A Mechanism Design Approach to Climate Agreements[*]

David Martimort[†] and Wilfried Sand-Zantman[‡]

Preliminary draft: September 21, 2011

Abstract: We analyze international environmental agreements in contexts with asymmetric information, voluntary participation by sovereign countries and possibly limited enforcement. Taking a mechanism design perspective, we study how countries can agree on effort levels and compensations to take into account multilateral externalities. We delineate conditions for efficient agreements and trace out possible inefficiencies to the different conjectures that countries might hold following disagreement. We show how optimal mechanisms admit simple approximations with attractive implementation properties. Finally, we also highlight how limits on commitment strongly hinder performances of optimal mechanisms.

Keywords: Asymmetric information, public goods, global warming.

JEL Codes: D82.

1 Introduction

Designing optimal climate-change policies is certainly the most challenging issue for today's generation. To give some clue on the size of the problem at stake, some scientists have indeed argued that, with emissions currently around 370 ppm of carbon dioxide, our goal should be to stabilize that number around 550 ppm rather than at the *business as usual* level that would bring us up to 750 ppm by the end of the century.[1]

Taking a Coasian perspective, some economists are prone to suggest a strikingly simple solution to this major problem. Under strong assumptions - complete information, absence of transaction costs, perfect enforceability of contractual arrangements-

[*] We thank workshop participants at Paris School of Economics and especially Gabrielle Demange, Jérôme Pouyet and Franois Salanié for helpful comments. We also thank Daniel Coublucq for outstanding research assistance. All errors are ours.

[†] Paris School of Economics-EHESS. Email: david.martimort@parisschoolofeconomics.fr

[‡] Toulouse School of Economics (GREMAQ-IDEI). Email: wsandz@cict.fr.

[1] See for instance Helm (2005, p.1), Cohen (1995), McNeill (2000).

efficient outcomes would emerge from environmental negotiations. If anything, the recent record of such negotiations from Montreal, to Kyoto, Copenhague and Cancun meetings and their repeated failures demonstrate that efficiency remains by and large out of reach.

To better fit the actual institutional context surrounding climate policy negotiations and give some hints on what sort of optimal treaty is actually feasible, we analyze in this paper environmental agreements in a more realistic context that entails asymmetric information, voluntary participation by sovereign countries and possibly limits on enforcement.

Main ingredients of the model. We consider a large number (in fact a continuum) of heterogenous countries which exert efforts to reduce pollution emissions. Countries have private information on their abatement costs (or, in an alternative interpretation on their opportunity costs of exerting depolluting efforts). The effort of a given country has both a local and a global positive impacts so that exerting effort is a public good. Environmental agreements- or mechanisms in the parlance of the incentives literature- consist in binding commitments to effort levels (for instance a commitment to some quantities for carbon emissions like in most real-world treaties) and monetary compensations contingent on those efforts.

Free-riding. In such multilateral externalities context, countries may free ride in providing efforts. As pointed out by Chander and Tulkens (2008), free-riding takes actually two different forms. First, countries may exaggerate their abatements costs, undersupply effort and leave most of the burden of cost abatements on others. Second, countries may also opt out of those negotiations and nevertheless enjoy the benefits of the agreements that others may have reached in the mean time.

The first form of free-riding has received much attention in information economics since the seminal works of Samuelson (1954), Groves and Ledyard (1977) and Green and Laffont (1977). This line of research has certainly culminated with the work of Laffont and Maskin (1982), Mailath and Postlewaite (1990) in general environments whereas Rob (1989), Neeman (1999) and Baliga and Maskin (2003) have developed specific applications targeted to environmental economics. In those papers, ensuring incentive compatibility and budget balance while satisfying no-veto constraints might generate underprovision (or even no provision at all of the public good).

The second form of free-riding results from the main specificities of environmental negotiations. In this standard mechanism design paradigm, each agent has the right to veto the mechanism with no provision being the fall-back option. However, imposing a no-veto constraint is by and large improper in the context of climate change agreements. Indeed, climate negotiations are voluntary agreements between sovereign countries. If any country refuses to participate, others may just band together in a smaller coalition or just adopt an individualistic behavior. Therefore, once it considers leaving the negotiation table, a country should form conjectures on how others will react. Incentives to free-ride by not participating certainly depend on those conjectures.

The standard mechanism design approach for public good provision inherited from the existing literature seems thus ill-equipped to tackle these specificities. In this paper, we revamp the conflict between individual incentives, ensuring budget balance for the mechanism and satisfying the agents' participation constraints when those participation constraints explicitly take into account conjectures about the behavior of non-deviating agents. This will give a fresh look at the various institutional problems that a good design of climate-change policies must solve.

In this respect, two conjectures are particularly attractive. First, once an agent decides leaving a negotiation, he may conjecture that others will agree on choosing inefficiently low levels of efforts; a *Worst-Punishments* scenario. Not participating in the coalition is thus very costly. In such contexts, allocative efficiency is always preserved even under asymmetric information.

Suppose instead that, whenever an agent refuses the agreement, the coalition fully disbands and non-deviating agents just play individual best responses which leads to the *business as usual* outcome. Of course, the *Bayesian-Nash* equilibrium so obtained gives to each individual country type-dependent reservation payoffs which are above those achieved under the *Worst-Punishments* scenario. It becomes harder to reconcile incentive compatibility, budget balance and participation.

Constrained mechanisms. With a Bayesian-Nash scenario as a fall-back option, the optimal agreement depends on the size of the global externality. For a sufficiently significant externality, efficiency remains feasible despite asymmetric information. If the externality decreases, some inefficiency arises. To limit free riding in preferences revelation, the effort levels of the least efficient countries are reduced while at the same time the most efficient ones are subsidized to better internalize the externality. Yet (almost) all countries are strictly better off joining the agreement. Decreasing further the size of the externality, a whole set of inefficient countries get the same payoff with the agreement than with the fall-back option and exert the same effort level. This group subsidizes the most efficient countries' efforts but benefits from the global impact of the latter's greater effort levels.[2]

Enforcement. Casual evidence suggest that international treaties may not come with all facilities and monitoring devices that are required to enforce the quotas committed by ratifying parties. From a more theoretical viewpoint, the feasibility of any agree-

[2] The characterization of such regime is made technically complex by the addition of a type-dependent participation constraint to a mechanism design problem under budget balance. We rely on and adapt techniques developed in Martimort and Stole (2011) to tackle those issues.

ment depends on the possibility to credibly enforce punishments following deviations by some agents. Those deviations may concern participation to the mechanism or its mere playing.

Suppose first that whenever a country deviates, others are still bound by the initial mechanism.[3] The treaty entails no punishment proviso. Free-riding in participation is exacerbated. Strikingly, no mechanism can then outperform the Bayesian-Nash outcome in such context.

Consider instead the possibility that non-deviating countries react following a deviation by themselves choosing effort levels to punish the non-participating country. Those effort levels should nevertheless be credible in the sense that non-deviating agents should find it optimal to carry those threats. The treaty incorporates only some compliance mechanism if it is itself credible. In this context, we again show that this "self-referential" criterion selects only the Bayesian-Nash outcome. The last two scenarios shows thus that moving away from the *business as usual* scenario may be very difficult without much commitment power.

Lastly, we investigate the countries' incentives to abide to an agreement after acceptance. Indeed, internal political pressures at reelections time, lobbying, incentives to foster short-term growth may all push governments to cheat on agreements. In other words, on top of assuming that countries can just not join the agreement, we thus look at their incentives to abide by the mechanism once accepted. Those considerations lead us to introduce an *enforcement constraint* which is harder to satisfy than the participation constraints considered earlier on. In those environments, inefficiencies are more pronounced and the set of countries which are just indifferent between joining in or not expands. International treaties thus suffer from the lack of credibility of joining countries.

Literature review. The existing literature on climate negotiations has stressed the possible failures in reaching global agreements. The focus is on conditions for reaching efficiency while at the same time requiring the grand-coalition to be robust to secessions. To tackle those issues, Chandler and Tulkens (1995, 1997) introduced the notion of γ-core for such economies and defined the worth of a coalition, assuming that countries outside the coalition play individual best responses. They demonstrated that the grand-coalition is feasible despite individual incentives to free-ride in participation. The take-away from this research is that, in those complete information environments, efficiency may be compatible with a grand-coalition forming. We borrow from this contribution the concern on the role played by conjectures on the level of participation constraints. However, in our context with private information, efficiency is far less

[3] Given our assumption of a large population of countries possibly joining in, this assumption is akin to assuming that individual deviations are just non-observable. This might be quite reasonable in a context of limited capabilities to audit actual emissions reductions.

easy to reach.

Another important line of research (Carraro and Siniscalco 1993, 1995; Barrett 1994) has instead focused on incentives to form coalitions by imposing external and internal stability criterions similar to those developed in cartel theory earlier on. Subsequent research in the field (Carraro, 2005) has then stressed the importance of various institutional rules to ensure participation, stability, and solve the free-riding problem. One important feature of this literature is that institutional constraints are imposed at the outset and not derived from primitives. Such approach stands thus in sharp contrast with the mechanism design literature that precisely derives optimal institutions from primitives - well-specified informational constraints and strategic behavior.[4] Our mechanism design approach is, by tradition, more normative and, by construction, does not leave much room for discussing the exact details of the negotiation process. Nevertheless, we show how the optimal mechanism can be approximated in practice by means of simple menus of contracts.

The commitment issue for environmental treaties has also received some (rather informal) attention in the literature. Barrett (2003) points out that the Kyoto Protocol did not incorporate any compliance mechanism and that parties could refuse to ratify without further being punished.[5] This description nicely echoes our modeling of the enforcement limits which suggests that very little beyond the *business as usual* outcome can be achieved in a world plagued with informational asymmetries. Helm, Hepburn and Mash (2005) are instead concerned with the credibility of domestic policies towards reducing pollution, especially with the incentives of governments to implement lax carbon policies in the short run for electoral concerns. These authors advocate for setting up an independent agency, most likely an international one, to whom policies enforcement would be delegated to solve a time-inconsistency problem.[6] Our mechanism design approach departs from such incentives problem due to time-inconsistency to put asymmetric information upfront as the source of inefficiency. It also relies implicitly on the use of a mediator (or and international external agency) who monitors and enforces, possibly under some observability constraints, depolluting efforts made by treaty members.

Organization of the paper. Section 2 presents the model and solves for the complete information benchmark. Section 3 first describes incentive feasible allocations. Second, it qualifies conditions ensuring that efficiency can be achieved even in a context with

[4] This stability program was developed in a complete information framework and often assumed away the possible heterogeneity between countries. On the difficulties in reaching agreements among heterogenous countries in a complete information setting, especially in view of the problem of ensuring participation, see Thoron (2008).

[5] See also Schelling (2002) and Victor (2001).

[6] Other authors, like for instance Guesnerie (2008) have proposed mechanisms that also rely on an International Bank for Emissions Allowance Acquisition.

private information and type-dependent participation constraints, allowing either for the Bayesian-Nash scenario or the Worst-Punishment outcome following a deviation. Focusing on the Bayesian-Nash scenario, Section 4 then develops the various kinds of inefficiencies that free-riding on preferences manipulation might induce. A particular attention is given to developing simple instruments that could be used in practice to implement this optimal allocation. Section 5 investigates various limits on the commitment of parties to the mechanism. Section 6 concludes and highlights a few alleys for further research. Proofs are in an Appendix.

2 The Model

Preferences and technology. We consider a continuum of countries of unit mass (sometimes referred to as "agents" in the sequel) who undertake activities mitigating pollution emissions. By exerting a pollution mitigating effort e_i, country i generates two kinds of benefits. The first ones of size αe_i (where $\alpha \in [0,1)$) are purely *local* and accrue only to country i. The second sort of benefits are *global*, worth $(1-\alpha)e_i$ and accrue to all countries worldwide. As α varies from zero to one, efforts go from having pure global to pure local consequences.

Countries differ according to their marginal cost of exerting effort. For tractability, we choose a quadratic formulation and assume that the disutility of effort writes as $C(e_i, \theta_i) = \frac{e_i^2}{2\theta_i}$. Cost convexity captures the fact that emissions cannot be reduced too much without impairing the basic functioning of the economy for instance by imposing technological changes and adjustments that are increasingly harder to implement. Countries with higher values of θ_i are the most efficient at undertaking those activities.

We can then write country i's payoff as:

$$U_i = t_i + \alpha e_i + (1-\alpha)\int_j e_j dj - \frac{e_i^2}{2\theta_i}.$$

The payment t_i stands for any financial compensation (taxes or subsidies) that this country may receive for undertaking the requested effort and $(1-\alpha)\int_j e_j dj$ represents the "aggregate" effort taken over the whole population of countries.[7] Given our normalization to a unit mass, this quantity is nothing else than the average effort worldwide.

Remark 1 *Monetary payments may in fact be given a broader interpretation and be viewed as*

[7] An a priori alternative formulation of the objective would be $t_i + \alpha e_i + \beta \int_j e_j dj - \frac{e_i^2}{2\theta_i}$ for some non-negative α and β. Normalizing by $\alpha+\beta$ and changing θ_i into $\theta_i(\alpha+\beta)$ gives us our posited formulation. The latter has the benefit of keeping the first-best unchanged as α changes making comparative statics significantly simpler.

the benefits or costs that countries withdraw when climate negotiations are linked to negotiations on other issues such as technology transfers, trade agreements and so on.[8]

Information. The efficiency parameters θ_i are independently drawn from the same cumulative distribution $F(\cdot)$ with support $\Theta = [\underline{\theta}, \bar{\theta}]$ (with $\underline{\theta} > 0$) and everywhere positive and atomless density $f(\theta) = F'(\theta)$. Let denote by $E_\theta(\cdot)$ the expectation operator with respect to the law of θ. We also impose the following monotonicity condition that will ensure monotonicity of effort at the optimal mechanism under asymmetric information.

Assumption 1

$$\frac{d}{d\theta}\left(\frac{1 - F(\theta)}{\theta f(\theta)}\right) \leq 0 \quad \forall \theta \in \Theta.^{9}$$

Country i has private information on its efficiency parameter θ_i although its effort in mitigating pollution is observable. Therefore, countries cannot receive payments conditional on the realization of this efficiency parameter although efforts can be contractually specified and rewarded.

Remark 2 *Our model can also be applied in complete information settings, i.e., when costs are common knowledge, but in contexts where no mechanism conditional on the countries' identity can be written.*[10]

Finally, the following assumption requires that the externality is not too strong. We will show below that when the externality is large enough, an efficient allocation can still be implemented even under asymmetric information.

Assumption 2

$$\alpha > \alpha_1 = \frac{\underline{\theta}}{2E_{\tilde{\theta}}(\tilde{\theta}) - \underline{\theta}} \in (0, 1).$$

[8] Barrett (2005) stresses the role of linking together various policies to achieve better treaties.

[9] Any distribution (uniform, exponential, truncated normal...) satisfying the more common monotonicity of the hazard rate $\frac{d}{d\theta}\left(\frac{1-F(\theta)}{f(\theta)}\right) \leq 0$ (see Bagnoli and Bergstrom, 2005) satisfies our weaker Assumption 1.

[10] This concern for a non-discriminatory design was pushed forward by the Bush administration to justify its withdrawal from the 2001 Kyoto protocol by calling the treaty unfair for industrialized countries vis-à-vis developing countries.

Assumption 2 certainly holds when the parameter α is close enough to one (weak externality) or when uncertainty on the productivity type θ is large enough.

Mechanisms and incentive compatibility. Because of asymmetric information, payments and effort levels must be *incentive compatible*. We turn now to the description of such incentive compatibility allocations. Consider thus direct revelation mechanisms of the form $\{t(\hat{\theta}), e(\hat{\theta})\}_{\hat{\theta}\in\Theta}$ that determine compensations and effort levels as a function of a country's announcement $\hat{\theta}$ on his cost parameter. By the Revelation Principle,[11] there is no loss of generality in considering such direct and truthful revelation mechanisms.[12]

Following a truthful strategy, a country with type θ exerts an effort $e(\theta)$. We may thus rely on the Law of Large Numbers and identify the average global benefits of the countries' efforts with its expected value, i.e., $(1-\alpha)\int_j e_j dj = (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta}))$. With that remark in mind, we define the equilibrium payoff $U(\theta)$ of a country with type θ as:

$$U(\theta) = t(\theta) + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\theta)}{2\theta}.$$

From incentive compatibility, we immediately get:

$$U(\theta) = \max_{\hat{\theta}\in\Theta} t(\hat{\theta}) + \alpha e(\hat{\theta}) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\hat{\theta})}{2\theta}. \tag{1}$$

In the sequel, we shall repeatedly rely on a more compact **(dual)** characterization of incentive compatibility by using the rent $U(\theta)$ instead of the payment $t(\theta)$. Together with an effort level, an allocation is thus a pair $(U(\theta), e(\theta))$.

Budget balance. On top of those incentive constraints, a mechanism must also satisfy some feasibility conditions. Assuming that no external source of funds is available, i.e., the mechanism is self-financed, the following budget balance condition must hold:

$$E_{\tilde{\theta}}(t(\tilde{\theta})) \leq 0.$$

It will be often useful to rewrite this constraint as:

$$E_{\tilde{\theta}}\left(e(\tilde{\theta}) - \frac{e^2(\tilde{\theta})}{2\tilde{\theta}}\right) \geq E_{\tilde{\theta}}\left(U(\tilde{\theta})\right). \tag{2}$$

This condition expresses the fact that the overall expected surplus generated by the countries' effort should be at least equal to their expected payoff. Of course, this condition turns out to be an equality (no waste of resources) for optimal mechanisms in all circumstances below.

[11] Myerson (1982).

[12] In particular, those mechanisms replace any nonlinear payment schedule $T(e)$ that would map observable effort levels into compensations. See Section 4.2 for the shape of those optimal nonlinear schedules.

Participation constraints. Finally, the mechanism must satisfy a set of participation constraints to ensure that all countries join in. Those participation constraints depend on the commitment ability of the coalition, i.e., to what extent a coalition can commit to actions in case any of its members deviates. Much of our analysis throughout the paper will consist in studying how the mechanism design problem changes as we consider alternative fall-back options.

Two alternative assumptions are analyzed in the remainder of the paper.

Bayesian-Nash outcome. Suppose that the whole coalition breaks down following an individual deviation where one agent refuses the mechanism. The corresponding fall-back option is thus the (symmetric) Bayesian-Nash equilibrium (thereafter BNE) where countries non-cooperatively choose their effort levels. Let denote by $U_N(\theta)$ the payoff of a country with type θ in such BNE. By definition, we have

$$U_N(\theta) = \max_e \alpha e - \frac{e^2}{2\theta} + (1-\alpha)E_{\tilde{\theta}}(e_N(\tilde{\theta}))$$

where the Bayesian-Nash level of effort $e_N(\tilde{\theta})$ is

$$e_N(\theta) = \arg\max_e \alpha e - \frac{e^2}{2\theta} + (1-\alpha)E_{\tilde{\theta}}(e_N(\tilde{\theta})) = \alpha\theta.^{13}$$

Because a given country does not internalize the impact of its own effort on other countries' welfare, efforts are under-provided at the BNE outcome: a well-identified instance of a positive externality.

Immediate computations lead to the following expression of payoffs in the BNE fall-back option:

$$U_N(\theta) = \frac{\alpha^2}{2}\theta + (1-\alpha)\alpha E_{\tilde{\theta}}(\tilde{\theta}).$$

Worst-punishment outcome. Alternatively, suppose that the coalition can still specify what non-deviating agents do whenever a given country deviates and refuses the mechanism. Choosing an effort level $e_N(\theta)$ is still a dominant strategy for that deviating country irrespectively of what the non-deviant ones can enforce. The *worst punishment* is however obtained when the non-deviating agents just exert no effort. This yields a lower payoff to the deviating agent:

$$U_W(\theta) = \frac{\alpha^2}{2}\theta.$$

Given that countries know their efficiency parameter at the time of deciding whether to join the treaty or not, we write the corresponding *ex post* participation constraints as:

$$U(\theta) \geq U_l(\theta) \quad \forall \theta \in \Theta \text{ and } l = N, W. \tag{3}$$

[13] Thanks to our separability assumption between returns from local and global benefits, non-deviating countries choose the same effort level whatever their beliefs on the deviant (and negligible) country as long as they revert to a non-cooperative behavior.

Complete information benchmark. Suppose that the countries' efficiency parameters are common knowledge. Type-dependent instruments can be used to fix efforts at their target levels and compensate countries for those efforts according to the exact cost they incur. Incentive compatibility constraints are not an issue in this ideal world.

We are interested in mechanisms that maximize aggregate welfare, i.e.,

$$(\mathcal{P}^F): \quad \max_{U(\cdot),e(\cdot)} E_{\tilde{\theta}}(U(\tilde{\theta})) \text{ subject to (2) and (3).}$$

The budget-balance condition (2) is binding and aggregate welfare is maximized for the first-best level of effort

$$e^{FB}(\theta) = \theta \quad \forall \theta \in \Theta.$$

Finally, to show that the ex post participation constraints (3) are satisfied, consider the following payment schedule:

$$\begin{aligned} \tilde{t}_l^{FB}(\theta) &= U_l(\theta) - \alpha e^{FB}(\theta) - (1-\alpha)E_{\tilde{\theta}}(e^{FB}(\tilde{\theta})) + \frac{(e^{FB}(\theta))^2}{2\theta} \\ &= \begin{cases} (1-\alpha)^2 \left(\frac{\theta}{2} - E_{\tilde{\theta}}(\tilde{\theta})\right) & \text{if } l = N \\ (1-\alpha)^2 \frac{\theta}{2} - (1-\alpha)E_{\tilde{\theta}}(\tilde{\theta}) & \text{if } l = W. \end{cases} \end{aligned}$$

These payments ensure that the type θ-country is just indifferent between its fall-back option (whether it is the Bayesian-Nash or the Worst-Punishment outcome) and receiving the payment $\tilde{t}^{FB}(\theta)$ while exerting the first-best effort $e^{FB}(\theta)$. This scheme is feasible since we have:

$$0 < -E_\theta(\tilde{t}_l^{FB}(\theta)) = \begin{cases} \frac{(1-\alpha)^2}{2} E_{\tilde{\theta}}(\tilde{\theta}) & \text{if } l = N \\ \frac{(1-\alpha^2)}{2} E_{\tilde{\theta}}(\tilde{\theta}) & \text{if } l = W. \end{cases}$$

Let us now construct a system of payments $t^{FB}(\theta)$ that are budget-balanced and implement the efficient effort levels as follows:

$$t_l^{FB}(\theta) \quad = \quad \tilde{t}_l^{FB}(\theta) - E_{\tilde{\theta}}(\tilde{t}_l^{FB}(\tilde{\theta}))$$

By construction, the budget-balance condition (2) is satisfied as an equality. Moreover, and also by construction, $t_l^{FB}(\theta) > \tilde{t}_l^{FB}(\theta)$ so that all countries now get a payoff which is strictly greater than their fall-back option:

$$U^{FB}(\theta) = U_l(\theta) - E_\theta(\tilde{t}_l^{FB}(\theta)) > U_l(\theta).$$

Lastly, note that $t_l^{FB}(\theta)$ is increasing with θ. Indeed, the most efficient countries are asked much higher levels of effort in the first-best situation than in the BNE or the Worst-Punishment situations. Without compensation, the most efficient countries are better off choosing their Nash effort levels than the first-best levels. To induce participation, it is thus necessary to compensate efficient types by taxing less efficient ones.

3 Asymmetric Information and Efficiency

Incentive compatibility. Consider now the case where the countries' efficiency parameters are private information Incentive compatibility constraints should then be added to characterize feasible allocations. Next lemma further characterizes those incentive constraints.

Lemma 1 *An allocation $(U(\theta), e(\theta))$ is incentive compatible if and only if:*

1. *$U(\theta)$ is absolutely continuous with at each differentiability point (i.e., almost everywhere)*
$$\dot{U}(\theta) = \frac{e^2(\theta)}{2\theta^2}. \tag{4}$$

2. *$e(\theta)$ is non-decreasing and thus almost everywhere differentiable with at each differentiability point*
$$\dot{e}(\theta) \geq 0. \tag{5}$$

It is standard to neglect the monotonicity condition (5) and obtain a relaxed optimization problem whose solution satisfies that extra condition. We can then rewrite the so relaxed second-best optimization problem as:

$$(\mathcal{P}^{SB}): \quad \max_{U(\cdot), e(\cdot)} E_{\tilde{\theta}}(U(\tilde{\theta})) \text{ subject to (2), (3) and (4).}$$

Conditions for efficiency. As a preliminary step, we investigate whether the efficient allocation can still be implemented under asymmetric information. Of course, the answer depends on how stringent the participation constraint is. We obtain below a less optimistic result with BNE than with the Worst-Punishment outcome.

Proposition 1 *Under asymmetric information, the first-best allocation $(U^{FB}(\theta), e^{FB}(\theta))$*

1. *cannot be implemented when the fall-back option is BNE and Assumption 2 holds;*

2. *can always be implemented when the fall-back option is the Worst-Punishment outcome.*

To understand the first item, one must analyze the impact of α, on both participation and incentive constraints. Consider first the participation problem. When the parameter α is small, positive externalities are significant and the cost of disagreement is high. This relaxes participation constraints and eases cooperation. However, on the incentives side, with small α, countries do not care much about the local impact of their effort provision. Avoiding such “free-riding” requires larger payments to stimulate provision. When α is small enough, there is enough gains from cooperation to compensate for the incentive cost. The first-best allocation can be implemented.

When α is instead large enough, the global effect is small. Countries choose efforts close to their efficient level even when behaving non-cooperatively. The gains from cooperation are small. Although there is less “free-riding”, the gains from cooperation are too small to allow first-best implementation.

Turning now to the case of the Worst-Punishments scenario, observe that the fall-back option entails zero effort by non-deviating agents. This makes the gains from cooperation very large allowing to implement the first best even with incentive constraints.[14] Note that the Worst-Punishments outcome relies on a strong ability to commit for non-deviating agents. It is not immune to further deviations by non-deviating agents who may prefer to enforce at least their Nash effort levels. We investigate those commitment issues in Section 5.

4 Second-Best Mechanisms

4.1 Characterization

First, we characterize second-best allocations when the first best is no longer feasible and the fall-back option is the BNE outcome. Inefficiencies depend on the tension between incentive compatibility, participation and budget balance.

We distinguish two scenarios. In the first one, all countries except the less efficient ones strictly gain from joining the mechanism. This arises when the inefficiency is rather mild and the gains from cooperation rather large. In the second scenario, only a strict subset of countries strictly prefer joining in. Inefficiencies are more pronounced.

To describe the properties of these scenarios, let us define the variable $\zeta^*(\alpha)$ as

$$\zeta^*(\alpha) = \frac{1}{1 - \frac{1-\alpha}{\alpha}\underline{\theta} f(\underline{\theta})}. \tag{6}$$

Observe that $\zeta^*(\alpha)$ is decreasing with α and that $1 - \frac{1-\alpha}{\alpha}\underline{\theta} f(\underline{\theta}) > 0$ (hence $\zeta^*(\alpha) > 1$ holds) when the externality is not too big, namely,

$$\alpha > \alpha_2 = \frac{1}{1 + \frac{1}{\underline{\theta} f(\underline{\theta})}}. \tag{7}$$

Furthermore, we impose:

[14] This result is reminiscent of other works in the mechanism design literature (Makowski and Mezzetti 1994, Williams 1999, Krishna and Perry 2000, and Schweizer 2006) although these papers study Bayesian environments with a finite number of players.

Assumption 3

$$\alpha_2 \leq \alpha_1 \Leftrightarrow E_{\tilde{\theta}}(\tilde{\theta}) \leq \underline{\theta} + \frac{1}{f(\underline{\theta})}.$$

Assumption 3 simplifies the number of cases relevant for our analysis without loss of economic insight.[15]

Let us also define an effort schedule $\bar{e}(\theta, \zeta)$ and a critical type $\theta^*(\zeta)$ which are both parameterized by some number $\zeta \geq 1$ such that:

$$\bar{e}(\theta, \zeta) = \frac{\theta}{1 + \frac{\zeta - 1}{\zeta} \frac{1 - F(\theta)}{\theta f(\theta)}} \tag{8}$$

and

$$\begin{cases} \frac{1-F(\theta^*(\zeta))}{\theta^*(\zeta) f(\theta^*(\zeta))} = \frac{1-\alpha}{\alpha} \frac{\zeta}{\zeta - 1} & \text{if } \zeta \geq \zeta^*(\alpha) \\ \theta^*(\zeta) = \underline{\theta} & \text{if } \zeta \in [1, \zeta^*(\alpha)). \end{cases} \tag{9}$$

Define now $\hat{\zeta}$ as the unique solution[16] to

$$\int_{\underline{\theta}}^{\theta^*(\zeta)} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta + \int_{\theta^*(\zeta)}^{\bar{\theta}} \left(\bar{e}(\theta, \zeta) - \frac{\bar{e}^2(\theta, \zeta)}{2\theta} \left(1 + \frac{1 - F(\theta)}{\theta f(\theta)} \right) \right) f(\theta) d\theta$$

$$= \int_{\underline{\theta}}^{\theta^*(\zeta)} U_N(\theta) f(\theta) d\theta + U_N(\theta^*(\zeta))(1 - F(\theta^*(\zeta))). \tag{10}$$

Constraint (10) is obtained by aggregating incentive, participation and budget-balance constraints altogether[17] and $\hat{\zeta}$ is in fact the multiplier of this constraint in our maximization problem. This condition states that total welfare has to be fully redistributed while keeping incentive compatibility and inducing participation.[18]

Incentive compatibility explains the extra informational distortion (proportional to $\frac{1-F(\theta)}{\theta f(\theta)}$ on the left-hand side of (10)). Inducing effort profiles closer to the first best is now costly because it exacerbates the incentives of the most efficient countries to pretend being less efficient.

Participation constraints impose that rent profiles are above the BNE outcome. Nevertheless, as those constraints bind on an interval $\Omega^c = [\underline{\theta}, \theta^*(\zeta)]$ (which might be reduced to a single point), the BNE effort and rent profiles are found respectively on the left-hand side of condition (10) which evaluates total welfare and on the right-hand side which measures expected payoffs.

[15] It is for instance satisfied by the uniform distribution to which we will refer later on.

[16] The proof of uniqueness is found in the Appendix.

[17] The techniques behind such approach of consolidating the various constraints that define incentive-feasible allocations into a single one are well-known in Bayesian environments from the works of Laffont and Maskin (1982), Myerson and Sattherwaite (1983) and Mailath and Postlewaite (1990) among others.

[18] From Assumption 2, this aggregate feasibility constraint is already known to be binding otherwise the first-best allocation could be implemented.

Distortion regimes. We are now ready to describe two regimes of distortions that depend on the value of the multiplier $\hat{\zeta}$ associated to condition (10).

Proposition 2 *Suppose that the fall-back option is BNE and that Assumption 3 holds. There exist $\alpha_3 \in (\alpha_1, 1)$ and $\alpha_4 \in (\alpha_1, \alpha_3)$ that define two different profiles of payoffs at the optimal mechanism.*

1. ***Weak distortions.*** *For $\alpha \in [\alpha_1, \alpha_4]$, $\hat{\zeta} \in (1, \zeta^*(\alpha)]$.*

2. ***Strong distortions.*** *For $\alpha \in (\alpha_3, 1)$, $\hat{\zeta} > \zeta^*(\alpha)$.*

The intuition for those distortions is better understood when thinking of α as being close enough to α_1, i.e., small enough while still keeping Assumption 2 satisfied. In that case, the efficiency gains from coordinating effort levels are rather strong but yet not large enough to allow implementation of the first-best. Nevertheless, we expect rather small distortions, or more formally, the parameter $\hat{\zeta}$ should be close to one.

When α increases, the gains from coordination are lower and asymmetric information has more bite. Distortions are stronger and the multiplier $\hat{\zeta}$ increases.

Rents profile. Depending on the scenario, the rents profile has different shapes which are described in the next proposition.

Proposition 3 *Suppose that the fall-back option is BNE and that Assumptions 1, 2 and 3 hold together. The second-best profile of rents $\bar{U}(\theta)$ is such that the participation constraint (3)*

1. *is binding only at $\underline{\theta}$ when $\hat{\zeta} \leq \zeta^*(\alpha)$;*

2. *is binding on a non-empty interior interval $\Omega^c = [\underline{\theta}, \theta^*(\hat{\zeta})]$ when $\hat{\zeta} > \zeta^*(\alpha)$.*

Efforts profile. Turning now to the characterization of effort levels, we get:

Proposition 4 *Suppose that the fall-back option is BNE and that Assumptions 1, 2 and 3 hold together. The second-best profile of effort levels $\bar{e}(\theta)$ is continuous, increasing in θ, weakly greater than the BNE outcome but downward distorted below its first-best level everywhere except at $\bar{\theta}$. More precisely:*

1. *If $\underline{\theta} = \theta^*(\hat{\zeta})$, then*
$$\bar{e}(\theta) = \bar{e}(\theta, \hat{\zeta}) > e_N(\theta) \quad \forall \theta \in \Theta; \tag{11}$$

2. *If $\underline{\theta} < \theta^*(\hat{\zeta})$, then*
$$\bar{e}(\theta) = \begin{cases} \bar{e}(\theta, \hat{\zeta}) > e_N(\theta) & \textit{if } \theta \in \Omega = (\theta^*(\hat{\zeta}), \bar{\theta}] \\ e_N(\theta) & \textit{if } \theta \in \Omega^c = [\underline{\theta}, \theta^*(\hat{\zeta})]. \end{cases} \tag{12}$$

When Assumption 2 holds, we already know that efficiency cannot be achieved. One cannot find incentive compatible payments that implement efficient effort levels and that give all types strictly more than their BNE payoffs. The participation constraint (3) is binding somewhere.

Under asymmetric information, the most efficient types (such that $\theta \in \Omega = (\theta^*(\hat{\zeta}), \bar{\theta}]$) have now some incentives to claim being less efficient and produce less effort than requested by the mechanism. Those types want to "free-ride" by producing less effort when playing the mechanism. By doing so, those efficient types get some rent above their BNE payoffs. That rent is the amount of money that these efficient types can save by producing the same effort as some less efficient types but at a lower marginal cost.

To limit those incentives, the optimal mechanism relies on two kinds of distortions. First, effort is reduced below the efficient level for all types except the most efficient one. This first distortion reduces how much can be saved by the most efficient types by mimicking a slightly less efficient one. Second, the mechanism taxes more the least efficient types to make their allocation less attractive for the most efficient ones. This second distortion might push the least efficient types to opt out of the mechanism and play non-cooperatively. Participation constraints are binding on the lower tail of the types distribution, possibly at a single point or on a whole interval.

In other words, solving the "free-riding" problem on information revelation for the most efficient types exacerbates "free-riding" on participation by the least efficient ones. There is thus a significant conflict between the most efficient types' incentives to truthful reveal and the least efficient types' incentives to participate.

Payments profile. Observe that at any differentiability point of the payment schedule, the incentive compatibility condition (1) also implies the following relationship between payments and efforts:

$$\dot{\bar{t}}(\theta) = \frac{\dot{\bar{e}}(\theta)}{\theta} \left(\bar{e}(\theta) - e_N(\theta)\right). \tag{13}$$

From Proposition 4, it follows that $\bar{t}(\cdot)$ is strictly increasing on $(\theta^*(\hat{\zeta}), \bar{\theta}]$ and constant on $[\underline{\theta}, \theta^*(\hat{\zeta})]$ if such interval has a non-empty interior. From the fact that the budget-balance constraint (2) is binding at the optimum, it also follows that

$$\bar{t}(\underline{\theta}) < 0 < \bar{t}(\bar{\theta}).$$

In other words, the least efficient countries always pay for joining the coalition even though they get the same payoff in and out. They are ready to pay exactly the benefit they receive from the greater effort exerted by the most efficient types.

In particular, for large inefficiencies (i.e., when $\hat{\zeta} > \zeta^*$), any type in the interval $[\underline{\theta}, \theta^*(\hat{\zeta})]$ pays a tax

$$\bar{t}(\theta) = -(1-\alpha) \int_{\theta^*(\hat{\zeta})}^{\bar{\theta}} (\bar{e}(\theta) - e_N(\theta)) f(\theta) d\theta < 0.$$

Indeed, when such country deviates and opts out of the coalition, the most efficient countries with types $\theta \in (\theta^*(\hat{\zeta}), \bar{\theta}]$ react by producing their BNE effort level which is strictly less than the effort requested by the mechanism. This punishment reduces the overall payoff of the deviating agent by an amount which is exactly the tax

$$(1-\alpha)\int_{\theta^*(\hat{\zeta})}^{\bar{\theta}} (\bar{e}(\theta) - e_N(\theta)) f(\theta) d\theta.$$

4.2 Implementation in Practice

Much has already been written on the kind of instruments that can be used in practice to implement good climate-change policies.[19] Pursing this tradition, this section investigates how the optimal mechanism found in Section 4 can be implemented in practice. Our analysis reveals that a simple two-items menu that specifies either a fixed contribution or a subsidy per unit of effort may perform pretty well to approximate the optimal mechanism.

Convexity of the optimal payment schedule. Our analysis above demonstrated that $\bar{e}(\theta)$ is an increasing function of θ when Assumption 1 holds. Hence, we may define the inverse mapping $\bar{\theta}(e)$ on the relevant interval and a nonlinear payment schedule that implements the optimal allocation as:

$$T(e) = \bar{t}(\bar{\theta}(e)) = \int_{\underline{\theta}}^{\bar{\theta}(e)} \frac{\bar{e}^2(x)}{2x^2} dx - \alpha e + \frac{e^2}{2\bar{\theta}(e)} - (1-\alpha) E_{\tilde{\theta}}(\bar{e}(\tilde{\theta})).$$

Proposition 5 *$T(e)$ is flat for $e \leq e_N(\theta^*(\hat{\zeta}))$, strictly increasing and convex for $e > e_N(\theta^*(\hat{\zeta}))$.*

Interestingly, one can check that $T'(\bar{e}(\bar{\theta})) = 1 - \alpha \geq T'(\bar{e}(\theta))$ for all θ. In other words, the marginal incentives to expand effort for the most efficient types make those types fully internalize the impact of their effort on global welfare. Less efficient types are less rewarded at the margin and do not expand effort as much.

Approximation. The convexity of $T(e)$ makes it a good candidate to be approximated by a pair of simple linear schemes.[20] With the first one, countries only pay up-front a fixed amount $\underline{T}$ and continue to exert their Bayesian-Nash effort level. The second linear payment entails both a greater fixed up-front contribution $\overline{T} > \underline{T}$ but also a subsidy $1-\alpha$ per unit of abatement so that the first-best effort is exerted by types who choose this scheme. Initial contributions cover the expected subsidies needed to ensure budget balance.

[19] See for instance Bradford (2008) and Guesnerie (2008) among others.

[20] For a similar insight in other contexts (respectively regulation and nonlinear pricing), see also Rogerson (2003) and Wilson (1993).

Let us denote by θ^* the cut-off type just indifferent between those two schemes. By incentive compatibility and single-crossing, types below θ^* choose their Bayesian-Nash effort while those above choose the efficient effort. This leads us to the following indifference condition for θ^*:

$$e^{FB}(\theta^*) - \frac{(e^{FB}(\theta^*))^2}{2\theta^*} - \overline{T} + (1-\alpha)\left(\int_{\underline{\theta}}^{\theta^*} e_N(\tilde{\theta})d\tilde{\theta} + \int_{\theta^*}^{\bar{\theta}} e^{FB}(\tilde{\theta})d\tilde{\theta}\right)$$

$$= \alpha e_N(\theta^*) - \frac{e_N^2(\theta^*)}{2\theta^*} - \underline{T} + (1-\alpha)\left(\int_{\underline{\theta}}^{\theta^*} e_N(\tilde{\theta})d\tilde{\theta} + \int_{\theta^*}^{\bar{\theta}} e^{FB}(\tilde{\theta})d\tilde{\theta}\right).$$

Simplifying, we obtain:

$$\overline{T} = \underline{T} + (1-\alpha^2)\frac{\theta^*}{2}. \tag{14}$$

To ensure participation of the least efficient types, the upfront contribution they make must just balance the externality gain created by the extra effort of types above θ^*. This extra effort being $e^{FB}(\theta) - e_N(\theta) = (1-\alpha)\theta$, the expected externality on types below θ^* becomes $(1-\alpha)^2 \int_{\theta^*}^{\overline{\theta}} \theta f(\theta)d\theta$ which gives the following expression for $\underline{T}$:

$$\underline{T} = (1-\alpha)^2 \int_{\theta^*}^{\overline{\theta}} \theta f(\theta)d\theta. \tag{15}$$

Finally, the menu must satisfy a budget-balance condition where the expenses are the subsidies per unit of effort given to the most efficient agents and the resources the lump-sum contributions paid by both groups, namely:

$$F(\theta^*)\underline{T} + (1-F(\theta^*))\overline{T} = (1-\alpha)\int_{\theta^*}^{\overline{\theta}} \theta f(\theta)d\theta. \tag{16}$$

Using the expressions of $\overline{T}$ and $\underline{T}$ drawn from (14) and (15) and inserting into (16), θ^* is implicitly defined as a solution to the following equation (for $\alpha < 1$):

$$\mathcal{J}(\theta^*) = \frac{\theta^*}{2}(1-F(\theta^*))(1+\alpha) - \alpha \int_{\theta^*}^{\overline{\theta}} \theta f(\theta)d\theta = 0. \tag{17}$$

Remark first that $\theta^* = \overline{\theta}$ is a solution and that $\mathcal{J}'(\overline{\theta}) < 0$. Moreover, Assumption 1 implies that $\mathcal{J}(\cdot)$ is quasi-concave and there are thus at most two solutions to (17). More precisely, note that $\mathcal{J}(\underline{\theta}) > 0$ if and only if $\alpha \leq \alpha_1$. Therefore, for $\alpha \leq \alpha_1$, $\theta^* = \underline{\theta}$, the first best is always implemented with a single linear contract of slope $1-\alpha$ and we recover our previous findings. On the contrary, for $\alpha > \alpha_1$, we then have $\theta^* \in (\underline{\theta}, \overline{\theta})$, and the type space is nicely split into two connected subsets taking different contracts.

Numerical application. One may now wonder how significant is the welfare loss from using the simple two-item menu above instead of the optimal nonlinear mechanism.

As the following numerical simulations show, the loss is surprisingly small and therefore the two-item menu turns to be a good approximation of the optimal mechanism.

Let us characterize the optimal contract and its two-item approximation for a uniform distribution on $\Theta = [1, 2]$. For this particular specification, we find $\alpha_1 = \alpha_2 = .5$. Moreover, tedious computations show that $\alpha_3 = \alpha_4 = .726$. Following the terminology of Proposition 2, we will take $\alpha = .65$ and $\alpha = .85$ to respectively illustrate the cases of *weak* and *strong distortions*.[21]

- For *weak distortions*, i.e., $\alpha = .65$, we know that $\theta^*(\hat{\zeta}) = \underline{\theta} = 1$. Moreover, computations lead to $\hat{\zeta} = 1.397$ so that the optimal effort is everywhere given by

  $$\bar{e}(\hat{\zeta}, \theta) = \frac{\theta^2}{.792\theta + .416}.$$

  From this, we compute that the aggregate welfare under the optimal mechanism is roughly equal to 0.367.

  If a two-item menu is instead offered, (17) yields $\theta^* = 1.300$, i.e., the thirty percent least efficient countries pay the lower amount $\underline{T}$. Equations (14) and (15) yield then

  $$\underline{T} = .190 \text{ and } \bar{T} = .565.$$

  Finally, the aggregate welfare achieved with such menu is roughly worth 0.328. Comparing with the optimal mechanism, the welfare loss from using the simple menu is 10.7 percent which is admittedly small and that loss must be put beside the significantly simpler design of the two-item menu.

- For *strong distortions*, i.e., $\alpha = .85$, we know that $\theta^*(\hat{\zeta}) > 1$. Computations lead to $\hat{\zeta} = 1.779$ and $\theta^*(\hat{\zeta}) = 1.425$. The optimal effort is everywhere given by

  $$\bar{e}(\hat{\zeta}, \theta) = \begin{cases} \frac{\theta^2}{.557\theta + .886} & \text{if } \theta \in (1.425, 2] \\ .85\theta & \text{if } \theta \in [1, 1.425]. \end{cases}$$

  This corresponds to a value of the aggregate welfare under the optimal mechanism which is roughly equal to 0.380.

  If a two-item is instead offered, (17) yields $\theta^* = 1.700$, i.e., the thirty percent most efficient countries pay the higher tax $\overline{T}$ and receive subsidies per unit of abatements. Equations (14) and (15) yields then

  $$\underline{T} = .012 \text{ and } \bar{T} = .247.$$

  The aggregate welfare achieved with such menu is approximatively equal to 0.373. Now, the welfare loss from using the menu instead of the optimal mechanism is less than 2 percent; a surprisingly small loss indeed.

[21] Because $\alpha_3 = \alpha_4 = .726$, we span here all possible configurations.

These numerical examples testify that even a complex mechanism design problem can be handled efficiently with coarse instruments whose implementability in practice is much simpler.

Even though our simple menu above does not perfectly fit any existing real-world mechanism, it lends itself into a nice and realistic interpretation. Suppose that developing countries face lower marginal opportunity costs of reducing pollution because they just do not produce as much as developed countries. Those countries, which are well-equipped to contribute to the public good provision, self-select on the higher powered incentive scheme. They exert effort above their *"business as usual"* level, get subsidized for that, but contribute to fund this program through significant ex ante lump sum taxes.*A contrario,* the more developed countries face higher opportunity costs of reducing pollution and do not expand effort beyond the *"business as usual"* level. *Per capita,* those countries contribute much less through ex ante contributions to global funding but, as our examples illustrate, the fraction of countries that self-select on the fixed payment scheme may be significant.

Our mechanism nevertheless bears some strong resemblance with other proposals, most noticeably the so-called *Global Public Good Purchase* pushed forward by Bradford (2008). In Bradford's mechanism, countries make a set of voluntary contributions to an international agency; this agency buys then any reduction below the *business as usual* allowances. In the mechanism we precisely describe below, countries choose between two possible initial contributions that are pocketed by an agency. Some of those countries choose a larger contribution but will also receive a subsidy for any effort made in reducing pollution. Those subsidies are themselves paid back by from the agency's budget. Note that, under asymmetric information, self-selection imposes that countries choosing different voluntary contributions receive different subsidies.

5 Credibility

In this section, we investigate the impact of various assumptions on the treaty members' commitment ability to punish countries which do not join in.

5.1 No Commitment

A first possibility is that the treaty (or mechanism) cannot credibly impose any threat to coordinate the behavior of non-deviating countries following a deviation. This typically arises when the identity of a deviating agent cannot be detected and the mechanism remains in force for all non-deviating agents as well.

The type-dependent participation constraint becomes:

$$U(\theta) \geq \frac{\alpha^2}{2}\theta + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})), \quad \forall \theta \in \Theta \tag{18}$$

where the expected effort level on the right-hand side is simply the one prescribed by the mechanism itself.

Here, all opportunities to punish deviations are lost and incentives to free-ride in not participating to the agreement are maximal. Leaving the agreement does not change the aggregate effort but avoids paying any contribution. We show that, in this limit setting, an incentive compatible mechanism cannot achieve anything beyond the Bayesian-Nash outcome.

Proposition 6 *When the mechanism designer cannot commit to any inefficient threat, the only feasible allocation is the BNE outcome.*

5.2 Credible Mechanisms

The BNE and the Worst-Punishment outcomes have some common features. They not only rely on the designer's ability to enforce threats but also on his ability to detect deviations. On the other hand, those two fall-back options have also very different features. First, the BNE outcome is credible as non-deviating agents choose effort levels that are not only best responses to the deviating agent but also best responses to each other. In other words, the BNE outcome is itself a Bayesian incentive compatible mechanism which, although very crude, is robust to further deviations.

Instead, in the Worst-Punishment scenario, non-deviating agents are committed to out-of equilibrium effort levels which are no longer best responses. The corresponding payoffs are excessively low compared to what non-deviating agents would get by further deviating and choosing their non-cooperative effort levels. This mechanism relies on incredible threats. We should of course cast serious doubts on the possibility of using such threats. This raises in turn the question of finding a mechanism relying on credible threats.

A credible effort profile $e_C(\cdot)$ should be the minimal expected effort that can be implemented by a mechanism given that agents playing such mechanism may themselves further deviate and be punished for such deviation by other agents still coordinating on the credible profile itself.[22] In particular, any agent playing that credible mechanism following an earlier deviation expects that the mechanism would still be enforced if he further deviates himself.

That consistency requirement can be expressed with the following type-dependent participation constraints:

$$U(\theta) \geq \frac{\alpha^2}{2}\theta + (1-\alpha)E_{\tilde{\theta}}(e_C(\tilde{\theta})), \quad \forall \theta \in \Theta. \tag{19}$$

[22] Our notion of credibility shares some features with other concepts developed in game theoretic contexts: for instance, the (non-cooperative) notion of *coalition-proofness* equilibrium pushed forward by Bernheim, Peleg and Whinston (1987) and the (more cooperative) notion of *binding agreements* due to Ray and Vohra (1997).

where $e_C(\cdot)$ must itself solve the following problem

$$(\mathcal{P}_C): \quad \max_{U(\cdot),e(\cdot)} -E_\theta(e(\theta)) \text{ subject to (2), (4) and (19).}$$

To obtain sharp predictions on this problem, we shall impose the following extra assumption.

Assumption 4

$$\frac{d}{d\theta}\left(\frac{F(\theta)}{\theta f(\theta)}\right) \geq 0 \quad \forall \theta \in \Theta.$$

Equipped with this assumption, we obtain the following characterization.

Proposition 7 *The only credible effort profile is the BNE outcome.*

Taken together, Propositions 6 and 7 provide a very pessimistic view of what can be achieved by environmental treaties under limited commitment. Only the *business as usual* outcome might emerge.

5.3 Limited Enforcement

The optimal mechanism characterized in Section 4 has some surprising features, especially when the participation constraint is binding on a non-empty interval $\Omega^c = [\underline{\theta}, \theta^*(\hat{\zeta})]$. Indeed, types in that interval produce the Bayesian-Nash effort both if they join the mechanism and if they don't. This makes the mechanism particularly sensitive to an enforcement problem as once those types have already chosen their effort level, they could just choose not to pay the tax and free-ride on the most efficient types.

We model this limited enforcement problem as in Laffont and Martimort (2002, Chapter 9) and view the agent's decision of paying back money to the mechanism as a moral hazard variable that can nevertheless partially controlled. If an agent does not contribute once he has already chosen his effort, he can be punished with some probability $\delta < 1$. Punishments following a deviation are non-monetary and consist for non-deviating agents in returning to their BNE effort levels.

An agent abides to the mechanism whenever the following moral hazard incentive constraint holds:

$$U(\theta) \geq (1-\delta)\left(-\frac{e^2(\theta)}{2\theta} + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta}))\right) + \delta U_N(\theta). \tag{20}$$

This *enforcement constraint* (20) can be also written as:

$$t(\theta) \geq \frac{\delta}{1-\delta}(U_N(\theta) - U(\theta)). \tag{21}$$

Payments cannot be too low without inducing a deviation. The enforcement constraint certainly holds for the most efficient agents who are subsidized by the mechanism, receive positive transfers and get more than their BNE payoff.

However, the *enforcement constraint* is harder to satisfy than the usual participation constraint. In other words, countries find it more attractive accepting the mechanism and then not paying taxes when necessary than refusing the mechanism right away. To see why consider the optimal mechanism that we characterized in Section 4. For such mechanism, a type $\underline{\theta}$ pays a tax, $\bar{t}(\underline{\theta}) < 0$, and $U_N(\underline{\theta}) = U(\underline{\theta})$ which leads to an immediate contradiction with (21).

Remark 3 *Although our analysis does not rely on a full-fledged modeling of the dynamics of the relationship, this enforcement constraint admits of course an interpretation in terms of repeated games. Everything happens as if parties were committed to a stationary mechanism that covers an infinite number of periods with a discount factor δ. Types are stationary and drawn once for all.*[23] *The mechanism defines a repeated game with per-period payoff $U(\theta)$ for the agents. Payments have to be enforceable and whenever an agent does not repay, non-deviating agents play trigger strategies and the BNE outcome in the continuation.*

From a technical viewpoint, the enforcement constraint (20) is complex because it is *mixed*, involving both the state variable $U(\theta)$, the control $e(\theta)$ and its average value $E_{\tilde{\theta}}(e(\tilde{\theta}))$. Next lemma simplifies the analysis by replacing the enforcement constraint with an *a priori* stronger constraint that only depends on $E_{\tilde{\theta}}(e(\tilde{\theta}))$.

Lemma 2 *Suppose that Assumption 1 holds and that the enforcement constraint (20) is binding on an interval Ω^c with a non-empty interior.*

- *Types in Ω^c exert efforts equal to their BNE level:*

$$e(\theta) = e_N(\theta) \quad \forall \theta \in \Omega^c. \tag{22}$$

- *The following constraint holds:*

$$U(\theta) \begin{cases} = U_N(\theta) + (1-\delta)(1-\alpha)(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))) & \textit{if } \theta \in \Omega^c \\ > U_N(\theta) + (1-\delta)(1-\alpha)(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))) & \textit{otherwise.} \end{cases} \tag{23}$$

This lemma allows us to replace (20) by the simpler constraint

$$U(\theta) \geq U_N(\theta) + (1-\delta)(1-\alpha)(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))). \tag{24}$$

Consider thus the case of a strong distortion as exemplified in Proposition 2. We now want to characterize the optimal mechanism when replacing (3) by the more stringent condition (24). We expect (24) to be binding on a lower tail interval. Next proposition summarizes the solution.

[23] See Baron and Besanko (1984) for instance

Proposition 8 *Assume limited enforcement and that α is large enough. There exists $\hat{\zeta} > 1$ such that the optimal mechanism such that (20) is binding on an interval $\theta \in \Omega^c = [\underline{\theta}, \theta^*(\hat{\zeta})]$. This mechanism implements an effort profile:*

$$\bar{e}(\theta) = \begin{cases} \left(1 - \frac{\hat{\zeta}-1}{\hat{\zeta}}(1-\delta)(1-\alpha)\right) \frac{\theta}{1+\frac{\hat{\zeta}-1}{\hat{\zeta}}\frac{1-F(\theta)}{\theta f(\theta)}} > e_N(\theta) & \textit{if } \theta \in \Omega = (\theta^*(\hat{\zeta}), \bar{\theta}] \\ e_N(\theta) & \textit{if } \theta \in \Omega^c = [\underline{\theta}, \theta^*(\hat{\zeta})] \end{cases} \tag{25}$$

where

$$\frac{1 - F(\theta^*(\hat{\zeta}))}{\theta^*(\hat{\zeta}) f(\theta^*(\hat{\zeta}))} = \frac{1-\alpha}{\alpha}\left(\frac{\hat{\zeta}}{\hat{\zeta}-1} - 1 + \delta\right). \tag{26}$$

Under limited enforcement, reducing the effort level of the most efficient agents and moving it closer to the BNE level relaxes the enforcement constraint (24). Comparing (25) with (11) shows that the effort is everywhere distorted downwards on the upper tail. Comparing (26) and (9) and using Assumption 1, we observe that $\theta^*(\hat{\zeta})$ is now greater. In other words, the area where the enforcement constraint binds is larger than with the weaker participation constraint. This captures again the increased inefficiency that arises due to limited enforcement.

6 Conclusion

This paper has investigated optimal environmental mechanisms in a context of asymmetric information, voluntary participation and limited enforcement. We have shown that the optimal mechanism has some simple features (indifference between joining in or not for countries facing the highest costs of effort, strict benefits for the most efficient countries) and can be implemented with strikingly simple menus of linear contracts. This simplicity at the implementation stage brings some optimistic view on how the climate-change problem could be solved in practice even under assumptions significantly less favorable than those used in the traditional Coasian perspective.

The least optimistic take-away from our analysis is that a strong commitment ability is also needed to achieve the welfare gains from an (eventually approximate) welfare maximizing mechanism. The compounding of asymmetric information and limited commitment may destroy all those gains and generates outcomes close to *business as usual*. This again suggests that setting up an International Agency with enough auditing and enforcement capabilities is an essential and unavoidable step towards solving the climate-change problem.

Equipped with the mechanism design methodology we have developed in this paper, a number of important other questions could be addressed in future research. Let us mention a couple of such problems. A first important extension should address the design of dynamic mechanisms. Under asymmetric information, dynamics introduces the well-known difficulties of the ratchet effect.[24] How incentive compatible mechanisms must be adapted on such contexts remains a fascinating question both from a theoretical and an applied viewpoint. In particular, one may want to assess the performance of menus of linear contracts in those dynamic environments. A second extension would be to go more deeply into the analysis of the relationship between local politics and international agreements. This requires to view the countries' objective functions no longer as those of unified entities but as the more complex outcome of domestic political games where lobbying and reelection concerns play a key role. The analysis of such two-tier mechanism design problem will be particularly fruitful to understand the climate-change problem.[25]

We hope to contribute to those extensions in the future.

References

Bagnoli, M. and T. Bergstrom, 2005, "Log-Concave Probability and its Applications," *Economic Theory*, 26: 445-469.

Baliga, S. and E. Maskin, 2003, "Mechanism Design for the Environment," in eds., *Handbook of Environmental Economics*, Elsevier.

Baron, D. and D. Besanko, 1984, "Regulation and Information in a Continuing Relationship," *Information Economics and Policy*, 84: 267-302.

Barrett, S., 1994, "Self-Enforcing International Environmental Agreements," *Oxford Economic Papers*, 46: 878-894.

Barrett, S., 2003, *Environment and Statecraft: The Strategy of Environmental Treaty-Making*, 25: 11-34.

Bernheim, D., B. Peleg, and M. Whinston, 1987, "Coalition-Proof Nash Equilibria I. Concepts," *Journal of Economic Theory*, 42: 1-12.

Bradford, D., 2008, "Improving on Kyoto: Greenhouse Gas Control as the Purchase of a Global Public Good", in R. Guesnerie and H. Tulkens eds. *The Design of Climate Policy*, MIT Press.

Carraro, C., 2005, "Institution Design for Managing Global Commons," in G. Demange and M. Wooders eds., *Group Formation in Economics*, Cambridge University Press.

[24] Freixas, Guesnerie and Tirole (1985).

[25] Those models could for instance borrow much to the framework for solving transnational public good problems which was initially proposed by Laffont and Martimort (2005).

Carraro, C. and D. Siniscalco, 1993, "Strategies for International Protection of the Environment," *Journal of Public Economics*, 52: 309-328.

Carraro, C. and D. Siniscalco, 1995, "International Coordination of Environmental Policies and Stability of Global Environmental Agreements," in L. Bovenberg and S. Cnossen eds., *Public Economics and the Environment in an Imperfect World*, Kluwer Academics.

Champsaur, P. and J.C. Rochet, 1989, "Multiproduct Monopolists," *Econometrica*, 57: 533-557.

Chander, P. and H. Tulkens, 1995, "A Core-Theoretic Solution for the Design of Cooperative Agreements and Transfrontier Pollution", *International Tax and Public Finance*, 2: 279-294.

Chander, P. and H. Tulkens, 1997, "The Core of an Economy with Multilateral Environment Exernalities", *International Journal of Game Theory*, 26: 379-401.

Chander, P. and H. Tulkens, 2008, "Cooperation, Stability, Self-Enforcement in Agreements", in R. Guesnerie and H. Tulkens eds., *The Design of Climate Policy*, MIT Press.

Cohen, J., 1995, *How Many People Can the Earth Support?*, Norton.

Cramton, P., R. Gibbons and P. Klemperer, 1987, "Dissolving a Partnership Efficiently," *Econometrica*, 55: 615-632.

Cramton, P. and T. Palfrey, 1995, "Ratifiable Mechanisms: Learning from Disagreement," *Games and Economic Behavior*, 10: 255-283.

Freixas, X., R. Guesnerie and J. Tirole, 1985, "Planning under Incomplete Information and the ratchet effect," *Review of Economic Studies*, 52: 173-191.

Galbraith, G and R. Vinter, 2004, "Regularity of Optimal Controls for State-Constrained Problems," *Journal of Global Optimization*, 28: 305-317.

Green, J. and J.J. Laffont, 1979, *Incentives in Public Decision-Making*, North Holland.

Groves, T. and J. Ledyard, 1977, "Optimal Allocation of Public Goods: A Solution to the Free-Riding Problem", *Econometrica*, 45: 783-810.

Guesnerie, R., 2008, "Design of Post-Kyoto Climate Schemes: Selected Questions in Analytical Perspective", in R. Guesnerie and H. Tulkens eds. *The Design of Climate Policy*, MIT Press.

Helm, D., 2005, "Introduction", in D. Helm, ed. *Climate-Change Policy*, Oxford University Press.

Helm, D., C. Helpburn and R. Mash, 2005, "Credible Carbon Policy", in D. Helm, ed. *Climate-Change Policy*, Oxford University Press.

Krishna, V. and M. Perry, 2000, "Efficient Mechanism Design," mimeo Penn State University.

Laffont, J.J. and D. Martimort, 2005, "The Design of Transnational Public Good Mechanisms for Developing Countries," *Journal of Public Economics*, 89: 159-196.

Laffont, J.J. and E. Maskin, 1982, "The Theory of Incentives: An Overview", in *Advances in Economic Theory*, ed. W. Hildenbrand. Cambridge University Press.

Makowski, L. and C. Mezzetti, 1995, "Bayesian and Weakly Robust First-Best Mechanisms: Characterizations", *Journal of Economic Theory*, 64: 500-519.

Martimort, D. and L. Stole, 2011, "Public Contracting in Delegated Agency Games," mimeo Paris School of Economics.

McNeill, J., 2000, *Something New under the Sun: An Environmental History of of the Twentieth Century*, Penguin Press.

Milgrom, P. and I. Segal, 2002, "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70: 583-601.

Myerson, R., 1982, "Optimal Coordination Mechanisms in Generalized Principal-Agent Models", *Journal of Mathematical Economics*, 10: 67-81.

Myerson, R. and M. Satterthwaite, 1983, "Efficient Mechanisms for Bilateral Trading", *Journal of Economic Theory*, 28: 265-281.

Neeman, Z., 1989, "Property Rights and Efficiency of Voluntary Bargaining under Asymmetric Information", *Review of Economic Studies*, 66: 679-691.

Ray, D. and R. Vohra, 1997, "Equilibrium Binding Agreements", *Journal of Economic Theory*, 73: 30-78.

Rob, R., 1989, "Pollution Claim Settlements under Private Information", *Journal of Economic Theory*, 47: 307-333.

Rogerson, W., 2003, "Simple Menus of Contracts in Cost-Based Procurement and Regulation," *American Economic Review*, 93: 919-926.

Schelling, S., 2002, "What Makes Greenhouse Sense?", *Foreign Affairs*, 81: 2-9.

Schweizer, U., 2006, "Universal Possibility and Impossibility Results", *Games and Economic Behavior*, 57: 73-85.

Thoron, S., 2008, "Heterogenity in Negotiations of International Agreements", in R. Guesnerie and H. Tulkens eds., *The Design of Climate Policy*, MIT Press.

Victor, D., 2001, *The Collapse of the Kyoto Protocol and the Struggle to Slow Global Warming*, Princeton University Press.

Williams, S., 1999, "A Characterization of Efficient, Bayesian Incentive Compatible Mechanisms", *Economic Theory*, 14: 155-180.

Wilson, R., 1993, *Nonlinear Pricing*, Oxford University Press.

Appendix

Proof of Lemma 1. From (1), we immediately obtain that $U(\theta)$ is the maximum of convex functions of θ and as such it is convex, absolutely continuous and thus almost everywhere differentiable.[26] Condition (4) follows at any such differentiability points.

Using simple revealed arguments, we get for $\theta \geq \hat{\theta}$

$$t(\theta) + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\hat{\theta})}{2\theta} \geq t(\hat{\theta}) + \alpha e(\hat{\theta}) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\hat{\theta})}{2\theta},$$

$$t(\hat{\theta}) + \alpha e(\hat{\theta}) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\hat{\theta})}{2\hat{\theta}} \geq t(\theta) + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\theta)}{2\hat{\theta}}$$

Summing on both sides and simplifying yields immediately

$$e(\theta) \geq e(\hat{\theta}).$$

Thus $e(\cdot)$ is monotonically increasing and thus a.e. differentiable.

Reciprocally, that $U(\cdot)$ is absolutely continuous implies that it can be written everywhere as:

$$U(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \frac{e^2(x)}{2x^2}dx = t(\theta) + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\theta)}{2\theta}.$$

From this, incentive compatibility immediately follows since:

$$t(\theta) + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\theta)}{2\theta} - \left(t(\hat{\theta}) + \alpha e(\hat{\theta}) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta})) - \frac{e^2(\hat{\theta})}{2\theta} \right)$$

$$= \int_{\hat{\theta}}^{\theta} \frac{e^2(x) - e^2(\hat{\theta})}{2x^2}dx \geq 0$$

when $e(\cdot)$ is weakly increasing. ■

[26] See Champsaur and Rochet (1989) or Milgrom and Segal (2002).

Proof of Proposition 1. An important step of the analysis consists in consolidating the incentive compatibility constraint (4) and the feasibility condition (2). In this respect, let define a *critical type* θ^* as:

$$\theta^* = \max \arg \min_{\theta \in \Theta} U(\theta) - U_l(\theta).$$

Of course, such critical type depends on the choice of the mechanism since it affects the profile of implementable rent $U(\theta)$. From continuity of $U(\theta) - U_l(\theta)$ and compacity of Θ, such θ^* necessarily exists for any implementable profile $U(\theta)$.

Note that satisfying the participation constraint (3) at θ^* is enough to have it satisfied for all θ. Hence a necessary and sufficient condition for (3) to hold is that

$$U(\theta^*) \geq U_l(\theta^*). \tag{A1}$$

Integrating (4) yields

$$U(\theta) = U(\theta^*) + \int_{\theta^*}^{\theta} \frac{e^2(x)}{2x^2} dx. \tag{A2}$$

Integrating by parts on each interval $[\underline{\theta}, \theta^*]$ and $[\theta^*, \bar{\theta}]$, we finally obtain the following expression of the average payoff of countries:

$$E_{\tilde{\theta}}(U(\tilde{\theta})) = U(\theta^*) + E_{\tilde{\theta}} \left(\frac{(1_{\tilde{\theta} \geq \theta^*} - F(\tilde{\theta}))e^2(\tilde{\theta})}{2\tilde{\theta}^2 f(\tilde{\theta})} \right)$$

where $1_{\tilde{\theta} \geq \theta^*} = \begin{cases} 1 & \text{if } \tilde{\theta} \geq \theta^* \\ 0 & \text{otherwise.} \end{cases}$

Finally, the feasibility condition can be rewritten as

$$E_{\tilde{\theta}} \left(e(\tilde{\theta}) - \frac{e^2(\tilde{\theta})}{2\tilde{\theta}} \right) \geq U(\theta^*) + E_{\tilde{\theta}} \left(\frac{(1_{\tilde{\theta} \geq \theta^*} - F(\tilde{\theta}))e^2(\tilde{\theta})}{2\tilde{\theta}^2 f(\tilde{\theta})} \right). \tag{A3}$$

Notice that any rent profile for a mechanism that implements the first-best effort level $e^{FB}(\theta)$ is such that $\underline{\theta}$ is the critical type since $U(\theta) - U_l(\theta)$ (for $l = N, W$) is increasing ($\dot{U}(\theta) - \dot{U}_l(\theta) = \frac{1-\alpha^2}{2} > 0$ when $\alpha < 1$). Hence, a necessary and sufficient condition for the participation constraint (3) to hold everywhere is that it holds at $\underline{\theta}$. That remark being made, the feasibility constraint and the critical type's participation constraint are altogether satisfied when:

$$E_{\tilde{\theta}} \left(e^{FB}(\tilde{\theta}) - \frac{(e^{FB}(\tilde{\theta}))^2}{2\tilde{\theta}} \right) \geq U_l(\underline{\theta}) + E_{\tilde{\theta}} \left(\frac{(1 - F(\tilde{\theta}))(e^{FB}(\tilde{\theta}))^2}{2\tilde{\theta}^2 f(\tilde{\theta})} \right).$$

This amounts to check

$$E_{\tilde{\theta}} \left(e^{FB}(\tilde{\theta}) - \frac{(e^{FB}(\tilde{\theta}))^2}{2\tilde{\theta}} \left(1 + \frac{1 - F(\tilde{\theta})}{\tilde{\theta} f(\tilde{\theta})} \right) \right) = \frac{1}{2} \int_{\underline{\theta}}^{\bar{\theta}} (\theta f(\theta) - 1 + F(\theta)) d\theta \geq U_l(\underline{\theta})$$

$$\Leftrightarrow \begin{cases} \frac{\bar{\theta}}{2} \geq \frac{\alpha^2}{2}\underline{\theta} + (1-\alpha)\alpha E_{\tilde{\theta}}(\tilde{\theta}) & \text{if } l = N \\ \frac{\bar{\theta}}{2} \geq \frac{\alpha^2}{2}\underline{\theta} & \text{if } l = W. \end{cases} \tag{A4}$$

Hence, when $l = N$, we get an impossibility if (2) holds. Instead, when $l = W$, (A4) holds and one can find budget-balanced transfers that ensure that the first-best is implemented. ■

Proofs of Propositions 3 and 4. We first characterize the optimal mechanism when Assumption 2 holds. The proof of Propositions 3 and 4 is a direct consequence of that characterization.

Neglecting the second-order condition (5) that will be checked ex post; we rewrite the so relaxed optimization problem under asymmetric information as:

$$(\mathcal{P}^{SB}) : \quad \max_{U(\cdot)\in W(\Theta), e(\cdot)} E_{\tilde{\theta}}(U(\tilde{\theta})) \text{ subject to (2), (3) and (4)}$$

where $W(\Theta)$ is the set of absolutely continuous arcs on Θ.

$(\mathcal{P}^{SB})$ is a generalized Bolza problem with an isoperimetric constraint (2) and a state-dependent constraint (3). We denote by ζ the non-negative multiplier of the former constraint. This allows us to write the Lagrangian for this problem as:

$$L(\theta, U, e, \zeta) = f(\theta)\left(U + \zeta\left(e - \frac{e^2}{2\theta} - U\right)\right).$$

Let then define the unmaximized Hamiltonian as

$$H(\theta, U, e, \zeta, q) = L(\theta, U, e, \zeta) + q\frac{e^2}{2\theta^2}.$$

This Hamiltonian is linear in U and strictly concave in e when

$$q \leq \xi\theta f(\theta). \tag{A5}$$

This latter condition is checked below for the optimal profile.
Necessity. Following Galbraith and Winter (2004), the necessary optimality conditions for this *State Maximum Problem* that are satisfied by a normal extremum $(\bar{U}(\theta), \bar{e}(\theta))$ can be written as follows.

Proposition A.1 ***Necessary conditions*** *(Galbraith and Winter, 2004). There exists an absolutely continuous function $p(\theta)$, a function $q(\cdot)$, and a non-negative measure μ which are all defined on Θ such that:*

$$-\dot{p}(\theta) = \frac{\partial H}{\partial U}(\theta, \bar{U}(\theta), \bar{e}(\theta), \zeta, q(\theta)), \tag{A6}$$

$$\bar{e}(\theta) \in \arg\max_{e\geq 0} H(\theta, \bar{U}(\theta), \bar{e}(\theta), \zeta, q(\theta)), \tag{A7}$$

$$q(\theta) = p(\theta) - \int_{\underline{\theta}}^{\theta^-} \mu(d\theta), \quad \forall \theta \in (\underline{\theta}, \bar{\theta}], \tag{A8}$$

$$supp\{\mu\} \subset \{\theta \; s.t. \; \bar{U}(\theta) = U_N(\theta)\} = \Omega^c, \tag{A9}$$

$$p(\underline{\theta}) = -p(\bar{\theta}) + \int_{\underline{\theta}}^{\bar{\theta}} \mu(d\theta) = 0. \tag{A10}$$

Sufficient conditions. *Moreover, those necessary conditions are also sufficient (Martimort and Stole, 2011, Appendix B).*

Condition (A6) describes how the costate variable $p(\cdot)$ evolves whereas (A7) is the optimality condition for the control. Some explanations for the other conditions are in order. From (A8), the left-side limit of $q(\cdot)$ at any θ is the costate variable deflated by a term related to the measure w.r.t. μ of the open interval $[\underline{\theta}, \theta)$.[27] This costate variable measures the distortions induced by asymmetric information. From (A9), the support of the measure μ is contained in the subset of types for which the participation constraint (3) is binding. Together, with (A7), it implies that distortions due to asymmetric information are less significant on intervals where the participation constraint is binding. Sufficiency is straightforwardly obtained by adapting the same Arrow-like argument as in Martimort and Stole (2011, Appendix B). The conditions (A6) to (A10) are also sufficient for $(\bar{U}(\theta), \bar{e}(\theta))$ to be an optimum.

Let us rewrite some of these optimality conditions. First, observe that (A6) can be transformed as

$$-\dot{p}(\theta) = f(\theta)(1 - \zeta). \tag{A11}$$

From (A10), we get

$$p(\bar{\theta}) = \int_{\underline{\theta}}^{\bar{\theta}} \mu(d\theta). \tag{A12}$$

We may rewrite (A11) as

$$p(\theta) = p(\bar{\theta}) + (1 - \zeta)(1 - F(\theta)). \tag{A13}$$

Second, (A7) yields the first-order condition

$$\zeta f(\theta) \left(1 - \frac{\bar{e}(\theta)}{\theta}\right) = -q(\theta) \frac{\bar{e}(\theta)}{\theta^2}. \tag{A14}$$

In the sequel, we study in turn two possibilities for where the participation constraint (A1) is binding. In **Case 1** below, this participation constraint is supposed to be

[27] Such formulation is made necessary to take into account the fact that μ may be singular at θ.

binding on an interval $\Omega^c = [\underline{\theta}, \theta^*]$ with non-zero measure. **Case 2** deals with the case where $\Omega^c = \{\underline{\theta}\}$.

Case 1. $\Omega^c = [\underline{\theta}, \theta^*]$**, with** $\theta^* > \underline{\theta}$**.**

Analysis of the set of types Ω^c *where the participation constraint (3) is binding.* We are looking for an optimal arc $\bar{U}(\theta)$ such that (3) is slack on some interval $\Omega = (\theta^*, \bar{\theta}]$ and binding on some complementary interval $\Omega^c = [\underline{\theta}, \theta^*]$ with a non empty-interior, i.e., $\underline{\theta} < \theta^*$.[28]

Several facts immediately follow from such inspection.

- Equation (A12) implies that

$$p(\bar{\theta}) = \int_{\underline{\theta}}^{\theta^*} \mu(dx). \tag{A15}$$

- Consider now an interval $\Omega = (\theta^*, \bar{\theta}]$ with non-zero measure where (A1) is slack, i.e., $\bar{U}(\theta) > U_N(\theta)$. On the interior of such interval, $\mu = 0$ and (A8) implies that

$$q(\theta) = p(\theta) - \int_{\underline{\theta}}^{\theta^*} \mu(dx). \tag{A16}$$

  Using (A15) and (A16), (A14) yields an optimal effort level $\bar{e}(\theta, \zeta)$ given by (8) (where we make the dependence on ζ explicit for further references).

- Consider now an interval $\Omega^c = [\underline{\theta}, \theta^*]$ with non-zero measure where (A1) is binding, i.e., $\bar{U}(\theta) = U_N(\theta)$. Differentiating with respect to θ in the interior of $\Omega^c = [\underline{\theta}, \theta^*]$ yields

$$\dot{\bar{U}}(\theta) = \dot{U}_N(\theta) \Leftrightarrow \bar{e}(\theta) = e_N(\theta).$$

  Therefore, (A14) becomes now:

$$q(\theta) = -\left(\frac{1-\alpha}{\alpha}\right)\zeta\theta f(\theta) \quad \forall \theta \in (\underline{\theta}, \theta^*). \tag{A17}$$

  From (A8), (A11) and (A17), we deduce that

$$\int_{\underline{\theta}}^{\theta^-} \mu(d\theta) = p(\bar{\theta}) + (1-\zeta)(1-F(\theta)) + \left(\frac{1-\alpha}{\alpha}\right)\zeta\theta f(\theta) \quad \forall \theta \in (\underline{\theta}, \theta^*)$$

  or, using (A15)

$$-\int_{\theta^-}^{\theta^*} \mu(d\theta) = (1-\zeta)(1-F(\theta)) + \left(\frac{1-\alpha}{\alpha}\right)\zeta\theta f(\theta) \quad \forall \theta \in (\underline{\theta}, \theta^*). \tag{A18}$$

  Let us look for a positive measure μ that is absolutely continuous with respect to the Lebesgue measure on $(\underline{\theta}, \theta^*]$ and so writes as $\mu(d\theta) = g(\theta)d\theta$ for some measurable and non-negative function g on this interval.

  Before studying further the properties of g, we prove the following Lemma:

[28] From the sufficiency conditions in Proposition A.1, finding a vector (p, q, e) that induces such allocation and satisfies the necessary conditions (A6) to (A10) validates this *"guess and try"* approach.

Lemma A.1 *Assume that Assumption 1 holds. Take* $k \leq \frac{1}{\underline{\theta} f(\underline{\theta})}$ *and define uniquely* $\theta^* \in [\underline{\theta}, \bar{\theta}]$ *as the solution to*

$$k = \frac{1 - F(\theta^*)}{\theta^* f(\theta^*)} > 0. \tag{A19}$$

Then, we have

$$\frac{d}{d\theta}\left(1 - F(\theta) - k\theta f(\theta)\right) \leq 0 \quad \forall \theta \in [\underline{\theta}, \theta^*]. \tag{A20}$$

Proof. Observe that Assumption 1 can be rewritten as

$$0 \geq \frac{d}{d\theta}\left(\frac{1 - F(\theta)}{\theta f(\theta)}\right) = -\frac{1}{\theta} - \frac{(1 - F(\theta))}{\theta^2 f^2(\theta)} \frac{d}{d\theta}\left(\theta f(\theta)\right) \Leftrightarrow -(1 - F(\theta)) \frac{d}{d\theta}\left(\theta f(\theta)\right) \leq \theta f^2(\theta).$$

From this, it follows that

$$\frac{d}{d\theta}\left(1 - F(\theta) - k\theta f(\theta)\right) = -f(\theta) - k\frac{d}{d\theta}\left(\theta f(\theta)\right) \leq f(\theta)\left(-1 + k\frac{\theta f(\theta)}{1 - F(\theta)}\right).$$

Using the definition of k from (A19) and again Assumption 1, we get:

$$k \leq \frac{1 - F(\theta)}{\theta f(\theta)} \quad \forall \theta \leq \theta^*$$

Therefore, we get

$$-f(\theta) - k\frac{d}{d\theta}\left(\theta f(\theta)\right) \leq 0 \quad \forall \theta \leq \theta^*$$

which yields (A20). ■

Consider now $k = \frac{\zeta(1-\alpha)}{(\zeta - 1)\alpha}$ and observe that $k \leq \frac{1}{\underline{\theta} f(\underline{\theta})}$ when $\zeta > \zeta^*$ where ζ^* is defined in (6).

From Lemma A.1, applied to such k, g is indeed non-negative on $[\underline{\theta}, \theta^*]$ if $\zeta > 1$. More precisely, when $\zeta > 1$, we get:

$$g(\theta) = (1 - \zeta)\frac{d}{d\theta}\left(1 - F(\theta) - k\theta f(\theta)\right) \geq 0 \quad \forall \theta \in (\underline{\theta}, \theta^*). \tag{A21}$$

Note that by construction, μ has no mass point at θ^*. Note also that putting altogether (A15) and (A21) implies that

$$p(\bar{\theta}) = \mu(\{\underline{\theta}\}) + (1 - \zeta)\int_{\underline{\theta}}^{\theta^*} \frac{d}{d\theta}\left(1 - F(\theta) - k\theta f(\theta)\right) d\theta$$

where $\mu(\{\underline{\theta}\})$ is the mass that the measure μ charges at $\underline{\theta}$. Using (A19), this latter equation can be rewritten as:

$$p(\bar{\theta}) = \mu(\{\underline{\theta}\}) - (1 - \zeta) - \frac{1 - \alpha}{\alpha}\zeta \underline{\theta} f(\underline{\theta}). \tag{A22}$$

But from (A10) and (A11), we get

$$p(\underline{\theta}) = p(\bar{\theta}) + 1 - \zeta = 0. \tag{A23}$$

Inserting into (A22) yields

$$\mu(\{\underline{\theta}\}) = \frac{1-\alpha}{\alpha}\zeta\underline{\theta}f(\underline{\theta}) > 0 \tag{A24}$$

which shows that μ has a mass point at $\underline{\theta}$.

Concavity of $H(\theta, U, e, \zeta, q)$ *in* e. Observe that, for $\theta \in \Omega^c$, $q(\theta)$ as defined by (A17) is negative and thus (A5) holds where $q = q(\theta)$.

For $\theta \in \Omega^c$, we deduce from (A11), (A16) and (A23) that $q(\theta) = (1-\zeta)(1-F(\theta)) < 0.$ and thus (A5) again holds.

Continuity of $\bar{e}(\cdot)$ *at* θ^*. It immediately follows from the fact that μ has no singularity at θ^*. This implies "smooth-pasting" of the rent profile with:

$$U(\theta^*) = U_N(\theta^*) \text{ and } \dot{U}(\theta^*) = \dot{U}_N(\theta^*).$$

Monotonicity of $\bar{e}(\cdot)$. It immediately follows from the fact that $\bar{e}(\cdot)$ is continuous and, trivially increasing on Ω^c and also so on Ω from Condition 1.

Case 2. $\Omega^c = \{\underline{\theta}\}$**.** Observe that $k = \frac{\zeta(1-\alpha)}{(\zeta-1)\alpha} > \frac{1}{\underline{\theta}f(\underline{\theta})}$ when $\zeta \leq \zeta^*$. In that case, the participation constraint (A1) is binding at $\underline{\theta}$ only. From (A24), the measure μ has a charge at $\underline{\theta}$ only. When $\zeta \geq 1$, we have

$$\mu(\{\underline{\theta}\}) = \frac{1-\alpha}{\alpha}\zeta\underline{\theta}f(\underline{\theta}) \geq (\zeta - 1)\frac{\zeta^*}{\zeta^* - 1} \geq 0. \tag{A25}$$

The optimal effort on Θ is still given by (8) on the whole interval $[\underline{\theta}, \bar{\theta}]$.

Proof that $\zeta > 1$. Observe that, when binding, (2) can be rewritten as:

$$\int_{\underline{\theta}}^{\theta^*(\zeta)} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta}\right) f(\theta)d\theta + \int_{\theta^*(\zeta)}^{\bar{\theta}} \left(\bar{e}(\theta,\zeta) - \frac{\bar{e}^2(\theta,\zeta)}{2\theta}\right) f(\theta)d\theta$$

$$= \int_{\underline{\theta}}^{\theta^*(\zeta)} U_N(\theta)f(\theta)d\theta + \int_{\theta^*(\zeta)}^{\bar{\theta}} \left(U_N(\theta^*(\zeta)) + \int_{\theta^*(\zeta)}^{\theta} \frac{\bar{e}^2(\xi,\zeta)}{2\xi^2}d\xi\right) f(\theta)d\theta \tag{A26}$$

where we make explicit the dependence of $\bar{e}(\cdot)$ and θ^* on ζ as specified in (8) and (9) to express the left-hand side and where we use (A2) to rewrite the right-hand side.[29]

Let denote respectively by $L(\zeta)$ and $R(\zeta)$ the left-hand and right-hand sides of (A26).

The following observations are readily made.

[29] Observe that this formula encompasses both **Case 1** which applies for $\zeta \geq \zeta^*$ and **Case 2** which applies for $\zeta \in [1, \zeta^*]$.

1. *$L(\zeta) - R(\zeta)$ is strictly increasing.* First, observe that

$$\frac{\partial \bar{e}}{\partial \zeta}(\theta, \zeta) = -\frac{\frac{1-F(\theta)}{f(\theta)}}{\left(\zeta + (\zeta - 1)\frac{1-F(\theta)}{\theta f(\theta)}\right)^2} < 0. \tag{A27}$$

   Using the fact that $\bar{e}(\theta, \zeta)$ is continuous at $\theta = \theta^*(\zeta)$, i.e., $\bar{e}(\theta^*(\zeta), \zeta) = e_N(\theta^*(\zeta))$, we have:

$$L'(\zeta) = \int_{\theta^*(\zeta)}^{\bar{\theta}} \frac{\partial \bar{e}}{\partial \zeta}(\theta, \zeta)\left(1 - \frac{\bar{e}(\theta, \zeta)}{\theta}\right) f(\theta) d\theta = (\zeta - 1) \int_{\theta^*(\zeta)}^{\bar{\theta}} \frac{\partial \bar{e}}{\partial \zeta}(\theta, \zeta) \frac{\frac{1-F(\theta)}{\theta f(\theta)}}{\zeta + (\zeta - 1)\frac{1-F(\theta)}{\theta f(\theta)}} f(\theta) d\theta. \tag{A28}$$

   Using the fact that $U_N(\theta, \zeta)$ is continuous at $\theta = \theta^*(\zeta)$, we have

$$R'(\zeta) = \dot{\theta}^*(\zeta) \int_{\theta^*(\zeta)}^{\bar{\theta}} \left(\dot{U}_N(\theta^*(\zeta)) - \frac{\bar{e}^2(\theta^*(\zeta), \zeta)}{2(\theta^*(\zeta))^2}\right) f(\theta) d\theta + \int_{\theta^*(\zeta)}^{\bar{\theta}} \int_{\theta^*(\zeta)}^{\theta} \frac{\partial \bar{e}}{\partial \zeta}(\xi, \zeta) \frac{\bar{e}(\xi, \zeta)}{\xi^2} f(\theta) d\xi d\theta.$$

   Using that $\dot{U}_N(\theta^*(\zeta)) = \frac{e_N^2(\theta^*(\zeta))}{2(\theta^*(\zeta))^2}$, and continuity of $\bar{e}(\cdot, \zeta)$ at $\theta = \theta^*(\zeta)$, i.e., $\bar{e}(\theta^*(\zeta), \zeta) = e_N(\theta^*(\zeta))$, we get

$$R'(\zeta) = \int_{\theta^*(\zeta)}^{\bar{\theta}} \left(\int_{\theta^*(\zeta)}^{\theta} \frac{\partial \bar{e}}{\partial \zeta}(\xi, \zeta) \frac{\bar{e}(\xi, \zeta)}{\xi^2} d\xi\right) f(\theta) d\theta.$$

   Integrating by parts yields

$$R'(\zeta) = \int_{\theta^*(\zeta)}^{\bar{\theta}} (1 - F(\theta)) \frac{\partial \bar{e}}{\partial \zeta}(\theta, \zeta) \frac{\bar{e}(\theta, \zeta)}{\theta^2} d\theta = \int_{\theta^*(\zeta)}^{\bar{\theta}} \frac{\partial \bar{e}}{\partial \zeta}(\theta, \zeta) \frac{\zeta \frac{1-F(\theta)}{\theta f(\theta)}}{\zeta + (\zeta - 1)\frac{1-F(\theta)}{\theta f(\theta)}} f(\theta) d\theta. \tag{A29}$$

   Using (A28) and (A29) we finally get

$$L'(\zeta) - R'(\zeta) = -\int_{\theta^*(\zeta)}^{\bar{\theta}} \frac{\partial \bar{e}}{\partial \zeta}(\theta, \zeta) \frac{\frac{1-F(\theta)}{\theta f(\theta)}}{\zeta + (\zeta - 1)\frac{1-F(\theta)}{\theta f(\theta)}} f(\theta) d\theta > 0.$$

2. Notice that when $\zeta = 1$, $\theta^*(\zeta) = \underline{\theta}$ and $L(1) < R(1)$ indeed amounts to (2).

3. We have

Lemma A.2

$$\lim_{\zeta \to +\infty} L(\zeta) - R(\zeta) > 0. \tag{A30}$$

Proof. Consider the following problem:

$$\mathcal{V}^M = \max_{e(\cdot), \theta^*} \int_{\underline{\theta}}^{\theta^*} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta}\right) f(\theta) d\theta + \int_{\theta^*}^{\bar{\theta}} \left(e(\theta) - \frac{e^2(\theta)}{2\theta}\left(1 + \frac{1 - F(\theta)}{\theta f(\theta)}\right)\right) f(\theta) d\theta$$

$$-\int_{\underline{\theta}}^{\theta^*} U_N(\theta)f(\theta)d\theta - U_N(\theta^*)(1-F(\theta^*)). \tag{A31}$$

First, observe that $\mathcal{V}^M \geq 0$. Indeed, taking $e(\theta) = e_N(\theta)$ and $\theta^* = \bar{\theta}$ obviously yields 0 for the maximand.

The above maximum is achieved for $(\bar{e}_\infty(\theta), \theta^*_\infty)$ where

$$\bar{e}_\infty(\theta) = \frac{\theta}{1+\frac{1-F(\theta)}{\theta f(\theta)}} \tag{A32}$$

and

$$\begin{cases} \frac{1-F(\theta^*_\infty)}{\theta^*_\infty f(\theta^*_\infty)} = \frac{1-\alpha}{\alpha} & \text{if } \frac{1-\alpha}{\alpha} < \frac{1}{\underline{\theta} f(\underline{\theta})} \\ \theta^*_\infty = \underline{\theta} & \text{if } \frac{1-\alpha}{\alpha} \geq \frac{1}{\underline{\theta} f(\underline{\theta})}. \end{cases} \tag{A33}$$

Condition 1 ensures that $\theta^*_\infty \in (\underline{\theta}, \bar{\theta})$ always exists whenever (7) holds. That $\mathcal{V}^M > 0$ immediately follows from observing that $\mathcal{V}^M$ is not achieved for $e_N(\theta)$ and $\theta^* = \bar{\theta}$. Finally, this strict inequality amounts to (A30). ■

From Items [1.], [2.] and [3.] above, we immediately obtain that there exists $\hat{\zeta} > 1$ such that

$$L(\hat{\zeta}) = R(\hat{\zeta}).$$

Integrating by parts and manipulating finally yields (10). ■

Proof of Proposition 2. Because Assumption 3 holds, we have $\zeta^*(\alpha) > 1$ for any $\alpha \geq \alpha_1$. A first implication is that, for $\zeta \leq \zeta^*(\alpha)$, we get $\theta^*(\zeta) = \underline{\theta}$. Because $L(\cdot) - R(\cdot)$ is strictly increasing as shown above, we have $\hat{\zeta} \leq \zeta^*(\alpha)$ if and only if

$$L(\zeta^*(\alpha)) \geq R(\zeta^*(\alpha)) \Leftrightarrow J(\alpha) \geq U_N(\underline{\theta}, \alpha) \tag{A34}$$

where

$$J(\alpha) = \int_{\underline{\theta}}^{\bar{\theta}} \left(\bar{e}(\theta, \zeta^*(\alpha)) - \frac{\bar{e}^2(\theta, \zeta^*(\alpha))}{2\theta} \left(1 + \frac{1-F(\theta)}{\theta f(\theta)} \right) \right) f(\theta) d\theta$$

and where, for future reference, we make explicit the dependence of $U_N(\cdot)$ on α.

We have

$$J'(\alpha) = \frac{\partial \zeta^*}{\partial \alpha}(\alpha) \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial \bar{e}}{\partial \zeta}(\theta, \zeta^*(\alpha)) \left(1 - \frac{\bar{e}(\theta, \zeta^*(\alpha))}{\theta} \left(1 + \frac{1-F(\theta)}{\theta f(\theta)} \right) \right) f(\theta) d\theta$$

$$= \int_{\underline{\theta}}^{\bar{\theta}} \frac{\underline{\theta} f(\underline{\theta})(1-F(\theta))^2}{\theta f(\theta)} \frac{((1-\alpha)\underline{\theta} f(\underline{\theta}) - \alpha)}{\left(\alpha + (1-\alpha)\frac{(1-F(\theta))\underline{\theta} f(\underline{\theta})}{\theta f(\theta)}\right)^3} d\theta.$$

We have $J'(\alpha) \leq 0$ (with equality only at $\alpha = \alpha_2$) and thus $J(\alpha)$ is almost everywhere strictly decreasing with α. Moreover, for $\alpha = 1$, we have $\zeta^*(1) = 1$ and $\bar{e}(\theta, \zeta^*(1)) =$

$e^{FB}(\theta)$. Therefore, we get:

$$J(1) = \int_{\underline{\theta}}^{\bar{\theta}} \left(e^{FB}(\theta) - \frac{(e^{FB}(\theta))^2}{2\theta} \left(1 + \frac{1-F(\theta)}{\theta f(\theta)} \right) \right) f(\theta) d\theta = \frac{\underline{\theta}}{2} = U_N(\underline{\theta}, 1). \qquad \text{(A35)}$$

We also find:

$$J'(1) = -\underline{\theta} f(\underline{\theta}) \int_{\underline{\theta}}^{\bar{\theta}} \frac{(1-F(\theta))^2}{\theta f(\theta)} d\theta.$$

From Assumption 1, we immediately derive the inequality

$$\frac{(1-F(\theta))^2}{\theta f(\theta)} \leq \frac{1-F(\theta)}{\underline{\theta} f(\underline{\theta})}$$

with an equality only at $\theta = \underline{\theta}$. Therefore, we get:

$$-J'(1) < \int_{\underline{\theta}}^{\bar{\theta}} (1-F(\theta)) d\theta = E_\theta(\theta) - \underline{\theta} = -\frac{dU_N}{d\alpha}(\underline{\theta}, \alpha)|_{\alpha=1}. \qquad \text{(A36)}$$

Therefore, it follows from $J(\cdot)$ continuity,that there exists $\alpha_3 < 1$ such that

$$J(\alpha) < U_N(\underline{\theta}, \alpha) \quad \forall \alpha \in (\alpha_3, 1). \qquad \text{(A37)}$$

Moreover, we also have

$$J(\alpha_1) > J(1) = U_N(\underline{\theta}, 1) = \frac{\underline{\theta}}{2}. \qquad \text{(A38)}$$

From (A39), we deduce that, necessarily, $\alpha_3 \in (\alpha_1, 1)$. From (A34) and (A37), we also deduce that:

$$\hat{\zeta} > \zeta^*(\alpha) \quad \forall \alpha \in (\alpha_3, 1).$$

A contrario, we also have

$$J(\alpha_1) > J(1) = U_N(\underline{\theta}, 1) = U_N(\underline{\theta}, \alpha_1) = \frac{\underline{\theta}}{2}. \qquad \text{(A39)}$$

From (A34), we deduce that there exists $\alpha_4 \in (\alpha_1, \alpha_3]$ such that

$$J(\alpha) \geq U_N(\underline{\theta}, \alpha) \quad \forall \alpha \in [\alpha_1, \alpha_4] \qquad \text{(A40)}$$

where the last inequality follows from the fact that α_1 solves (A4) as an equality. Finally, we get

$$\hat{\zeta} \leq \zeta^*(\alpha) \quad \forall \alpha \in [\alpha_1, \alpha_4].$$

■

Proof of Proposition 5. From (13), we immediately get:

$$T'(e) = \frac{\bar{e}(\theta)}{\theta} - \alpha = \begin{cases} \frac{1}{1+\frac{\hat{\zeta}-1}{\hat{\zeta}}\frac{1-F(\theta)}{\theta f(\theta)}} - \alpha & \text{if } e > e_N(\bar{\theta}(e)) \\ 0 & \text{if } e \leq e_N(\bar{\theta}(e)). \end{cases}$$

where the last equality follows from (8). Note that $T'(e)$ is continuous at $\theta^*(\hat{\zeta})$ if it is interior.

Differentiating once more, we get

$$T''(e) = \begin{cases} -\frac{\frac{\hat{\zeta}-1}{\hat{\zeta}}\frac{d}{d\theta}\left(\frac{1-F(\theta)}{\theta f(\theta)}\right)|_{\bar{\theta}(e)}\dot{\bar{\theta}}(e)}{\left(1+\frac{\hat{\zeta}-1}{\hat{\zeta}}\frac{1-F(\theta)}{\theta f(\theta)}\right)^2} > 0 & \text{if } \bar{e}(\theta) > e_N(\theta) \\ 0 & \text{if } e \leq e_N(\bar{\theta}(e)). \end{cases}$$

Hence , $T(e)$ is convex and strictly so if and only if $e > e_N(\bar{\theta}(e))$. It is flat when $e \leq e_N(\bar{\theta}(e))$. ■

Proof of Proposition 6. Observe that the budget balance condition (2) altogether with the participation constraints (18) yield the following simpler inequality:

$$\int_{\underline{\theta}}^{\bar{\theta}} \left(\alpha e(\theta) - \frac{e^2(\theta)}{2\theta}\right) f(\theta)d\theta \geq \frac{\alpha^2}{2}\int_{\underline{\theta}}^{\bar{\theta}} \theta f(\theta)d\theta. \tag{A41}$$

The pointwise maximum of the left-hand side is $e_N(\theta) = \alpha\theta$ and then the left- and right-hand sides of (A41) are both equal. Therefore, the optimal mechanism robust to any individual deviation consists in proposing the BNE outcome which is, by definition, also incentive compatible. ■

Proof of Proposition 7. First, fix $\gamma \in \Gamma = [0, E_{\tilde{\theta}}(e_N(\tilde{\theta}))]$ and define the new type-dependent participation constraint:

$$U(\theta) \geq \frac{\alpha^2}{2}\theta + (1-\alpha)\gamma, \quad \forall \theta \in \Theta. \tag{A42}$$

Consider now the non-negative mapping $\Phi(\gamma)$ defined over Γ such that:

$$-\Phi(\gamma) = \arg\max_{U(\cdot),e(\cdot)} \left\{-E_{\tilde{\theta}}(e(\tilde{\theta})) \text{ subject to (2), (4) and (A42)}\right\}.$$

A credible effort profile $e^C(\cdot)$ is thus such that

$$\Phi(E_{\tilde{\theta}}(e_C(\tilde{\theta}))) = E_{\tilde{\theta}}(e_C(\tilde{\theta})).$$

Clearly, $\Phi(\cdot)$ is continuous. Moreover, $\Phi(0) > 0$ and $\Phi(E_{\tilde{\theta}}(e_N(\tilde{\theta}))) \leq E_{\tilde{\theta}}(e_N(\tilde{\theta}))$. Hence, there exists a minimal non-negative fixed-point γ^* for the mapping $\Phi(\cdot)$. For such value, the participation constraints become:

$$U(\theta) \geq \frac{\alpha^2}{2}\theta + (1-\alpha)\gamma^*, \quad \forall \theta \in \Theta. \tag{A43}$$

Neglecting the second-order condition (5) that will be checked ex post; we rewrite the so relaxed optimization problem for this particular γ^* as:

$$(\mathcal{P}_C(\gamma^*)): \quad \max_{U(\cdot)\in W(\Theta),e(\cdot)} -E_{\tilde{\theta}}(e(\tilde{\theta})) \text{ subject to (2), (4) and (A43).}$$

$(\mathcal{P}_C(\gamma^*))$ is again a generalized Bolza problem with an isoperimetric constraint (2) and a state-dependent constraint (A43). We use therefore the same techniques as in the Proofs of Propositions 3 and 4. Denoting again by ζ the non-negative multiplier of (2), we write the Lagrangian for this problem as:

$$L_C(\theta, U, e, \zeta) = f(\theta)\left(-e + \zeta\left(e - \frac{e^2}{2\theta} - U\right)\right).$$

Let then define the unmaximized Hamiltonian as

$$H_C(\theta, U, e, \zeta, q) = L_C(\theta, U, e, \zeta) + \frac{e^2}{2\theta^2}.$$

This Hamiltonian is linear in U and strictly concave in e when (A5) again holds. This latter condition is checked below for the optimal effort profile.
Necessary conditions.[30] A normal extremum $(U_C(\theta), e_C(\theta))$ is such that there exists an absolutely continuous function $p(\theta)$, a function $q(\cdot)$, and a non-negative measure μ which are all defined on Θ such that:

$$-\dot{p}(\theta) = \frac{\partial H_C}{\partial U}(\theta, U_C(\theta), e_C(\theta), \zeta, q(\theta)), \tag{A44}$$

$$e_C(\theta) \in \arg\max_{e \geq 0} H_C(\theta, U_C(\theta), e_C(\theta), \zeta, q(\theta)), \tag{A45}$$

$$q(\theta) = p(\theta) - \int_{\underline{\theta}}^{\theta^-} \mu(d\theta), \quad \forall \theta \in (\underline{\theta}, \bar{\theta}], \tag{A46}$$

$$supp\{\mu\} \subset \left\{\theta \text{ s.t. } U_C(\theta) = \frac{\alpha^2}{2}\theta + (1-\alpha)\gamma^*\right\} = \Omega_C^c, \tag{A47}$$

$$p(\underline{\theta}) = -p(\bar{\theta}) + \int_{\underline{\theta}}^{\bar{\theta}} \mu(d\theta) = 0. \tag{A48}$$

Let us rewrite some of these optimality conditions. First, observe that (A6) can be transformed as

$$\dot{p}(\theta) = \zeta f(\theta). \tag{A49}$$

From (A48), we may rewrite (A49) as

$$p(\theta) = \zeta F(\theta) \tag{A50}$$

and obtain also

$$\int_{\underline{\theta}}^{\bar{\theta}} \mu(d\theta) = \zeta. \tag{A51}$$

[30] Sufficient conditions again follow from Martimort and Stole (2010, Appendix B).

Second, (A45) yields the first-order condition

$$f(\theta)\left(-1+\zeta\left(1-\frac{e_C(\theta)}{\theta}\right)\right)=-q(\theta)\frac{e_C(\theta)}{\theta^2}. \tag{A52}$$

Consider the possibility that the participation constraint (A43) is binding on an interval with non-empty interior. On such interval, it must be that $\dot{U}_C(\theta)=\frac{\alpha^2}{2}$ and so $e_C(\theta)=e_N(\theta)$. Inserting into (A52) we get the following condition

$$\int_{\underline{\theta}}^{\theta}\mu(d\theta)=\frac{\theta f(\theta)}{\alpha}(-1+\zeta(1-\alpha))+\zeta F(\theta). \tag{A53}$$

Next lemma is the dual of Lemma A.1.

Lemma A.3 *Assume that Assumption 4 holds. Take* $k=1-\frac{\zeta-1}{\alpha\zeta}\in(0,1)$ *and define uniquely* $\theta^*\in[\underline{\theta},\bar{\theta}]$ *as the solution to*

$$k=\frac{F(\theta^*)}{\theta^* f(\theta^*)} \tag{A54}$$

if it exists and $\theta^*=\bar{\theta}$ *otherwise. Then, we have*

$$\frac{d}{d\theta}\left(F(\theta)-k\theta f(\theta)\right)\geq 0\quad\forall\theta\in[\theta^*,\bar{\theta}]. \tag{A55}$$

Proof. Observe that Assumption 4 can be rewritten as

$$0\leq\frac{d}{d\theta}\left(\frac{F(\theta)}{\theta f(\theta)}\right)=\frac{1}{\theta}-\frac{F(\theta)}{\theta^2f^2(\theta)}\frac{d}{d\theta}\left(\theta f(\theta)\right)\Leftrightarrow F(\theta)\frac{d}{d\theta}\left(\theta f(\theta)\right)\leq\theta f^2(\theta).$$

From this, it follows that

$$\frac{d}{d\theta}\left(F(\theta)-k\theta f(\theta)\right)\geq\zeta f(\theta)\left(1-k\frac{\theta f(\theta)}{F(\theta)}\right).$$

Using the definition of k from (A19) and again Assumption 4, we get:

$$f(\theta)-k\frac{d}{d\theta}\left(\theta f(\theta)\right)\geq 0\quad\forall\theta\geq\theta^*$$

which yields (A55). ■

Let us look for a measure μ absolutely continuous with respect to the Lebesgue measure and let us thus write $\mu(d\theta)=g(\theta)d\theta$ where $g(\cdot)$ is non-negative on $\sup\mu=\Omega^c=[\theta^*,\bar{\theta}]$. We show below that all necessary conditions for optimality are satisfied with such scheme. From Proposition A.1, those conditions are also sufficient and thus characterize the optimum.

Coming back to (A53) and using Lemma, it is immediate that $g(\cdot)$ is non-negative on the interval $\Omega^c = [\theta^*, \bar{\theta}]$.

On the complementary interval $\Omega = [\underline{\theta}, \theta^*]$, we have

$$q(\theta) = \zeta F(\theta) \quad \forall \theta \in \Omega. \tag{A56}$$

For further reference, it is useful to define $X = \frac{\zeta - 1}{\zeta} \in (0, 1)$ and make explicit the dependence of θ^* on X. In particular, observe that $\theta^*(\alpha) = \underline{\theta}$.

Similarly, the optimality condition for the effort (A53) yields also

$$e_C(\theta, X) = X \frac{\theta}{1 - \frac{F(\theta)}{\theta f(\theta)}}. \tag{A57}$$

Observe that, whenever $\theta^* \in (\underline{\theta}, \bar{\theta})$, $e_C(\cdot)$ is continuous at θ^* with $e_C(\theta^*) = e_N(\theta^*)$.

Note also that $e_C(\cdot, X)$ remains positive on the interval $\Omega = [\underline{\theta}, \theta^*(X)]$ as long as Assumption 4 holds and $\zeta > 1$, a condition that we check below. Finally, Assumption 4 implies also that $e_C(\cdot, X)$ is monotonically increasing on the interval $\Omega = [\underline{\theta}, \theta^*(X)]$, so that the neglected second-order conditions (5) holds.

The value of the multiplier ζ, or alternatively the value of X, is obtained when (2) is binding. This condition can be rewritten as:

$$\int_{\underline{\theta}}^{\theta^*(X)} \left(e_C(\theta, X) - \frac{e_C^2(\theta, X)}{2\theta} \right) f(\theta) d\theta + \int_{\theta^*(X)}^{\bar{\theta}} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta$$

$$= \int_{\underline{\theta}}^{\theta^*(X)} \left(\frac{\alpha^2}{2} \theta^*(X) + \int_{\underline{\theta}}^{\theta} \frac{e_C^2(y, X)}{2y^2} dy \right) f(\theta) d\theta + \int_{\theta^*(X)}^{\bar{\theta}} \frac{\alpha^2}{2} \theta f(\theta) d\theta + (1 - \alpha)\gamma^* \tag{A58}$$

where

$$\gamma^* = \int_{\underline{\theta}}^{\theta^*(X)} e_C(\theta, X) f(\theta) d\theta + \int_{\theta^*(X)}^{\bar{\theta}} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta.$$

Simplifying further, we get:

$$\left(\alpha X - \frac{X^2}{2} \right) = \frac{\alpha^2}{2} \varpi(X) \tag{A59}$$

where

$$\varpi(X) = \frac{\theta^*(X) F(\theta^*(X))}{\int_{\underline{\theta}}^{\theta^*(X)} \frac{\theta}{1 - \frac{F(\theta)}{\theta f(\theta)}} f(\theta) d\theta}.$$

Taking into account that $\theta^*(\alpha) = \underline{\theta}$, Lhôspital's rule finally yields

$$\varpi(\alpha) = \lim_{\theta \to \underline{\theta}} \frac{\theta f(\theta) + F(\theta)}{\frac{\theta f(\theta)}{1 - \frac{F(\theta)}{\theta f(\theta)}}} = 1.$$

Inserting into (A59), we finally obtain that the solution in X for this latter equation is $X = \alpha$. From the fact that the Hamiltonian $H_C(\cdot)$ is strictly concave in e, we know that the solution is unique. This ends the proof that the only credible effort profile is $e_N(\theta)$. ■

Proof of Lemma 2. Under Assumption 1, we know that (3) is necessarily binding. This is also the case for (20) which is harder to satisfy as we showed in the text. Take now Ω^c an interval with non-empty interior where (20) binds. On such interval, we have

$$U(\theta) = (1-\delta)\left(-\frac{e^2(\theta)}{2\theta} + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta}))\right) + \delta U_N(\theta). \tag{A60}$$

Differentiating w.r.t θ and taking into account (1), the following condition holds a.e.:

$$\delta\left(\frac{e^2(\theta)}{2\theta^2} - \frac{e_N^2(\theta)}{2\theta^2}\right) = (1-\delta)\left(\alpha - \frac{e(\theta)}{\theta}\right)\dot{e}(\theta)$$

which admits the trivial solution given by (22) and this solution is the unique one that is increasing on I as requested by second-order conditions for the incentive compatibility problem.[31] Inserting into (A60) then yields the first part of (24).

Consider now $\theta \in \Omega$. By definition, we have:

$$U(\theta) > (1-\delta)\left(-\frac{e^2(\theta)}{2\theta} + \alpha e(\theta) + (1-\alpha)E_{\tilde{\theta}}(e(\tilde{\theta}))\right) + \delta U_N(\theta)$$

$$= U_N(\theta)+(1-\delta)(1-\alpha)\left(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))\right)+(1-\delta)\left(-\frac{e^2(\theta)}{2\theta} + \alpha e(\theta) + \frac{e_N^2(\theta)}{2\theta} - \alpha e_N(\theta)\right).$$

$$> U_N(\theta) + (1-\delta)(1-\alpha)\left(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))\right)$$

where the last strict inequality follows from the definition of $e_N(\theta)$. This proves the second part of (24). ■

Proof of Proposition 8. Neglecting as usual the second-order condition (5) that will be checked ex post; we now rewrite the mechanism design problem as:

$$(\mathcal{P}_E): \quad \max_{U(\cdot)\in W(\Theta), e(\cdot)} E_{\tilde{\theta}}(U(\tilde{\theta})) \text{ subject to (2), (4) and (24).}$$

We are now looking for optimal mechanism such that (24) is binding on an interval $\Omega = [\underline{\theta}, \theta^*]$ for some $\theta^* \in \Theta$. Indeed, using the techniques of the previous Appendices, we could exhibit the non-negative measure for the type-dependent constraint (24) that allows us to check the necessary conditions for optimality as before. In that appendix, we proceed more directly.

For such structure of the solution, we have:

$$U(\theta) = U_N(\theta) + (1-\delta)(1-\alpha)(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))) \quad \forall \theta \in \Omega^c. \tag{A61}$$

[31] The other solution is indeed such that $\dot{e}(\theta) = -\frac{\delta}{2(1-\delta)}\left(\frac{e(\theta)}{\theta} + \alpha\right) < 0$.

Instead, on $\Omega = (\theta^*, \bar{\theta}]$, we have:

$$U(\theta) = U(\theta^*) + \int_{\theta^*}^{\theta} \frac{e^2(x)}{2x^2} dx. \tag{A62}$$

Using (A61) and (A62), we compute after an integration by parts:

$$E_{\tilde{\theta}}(U(\tilde{\theta})) = \int_{\underline{\theta}}^{\theta^*} U_N(\theta) f(\theta) d\theta + U_N(\theta^*)(1 - F(\theta^*)) + (1-\delta)(1-\alpha)(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))).$$

Using (22) to express effort on Ω^c, the feasibility condition (2) can thus be written as:

$$\int_{\underline{\theta}}^{\theta^*} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta + \int_{\theta^*}^{\bar{\theta}} \left(e(\theta) - \frac{e^2(\theta)}{2\theta} \left(1 + \frac{1 - F(\theta)}{\theta f(\theta)} \right) \right) f(\theta) d\theta$$

$$\geq \int_{\underline{\theta}}^{\theta^*} U_N(\theta) f(\theta) d\theta + U_N(\theta^*)(1 - F(\theta^*)) + (1-\delta)(1-\alpha)(E_{\tilde{\theta}}(e(\tilde{\theta})) - E_{\tilde{\theta}}(e_N(\tilde{\theta}))),$$

or

$$\int_{\underline{\theta}}^{\theta^*} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta + \int_{\theta^*}^{\bar{\theta}} \left((1 - (1-\delta)(1-\alpha)) e(\theta) - \frac{e^2(\theta)}{2\theta} \left(1 + \frac{1 - F(\theta)}{\theta f(\theta)} \right) \right) f(\theta) d\theta$$

$$\geq \int_{\underline{\theta}}^{\theta^*} U_N(\theta) f(\theta) d\theta + U_N(\theta^*)(1 - F(\theta^*)). \tag{A63}$$

The mechanism design problem can thus be written as:

$$(\mathcal{P}_E): \quad \max_{\theta^*, e(\cdot)} \int_{\underline{\theta}}^{\theta^*} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta + \int_{\theta^*}^{\bar{\theta}} \left(e(\theta) - \frac{e^2(\theta)}{2\theta} \right) f(\theta) d\theta$$

subject to (A63).

Let us suppose that the constraint (A63) is slack. Then, the solution is $\theta^* = \underline{\theta}$ and $e(\theta) = e^{FB}(\theta)$ for all θ. But then, (A63) does not hold when α is large enough and more precisely:

$$\alpha > \alpha_1 - \frac{2(1-\delta) E_{\tilde{\theta}}(\tilde{\theta})}{2 E_{\tilde{\theta}}(\tilde{\theta}) - \underline{\theta}}. \tag{A64}$$

When (A64) holds, necessarily (A63) is binding at the optimum of $(\mathcal{P}_E)$. Let denote by $\zeta - 1$ (where $\zeta > 1$) the positive multiplier of (A63). We form the Lagrangean

$$\int_{\underline{\theta}}^{\theta^*} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta + \int_{\theta^*}^{\bar{\theta}} \left(e(\theta) - \frac{e^2(\theta)}{2\theta} \right) f(\theta) d\theta$$

$$+(\zeta - 1)[\int_{\underline{\theta}}^{\theta^*} \left(e_N(\theta) - \frac{e_N^2(\theta)}{2\theta} \right) f(\theta) d\theta$$

$$+\int_{\theta^*}^{\bar{\theta}}\left((1-(1-\delta)(1-\alpha))e(\theta)-\frac{e^2(\theta)}{2\theta}\left(1+\frac{1-F(\theta)}{\theta f(\theta)}\right)\right)f(\theta)d\theta$$
$$-\int_{\underline{\theta}}^{\theta^*}U_N(\theta)f(\theta)d\theta+U_N(\theta^*)(1-F(\theta^*))].$$

Optimizing pointwise with respect to $e(\theta)$ on the interval $\Omega^c=(\theta^*,\bar{\theta}]$ gives a solution

$$\bar{e}(\theta,\zeta)=\left(1-\frac{\zeta-1}{\zeta}(1-\delta)(1-\alpha)\right)\frac{\theta}{1+\frac{\zeta-1}{\zeta}\frac{1-F(\theta)}{\theta f(\theta)}}$$

where we make explicit the dependence on ζ. The first part of (25) is the effort level obtained when the multiplier is $\hat{\zeta}$.

Optimizing with respect to θ^* yields $e(\theta^*)=e_N(\theta^*)$. Simplifying further, we get (26). Denote $\theta^*(\zeta)$ such solution making again its dependence on ζ explicit.

Finally, the value $\hat{\zeta}$ of the multiplier is obtained so that

$$\int_{\underline{\theta}}^{\theta^*(\hat{\zeta})}\left(e_N(\theta)-\frac{e_N^2(\theta)}{2\theta}\right)f(\theta)d\theta$$
$$+\int_{\theta^*(\hat{\zeta})}^{\bar{\theta}}\left((1-(1-\delta)(1-\alpha))\bar{e}(\theta,\hat{\zeta})-\frac{\bar{e}^2(\theta,\hat{\zeta})}{2\theta}\left(1+\frac{1-F(\theta)}{\theta f(\theta)}\right)\right)f(\theta)d\theta$$
$$=\int_{\underline{\theta}}^{\theta^*(\hat{\zeta})}U_N(\theta)f(\theta)d\theta+U_N(\theta^*(\hat{\zeta}))(1-F(\theta^*(\hat{\zeta}))). \tag{A65}$$

■