *If $\varphi^*\notin[a,b]$, then*

$$0\geq a\int_a^b \mathcal{S}(t,\mathcal{U})\, dt \geq \int_a^b t\mathcal{S}(t,\mathcal{U})\, dt \geq b\int_a^b \mathcal{S}(t,\mathcal{U})\, dt.$$

Moreover, if $\mathcal{U}$ has kink at a (at b), then $\int_a^b (t-a)\mathcal{S}(t,\mathcal{U})dt=0$ ($\int_a^b (t-b)\mathcal{S}(t,\mathcal{U})dt=0$).

- *If $a=t^*$ and $b\geq\varphi^*$, then*

$$\int_{t^*}^b \mathcal{S}(t,\mathcal{U})dt + S^*(\mathcal{U})\, F_1(t^*,\varphi^*)\leq 0 \text{ and } \int_{t^*}^b (t-\varphi^*)\mathcal{S}(t,\mathcal{U})dt\leq 0.$$

Moreover, if $\mathcal{U}$ has kink at b, then

$$\int_{t^*}^b \mathcal{S}(t,\mathcal{U})dt + S^*(\mathcal{U})\, F_1(t^*,\varphi^*) = 0 \text{ and } \int_{t^*}^b (t-\varphi^*)\mathcal{S}(t,\mathcal{U})dt = 0.$$

Recall that $\mathcal{S}(t,\mathcal{U})$ gives the marginal gain from a small perturbation to the projected rent function $\mathcal{U}$. Whenever it differs from zero in an interval where $\mathcal{U}$ is strongly convex, it is possible to find a small enough perturbation that respects convexity and yields a gain to the principal. Therefore, in such interval, the marginal virtual surplus $\mathcal{S}(t,\mathcal{U})$ has to equal zero.

Part 2 are the bunching conditions. In one-dimensional models, the optimal bunching condition is determined by the ironing principle, which can be obtained by studying perturbations to the region of pooled types. Because our model has two-dimensional types, there are two admissible perturbation directions that retain the convexity of $\mathcal{U}$: translations and rotations. The two bunching conditions are precisely the necessary conditions that state that perturbing the rent projection in either of these directions does not increase the principal's payoff.

Recall that Proposition 3 showed that in any optimal mechanism an interval of types with low probability of success given both high and low efforts get a fixed payment with utility equal to their cost of effort. The next proposition shows that there exists an adjacent interval in the low effort region where all types get a uniform contract featuring a fixed payment $w < c_0$ and a bonus between the incremental cost of effort Δc and the incremental output Δx:

Proposition 5 (**Two Contracts at the Bottom**). *Let (w,b,e) be an optimal nontrivial mechanism. There exists $\varphi^* > 0$ and $t^* \in (0,\varphi^*)$ such that*

- *All types $\mathbf{p} \in [0,t^*) \times [0,\varphi^*) \cap \overline{\Delta}$ get the same contract $w = c_0$ and $b = 0$, and*
- *All types $\mathbf{p} \in [t^*,\varphi^*] \times [0,1] \cap \Delta_0$ get the same contract $w < c_0$ and $b \in (\Delta c, \Delta x]$.*

Moreover, types in both regions exert low effort.

The result from of Proposition 5 is depicted in Figure 6. Types with sufficiently low probability of success given low and high efforts, $p_0 \leq t^*$ and $p_1 < \varphi^*$, receive a fixed payment equal to the cost of low effort c_0 and exert low effort. The next set, labeled B, comprises types with intermediate probabilities of success given low efforts. All types in this set are offered the same contract, involving a payment with a lower fixed part $w < c_0$ and a bonus greater than the incremental cost of effort $b > \Delta c$.

The intuition for this result is the following. All types projected into diagonal types $t < \varphi^*$ exert low effort. Therefore, slightly increasing their informational rents does not affect the effort region associated with them. However, it generates an incentive for types above them to reduce their effort (thereby reducing the effort region associated with types projected above φ^*) and requires more rents to be left for types that exert high effort and get the same contracts as them: (p_0,p_1) with $p_1 > \varphi(p_0)$ and $p_0 \in [\varphi^*, \varphi(\varphi^*)]$. Because both effects reduce the principal's payoff, she will want to leave as little informational rents as possible while preserving the condition that the effort frontier starts at φ^*. This is obtained by paying the zero bonus for all diagonal types that are not associated with anyone who exerts high effort (region A). For diagonal type t^*, the principal needs to pay a bonus greater than the incremental cost of effort in order to incentivize types $\{(t,\varphi^*) : t \leq t^*\}$ to exert high effort. The principal then reduces the informational rents left in this region by paying the same bonus to all those types.

We now examine the effort distortion relative to the first-best benchmark. Recall that, by equation (9) the first-best allocation recommends high effort from all types that satisfy

$$p_1 \geq p_0 + \frac{\Delta c}{\Delta x}. \tag{16}$$

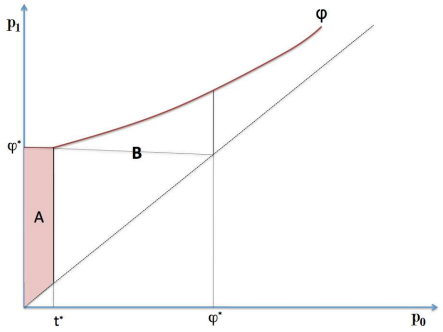

Figure 6: Two Contracts at the Bottom: Types in Region A receive the same constant payment ($w = c_0$, $b = 0$); types in Region B receive the same contract ($w < c_0$, $b > \Delta c$).

This condition can be rewritten as $(p_1 - p_0)\,\Delta x \geq \Delta c$. That is, a type should exert high effort if the incremental benefit from effort (i.e., the incremental effect of effort on outcomes $p_1 - p_0$ times the incremental output Δx) exceeds the incremental cost Δc.

Under risk neutrality, the first-best allocation is implemented by making the agent a residual claimant ($b = \Delta x$). Our first result establishes that, because the principal will never pay a bonus greater than the incremental output, the effort region in any optimal mechanism is contained in the first-best effort region.

Formally, let (w, b, e) be an optimal mechanism and consider a type (p_0, p_1) in the high effort region. By incentive compatibility, deviating to a low effort while reporting the same (true) type must yield a lower payoff:

$$w(p_0, p_1) + p_1 b(p_0, p_1) - c_1 \geq w(p_0, p_1) + p_0 b(p_0, p_1) - c_0.$$

Subtracting $w(p_0, p_1)$ from both sides and rearranging, yields

$$p_1 \geq p_0 + \frac{\Delta c}{b(p_0, p_1)} \geq p_0 + \frac{\Delta c}{\Delta x},$$

where the last inequality follows from the fact that bonuses are bounded above by Δx (Lemma 5). Therefore, the type must also exert high effort in the first-best benchmark. Recall that all types (p_0, p_1) with $\varphi(p_0) = 1$ exert low effort. Thus, we have thus established the following lemma:

Lemma 8. *Let (w, b, e) be an optimal mechanism and let φ be the effort frontier associated with it. Then $\varphi(t) \geq t + \frac{\Delta c}{\Delta x}$ whenever $\varphi(t) < 1$.*

Lemma 8 does not rule out the possibility of implementing the first-best effort frontier for some diagonal types t, which would be the case if the optimal mechanism left some types as residual claimants by offering them a bonus equal to the incremental output ($\dot{\mathcal{U}}(t) = \Delta x$). Offering such a large bonus eliminates the distortion but leaves large informational rents to the agents.

We say that the a mechanism *partially sells the firm* if it consists of the following two contracts to all agent types: $(c_0, 0)$ and $(w, \Delta x)$, for some $w \leq c_0$. Under this mechanism, agents self-select into two categories: "employees" who work for the fixed wage contract and get zero rents, and "capitalists" who buy the firm for the price w and keep the additional informational rents.

Our next result establishes that the optimal mechanism either partially sells the firm, or it features "strict distortion at all points" in the sense that the low-effort region in the first-best benchmark is contained in the interior of the low-effort region of any optimal mechanism (see Figure 7):

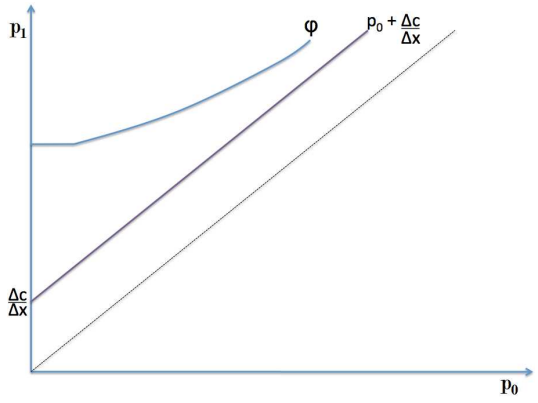

Figure 7: Strict distortion at all points.

Definition 2. Let (w, b, e) be an optimal mechanism and let φ be the associated effort frontier. We say that there is *strict distortion at all points* if $\varphi(t) > t + \frac{\Delta c}{\Delta x}$ for all t such that $\varphi(t) < 1$.

Proposition 6 (**Strict Distortion**). *Let (w, b, e) be an optimal mechanism. Then, either there is strict distortion at all points, or the principal partially sells the firm.*

The intuition behind Proposition 6 is the following. Because distortions close to the optimum have second-order costs, it can only be desirable to implement zero-distortion for some type if there are no other types with positive distortions and rents (otherwise, the principal can improve by rebalancing the amounts of distortion between these two types). Furthermore, if the optimal mechanism sells the firm to a certain projected type $\hat{t}$, it must also sell the firm to all types with projections above $\hat{t}$. Then, all types with projections above $\hat{t}$ are also undistorted.

Proposition 6 contrasts starkly with the usual no-distortion at the top result from one-dimensional models. In standard one-dimensional models, all agent's allocations except for the highest ones are distorted. Here, either the allocations of all projected types (including the highest type, $t = 1$) are distorted, or only projected types who get zero rends ($t \leq t^*$) get distorted allocations. Strict distortion at all points is a consequence of the global incentive constraint at $t = 1$, which implies that the principal must leave informational rents even for the highest types.

A natural question is whether partially selling the firm may be optimal. The next result shows that, for a certain family of distributions, partially selling the firm is generically not optimal and, therefore, the optimal mechanism features a strict distortion at all points:

Proposition 7 (**Genericity of Strict Distortion**). *Suppose that the distribution of types satisfies*

$$f(p_0, p_1) + \frac{\partial F_1}{\partial p_1}(p_0, p_1) \geq 0,$$

for all $(p_0, p_1) \in \bar{\Delta}$. For every $\epsilon > 0$, there exists a density $\tilde{f}$ such that $||f - \tilde{f}||_\infty < \epsilon$ and there is strict distortion at all points for $\tilde{f}$.[23]

Remark 1. Note that when types are uniformly distributed in $\bar{\Delta}$, the condition from Proposition 3 holds.

[23] $||\cdot||_\infty$ denotes the supreme norm in the space of continuous functions $f : \bar{\Delta} \to \mathbb{R}$.

Finite Mechanisms

A central message from models of multidimensional screening in the context of nonlinear pricing is the generality of bunching (Rochet and Choné, 1998). Obviously, since types are two-dimensional while instruments are one-dimensional, there has to be some bunching in our model. The interesting issue here is whether a *positive mass* of types get the same contract.

In the absence of adverse selection (i.e., when types are observable but effort is not), each contract is taken by at most two types. If the solution of the principal's program (P') consisted of a strictly convex rent projection $\mathcal{U}$, each contract would be taken by the vertical and horizontal projections from Figure 3, which also has zero mass. However, Proposition 3 shows that the convexity constraint has to bind at the optimum mechanism. As a result, regions of type with positive mass are offered the same contract. The intuition is reminiscent of Rochet and Choné (1998): type multidimensionality makes it hard to satisfy the local second-order condition from incentive compatibility (non-decreasing allocations) so that the solution involves bunching.

In this subsection, we show that under certain conditions the optimal mechanism can be implemented with a menu of *finitely many contracts*. Therefore, the additional incentive-constraints introduced by the multidimensionality of types imposes such a cost of leaving informational rents by preventing too many possible deviations that the principal will, in some cases, prefer to offer a very reduced number of contracts.

Let us define the *generalized hazard rate* function:

$$R(p_0, p_1) \equiv \frac{F_0(p_1, 1) + F_1(p_0, p_1)}{f(p_0, p_1)}.$$

We say that the generalized hazard rate satisfies the *increasing rents condition* if

$$\frac{\partial R}{\partial p_0}(p_0, p_1) > 0 \text{ and } \frac{\partial R}{\partial p_0}(p_0, p_1) + \frac{\partial R}{\partial p_1}(p_0, p_1) \geq 0.$$

Increasing rents is weaker than strict monotonicity since R may decrease in p_1 as long as it is sufficiently increasing in p_0. The term $\frac{F_0(p_1,1)}{f(p_0,p_1)}$ is the ratio between the mass of types above the diagonal point (p_1, p_1) and the mass at (p_0, p_1). The term $\frac{F_1(p_0,p_1)}{f(p_0,p_1)} = \frac{\int_0^{p_0} f(x,p_1)dx}{f(p_0,p_1)}$ is the ratio between the mass of types to the left of (p_0, p_1) and the mass at (p_0, p_1). Note that the uniform distribution satisfies increasing rents.

An implication of the increasing rents condition is that the informational rents associated with types along the diagonal are non-decreasing. The following lemma establishes that, under increasing rents, any optimal mechanism (w, b, e) can be implemented by offering at most two contracts to all types (p_0, p_1) with $\varphi(p_0) = 1$:[24]

Lemma 9. *Suppose that the distribution of types satisfies increasing rents. Let (w, b, e) be an optimal mechanism with associated rent projection $\mathcal{U}$. Then, $\mathcal{U}$ is a piecewise linear function with at most two pieces on $[\xi^*, 1]$.*

The intuition behind Lemma 9 is the following. When the generalized hazard rate has increasing rents, the marginal virtual surplus is a strictly decreasing function of bonus. Thus, the marginal virtual surplus will be different from zero in every interval where the bonus is strictly increasing (except for at most one point). There are three possible cases: the virtual surplus may

[24]Recall that ξ^* was defined as the type where the effort frontier reaches one, $\inf\{t : \varphi(t) = 1\}$.

be always positive, always negative, or initially positive and then negative. In all of these cases, it is possible to increase the virtual surplus by replacing the original strictly increasing bonus by a piecewise linear one that preserves incentive compatibility. For example, if the marginal virtual surplus is negative in the entire interval $[\xi^*, 1]$, replacing the rent projection by the piecewise linear function consisting of the maximum of the tangents of original rent projection at ξ^* and 1 preserves incentive compatibility and strictly increases the virtual surplus.

Recall that all contracts are identified by the contracts offered to types in the 45^o line. By Proposition 5, diagonal types (t, t) with $t \leq \varphi^*$ are offered exactly two contracts. Moreover, by Lemma 9, when the generalized hazard rate has increasing rents, all diagonal types (t, t) with $t \geq \xi^*$ are offered at most two contracts. In general, types with projections $t \in (\varphi^*, \xi^*)$ can be offered any number of contracts. The next proposition establishes that when the incremental output Δx is "not too large" relative to the incremental cost Δc of effort, this region is empty and, therefore, the optimal mechanism features at most three contracts:

Proposition 8 (**Three Contracts**). *Suppose that the distribution of types satisfies increasing rents and suppose that* $\frac{\Delta x}{\Delta c} \leq 2$. *Then:*
(i) *the rent projection* $\mathcal{U}$ *associated with any optimal mechanism is piecewise linear with at most three pieces; and*
(ii) *the optimal mechanism can be implemented by at most three pairs of conditional payments:* $\#B\left(\bar{\Delta}\right) = \#W\left(\bar{\Delta}\right) = 3$.

In particular, when the distribution is uniform the finiteness of contracts under the optimal mechanism holds for a slightly larger set of parameter values:

Corollary 1 (**Uniform Distribution**). *Suppose that types are uniformly distributed on* $\overline{\Delta}$ *and suppose that* $\frac{\Delta x}{\Delta c} \leq 3$. *Then:*
(i) *the rent projection* $\mathcal{U}$ *associated with any optimal mechanism is piecewise linear; and*
(ii) *the optimal mechanism can be implemented by a finite number of payments:* $B\left(\bar{\Delta}\right)$ *and* $W\left(\bar{\Delta}\right)$ *are finite.*

Finite optimal mechanisms also arise under different supports for the type distribution. For our next proposition, we depart from the full support assumption and assume that the conditional probability of a high outcome is bounded away from zero. Formally, for this proposition we consider following modified type space:

$$\bar{\Delta}\left(\underline{\mathrm{p}}\right) = \left\{(p_0, p_1) : \ \underline{\mathrm{p}} \leq p_0 \leq p_1\right\},$$

where $\underline{\mathrm{p}} > 0$. It is straightforward to adapt the characterization results previously derived for the type space $\bar{\Delta}$ to this case. Then, we obtain the following result:

Proposition 9 (**Two Contracts**). *Suppose the density* $f(p_0, p_1)$ *on* $\bar{\Delta}(\underline{p})$ *is non-increasing in* p_0, *and suppose that* $\underline{p} \geq \frac{\Delta x - \Delta c}{\Delta x + \Delta c}$. *Then,*
(i) *the rent projection* $\mathcal{U}$ *associated with any optimal mechanism is piecewise linear with at most two pieces; and*
(ii) *the optimal mechanism can be implemented by at most two contracts:* $\#b\left(\bar{\Delta}\right) = \#w\left(\bar{\Delta}\right) = 2$.

Note that Propositions 8 and 9 as well as Corollary 1 all required the incremental output Δx not to be "too large" relative to the incremental cost Δc of effort. This highlights the trade-off between the incentives for effort provision and rent extraction faced by the principal. When the benefit of effort is not "too high," the optimal mechanism involves offering a limited number of contracts, which reduces the informational rents that have to be left to the agent.

5 Risk Aversion

In this section, we generalize the characterization of optimal mechanisms obtained in the risk-neutral case for weakly concave utility functions. Let $\mathcal{U}$ be a feasible information projection function and let φ be the effort frontier associated with it. The generalizations of the marginal and inframarginal virtual surpluses when the utility function is weakly concave are:[25]

$$S_0(t,\mathcal{U}) \equiv \begin{cases} -\frac{(\varphi-t)\Delta x-(G(\varphi)-G)}{\dot{\mathcal{U}}(\varphi)} - \frac{\partial G}{\partial \mathcal{U}}\frac{F_0(t,\varphi)}{f(t,\varphi)} \text{ if } t < \xi^* \\ -\frac{\partial G}{\partial \mathcal{U}}\frac{F_0(t,1)}{f(t,1)} \text{ if } t \geq \xi^* \end{cases},$$

$$S_1(t,\mathcal{U}) \equiv \begin{cases} 0 \text{ if } t \leq \varphi^* \\ \frac{(t-\varphi^{-1})\Delta x-(G-G(\varphi^{-1}))}{\dot{\mathcal{U}}(\varphi^{-1})} - \frac{\partial G}{\partial \mathcal{U}}\frac{F_1(\varphi^{-1},t)}{f(\varphi^{-1},t)} \text{ if } t > \varphi^* \end{cases}, \text{ and}$$

$$S^*(\mathcal{U}) \equiv \frac{(\varphi^* - E[t|t \leq t^*, \varphi^*])\Delta x - G(\varphi^*)}{\dot{\mathcal{U}}(\varphi^*)} F_1(t^*, \varphi^*)$$

where we are using the notation $G = G(\mathcal{U},\dot{\mathcal{U}},t)$, $G(\varphi) = G(\mathcal{U}(\varphi),\dot{\mathcal{U}}(\varphi),\varphi)$ and $G(\varphi^{-1}) = G(\mathcal{U}(\varphi^{-1}),\dot{\mathcal{U}}(\varphi^{-1}),\varphi^{-1})$. The marginal virtual surpluses S_0 and S_1 differ from their risk-neutral counterparts (14) and (15) in that now the hazard rates are multiplied by the partial derivative $\partial G/\partial \mathcal{U}$. In the risk neutral case, each unit of utility left to the agent costed one dollar to the principal. Therefore, the informational rents left to the agent were determined solely by the hazard rates that specified the mass of types that received rents relative to the type on the effort frontier. Under risk aversion, each unit of utility left to the agent costs $\partial G/\partial \mathcal{U}$ to the principal. Since the principal cares about informational rents in dollars rather than in utility units, the hazard rate is multiplied by the "conversion rate" $\partial G/\partial \mathcal{U}$. Since there are no informational rents left for the inframarginal types, S^* remains unchanged relative to the risk neutral case. As in the risk neutral case, let $\mathcal{S}(t,\mathcal{U}) \equiv S_0(t,\mathcal{U}) f(t,\varphi) + S_1(t,\mathcal{U}) f(\varphi^{-1},t)$ denote the marginal virtual surplus weighted by its probability density.

When the agent is risk averse, the cost of providing a certain rent $\mathcal{U}$ is also a function of the power of the contract $\dot{\mathcal{U}}$. Thus, the relative costs of increasing the power of a diagonal type t equals the cost of providing power $\partial G/\partial \dot{\mathcal{U}}$ times the hazard rate of types who get the same contract on the high-effort region (horizontal effect) and the hazard rate of types who get the contract on the low-effort region (vertical effect). It is, therefore, useful to define each of these marginal costs as:

$$C_0(t,\mathcal{U}) \equiv \begin{cases} \frac{\partial G}{\partial \dot{\mathcal{U}}}\frac{F_0(t,\varphi)}{f(t,\varphi)} \text{ if } t < \xi^* \\ \frac{\partial G}{\partial \dot{\mathcal{U}}}\frac{F_0(t,1)}{f(t,1)} \text{ if } t \geq \xi^* \end{cases},$$

$$C_1(t,\mathcal{U}) \equiv \begin{cases} 0 \text{ if } t \leq \varphi^* \\ \frac{\partial G}{\partial \dot{\mathcal{U}}}\frac{F_1(\varphi^{-1},t)}{f(\varphi^{-1},t)} \text{ if } t > \varphi^* \end{cases}.$$

As in the risk-neutral case, it is convenient to define the marginal cost of providing power weighted by its probability density:

$$\mathcal{C}(t,\mathcal{U}) \equiv C_0(t,\mathcal{U}) f(t,\varphi) + C_1(t,\mathcal{U}) f\left(\varphi^{-1},t\right).$$

The following lemma establishes that any optimal mechanism must maximize the expected virtual surplus among the class of feasible mechanisms:

[25] To simplify the notation, the dependence of the derivatives $\partial G/\partial \mathcal{U}$ and $\partial G/\partial \dot{\mathcal{U}}$ on $(\mathcal{U},\dot{\mathcal{U}},t)$ is omitted.

Lemma 10. *Let $\mathcal{U}$ be the rent projection associated with an optimal mechanism. Then, for any feasible $\mathcal{V}:[0,1]\to\mathbb{R}$,*

$$\int_0^1 \left[\mathcal{U}(t)-\mathcal{V}(t)\right]\mathcal{S}(t,\mathcal{U})\,dt-\int_0^1\left[\dot{\mathcal{U}}(t)-\dot{\mathcal{V}}(t)\right]\mathcal{C}(t,\mathcal{U})dt+\left[\mathcal{U}(\varphi^*)-\mathcal{V}(\varphi^*)\right]S^*(\mathcal{U})\geq 0.$$

The following theorem gives the necessary optimality conditions:

Theorem 3 (**Optimal Mechanisms under Risk Aversion**). *Let $\mathcal{U}$ be an optimal rent projection. Then:*

1. *(pointwise condition) If $\mathcal{U}$ is strongly convex in a non-degenerated interval $[a,b]\subset[0,1]$ such that $\varphi^*\notin[a,b]$, then*
$$\mathcal{S}(t,\mathcal{U})+\frac{d}{dt}\left\{\mathcal{C}(t,\mathcal{U})\right\}=0,$$
for almost all $t\in[a,b]$.

2. *(bunching condition) Let $[a,b]\subset[0,1]$ be a maximal interval where $\mathcal{U}$ is affine.*

- *If $\varphi^*\notin[a,b]$, then*
$$0\geq a\int_a^b\mathcal{S}(t,\mathcal{U})\,dt\geq\int_a^b t\mathcal{S}(t,\mathcal{U})\,dt\geq b\int_a^b\mathcal{S}(t,\mathcal{U})\,dt.$$
Moreover, if $\mathcal{U}$ has kink at a (at b), then $\int_a^b(t-a)\mathcal{S}(t,\mathcal{U})dt=0$ ($\int_a^b(t-b)\mathcal{S}(t,\mathcal{U})dt=0$).

- *If $a=t^*$ and $b\geq\varphi^*$, then*
$$\int_{t^*}^b\mathcal{S}(t,\mathcal{U})dt+S^*(\mathcal{U})F_1(t^*,\varphi^*)\leq 0,\ \text{and}\ \int_{t^*}^b(t-\varphi^*)\mathcal{S}(t,\mathcal{U})dt\leq 0.$$
Moreover, if $\mathcal{U}$ has kink at b, then
$$\int_{t^*}^b\mathcal{S}(t,\mathcal{U})dt+S^*(\mathcal{U})F_1(t^*,\varphi^*)=0,\ \text{and}\ \int_{t^*}^b(t-\varphi^*)\mathcal{S}(t,\mathcal{U})dt=0.$$

In the next section, we apply the characterization derived in our basic framework to other settings.

6 Applications

6.1 Insurance

This subsection adapts our basic framework to study the optimal provision of insurance by a monopolist. The main difference between the model considered in Section 5 and a standard insurance framework is the presence of type-dependent participation constraints (c.f., Stiglitz, 1977; Chade and Schlee, 2012), since riskier types have a lower opportunity cost of remaining uninsured.

The model features a monopolistic insurance firm (principal) offering insurance to risk averse consumers (agents), with a strictly concave utility function u. Consumers have initial wealth

$I > 0$ and face a potential loss $L \in (0, I)$. They exert a preventive effort $e \in \{0, 1\}$, which affects the loss probability but is unobservable by the firm. We let e_i denote the probability of *not* suffering the loss L conditional on effort e_i, where $i = 0, 1$.

Consumers have private information about the loss probabilities conditional on each effort level. Therefore, their types are identified by a vector $(p_0, p_1) \in \bar{\Delta}$, satisfying MLRP. The insurance firm has a continuous prior distribution f over types with full support on $\bar{\Delta}$. A type-(p_0, p_1) consumer who does not purchase insurance gets expected utility

$$V(p_0, p_1) := \max_{i \in \{L,H\}} p_i u(I) + (1 - p_i) u(I - L) - c_i.$$

As in Section 2, we assume that consumers have access to a *free disposal* technology. In the insurance context, free disposal states that consumers can costlessly generate the loss L. As a result, the insurer will not offer policies in which the indemnity exceeds the loss L.

Writing mechanisms in terms of the consumer's utility as in Section 2 (equation 10), we obtain the following individual-rationality constraint for the insurance model:

$$U(p_0, p_1) \geq V(p_0, p_1), \quad \text{for all } (p_0, p_1) \in \bar{\Delta}. \tag{IR INS}$$

Thus, a mechanism is *feasible* if it satisfies incentive-compatibility (IC), individual-rationality (IR INS), and free disposal (FD). The firm's problem is to pick a feasible mechanism that maximizes its expected profits (2).

Note that any mechanism in which some types are excluded is equivalent to a mechanism in which the principal offers the *zero coverage contract*: $W = I - L$, $B = L$. In this contract, the agent pays zero in both states. Therefore, we say that a mechanism excludes a certain type if that type is offered the zero coverage contract. Our first result establishes that it is always optimal to exclude a non-degenerate region of safer types:

Proposition 10 (**Robust Exclusion in Insurance**). *The optimal exclusion region is*

$$\left\{(p_0, p_1) \in \bar{\Delta}; p_0 \geq \bar{p}_0 \text{ or } p_1 \geq \bar{p}_0 + \frac{\Delta c}{u(I) - u(I - L)}\right\}$$

for some $\bar{p}_0 < 1$.

The optimality of exclusion is a consequence of the interaction between multidimensional types and type-dependent participation constraints. With pure adverse selection and one-dimensional types, Chade and Schlee (2012, Proposition 2) showed that no type is excluded if there are enough low types in the population or if agents are sufficiently risk averse. Moreover, we have shown in Section 3 that when reservation utilities are not type-dependent, exclusion is not optimal (as long as there is no exclusion in the first best). Proposition 10 contrasts with both of these results in establishing that that exclusion is always optimal in this multidimensional model. In insurance, exclusion happens "at the top" in the sense that the safest types are the ones who do not purchase any coverage.

The intuition for our "exclusion at the top" result is the following. Starting from a situation in which all risk types participate, a reduction in informational rents excludes the types with the highest outside options. When the reduction is small enough, this set only includes the highest possible types (i.e., those with p_0 close enough to 1), who never find it beneficial choose to exert high effort. Therefore, excluding those types reduces the informational rents left to all other types and does not affect the region of effort.

Next, we establish that the presence of moral hazard shrinks the effort region among types who participate relative to a situation in which insurance is not available. In the absence of insurance, a type (p_0, p_1) chooses to exert high effort if

$$p_1 \geq p_0 + \frac{\Delta c}{u(I) - u(I - L)}. \tag{17}$$

Recall from equation (3) that any optimal mechanism identifies an effort frontier function φ, which partitions the type space into types who exert low and high efforts: $p_1 \leq \varphi(p_0)$ and $p_1 > \varphi(p_0)$, respectively. Among types who are excluded from the mechanism (and, therefore, do not get any coverage), the effort frontier has to coincide with the uninsured effort frontier (17). The next proposition establishes that, in the region of types who participate, the effort frontier lies strictly above the uninsured effort frontier. Therefore, types who participate exert "less effort" than if they were uninsured:

Proposition 11 (**Strict Distortion Relative to No Insurance**). *Let φ be the effort frontier associated with an optimal mechanism, and let $\bar{p}_0$ be the first projected type to be excluded as defined in Proposition 10. Then, $\varphi(p_0) > p_0 + \frac{\Delta c}{u(I) - u(I-L)}$ for all $p_0 < \bar{p}_0$.*

Remark 2. Because utility is non-transferable, principal and agent generally disagree over the first-best effort level. As seen above, high effort is efficient *from the agent's perspective* if condition (17) holds. On the other hand, high effort is efficient *from the principal's perspective* if $p_1 \geq p_0 + \frac{\Delta c}{L}$. The later corresponds to the first-best frontier in our model, since we are assuming that the principal has all bargaining power.

When the agent's incremental utility from the loss is lower than the loss itself (i.e., the principal's incremental utility from the loss) – i.e., $u(I) - u(I - L) \leq L$ – the agent picks a (weakly) lower effort level than the firm would prefer when effort is observable. Combining with Proposition 6, this implies that the second-best effort frontier lies above the first-best effort frontier. Note, however, that the second-best effort frontier is *not* above the first-best frontier when the agent's incremental utility from the loss exceeds the loss: $u(I) - u(I - L) > L$. For example, agents who are excluded from the mechanism will choose effort according to the frontier (17), which lies below the first-best frontier.

6.2 Regulation

In this section, we adapt our basic framework to a model of procurement and regulation. We follow the general setup from Laffont and Tirole (1986, 1993), except that we allow the firm's cost-reducing effort to affect firm costs stochastically. This modification implies that the model cannot be reduced to a pure adverse selection model anymore, and generates different results.

A regulated firm produces a single unit of a project at a random monetary cost, which can be either low C_L or high C_H, where $C_H > C_L$. The firm's manager exerts a cost-reducing effort which can be either high ($e = 1$) or low ($e = 0$), and is not observed by the regulator. The cost-reducing effort stochastically affects the firm's monetary cost: The firm faces a low cost C_L with probability p_e, and a high cost C_H with probability $1 - p_e$. Exerting effort increases the likelihood of a low cost realization: $p_1 \geq p_0$. Therefore, conditional probabilities satisfy MLRP: $(p_0, p_1) \in \bar{\Delta}$. The firm's manager gets utility c_e from exerting effort e, where $c_1 > c_0$ and $\Delta c \equiv c_1 - c_0$.

The project generates a consumer surplus of $S > 0$. The regulator observes the monetary cost incurred by the firm (but not the cost-reducing effort). Following a standard accounting

convention in the literature, we assume that the regulator reimburses the firm's monetary costs in addition to a payment w in case of high cost, and $w+b$ in case of low cost. Thus, b denotes the power of the regulated firm's contract. The expected utility of the firm's (risk-neutral) manager is then

$$U = w + p_e b - c_e, \tag{18}$$

where $e = 0, 1$. We assume that the firm manager has access to a free disposal technology and, therefore, can freely inflate costs. As a result, the regulator will not offer contracts with negative power. Moreover, the manager has an outside option with payoff normalized to zero.

Conditional on effort e, the regulator pays the firm an expected amount $w + p_e b + C_H - p_e (C_H - C_L)$. As in Laffont and Tirole (1986, 1993), we assume that the government has to revert to distortionary taxation in order to raise funds and, therefore, the regulator faces a shadow cost of public funds $\lambda > 0$. As a result, the net surplus of consumers/taxpayers is

$$S - (1+\lambda)\left[w + p_e b + C_H - p_e (C_H - C_L)\right].$$

A utilitarian regulator maximizes the sum of the consumers' net surplus and the expected utility of the firm's manager (18):

$$S - (1+\lambda)\left[w + p_e b + C_H - p_e (C_H - C_L)\right] + U. \tag{19}$$

In order to rewrite this model in terms of our basic framework, let us introduce the variables x_H and x_L, which denote the taxpayers' surplus net of the utility left to the firm's manager:

$$x_H := S - (1+\lambda)C_L, \quad x_L := S - (1+\lambda)C_H.$$

Note that a high output x_H corresponds to a low cost realization C_L and vice versa. Moreover, we let $\Delta x := x_H - x_L > 0$ denote the net gain from a low cost relative to a high cost realization. Rearranging expression (19), we can rewrite the regulator's objective function as

$$x_L + p_e \Delta x - (1+\lambda)c_e - \lambda U.$$

Because the shadow cost public funds λ is positive, the regulator would like to avoid leaving rents to the firm's manager.

In the first-best benchmark where both effort and the firm's type (p_0, p_1) are observable, the regulator solves

$$\max_{(U,e)} x_L + p_e \Delta x - (1+\lambda)c_e - \lambda U$$

subject to

$$U \geq 0.$$

There are two differences between this setup and our basic framework from Section 4. First, the principal's cost from leaving one unit of informational rent to the agent is λ rather than 1. Because the regulator's payoff consists of the sum between the manager's and the taxpayers' utility, and each dollar left to the manager costs $1+\lambda$ to taxpayers, the total effect on the regulator's payoff is the shadow cost λ. Second, the regulator takes into account the additional effect of compensating the manager's disutility of effort through the requirement of raising public funds. Therefore, instead of subtracting the total surplus by c_e as in the standard model, the principal subtracts $(1+\lambda)\, c_e$.

The solution entails $U = 0$ and $e = 1 \iff p_1 \geq p_0 + (1+\lambda)\frac{\Delta c}{\Delta x}$. Therefore, the first-best mechanism leaves zero rents to the firm's manager and implements the effort frontier function:

$$\varphi_f(p_0) := p_0 + (1+\lambda)\frac{\Delta c}{\Delta x}.$$

We now consider the situation where the regulator does not observe either the firm manager's cost-reducing effort e or the firm's effectiveness in reducing costs (p_0, p_1). The regulator has a prior distribution about the firm's type (p_0, p_1) with full support on the set of conditional distributions that satisfy MLRP $\bar{\Delta}$ described by the continuous density $f : \overline{\Delta} \to \mathbb{R}_{++}$.

Following the same steps as in our basic framework, we can establish the existence of an effort frontier function φ. The firm's rents associated to a feasible mechanism are determined by

$$U(\boldsymbol{p}) = \begin{cases} w(\boldsymbol{p}) + p_0 b(\boldsymbol{p}) - c_0, \text{ if } p_1 \leq \varphi(p_0) \\ w(\boldsymbol{p}) + p_1 b(\boldsymbol{p}) - c_1, \text{ if } p_1 > \varphi(p_0) \end{cases}.$$

Let $\mathcal{U}(t) := w(t,t) + tb(t,t) - c_0$ denote the projected rent function. By Lemma 4, we have

$$U(t,s) = \begin{cases} \mathcal{U}(t), \text{ if } s \leq \varphi(t) \\ \mathcal{U}(s) - \Delta c \text{ if } s > \varphi(t) \end{cases}.$$

Therefore, the regulator's problem is then

$$\max_{\varphi, \mathcal{U}} \quad x_L - \lambda c_0 + \int_0^1 \int_t^{\varphi(t)} (t\Delta x - \lambda \mathcal{U}(t) - c_0) f(t,s) ds dt + \int_0^1 \int_{\varphi(t)}^1 (s\Delta x - \lambda \mathcal{U}(s) - c_1) f(t,s) ds dt,$$

subject to $\mathcal{U}$ increasing and convex, $\mathcal{U}(0) = 0$, and $\mathcal{U}(\varphi(t)) = \mathcal{U}(t) + \Delta c$ if $\varphi(t) \in (\varphi(0), 1)$.

Adapting the results derived in Section 4, we summarize our main findings in the following proposition:

Proposition 12 (**Optimal Regulation**). *There exists an optimal mechanism with the following properties:*

- *Two contracts at bottom: There exists $\varphi^* > 0$ and $t^* \in (0, \varphi^*)$ such that*
  - *All types $\mathbf{p} \in [0, t^*) \times [0, \varphi^*) \cap \overline{\Delta}$ get a cost-plus contract ($w = c_0$, $b = 0$), exert zero effort, and get zero rents;*
  - *All types $\mathbf{p} \in [t^*, \varphi^*] \times [0,1] \cap \Delta_0$ get a contract with positive power ($w < c_0$, $b \in (\Delta c, \Delta x]$), exert zero effort, and get positive rents.*
- *The power of the contract does not exceed the cost reduction ($b \leq C_H - C_L$);*
- *The effort frontier lies above the first-best effort frontier. Either there is a strict distortion at all points, or the regulator offers only a cost-plus ($b = 0$) and a fixed-price contract ($b = C_H - C_L$); and*
- *Exclusion is optimal if and only if exclusion is first-best optimal.*

The characterization of the optimal mechanism in terms of the expected virtual surplus (Lemma 7 and Theorem 2) and the results on finite mechanisms can also be easily adapted for the regulation model.

6.3 Optimal Taxation

In this subsection, we apply our model to an optimal taxation framework. The model consists of a Rawlsian tax agency (principal) who wishes to design a tax system for a population of taxpayers (agents). Taxpayers generate an output that can be either high, x_H, or low, x_L. They choose effort $e \in \{0, 1\}$, which is unobserved by the tax agency and stochastically affects their output. Taxpayers are also privately informed about the effectiveness their effort. Thus, each taxpayer is represented by a type vector (p_0, p_1) representing the probability of a high output given each effort level. Types have full support on the set of probabilities that satisfy MLRP. We assume that taxpayers have access to a free disposal technology and, therefore, cannot be charged incremental taxes that exceed 100%.[26]

One natural interpretation of our model is in terms of optimal unemployment insurance. In this interpretation, unemployed workers (taxpayers) can either find or not find a job. The high output x_H corresponds to the income of a worker who finds a job and the low output x_L corresponds to the income of someone who does not find a job (possibly zero).

Our model can also be interpreted more generally as the optimal design of an income tax in the spirit of Mirrlees (1971), although the assumption that there are only two possible outcomes may be harder to justify in practice.[27] In the Mirrleesian framework, taxpayers also have an unobservable productivity type and choose an effort level. However, because the mapping from types and effort to total income is deterministic, the model can be reduced to a standard screening problem with adverse selection only.[28] Here, because effort affects income in a stochastic manner, the model does not reduce to a pure adverse selection model. Moreover, because taxpayers have private information about the probabilities of outputs given each effort level, their types are multi-dimensional.

We follow Piketty (1997) and Saez (2001) in assuming that the tax agency is Rawlsian and, therefore, maximizes the utility of the least favored individual.[29] By Property (a) from Lemma 2, incentive compatibility implies that taxpayers' utilities are increasing in their types. As a result, the least favored individual is the lowest type: $(0, 0)$. As defined in Section 2, a mechanism $(w, b, e) : \bar{\Delta} \to \mathbb{R}^2 \times \{0, 1\}$ specifies the agent's utility in case of low output w, the power of the contract b (utility gain from a high output relative to a low output), and an effort recommendation e. The tax agency's problem is then to design a mechanism that maximizes the utility of the lowest type, $w(0, 0) - c_0$, among mechanisms that satisfy incentive compatibility (IC), free disposal (FD), and the resource constraint

$$\int_{\bar{\Delta}} \left\{ \; x_L - u^{-1}(w(\mathbf{p})) + p_{e(\mathbf{p})} \left\{ \Delta x - [u^{-1}(w(\mathbf{p}) + b(\mathbf{p})) - u^{-1}(w(\mathbf{p}))] \right\} \; \right\} f(\mathbf{p}) d\mathbf{p} \geq R,$$

where the parameter $R \in \mathbb{R}$ denotes the total resources (possibly negative) that need to be financed by the tax program.

[26] There is a large literature on optimal taxation that assumes free disposal, starting with Diamond and Mirrlees (1971) and Mirrlees (1972).

[27] Stiglitz (1982) studies optimal taxation in a model with two types. As a result, only two outputs are observed. Our model features a (two-dimensional) continuum of types. However, because types affect outputs stochastically (whereas Stiglitz, as well as in most of the optimal taxation literature, assumes a deterministic relationship), we also obtain only two possible outcomes.

[28] Mirrlees (1990) studied optimal taxation in a model where incomes are uncertain, although he restricted the analysis to linear taxes.

[29] Saez (2001) considers both Rawlsian and utilitarianist tax agencies. Our approach can be extended to the utilitarianist case, although it requires considering an ex-ante participation constraint in our general framework.

In the principal-agent framework described in Section 2, the principal wanted to extract the largest amount of expected resources from agents subject to the lowest possible type obtaining a utility above a certain reservation utility (normalized to zero). Here, the tax agency wants to maximize the utility of the lowest possible type subject to expected resources left to agents not exceeding a certain level. Hence, the tax agency's problem is the dual of the principal's problem from our general framework. It is then straightforward to adapt the analysis from the previous sections to obtain several new results for optimal taxation in the presence of joint moral hazard and adverse selection. Theorem 3 derives local optimality conditions.

Adapting Proposition 3, it follows that *types in a non-degenerate region at the bottom of the distribution get a 100% tax rate*. Formally, there exists $\bar{p}_0 > 0$ and $\bar{p}_1 > 0$ such that $b(p_0, p_1) = 0$ for all $(p_0, p_1) \leq (\bar{p}_0, \bar{p}_1)$.

This conclusion is related to results on optimal tax rates at the bottom of the distribution from the one-dimensional type model. Under a utilitarianist welfare function, the tax rate at the bottom of the earnings distribution is *zero* if and only if earnings are bounded away from zero (Seade, 1977; Ebert, 1992). Under a Rawlsian welfare function, the optimal tax rate at the bottom should be strictly lower than 100% if earnings are bounded away from zero and 100% if they are not.[30] Note, however, that the optimality of the 100% tax rate in our model does not rely on the expected earnings of lowest types. Moreover, the difference between the after-tax income in case of high and a low earnings, B, is a non-decreasing function of types.

Following Diamond (1998, 2005) and Piketty (1997), suppose that taxpayers have a quasi-linear utility function: $W - c_e$, where $e \in \{0, 1\}$.[31] We can then adapt the results from Section 4 to obtain additional results. An adaptation of Proposition 5 establishes that only two contracts are offered to types "at the bottom". Types in the lowest region, $\mathbf{p} \in [0, t^*) \times [0, \varphi^*) \cap \bar{\Delta}$, are all offered the same after-tax income ($b = 0$) and exert low effort. Therefore, the tax agency guarantees a constant after-tax income to these workers, regardless of their outputs (100% tax rate). All types outside of this lower region face tax rates that are strictly lower than 100%. Types in the intermediate region, $\mathbf{p} \in [t^*, \varphi^*) \times [0, 1] \cap \bar{\Delta}$, face the same tax rate.

Proposition 6 establishes that either there second-best effort region is contained in the interior of the first-best effort region ("strict distortion at all points"), or the optimal tax system features only two tax brackets: one with a 100% tax rate, and another one with a 0% tax rate. Proposition 7 obtains sufficient conditions on the distribution of types to ensure strict distortion at all points generically. Strict distortion at all points, which contrasts with the famous efficiency-at-the-top result from models with one-dimensional types, is caused by the global incentive constraints that are binding. Additionally, Propositions 8 and 9 and Corollary 1 determine conditions under which optimal tax system can be implemented using a finite number of tax brackets.

7 Conclusion

Contracting situations typically combine elements of both adverse selection and moral hazard. Most of the literature, however, has focused on models in which only one of them is present. In this paper, we showed that adverse selection and moral hazard are not separable issues, and the

[30] Since, in practice, the most disadvantaged individuals have zero earnings, the optimal income taxes at the bottom are strictly positive under a utilitarian welfare function and 100% under a Rawlsian welfare function (c.f. Saez, 2001, Piketty and Saez, 2012).

[31] Quasi-linearity is often justified empirically by the fact that income elasticities of primary earners is close to zero (although income effects are important for secondary earners). Theoretically, optimal income taxes in the Mirrleesian framework are much simpler under quasi-linear utilities.

interaction between them can generate contracts that are fundamentally different from environments featuring only one of them.

In our model, the principal always extracts all agents' surpluses when there is either pure moral hazard or pure adverse selection. Moreover, she implements the first best in the case of pure adverse selection by offering a payment equal to the agent's effort cost. Under pure moral hazard, the principal offers a fixed wage to types who exert low effort, and a positive bonus to those that exert high effort. Agents do not get positive rents, although the outcome is no longer efficient if agents are risk averse.

Optimal mechanisms are quite different when both adverse selection and moral hazard are simultaneously present. The principal has to leave rents to some agents. As a result, she faces a trade-off between rent extraction and effort distortion (via local incentive-compatibility constraints). Moral hazard introduces new features through binding global incentive compatibility constraints. Some agents who exert low effort get positive bonuses because of their ability to mimic types who exert high effort. Moreover, because even some types at the boundary have binding global incentive compatibility constraints, the optimal mechanism typically features distortion at the boundary. This result contrasts with the "no distortion at the boundary" result from multidimensional screening when local incentive constraints are sufficient.

Our analysis can be extended in two ways. First, the dual approach used on the optimal taxation model naturally leads to a Rawlsian planner. In order to work with a utilitarianist planner, one needs to consider an ex-ante participation constraint, which we leave for further work. Second, since the principal's program is not concave and involves a continuum of intermediate constraints, it is unlikely that a solution will in general be attainable without applying a numerical method. We believe that developing such a method could provide additional insights into the properties of optimal mechanisms.

Appendix

A Pure Moral Hazard and Pure Adverse Selection

In this appendix, we study the mechanisms when either effort or conditional probabilities are observable. We refer to the situation where effort is observable as *pure adverse selection*, and to the one where conditional probabilities are observable as *pure moral hazard*. The main result is that the first best can be implemented under pure adverse selection but not under pure moral hazard (unless having all types exert the lowest effort is first-best efficient or agents are risk neutral). Moreover, the principal's payoff under joint adverse selection and moral hazard is strictly lower than under pure moral hazard. Therefore, adverse selection alone does not entail any payoff loss for the principal, although combining it with moral hazard further reduces the principal's payoff.[32]

[32]Our results contrast with the ones from Caillaud, Guesnerie, and Rey (1992) and Picard (1987), who study a model in which risk-neutral agents have (one-dimensional) private information about their cost of effort. In their setting, the principal can achieve the same utility as in the absence of noise (pure adverse selection). Therefore, the moral hazard dimension does not entail any additional loss for the principal in their model, whereas pure adverse selection does.

A.1 Pure Moral Hazard

There is a continuum of agents in the population with different productivities: $\mathbf{p} \in \bar{\Delta}$ is distributed according to the probability distribution function f with full support. Unlike the model from Section 2, assume that the principal observes the agents' productivities but still cannot monitor their efforts.

Assume that if the principal could monitor the agents' types, it would be optimal to have a non-empty set of agents exerting high effort:[33]

$$\Delta x > u^{-1}(c_1) - u^{-1}(c_0). \tag{20}$$

Following Grossman and Hart (1983), it is straightforward to characterize the optimal mechanism. In the optimal mechanism, types that exert high effort and have different conditional probability of success p_1 get different contracts (since the principal extracts the full surplus). All types who exert low effort get the same contract which gives them utility $u^{-1}(c_0)$. The high-effort region is non-empty under condition (20), since the principal recommends high effort from types in a neighborhood of $\mathbf{p} = (0, 1)$.

Since the optimal mechanism in the case of simultaneous moral hazard and adverse selection is also feasible under pure moral hazard (but it is not optimal), the principal obtains a strictly higher profit under pure moral hazard than under simultaneous moral hazard and adverse selection (as long as the high effort region is non-empty). Moreover, as long as the agent is risk averse, the principal's expected payoff is strictly lower in the pure moral hazard model than in the first best model.

A.2 Pure Adverse Selection

This subsection considers the case of pure adverse selection. We assume that the principal is able to monitor the agent's effort but cannot observe his conditional probability of each outcome given effort. In order to stress that the implementability of the first-best under pure adverse selection does not rely on the assumptions of two effort levels or two outcomes, we will consider a framework that generalizes of the model from Section 2.

A risk-neutral principal faces an agent who may be either risk-neutral or risk-averse. The agent exerts effort $e \in \mathbf{E}$, which is *observable* by the principal. The principal also observes output $x \in \mathbf{X}$. The effort and output spaces $\mathbf{E}$ and $\mathbf{X}$ are compact and non-empty subsets of the Euclidean spaces $\mathbb{R}^N$ and $\mathbb{R}^M$. Let $c(e)$ denote the agent's cost of effort e.

Each agent's type is a set of conditional distributions of outcomes given efforts $\{p(.|e) : \mathbf{X} \to \mathbb{R} | e \in \mathbf{E}\}$. This formulation allows for infinite-dimensional types. However, when there are two outcomes and two effort levels, the framework becomes the two-dimensional model of Section 2. More generally, when $\mathbf{E}$ and $\mathbf{X}$ are both finite, a type can be represented by a matrix of conditional probabilities. In this case, types have dimension $(m-1) \times n$, where m is the number of outcomes and n is the number of effort levels. Let $\mathbf{P}$ denote the space of possible types. The principal's beliefs about the agent's private information are represented by the cumulative distribution function F on $\mathbf{P}$.[34]

[33] If this condition did not hold, the first-best and the second-best solutions would coincide and all agents would exert low effort. Moreover, if agents are risk averse, the unique solution would involve paying a constant salary in both states of the world.

[34] Note that we are not imposing MLRP or full support, although the results are still true under these assumptions.

A direct mechanism $\{w_{\mathbf{p}}(x), e(\mathbf{p}) : \mathbf{p} \in \mathbf{P}, x \in \mathbf{X}\}$ specifies a payment function $w_{\mathbf{p}}(.) : \mathbf{X} \to \mathbb{R}$ and a recommended effort $e(\mathbf{p})$ for each type $\mathbf{p}$. The participation and free disposal constraints (IR) and (FD) are analogous to the ones from Section 2:

$$\int_{\mathbf{X}} u\left(w_{\boldsymbol{p}}(x)\right) p\left(dx|e\right) - c\left(e\left(\mathbf{p}\right)\right) \geq 0, \tag{IR}$$

$$x \geq \hat{x} \implies w_{\boldsymbol{p}}(x) \geq w_{\boldsymbol{p}}(\hat{x}), \tag{FD}$$

for all $\mathbf{p}, \hat{\mathbf{p}} \in \mathbf{P}$ and $x, \hat{x} \in \mathbf{X}$, where the first inequality in (FD) represents vector inequality.

The incentive-compatibility constraints require each agent type to take his own contract. However, since effort is observable, the agent cannot exert a different effort than the one recommended by the principal for the type for which the contract is designed. Thus, the incentive-compatibility constraints in the pure adverse selection model are:

$$\int_{\mathbf{X}} u\left(w_{\boldsymbol{p}}(x)\right) p\left(dx|e\right) - c\left(e\left(\mathbf{p}\right)\right) \geq \int_{\mathbf{X}} u\left(w_{\hat{\boldsymbol{p}}}(x)\right) \hat{p}\left(dx|e\right) - c\left(e\left(\hat{\mathbf{p}}\right)\right), \tag{IC AS}$$

for all $\mathbf{p}, \hat{\mathbf{p}} \in \mathbf{P}$.

The principal's expected utility equals expected output minus payments:

$$\int_{\mathbf{P}} \int_{X} \left[x - w_{\mathbf{p}}(x)\right] p\left(dx|e\right) dF\left(\mathbf{p}\right).$$

A mechanism satisfying (IC AS), (IR), and (FD) is called a *feasible mechanism for the pure adverse selection model*. A mechanism is *first-best optimal* if it maximizes the principal's expected utility subject to (IR). A mechanism is *optimal for the pure adverse selection model* if it maximizes the principal's expected utility within the class of feasible mechanisms for the pure adverse selection model. The following proposition establishes that the principal is able to obtain the first-best payoff when effort is observable:

Proposition 13. *Any optimal mechanism for the pure adverse selection model is equivalent to a first-best optimal mechanism.*

Proof. In any first-best optimal mechanism, the participation constraint must bind for almost every type. Therefore, for any first-best optimal mechanism there exists an equivalent mechanism in which the participation constraint binds for all types. Fix one such mechanism and let $e(\mathbf{p})$ denote the effort exerted by type $\mathbf{p}$ in this mechanism.

Consider the mechanism $(\tilde{w}, e)$ where $\tilde{w}_{\boldsymbol{p}}(x) = c\left(e\left(\mathbf{p}\right)\right)$ for all $\mathbf{p}$. This mechanism satisfies (IC AS) and satisfies (IR) with equality. Moreover, since the payments are constant in outcomes, it also satisfies (FD). Therefore, it implements the first best. □

Therefore, we can rank the principal's and agent's payoffs in the models of the pure adverse selection, pure moral hazard and simultaneous moral hazard and adverse selection considered in the text. The principal attains the first-best payoff under pure adverse selection, which is the highest attainable profit. She attains a strictly lower payoff in the case of pure moral hazard as long as the first-best contract does not implement low effort for all types (condition 20) and agents are risk averse, and an even lower payoff in the case of joint moral hazard and adverse selection.

The agent obtains the same payoff under both pure adverse selection and moral hazard (his reservation utility). However, all types with projections above t^* obtain payoffs strictly above their reservation utilities in the case of joint adverse selection and moral hazard (see Figure 5).

References

Acemoglu, D. (1998): "Credit market imperfections and the separation of ownership from control," *Journal of Economic Theory*, 78(2), 355–381.

Armstrong, M. (1996): "Multiproduct Nonlinear Pricing," *Econometrica*, pp. 51–75.

Bajari, P., H. Hong, and A. Khwaja (2011): "A Semiparametric Analysis of Adverse Selection and Moral Hazard in Health Insurance Contracts," Discussion paper, Working Paper,.

Boadway, R., M. Marchand, P. Pestieau, and M. del Mar Racionero (2002): "Optimal Redistribution with Heterogeneous Preferences for Leisure," *Journal of Public Economic Theory*, 4(4), 475–498.

Caillaud, B., R. Guesnerie, and P. Rey (1992): "Noisy Observation in Adverse Selection Models," *Review of Economic Studies*, 59(3), 595–615.

Carroll, G. (2013): "Robustness and Linear Contracts," Working paper, Microsoft Research and Stanford University.

Chade, H., and E. Schlee (2012): "Optimal Insurance with Adverse Selection," *Theoretical Economics*.

Chassagnon, A., and P.-A. Chiappori (1997): "Insurance under Moral Hazard and Adverse Selection: The Case of Pure Competition," *DELTA-CREST Working Paper*.

Chassang, S. (2011): "Calibrated Incentive Contracts," *Princeton University Economic Theory Center Working Paper*, (013-2011).

Chiu, W. H., and E. Karni (1998): "Endogenous Adverse Selection and Unemployment Insurance," *Journal of Political Economy*, 106(4), 806–827.

Choné, P., and G. Laroque (2010): "Negative Marginal Tax Rates and Heterogeneity," *American Economic Review*, 100(5), 2532–47.

Cremer, H., P. Pestieau, and J.-C. Rochet (2001): "Direct versus Indirect Taxation: the Design of the Tax Structure Revisited," *International Economic Review*, 42(3), 781–800.

De Meza, D., and D. C. Webb (2001): "Advantageous Selection in Insurance Markets," *RAND Journal of Economics*, pp. 249–262.

Diamond, P. A. (1998): "Optimal Income Taxation: an example with a U-shaped pattern of optimal marginal tax rates," *American Economic Review*, pp. 83–95.

——— (2005): *Taxation, Incomplete Markets, and Social Security*. MIT press.

Diamond, P. A., and J. A. Mirrlees (1971): "Optimal Taxation and Public Production I: Production Efficiency," *American Economic Review*, pp. 8–27.

Diamond, P. A., and J. Spinnewijn (2011): "Capital Income Taxes with Heterogeneous Discount Rates," *American Economic Journal: Economic Policy*, 3(4), 52–76.

Ebert, U. (1992): "A Reexamination of the Optimal Nonlinear Income Tax," *Journal of Public Economics*, 49(1), 47–73.

Edmans, A., and X. Gabaix (2011): "Tractability in Incentive Contracting," *Review of Financial Studies*, 24(9), 2865–2894.

Einav, L., A. Finkelstein, S. P. Ryan, P. Schrimpf, and M. R. Cullen (2013): "Selection on Moral Hazard in Health Insurance," *American Economic Review*, 103(1).

Grossman, S. J., and O. D. Hart (1983): "An Analysis of the Principal-Agent Problem," *Econometrica*, 51(1), 7–45.

Holmstrom, B., and P. Milgrom (1987): "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica*, pp. 303–328.

Innes, R. D. (1990): "Limited Liability and Incentive Contracting with Ex-Ante Action Choices," *Journal of Economic Theory*, 52(1), 45–67.

Judd, K., and C.-L. Su (2006): "Optimal Income Taxation with Multidimensional Taxpayer Types," Discussion paper, Working Paper.

Jullien, B., B. Salanie, and F. Salanie (2007): "Screening Risk-Averse Agents under Moral Hazard: Single-crossing and the CARA case," *Economic Theory*, 30(1), 151–169.

Kadan, O., P. J. Reny, and J. M. Swinkels (2011): "Existence of Optimal Mechanisms in Principal-Agent Problems," Discussion paper, Working Paper.

Karlan, D., and J. Zinman (2009): "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment," *Econometrica*, 77(6), 1993–2008.

Kleven, H. J., C. T. Kreiner, and E. Saez (2009): "The Optimal Income Taxation of Couples," *Econometrica*, 77(2), 537–560.

Laffont, J.-J., and D. Martimort (2002): *The Theory of Incentives*. Princeton University Press.

Laffont, J.-J., E. Maskin, and J.-C. Rochet (1987): *Optimal Nonlinear Pricing with Two-Dimensional Characteristics*pp. 256–266. University of Minnesota Press, Minneapolis.

Laffont, J.-J., and J. Tirole (1986): "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, pp. 614–641.

——— (1993): *A Theory of Incentives in Procurement and Regulation*. MIT press.

Maskin, E., and J. Riley (1984): "Monopoly with Incomplete Information," *The RAND Journal of Economics*, 15(2), 171–196.

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

——— (1972): "On Producer Taxation," *Review of Economic Studies*, pp. 105–111.

——— (1990): "Taxing Uncertain Incomes," *Oxford Economic Papers*, pp. 34–45.

Mussa, M., and S. Rosen (1978): "Monopoly and Product Quality," *Journal of Economic Theory*, 18(2), 301–317.

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

——— (1982): "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," *Journal of Mathematical Economics*, 10(1), 67–81.

Picard, P. (1987): "On the Design of Incentive Schemes under Moral Hazard and Adverse Selection," *Journal of Public Economics*, 33(3), 305–331.

Piketty, T. (1997): "La Redistribution Fiscale Face au Chômage," *Revue française d'économie*, 12(1), 157–201.

Piketty, T., and E. Saez (2012): "Optimal Labor Income Taxation," in *Handbook of Public Economics*, ed. by A. Auerbach, R. Chetty, and M. S. Feldstein, vol. 5. Amsterdam: Elsevier-North Holland.

Poblete, J., and D. Spulber (2012): "The Form of Incentive Contracts: Agency with Moral Hazard, Risk Neutrality, and Limited Liability," *RAND Journal of Economics*, 43(2), 215–234.

Rochet, J.-C. (1987): "A necessary and sufficient condition for rationalizability in a quasi-linear context," *Journal of Mathematical Economics*, 16(2), 191–200.

Rochet, J.-C., and P. Choné (1998): "Ironing, Sweeping, and Multidimensional Screening," *Econometrica*, pp. 783–826.

Rochet, J.-C., and L. A. Stole (2002): "Nonlinear Pricing with Random Participation," *The Review of Economic Studies*, 69(1), 277–311.

——— (2003): *Advances in Economics and Econometrics: Theory and Applications - Volume 1*chap. The Economics of Multidimensional Screening. Econometric Society Monographs.

Rothschild, C., and F. Scheuer (2012): "Optimal Taxation with Rent-Seeking," Working paper, Middlebury College and Stanford University.

——— (2013): "Redistributive Taxation in the Roy Model," *Quarterly Journal of Economics*, Forthcoming.

Rothschild, M., and J. Stiglitz (1976): "Equilibrium in Competitive Insurance Markets: An essay on the economics of imperfect information," *Quarterly Journal of Economics*, pp. 629–649.

Saez, E. (2001): "Using Elasticities to Derive Optimal Income Tax Rates," *Review of Economic Studies*, 68(1), 205–229.

Seade, J. K. (1977): "On the Shape of Optimal Tax Schedules," *Journal of Public Economics*, 7(2), 203–235.

Stewart, J. (1994): "The Welfare Implications of Moral Hazard and Adverse Selection in Competitive Insurance Markets," *Economic inquiry*, 32(2), 193–208.

Stiglitz, J. E. (1977): "Monopoly, Non-linear Pricing and Imperfect Information: the Insurance Market," *The Review of Economic Studies*, pp. 407–430.

——— (1982): "Self-Selection and Pareto Efficient Taxation," *Journal of Public Economics*, 17(2), 213–240.

Tarkiainen, R., and M. Tuomala (1999): "Optimal Nonlinear Income Taxation with a Two-Dimensional Population: A computational approach," *Computational Economics*, 13(1), 1–16.

Tenhunen, S., and M. Tuomala (2010): "On Optimal Lifetime Redistribution Policy," *Journal of Public Economic Theory*, 12(1), 171–198.