

On the Polluter-Pays Principle

Stefan Ambec* Lars Ehlers†

February 9, 2010

Preliminary and incomplete version.

Abstract

We consider the problem of regulating an economy with environmental pollution. A regulation mechanism defines payments contingent on pollution emissions. The polluter-pays mechanism requires that any agent compensates all other agents for the damages caused by his or her own emissions. It is budget-balanced and efficient. We show that it implements the unique welfare distribution that assigns non-negative individual welfare and renders each agent responsible for his or her pollution impact. Next we examine its acceptability by agents. A mechanism is acceptable if no group of agents can be made better-off with another mechanism. The polluter-pays principle might not be acceptable. It is acceptable if pollution externalities are multilateral among homogenous agents. Last we posit a more general mechanism that implements the polluter-pays welfare distribution under asymmetric information on pollution emissions and damages.

*Toulouse School of Economics (INRA-LERNA-IDEI), France stefan.ambec@toulouse.inra.fr

†Département de Sciences Économiques and CIREQ, Université de Montréal, Canada,
lars.ehlers@umontreal.ca

1 Introduction

From water management to air pollution, managing environmental problems efficiently requires well-designed public policies or coordination among stakeholders. Environmental policies and international environmental agreements are launched to mitigate the failure of market economy due to the presence of negative externalities . Yet public intervention impacts not only the welfare of the economies at a whole but also the distribution of this welfare. This paper addresses the efficiency and distributional impact of environmental policies in an economy with externalities. The model allows for a variety of externalities including unilateral or multilateral ones, heterogenous impacts due to distance or mitigation. It formalizes many complex environmental issues such as water quality management in a river or the reduction of sulfur dioxide or greenhouse gas emissions in an international setting. To that respect, it is as rich as the seminal model of Montgomery (1972).

In this framework, we define a regulation mechanism as individual transfers contingent on pollution emissions. In particular, we consider the mechanism inspired by a literal interpretation of the polluter-pays (PP) principle. It states that *the costs of pollution should be borne by the entity which profits from the process that causes pollution*. Strictly speaking, it requires that any agent compensates all agents who suffer from his or her pollution emissions for the damage he or she causes. The PP mechanism is by construction budget-balanced. It also efficient in the sense that it implements the allocation of pollution emissions that maximizes total welfare. It leads to the unique distribution of the (maximized) total welfare that satisfies two criteria. The first one is that individual's welfare is non-negative for all agents. It is a minimal acceptability requirement since an agent who obtains a negative welfare does not benefit from the economy activities exhibit pollution. It might pays too much from his or her emissions despite that they are at their efficient level. Or it might suffer too much from pollution damage which is efficient from a total welfare point of view. The second criteria relies on the concept of responsibility in axiomatic theory of justice (Fleurbaey, 2008). It states that any agent is responsible for a change of its pollution impact. More precisely, if any agent modifies the environmental impact of its own emissions in the economy, he or she should get the full return or lose due to this change in the economy. For instance, a firm who filter

its own emissions to reduce their sulfur content should get the full benefit for the economy of this cleaning investment. A farmer who uses more pesticide and fertilizers leading to dirtiest waste water should pay the social cost associated to this pesticide and fertilizer increase. We show that the welfare distribution implemented by the PP mechanism is the only one that satisfies the two above criteria: non-negativity and responsibility for pollution impact.

We are also concerned by the acceptability of the PP principle. It is an important issues since environmental policies emerge as a collective choice among citizen or negotiation among stakeholders at least in democratic society. Similarly, international environmental agreements such as the Kyoto protocol are designed by sovereign countries. Each country is free to refuse any agreement that is worse than the status-quo or to any other agreement. A regulation mechanism is acceptable if the welfare of any agent or group of agents is higher with any other mechanism (including no mechanism at all). The mechanism inspired by the PP principle fails to be acceptable in general. For instance, the most upstream polluter of a river prefers the *laisser-faire* to the application of the PP principle. We nevertheless show that it is acceptable if externalities are multilateral among homogeneous agents. This is for instance the case in most the theoretical models examining international agreements for greenhouse gas emission reduction including Chandler and Tulkens (1992), Carraro and Siniscalco (1993) and Barrett (1994).

We then examine the problem of applying the polluter pays principle when emissions and damages are private information. We posit a direct revelation mechanism that implements the PP welfare distribution. It matches reports on emissions and damages to induce truth-telling in Nash equilibrium and by iterative elimination of dominated strategy. To that respect our mechanism differs from the ones in Dugan and Robert (2002) and Montero (2008) since they assume private information only on the polluters side not on the victim side. Here damages are also private information. In is worth to note that our interpretation of the PP principle and our characterization result of the PP welfare distribution relies on the assumption of constant marginal damage. In the concluding section we discuss on the application of the PP principle to increasing marginal damage, that is with a damage convex

2 A model with negative externalities

Consider a set $N = \{1, \dots, n\}$ of agents (countries, cities, farmers, firms, consumers,...). Each agent $i \in N$ pollutes or is polluted or both. Agent i enjoys a benefit $b_i(e_i)$ from producing and/or consuming where $e_i \geq 0$ is the level of economic activity hereafter called “emissions”. Assume b_i both strictly concave and strictly increasing from 0 to a maximum \hat{e}_i with $b'_i(\hat{e}_i) = 0$ for every $i \in N$.¹ and twice continuously differentiable (for all $0 \leq e_i < \hat{e}_i$, both $b'_i(e_i) > 0$ and $b''_i(e_i) < 0$) for every $i \in N$. We normalize $b_i(0) = 0$. We also assume that the marginal benefit at $e_i = 0$ is high enough (say infinite) so it is optimal that all agents produce and/or consume.

The pollution originated from i causes a marginal damage a_{ij} to agent j at least. We first assume constant marginal damage before extending to convex damage and thus increasing marginal damage from emissions. Let $Ri = \{j \in N | a_{ij} > 0\}$ denote the receptors of i 's pollution: the set of agents which are polluted by i . Let $R^0i = \{j \in N \setminus \{i\} | a_{ij} > 0\}$ the receptor of i 's pollution excluding i . Symmetrically, denotes by $Si = \{j \in N | a_{ji} > 0\}$ be the set of agents who pollute i for any $i \in N$. Let $S^0i = \{j \in N \setminus \{i\} | a_{ji} > 0\}$ be the set of agents who pollute i excluding i . The environmental damage caused to i by the emission vector $e = (e_i)_{i \in N}$ is thus

$$d_i = \sum_{j \in Si} a_{ji} e_j.$$

The welfare of agent i with emissions $e = (e_i)_{i \in N}$ is:

$$b_i(e_i) - d_i = b_i(e_i) - \sum_{j \in Si} a_{ji} e_j. \quad (1)$$

The first term in (12) is i 's benefit from its own emissions whereas the second term is i 's welfare loss due to pollution. For computational simplicity and without loss of generality, we assume $a_{ii} > 0$: any agent's activity generates some pollution to himself (e.g. some waste that it has to get ride of) or simply some cost.²

A (negative) externality or pollution problem (N, b, a) is defined by a set of agents N , a profile benefit functions $b = (b_i)_{i \in N}$, and a matrix of externality/pollution marginal impacts $a = [a_{ij}]_{ij \in N \times N}$.

¹This is without loss of generality since the maximum could be $\hat{e}_i = +\infty$ for some $i \in N$.

²This assumption is made to exclude corner solutions. It does not change qualitatively our results.

The externality problem (N, b, a) is with only multilateral externalities if $S_i = R_i$ for any $i \in N$. It is with only unilateral externalities if $S^0_i \cap R^0_i = \emptyset$ for any $i \in N$. Let us denote by $V \subset N$ the set of only victims of pollution. Any agent $i \in V$ does not pollute another agent and suffer from pollution due to other agent's activities. Formally, for every $i \in V$, $a_{ij} = 0$ for every $j \neq i$ and $a_{ji} > 0$ for at least one $j \neq i$, or equivalently $R^0_i = \emptyset$ and $S^0_i \neq \emptyset$. Similarly let $P \subset N$ the set of only polluter agents. Any agent $i \in P$ pollutes at least another agent and does not suffer from other agents' pollution, that is $a_{ij} > 0$ for one $j \neq i$ at least and $a_{ji} = 0$ for every $j \neq i$, that is $R^0_i \neq \emptyset$ and $S^0_i = \emptyset$.

An *efficient* emissions plan $e^* = (e_i^*)_{i \in N}$ maximizes total welfare $\sum_{i \in N} [b_i(e_i) - d_i] = \sum_{i \in N} b_i(e_i) - \sum_{i \in N} \sum_{j \in S_i} a_{ji} e_i$. It satisfies the following first-order conditions for every $i \in N$:

$$b'_i(e_i^*) = \sum_{j \in R_i} a_{ij}, \quad (2)$$

The marginal benefit of pollution emitted by i should be equal to its marginal damage for society.

Example 1: The river pollution problem.

Agents are countries, cities, factories located along a river. The set of predecessors of i in the river is S^0_i while the set of followers of i is R^0_i . Each agent i emits e_i units of pollution which impact its followers downstream: one unit emitted in i causes a marginal damage a_{ij} in j . Symmetrically, agent i suffer from pollution emitted upstream by agents in P^0_i and by himself.³ It is a case of unilateral externalities: if we take two agents i and j , either i is upstream j or i is downstream j , i.e $i \in S_j$ or $i \in R_j$. In a single canal or one-tributary river, agents can be ordered according to their position from upstream to downstream. In this case, if $N = \{1, \dots, n\}$ and if agents suffer from their own pollution (e.g. countries) then $R_1 = N$, $S_1 = \{1\}$, $R_n = \{n\}$ and $S_n = N$. Moreover, for any i and j , if $j \in P_i$ then $P_j \subset P_i$. Symmetrically, if $j \in R_i$ then $R_j \subset R_i$. The later properties might not hold in more general rivers. With several tributaries than end up on the same main course, then for any agent i

³In the case of a river, "linearity is a good approximation up to the point at which the river becomes so overloaded with organic material that oxygen (needed for aerobic bacteriological decomposition) is depleted. At that point, [refereed as the river carrying capacity] the river's capacity to clean itself is greatly diminished." from Kolstad, footnote 2 page 177.

there might be $k, j \in Si$ but $k \notin Sj$ and $j \notin Sk$. Symmetrically, for river deltas or irrigations ditches originated from a source or weed or reservoir, we have the reverse: for any agent i there might be $k, j \in Ri$ but $k \notin Rj$ and $j \notin Rk$.

Example 2: The international Greenhouse gas emissions game

Players are countries. Each country i enjoys a benefit b_i from its own greenhouse gas emissions e_i . Greenhouse gases emitted on atmosphere causes global warming that damages countries' economies. The magnitude of global warming depends on total emissions on the earth surface $\sum_{j \in N} e_j$. Suppose that it causes a constant marginal damage of δ_i to country i . In this example, $Si = Ri = N$ and $a_{ii} = a_{ij} = \delta_i$ for every $i \in N$: all countries exert multilateral externalities on all other countries of the same magnitude. Yet countries differ on the damage that externalities cause on their economy. Seminal papers on international agreements for greenhouse emission reduction (Chandler and Tulkens, 1992, Carraro and Siniscalco, 1993, Barrett, 1994) rely on these assumption except that they consider convex damage (or concave benefit of emission abatements) and, therefore, increasing marginal damage.

Example 3: The international acid rain game.

Agents are countries. They emits sulfur dioxide by burning coal for power production. This causes acid rain which damages forests and ecosystems in neighboring countries. The parameter a_{ij} captures the marginal impact of country i 's SO2 emissions to acid rain in country j . It depends on the fraction of emissions from i that is deposited in j and its marginal damage on j . Mäler and De Zeeuw (1998) provides estimations on those parameters for 1990 and 1991 in Europe. For instance, among the SO2 emissions from Belgium, 19.4% ended up in Belgium, 13.3% in Germany, 9% in France, 4.8% in Netherland and so on. Mäler (1989,1994) consider a acid rain gain with theses heterogeneous "transportation" parameters and constant marginal damage. It has been extended by Finus and Tjøtta (2003) and Mäler and De Zeeuw (1998) with convex marginal damage.

Example 4: Polluters versus victims

Agents in V are consumers and those in $N \setminus V$ are firms. Firms emit pollution without incur-

ring any damage: $a_{ij} = 0$ for every $j \in N \setminus V$. In contrast, consumers do not emit pollution but suffer from pollution: $\hat{e}_i = 0$ for every $i \in V$ and $a_{ji} > 0$ for one $j \in N \setminus V$ at least. In this case, a_{ji} can be interpreted the marginal damage of each unit of pollution originated from firm j causes to consumer i in term of health or environmental impact. It depends on technologies, distance between firms and consumers, climatic condition, and so on. The victim of pollution might also be firms involved in different sectors than the polluter ones; for instance hotel and restaurants located close to a lake or sea shore that might be polluted by local factories. The main difference with the previous example is that emitter and victims of pollution are distinct agents.

We examine regulation mechanisms in pollution problems. A regulation mechanism or simply a mechanism is a vector of payments contingent on emissions $t(e) = (t_i(e))_{i \in N}$. It assigns to agent i a payment $t_i(e)$ for any emissions plan $e = (e_i)_{i \in N}$. The mechanism $t(e)$ with the emission plan e yields to agent i a welfare of

$$b_i(e_i) - d_i + t_i(e) = b_i(e_i) - \sum_{j \in S_i} a_{ji} e_j + t_i(e_i). \quad (3)$$

In the non-cooperative Nash equilibrium of the externality problem with the mechanism $t(e)$, each player i maximizes (3) with respect to e_i given e_j for every $j \neq i$. Let us denote e^e a Nash equilibrium plan. Agent i 's equilibrium welfare with $t(e)$ is:

$$z_i = b_i(e_i^e) - d_i^e + t_i(e^e),$$

with $d_i^e = \sum_{j \in S_i} a_{ji} e_j^e$. The total welfare is

$$W = \sum_{i \in N} z_i = \sum_{i \in N} [b_i(e_i^e) - d_i^e + t_i(e^e)] = \sum_{i \in N} b_i(e_i^e) - \sum_{i \in N} d_i^e + \sum_{i \in N} t_i(e^e)$$

The first right-hand term is total benefit from emission, the second is total damage and the third is the regulation mechanism surplus (or deficit if negative).

A particular regulation mechanism is the *laissez-faire* defined by $t_i(e) = 0$ for all $i \in N$ and any $e \in \mathbb{R}_+^n$. It implements the emissions plan $e^{lf} = (e_i^{lf})_{i \in N}$ that satisfies the following first-order conditions,

$$b'_i(e_i^{lf}) = a_{ii},$$

for every $i \in N$. In contrast to the efficient solution, under laissez-faire each agent i considers the impact of its emissions only on its own welfare. As long as $a_{ij} > 0$ for some $j \neq i$, i.e. i 's emissions have an impact on another agent j , then $e_i^{lf} > e_i^*$ and therefore $p_j^{lf} > p_j^*$ for every $j \in R_i$.

Another mechanism is the emission norm. It defines upper bounds of emission \bar{e}_i and penalties for exceeding these bounds, formally,

$$t_i(e) = \begin{cases} 0 & \text{if } e_i \leq \bar{e}_i \\ -F_i(e_i - \bar{e}_i) & \text{if } e_i > \bar{e}_i \end{cases}$$

for every $i \in N$ where F_i is the fine in case of excess pollution (which can be infinite or lump-sum). In case of an uniform norm $\bar{e}_i = \bar{e}$ and $F_i = F$ for every i . If the fine is high enough to be persuasive and the norm is binding in the sense that $e_i^{nc} > \bar{e}_i$, the emission plan implemented in Nash equilibrium are $e_i^e = \bar{e}_i$ for every polluter $i \in N$.

The emission fee defines $t_i(e) = -\tau_i e_i$ where $\tau_i > 0$ is the level of tax assigned to polluter i . Similarly, a fee on pollution charges $\tau_j > 0$ per unit of pollution for each receptor j . It translates into a scheme $t_i(e) = -\sum_{j \in R_i} a_{ij} e_i \tau_j$ to be paid to agent i . The Pigouvian fee is $\tau_i = \sum_{j \in R_i^0} a_{ij}$ for every polluter $i \in N$. It implement first-best emissions e^* in a Nash equilibrium. The money collected by the regulator might be redistributed through lump-sum transfers or to compensate for environmental damage d_i .

Other schemes: emissions or pollution tradable permits with different initial allocation of permits (auctioned, grandfathering,...), etc.

We want a regulation mechanism scheme to satisfy the following two axioms.

The first one requires that the first-best outcome is implemented.

Efficiency: $t(e)$ is efficient if $e^e = e^*$.

The second axiom imposes budget-balancing.

Budget balance: A regulation mechanism $t(e)$ is budget balanced if $\exists e^e : \sum_{i \in N} t_i(e^e) \leq 0$.

In addition to an emission plan e , a regulation mechanism implements a distribution z of the total welfare W . We want that the welfare distribution z implemented by the regulation mechanism z to satisfy two axioms. The first axiom requires that any agent should receive a

non-negative payoff.

Non-negativity: For all $i \in N$, $z_i \geq 0$.

If non-negativity is not met, then some agents receive a negative payoff even if they decide not to pollute, i.e. with zero emissions. As long as $W > 0$, since $\sum_{i \in N} z_i = W$, some agents obtains a positive welfare with polluting activities.

The second axiom renders the polluter responsible to any change of its pollution impacts on the economy.

Responsibility for pollution impact (RPI) Consider any arbitrary agent $i \in N$. Suppose that agent i 's pollution impact is reduced or augmented from $(a_{ij})_{j \in Ri}$ to $(a'_{ij})_{j \in Ri}$ with $(a_{ij})_{j \in Ri} \neq (a'_{ij})_{j \in Ri}$, every other pollution impacts $(a_{lj})_{j \in Ri}$ being unchanged for every $l \in N \setminus i$. The regulation mechanism $t(e)$ renders agents responsible for their pollution impact if for any i , any modification $(a'_{ij})_{j \in Ri}$ of i 's pollution impact and corresponding equilibrium individual welfare and social welfare z'_i and W'

$$z'_i - z_i = W' - W.$$

The responsibility for pollution impact axiom assigns to any agent the full return or loss to any change of its own pollution impact. In addition to be a fairness principle, the RPI axiom has attractive incentive properties. Suppose that an agent is able to reduce its impact on pollution at a cost by say switching to a greener technology, reducing or cleaning its wastes, improving energy efficiency or using less toxic inputs. By assigning the full return to this pollution reduction, the RPI axioms provides efficient incentives to invest in pollution impact reduction. Symmetrically, if an agent benefit from increasing its pollution impact per unit of emissions (e.g. using higher sulfur content coal), the RPI assigns to this agent the economic cost of this extra pollution.

Among the above regulations, the Pigouvian fee is efficient. It is budget balanced if the revenue collected is redistributed to agents. The welfare distribution it implements does not satisfy non-negativity since victim-only agents (i.e. agents $i \in V$) are not compensated for the environmental damage they incurs. The welfare distribution with emission Pigouvian fee also satisfies RPI. A emission norm $\bar{e}_i = e_i^*$ with a persuasive fine (e.g. infinite) is efficient

and budget balanced but its welfare distribution does not satisfy RPI and non-negativity. A cap-and-trade system (tradable pollution allowances) for pollution at each receptor with grandfathering is efficient, budget balanced but the welfare distribution it leads to does not satisfy non-negativity since victim only agents are not compensated entirely. It might or might not satisfy RPI depending on the initial allocation of permits. A similar cap-and-trade system where permits are auctioned is efficient and satisfies RPI (to check) but is not budget balanced unless the money collected is redistributed.

3 A characterization of the polluter-pays principle

Many countries have adopted the “polluter pays” (PP) principle. It basically renders the polluter responsible for the damage it causes to the environment. It requires that *the costs of pollution should be borne by the entity which profits from the process that causes pollution*. In order to satisfy the polluter-pays principle, the entity who pollutes should compensate those who suffer from this pollution for the damages it causes. If not the case, someone else pays for it, either the victim of pollution (e.g. under *laissez-faire*) or other agents. In our model, an arbitrary agent i should compensate every agent $j \in R_i^0$ from the damages $a_{ij}e_i$ causes. Agent i pays $a_{ij}e_i$ to every $j \in R_i^0$. It also receives compensations $a_{ji}e_j$ from all agents $j \in S_i^0$ who pollute it. Summing up all these side-payments, the polluter-pays principle defines a regulation mechanism (hereafter called the polluter-pays or PP mechanism) denoted $t^{PP}(e)$ such that for every agent $i \in N$:

$$t_i^{PP}(e) = \sum_{j \in S_i^0} a_{ji}e_j - \sum_{j \in R_i^0} a_{ij}e_i = d_i - a_{ii}e_i - \sum_{j \in R_i^0} a_{ij}e_i = d_i - \sum_{j \in R_i} a_{ij}e_i. \quad (4)$$

Agent i receives the cost of pollution it suffers from net of the cost of pollution it causes to society. Since the polluter-pays principle involves side-payments among agents, the PP mechanism sums-up to zero. It is therefore budget-balanced. Agent i 's welfare under the PP mechanism $t^{PP}(e)$ with emission plan e is:

$$b_i(e_i) - \sum_{j \in R_j} a_{ij}e_i \quad (5)$$

Since agent i pays for the marginal damage caused to others and is compensated from the marginal damage caused by others, its welfare under the PP mechanism in (5) is the social

benefit from its economic activity. It therefore has incentive to emit the efficient level e_i^* for any given emissions emitted by other agents. Indeed maximizing (5) with respect with e_i leads to the first-order condition (2) which implies $e_i^e = e_i^*$ for every $i \in N$. This implies that the PP regulation mechanism implements the efficient emission plan e^* in Nash equilibrium. Note that this individual's payoffs depend only on the agent's own choice (no externality), the efficient emission plan is a dominant strategy equilibrium, which is an equilibrium concept which is less demanding in term of cognitive skills of agents than the Nash Equilibrium. Agent i 's equilibrium welfare under the PP mechanism is:

$$z_i^{PP} = b_i(e_i^*) - \sum_{j \in R_j} a_{ij} e_i^* \quad (6)$$

Theorem 1 *Among the efficient and budget-balanced regulation mechanisms, the polluter-pays principle implements the unique welfare distribution that satisfies non-negativity and responsibility for pollution impact.*

Proof. First, we show that if a welfare distribution satisfies non-negativity and responsibility for pollution impact, then it must be the polluter-pays welfare distribution z^{PP} . Consider another $\tilde{t}(e)$ with corresponding equilibrium individual welfare \tilde{z}_i for every $i \in N$ and total welfare \tilde{W} with $\tilde{z} \neq z^{PP}$. Since $t^{PP}(e)$ is efficient and sums-up to zero, $\tilde{W} \leq W^{PP}$ with $W^{PP} \equiv \sum_{i \in N} z_i^{PP}$. By definition of W , it implies $\sum_{i \in N} \tilde{z}_i \leq \sum_{i \in N} z_i^{PP}$ which, combined $\tilde{z} \neq z^{PP}$ forces $\tilde{z}_i < z_i^{PP}$ for one $i \in N$ at least. Consider a pollution impact decrease a' such that $a'_{ii} < a_{ii}$ and everything else remains identical, i.e. $a'_{lj} = a_{lj}$ for all $l, j \in N$ such that $lj \neq ii$. Pick a'_{ii} sufficiently low such that $a'_{ii} < z_i^{PP} - \tilde{z}_i$. Since $b'(e^{lf}) = a_{ii}$, we have:

$$b'(e^{lf}) < z_i^{PP} - \tilde{z}_i \quad (7)$$

Let $z_j'^{PP}$ and \tilde{z}'_j denote individual welfare implemented by the regulation mechanisms $t^{PP}(e)$ and $\tilde{t}(e)$ of the new externality problem (N, a', b) for any any $j \in N$. By responsibility for pollution impact,

$$\tilde{z}'_i - \tilde{z}_i = z_i^{PP} - z_i'^{PP}$$

Rearranging terms and using the definition of $z_i'^{PP}$ leads to

$$z_i'^{PP} - \tilde{z}_i = b(e_i^{I*}) - \sum_{j \in Fi} a'_{ij} e_i^{I*} - \tilde{z}'_i,$$

where e^{l*} denotes the efficient emission plan for (N, a', b) . Now since $b_i(e_i^{lf}) > b_i(e_i^{l*})$, $a'_{ij} \geq 0$ for all $j \in Fi$, and $z'_i \geq 0$, the above equality contradicts (7).

Second, we already established that $t^{PP}(e)$ is efficient and budget-balanced. We now show that z^{PP} satisfies non-negativity and responsibility for pollution impact. For non-negativity,

$$z_i^{PP} = b_i(e_i^*) - \sum_{j \in Ri} a_{ij} e_i^* = \max_{e_i} \left\{ b_i(e_i) - \sum_{j \in Ri} a_{ij} e_i \right\} \geq b(0) - \sum_{j \in Ri} a_{ij} \times 0 = 0.$$

where the inequality follows from the fact that agent i can always choose $e_i = 0$ (no emission or production).

For efficient incentives for pollution reduction, for any agent i , consider any change of pollution impact from a to a' for agent i 's pollution only: $(a_{ij})_{j \in N} \neq (a'_{ij})_{j \in N}$ and $(a_{kj})_{j \in N} = (a'_{kj})_{j \in N}$ for any $k \neq i$. Let z_i^{PP} and $z'_i{}^{PP}$ be i 's equilibrium welfare under the PP mechanism in (N, a, b) and (N, a', b) respectively and by W^{PP} and W'^{PP} the corresponding total welfare. Similarly, denote by e^* and e'^* the efficient emission plan of (N, a, b) and (N, a', b) respectively. By definition,

$$z'_i{}^{PP} - z_i^{PP} = b(e'^*_i) - \sum_{j \in Ri} a'_{ij} e'^*_{ij} - \left(b(e^*_i) - \sum_{j \in Ri} a_{ij} e^*_{ij} \right), \quad (8)$$

Since $a_{kj} = a'_{kj}$ for every $k \neq i$, the efficient emission levels are not affected by the change of matrix of pollution impacts from a to a' which implies $e^*_k = e'^*_k$ for every $k \in N \setminus i$. Therefore, we have:

$$W'^{PP} - W^{PP} = b(e'^*_i) - \sum_{j \in Ri} a'_{ij} e'^*_{ij} - \left(b(e^*_i) - \sum_{j \in Ri} a_{ij} e^*_{ij} \right)$$

which, combined with (8), leads to $z'_i{}^{PP} - z_i^{PP} = W'^{PP} - W^{PP}$. \square

4 Acceptability of the polluter-pays principle

In modern society, environmental regulations emerge from negotiation among stakeholders. Sovereign countries negotiate to design environmental international agreements such as the Kyoto protocol. Firms, public authorities and NGO are involved in the debates on the design of regulations. In this section, we analyze such negotiations using a cooperative game theory

approach. We examine whether the polluter-pays mechanism is the preferred mechanism for any possible coalition of agents. That is if a group of agents can be better-off by agreeing on another way to regulate externalities among them. For instance, in the case of international agreement, a group of countries would refuse to agree to apply the polluter-pays mechanism if it can achieve a higher welfare with another agreement among them.

Our analysis requires new notation from cooperative game theory. A coalition is a non-empty subset of N . For any coalition $T \subset N$, we denote $ST = \{j \in N : \text{for some } i \in T, a_{ji} > 0\} = \cup_{i \in T} Si$ the set of agents which pollute some of the members in T and $RT = \{j \in N : \text{for some } i \in S, a_{ij} > 0\} = \cup_{i \in T} Ri$ the set of victims of coalition S 's members. Similarly, $S^0T = ST \setminus T$ the set of agents outside of T which pollute some of the members of T and $R^0T = RT \setminus T$ denote the set of agents outside of T which are polluted by some of the members in T .

For any coalition T , let $e_T = (e_i)_{i \in T}$. We need to define the welfare that a coalition T can achieve by agreeing on a regulation mechanism. This welfare depends on the behavior of agents outside the coalition. We assume that if a coalition T form, they agree to implement the mechanism that maximizes their joint payoff given the behavior of agents outside T . For our purpose of computing T 's welfare, it is equivalent to agree on an emission plan e_T among members of T . Agents outside T behave the same way: they pick the emission plan that maximizes the total welfare of the coalition they belong in given the behavior of outsiders. Therefore, the welfare that a coalition can achieve depends mainly the cooperative behavior of agents in N which is summarized by a partition \mathcal{P} of N .

More precisely, for any coalition T and emissions e'_j for agents $j \in N \setminus T$ outside T , members of coalition T would implements the emission plan solution to:

$$\max_{e_T} \sum_{i \in T} \left(b_i(e_i) - \left(\sum_{j \in T} a_{ji} e_j + \sum_{j \in N \setminus T} a_{ji} e'_j \right) \right).$$

The first-order conditions defines the emissions e_i^T of members of coalition T :

$$b'_i(e_i^T) = \sum_{j \in Ri \cap T} a_{ij}, \tag{9}$$

for every $i \in T$. Each agent $i \in T$ internalizes his impact on the environmental damage only to members of T . Therefore e_i^T weakly decreases as T expands. It strictly decreases if T expands

by including new members in R^0T . Notice that the linearity of the damage function

For any given partition \mathcal{P} let denote by $e^{\mathcal{P}}$ the emission implemented by agents given that each coalition $T \in \mathcal{P}$ chooses e_T^T defined by (9). Given a coalition structure \mathcal{P} , (9) uniquely determines a Nash equilibrium were all coalitions in \mathcal{P} play non-cooperatively against each other. The welfare that any coalition $T \in \mathcal{P}$ can achieve is:

$$v(T, \mathcal{P}) = \max_{e_T} \sum_{i \in T} \left(b_i(e_i) - \left(\sum_{j \in T} a_{ji} e_j + \sum_{j \in N \setminus T} a_{ji} e_j^{\mathcal{P}} \right) \right),$$

That is,

$$v(T, \mathcal{P}) = \sum_{i \in T} \left(b_i(e_i^{\mathcal{P}}) - \left(\sum_{j \in Ri} a_{ji} e_j^{\mathcal{P}} \right) \right)$$

Importantly, with constant marginal damage, coalition T 's emission allocation e_T^T does not depend on the outsider emissions $e_i^{\mathcal{P}}$ for $i \in N \setminus T$. In particular, the members of T choose the same emissions if the others agents coordinate their emissions (by forming coalitions) or not. Hence, although $v(T, \mathcal{P})$ depends on what is emitted outside T (by members of S^0T), the behavior of members of T is not affected by the one of outsiders. The latter property implies that, since for any $j \in N$, condition (9) with $T = Ri$ is equivalent to (2) for $i \in N$, we have $e_i^{Ri} = e_i^N$. More generally, for any $T \subseteq N$, we have $e_{RT}^{RT} = e_{RT}^N$.

For any two coalition structures \mathcal{P} and \mathcal{P}' , we say that \mathcal{P}' is coarser than \mathcal{P} if for all $T \in \mathcal{P}'$ there exist $S_1, \dots, S_k \in \mathcal{P}$ such that $T = \cup_{l=1}^k S_l$.

Lemma 1 *For all coalition structures \mathcal{P} and \mathcal{P}' and all $T \in \mathcal{P} \cap \mathcal{P}'$, if \mathcal{P}' is coarser than \mathcal{P} , then $v(T, \mathcal{P}) \leq v(T, \mathcal{P}')$.*

Proof. Note that by (9), $e_T^{\mathcal{P}} = e_T^{\mathcal{P}'}$. Let $U \in \mathcal{P}$ and $U' \in \mathcal{P}'$ be such that $U \subseteq U'$. Then for all $i \in U$, $Ri \cap U \subseteq Ri \cap U'$. Thus, for all $i \in U$,

$$\sum_{j \in Ri \cap U} a_{ij} \leq \sum_{j \in Ri \cap U'} a_{ij}$$

and $b'_i(e_i^{\mathcal{P}}) \leq b'_i(e_i^{\mathcal{P}'})$, which implies $e_i^{\mathcal{P}} \geq e_i^{\mathcal{P}'}$. Since $e_T^{\mathcal{P}} = e_T^{\mathcal{P}'}$ and $e_i^{\mathcal{P}} = e_i^{\mathcal{P}'}$ for every $i \notin T$, we now obtain $v(T, \mathcal{P}) \leq v(T, \mathcal{P}')$, the desired conclusion. \square

Lemma 1 shows that starting from any coalition structure, any coalition is better off from more cooperation, i.e. from a more global agreement among the other coalitions.

For any $T \subseteq N$, let $\bar{v}(T) = v(T, \{T, N \setminus S\})$. Now by Lemma 1, for any coalition structure \mathcal{P} such that $T \in \mathcal{P}$, we have $v(T, \mathcal{P}) \leq \bar{v}(T)$. Note that this is in contrast to the river sharing problem with satiable benefit functions.

Core lower bounds: A welfare distribution z satisfies all core lower bounds if for every $T \subseteq N$, $\sum_{i \in T} z_i \geq \bar{v}(T)$.

Let $\mathcal{P}^* = \{\{i\} : i \in N\}$ and for all $T \subseteq N$, let $\mathcal{P}_T^* = \{\{i\} : i \in N \setminus T\} \cup \{T\}$ and $\underline{v}(T) = v(T, \mathcal{P}_T^*)$.

Non-cooperative core lower bounds: A welfare distribution z satisfies non-cooperative core lower bounds if for every $T \subseteq N$, $\sum_{i \in T} z_i \geq \underline{v}(T)$.

Proposition 1 *The polluter-pays welfare distribution z^{PP} might not satisfy non-cooperative core lower bounds.*

Proof. First, note that for any agent i ,

$$y_i^{PP} = b_i(e_i^*) - \sum_{j \in F^i} a_{ij} e_j^* = b_i(e_i^*) - a_{ii} e_i^* - \sum_{j \in F^{0i}} a_{ij} e_j^*$$

and the non-cooperative core lower bound for $S = \{i\}$ is

$$\underline{v}(i) = b_i(e_i^{lf}) - a_{ii} e_i^{lf} - \sum_{j \in P^{0i}} a_{ji} e_j^{lf}$$

Since e_i^{lf} maximizes $b_i(e_i) - a_{ii} e_i$, as long as agent i exerts some externalities i.e. $F^{0i} \neq \emptyset$, a sufficient condition for the core lower bound for coalition $\{i\}$ to be violated, i.e. for $z_i^{PP} < \underline{v}(i)$, is

$$\sum_{j \in R^{0i}} a_{ij} e_j^* \geq \sum_{j \in S^{0i}} a_{ji} e_j^{lf}. \quad (10)$$

Obviously, (10) holds if agent i is not polluted, i.e. $P^0i = \emptyset$. More generally, it holds when the damage from others at the laissez-faire is lower than the damage to others at the first-best. In this case, the payment imposed to agents i by the polluter-pay principle exceed the non-cooperative cost of pollution for i . One can find examples where the left-hand term in (10) is high enough while the right-hand term is low enough with unilateral externalities. With multilateral externalities $S^0i = R^0i$ and thus, (10) becomes:

$$\sum_{j \in S^0i} a_{ij}e_i^* \geq \sum_{j \in S^0i} a_{ji}e_j^{lf}.$$

The above condition holds if i has a large marginal impact on the others while the others have a low impact on i . Then i is required to pay a lot when his gain is low. We can find examples where (10) is violated. \square

Interestingly, the polluter pays welfare distribution satisfies all non-cooperative core lower bounds in a symmetric pollution environment, i.e. where all benefit functions are identical and pollution impacts are the same across all agents.

Proposition 2 *Under multilateral externalities and homogenous agents where for all $i, j \in N$, $b_i = b_j$ and $a_{ij} = a_{ji}$, the polluter-pay welfare distribution z^{PP} satisfies all non-cooperative core lower bounds.*

Proof. For all $i, j \in N$, let $b = b_i = b_j$ and $a = a_{ij} = a_{ji}$. Since $a_{ii} > 0$ for all $i \in N$, we have $S_i = R_i = N$ for all $i \in N$. Let $\emptyset \neq T \subseteq N$ and $e_N^T = e_N^{\{T, N \setminus T\}}$.

Now for all $i \in T$,

$$\begin{aligned} z_i^{PP} &= b(e_i^*) - \sum_{j \in N} a e_j^* \\ &= b(e_i^*) - e_i^* \sum_{j \in N} a \\ &= \max_{e_i \geq 0} \{b(e_i) - e_i \sum_{j \in N} a\} \\ &\geq b(e_i^T) - e_i^T \sum_{j \in N} a \\ &= b(e_i^T) - \sum_{j \in T} a e_j^T - \sum_{j \in N \setminus T} a e_j^T \end{aligned}$$

$$\geq b(e_i^T) - \sum_{j \in T} ae_j^T - \sum_{j \in N \setminus T} ae_j^T.$$

where the first equality is the definition of the polluter-pay welfare distribution, the second follows from the fact that in the homogenous case, for all $j \in N$, $e_i^* = e_j^*$, and the last inequality follows again from the fact that in the homogenous case, for all $j \in T$, $e_i^T = e_j^T$, and that for all $j \in N \setminus T$, $e_j^T \geq e_i^T$.

Now the above implies

$$\sum_{i \in T} z_i^{PP} \geq \sum_{i \in T} \left(b(e_i^T) - \sum_{j \in T} ae_j^T - \sum_{j \in N \setminus T} ae_j^T \right) = \underline{v}(T).$$

Hence, z^{PP} satisfies non-cooperative core lower bounds, the desired conclusion. \square

5 Implementation under asymmetric information

The regulator cannot observe emissions e_i neither damage due to pollution d_i (or the level of pollution at i) which is agent i 's private information for every $i \in N$. Yet she can try to elicit information by asking agents to report their emissions and pollution levels. Let denote agent i 's report on its own emissions by \hat{e}_i and on the damages due to pollution by \hat{d}_i . A regulation mechanism under asymmetric information is a vector of transfers contingent on reports $\tau(\hat{e}, \hat{d}) = (\tau(\hat{e}, \hat{d}))_{i \in N}$. Sequence of moves is as follow. First, the regulator imposes a regulation mechanism $\tau(\hat{e}, \hat{d})$. Second, agents chooses their emission levels e simultaneously and non-cooperatively. Third, each agent reports how much it emitted \hat{e}_i and what is the damage from pollution \hat{d}_i . Fourth, the regulator collects all reports, compute payments $\tau(\hat{e}, \hat{d})$ and pays or charges agents accordingly.

Consider the following mechanism. The regulator applies the polluter-pays principle based and the following computed emissions and damages based on the agents' reports \hat{e}_i and \hat{d}_i for every $i \in N$:

$$\underline{d}_i = \min\{\hat{d}_i, \sum_{j \in S_i} a_{ji} \hat{e}_j\},$$

$$\bar{e}_i = \max\{\hat{e}_i, \sum_{j \in R_i} \alpha_{ij} \hat{d}_j\},$$

where $\alpha = (\alpha_{ij})_{i,j \in N \times N}$ is a $N \times N$ matrix such that $\alpha a = I$ where I is the identity matrix.⁴ That is it pays the computed damage \underline{d}_i and charges the computed impact $\sum_{i \in Ri} a_{ij} \bar{e}_i$ to every agent $i \in N$. The net transfer to agent i is thus $\tau_i(\hat{e}, \hat{d}) = \underline{d}_i - \sum_{i \in Ri} a_{ij} \bar{e}_i = \min\{\hat{d}_i, \sum_{j \in Si} a_{ji} \hat{e}_j\} - \sum_{i \in Ri} a_{ij} \max\{\hat{e}_i, \sum_{j \in Ri} \alpha_{ij} \hat{d}_j\}$.

Proposition 3 *If $P = V = \emptyset$, truth-telling $\hat{e} = e$ and $\hat{d} = d$ is the unique equilibrium by iterative elimination of dominated strategies in the report subgame.*

Proof.

Step 1: We show that reporting higher emissions $\hat{e}_i > e_i$ and lowest damages $\hat{d}_i < d_i$ is a dominated strategy for any arbitrary agent i . We start by showing that i 's welfare $b_i(e_i) - d_i + \tau_i(\hat{e}, \hat{d})$, is never lower and can be sometime higher by reporting truthfully d_i rather than $d'_i < d_i$ for any \hat{e} and $i \in N \setminus P$. Underreporting $d'_i < d_i$ leads to a computed damage level $\underline{d}_i = \min\{d'_i, \sum_{j \in Si} a_{ji} \hat{e}_j\} < d_i$ which implies a transfer $\tau_i(\hat{e}, \hat{d})$ weakly lower than with truthfully reporting d_i . Indeed it is equal if the reported emissions \hat{e} are such that $\sum_{j \in Si} a_{ji} \hat{e}_j \leq d'_i$ and strictly lower if $d'_i < \sum_{j \in Si} a_{ji} \hat{e}_j$. Therefore i 's payoff is weakly lower if i reports $d'_i < d_i$ instead of d_i for any $i \in N \setminus P$. The same argument holds for emissions reported for any agent $i \in N \setminus V$. Suppose that i reports $e'_i > e_i$. Then, for a given reported damage \hat{d}_i , i ' transfer $\tau_i(\hat{e}, \hat{d})$ will be the same than by reporting truthfully if $\sum_{j \in Ri} \alpha_{ij} \hat{d}_j \geq e'_i$ and strictly lower if $\sum_{j \in Ri} \alpha_{ij} \hat{d}_j < e'_i$. Therefore so is i 's welfare which shows that over-reporting emissions is weakly dominated by reporting truthfully.

Step 2: We show that, assuming that all agents report undominated strategy, truth telling is a dominant strategy for any arbitrary agent $i \notin V \cup P$. By assumption $\hat{e}_j \leq e_j$ and $\hat{d}_j \geq d_j$ for every $j \in N$ which implies for any agent i :

$$\begin{aligned} \sum_{j \in Si} a_{ji} \hat{e}_j &\leq \sum_{j \in Si} a_{ji} e_j, \\ \sum_{j \in Ri} \alpha_{ij} \hat{d}_j &\geq \sum_{j \in Ri} \alpha_{ij} d_j. \end{aligned}$$

By definition, the two above equations imply:

$$\sum_{j \in S^0 i} a_{ji} \hat{e}_j + a_{ii} \hat{e}_i \leq d_i,$$

⁴ Since $ae = d$, $\alpha ae = \alpha d \iff Ie = e = \alpha d$.

$$\sum_{j \in R^0 i} \alpha_{ij} \hat{d}_j + \alpha_{ii} \hat{d}_i \geq e_i.$$

Given the above inequalities, for any undominated reports $\hat{e}_i \leq e_i$ and $\hat{d}_i \geq d_i$, the computed emission level and damage for i are $\bar{e}_i = \sum_{j \in R^0 i} \alpha_{ij} \hat{d}_j + \alpha_{ii} \hat{d}_i$ and $\underline{d}_i = \sum_{j \in S^0 i} a_{ji} \hat{e}_j + a_{ii} \hat{e}_i$. It leads to the following transfer for i :

$$\tau_i(\hat{e}, \hat{d}) = \sum_{j \in S^0 i} a_{ji} \hat{e}_j + a_{ii} \hat{e}_i - \sum_{i \in R i} a_{ij} \left(\sum_{j \in R^0 i} \alpha_{ij} \hat{d}_j + \alpha_{ii} \hat{d}_i \right)$$

Maximizing $\tau_i(\hat{e}, \hat{d})$ with respect to \hat{e}_i and \hat{d}_i subject to $\hat{e}_i \leq e_i$ and $\hat{d}_i \geq d_i$ leads to $\hat{e}_i = e_i$ and $\hat{d}_i = d_i$. \square

The intuition of the result is quite straightforward. First, over-reporting emissions and underreporting damages is definitively a bad idea since it increases the chance of paying more to compensate the others agents and getting less from the other agents. It is definitively a (weakly) dominated strategy. On the other hand, under-reporting emissions and over-reporting damages is tempting to reduce the compensation paid to others and increase the compensation received from others. Yet it pays for an agent i only if the other agents over-report emissions and underreport damages. In this case, the computed emission level \bar{e}_i which defines the compensation of i to others is based on under-reported damages by others whereas the computed damage \underline{d}_i which defines the compensation received by i is based on over-reported emissions by others. But, since it is a weakly dominated strategy, the agent i anticipates that this would never happen. That is all agents (including agent i) will report their right emissions or less and their right damage or more. It implies that the computed emission \bar{e}_i will be based on reported damages whereas the computed damage \underline{d}_i will be based on reported emissions. By under-reporting its own emissions e_i , agent i reduces its own responsibility on its own damage. It indeed decreases the computed damage \underline{d}_i and, therefore, the compensation it receives. Similarly, by over-reporting its own damage d_i , it increases its responsibility of its own emission on its own damage and thus the computed emission \bar{e}_i which increases the compensation it has to pay to others.

Proposition 4 *The mechanism $\tau(\hat{e}, \hat{d})$ is incentive-compatible: each agent i reports truthfully its own emission e_i and damage d_i at the Nash equilibrium in the message subgame (given*

that all other agents report the truth). It implements z^{PP} as a welfare distribution of the asymmetric information game.

Proof. TO BE COMPLETED.

6 Concluding remark: convex damages

We now consider the polluter-pays principle with convex damages which requires a slight modification of the model. We differentiate emissions from pollution and damage. The emission plan e generates a pollution level p_i at i 's location or to agent i for every $i \in N$ defined by:

$$p_i = \sum_{j \in S_i} a_{ji} e_j. \quad (11)$$

Pollution at level p_i causes damages $d_i(p_i)$ to i with d_i increasing and convex: $d_i'(p_i) > 0$ and $d_i''(p_i) \geq 0$ for every $p_i \in \mathbb{R}^+$ and $i \in N \setminus P$.⁵ The welfare of agent i with emissions $e = (e_i)_{i \in N}$ is:

$$b_i(e_i) - d_i(p_i), \quad (12)$$

with p_i defined by (11).

The first-best plan e^* is defined by the first-order conditions for every $i \in N$:

$$b_i'(e_i^*) = \sum_{j \in R_i} a_{ij} d_j'(p_j^*), \quad (13)$$

for every $i \in N$ with $p_j^* = \sum_{l \in S_j} a_{lj} e_l^*$ for every $j \in R_i$. The marginal benefit of emissions by i should be equal to its marginal cost for society which depends on its marginal impact on pollution and the marginal damage of pollution at each receptor. For each unit of emissions, a_{ij} units ends up at receptor j which causes marginal damages evaluated to $a_{ij} d_j'(p_j^*)$.

A regulation mechanism $t(e)$ implements an emission plan e^e such that e_i^e maximizes i 's welfare $b_i(e_i) - d_i(a_{ii} e_i + \sum_{j \in R_i^0} a_{ij} e_j^e) + t(e_i, e_{-i}^e)$ given the other agent's choices e_{-i}^e for every $i \in N$. Agent i 's equilibrium welfare under the mechanism $t(e)$ is:

$$z_i = b_i(e_i^e) - d_i(p_i^e) + t_i(e^e), \quad (14)$$

⁵Recall that P is the set of only polluter agents.

with $p_i^e = \sum_{j \in S_i} a_{ji} e_j^e$.

The *laissez-faire* emission plan e^{lf} maximizes $b_i(e_i) - d_i(a_{ii}e_i + \sum_{j \in S_i^0} a_{ji}e_j^{lf})$ with respect to e_i which yields to the following first-order conditions for every $i \in N$:

$$b'_i(e_i^{lf}) = a_{ii}d'_i(p_i^{lf}),$$

with $p_i^{lf} = \sum_{j \in S_i} a_{ji}e_j^{lf}$.

The definition of the four axioms that are efficiency, budget-balance and non-negativity and responsibility for pollution, are unchanged modulo the new definition for z_i for every $i \in N$. To define a regulation mechanism inspired by the polluter-pays principle we need to clarify the responsibility of a polluter agent i to the damage incurred by the polluted agent j . By definition, i 's emissions e_i increases ambient pollution at j by $a_{ij}e_i$. Yet, since the damage function is convex which implies that marginal damage is increasing, the damage due to the $a_{ij}e_i$ units of pollution out of a total of p_j is higher if applied on the first units than on the last unit. Formally, in the first case, i would have to pay $d_j(a_{ij}e_i)$ to j whereas in the second case, it pays $d_j(p_j) - d_j(p_j - a_{ij}e_i)$ which is strictly higher as long as $p_j > a_{ij}e_i$.⁶

Efficiency can be achieved by assigning responsibility to the last units of pollution: Any agent i is required to pay the last unit of damages which departs from the efficient level of pollution $d_j(\sum_{l \in S_j \setminus i} a_{lj}e_l^* + a_{ij}e_i) - d_j(p_j^* - a_{ij}e_i^*)$ with $p_i = \sum_{j \in S_i} a_{ji}e_j^*$ for every victim of i 's pollution $j \in R^0i$. The net transfer received by i with the emission plan e and corresponding pollution levels $(p_i)_{i \in N}$ is

$$t_i^{PP}(e) = \sum_{j \in S^0i} \left\{ d_i \left(\sum_{l \in S_j \setminus j} a_{li}e_l^* + a_{ji}e_j \right) - d_i(p_j^* - a_{ji}e_j^*) \right\} \\ - \sum_{j \in R^0i} \left\{ d_j \left(\sum_{l \in S_j \setminus i} a_{lj}e_j^* + a_{ij}e_i \right) - d_j(p_j^* - a_{ij}e_i^*) \right\},$$

with $p_j^* = \sum_{l \in S_j} a_{lj}e_l^*$. First, maximizing agent i 's welfare defined in (14) under the mechanism t^{PP} assuming that the other agents j emit e_j^* for every $j \in N \setminus i$ leads to the first-order condition (13) for every $i \in N \setminus V$. Therefore $t^{PP}(e)$ is efficient. Second, t^{PP} is also budget-balanced since, as before, it involves side-payments among agents: transfers $t_i^{PP}(e)$ sum-up to zero

⁶By convexity, $d_j(p_j) > d_j(p_j - a_{ij}e_i) + d_j(a_{ij}e_i)$ with $a_{ij}e_i < p_j$.

for every emission plan e . Agent i 's equilibrium welfare is:

$$z_i^{PP} = b_i(e_i^*) - d_i(p_i^*) + \sum_{j \in S^0 i} \{d_i(p_i^*) - d_i(p_j^* - a_{ji}e_j^*)\} - \sum_{j \in R^0 i} \{d_j(p_j^*) - d_j(p_j^* - a_{ij}e_i^*)\},$$

Third, $t^{PP}(e)$ satisfies non-negativity because

$$z_i^{PP} \geq b_i(0) - d_i(p_i^* - a_{ii}e_i^*) + \sum_{j \in S^0 i} \{d_i(p_i^*) - d_i(p_j^* - a_{ji}e_j^*)\} \geq 0.$$

where the first inequality follows from the fact that agent i can always choose $e_i = 0$ (no emission or production) in which case it pays nothing, produce and/or consume nothing (which yields $b_i(0) = 0$) and is compensated at least for its damage du to convexity of d_i . non-negativity and responsibility for pollution impact in addition to being efficient and budget-balanced.

Given than d_i is convex the sum of i 's compensations with the efficient plan e^* exceed i 's damage $d_i(p_i^*)$ which implies that any agent i is more than compensated for its damage $d_i(p_i^*)$. In particular, a victim only agent $i \in V$ obtains $\sum_{j \in S^0 i} (d_i(p_i^*) - d_i(p_j^* - a_{ji}e_j^*)) - d_i(p_i^*) > 0$ which is strictly positive if $|S_i^0| > 1$ (for instance if $S_i^0 = \{j, l\}$ and $a_{ji}e_j^* = a_{li}e_l^* = p_i^*/2$ i 's equilibrium welfare $z_i^{PP} = d_i(p_i^*) - 2d_i(p_i^*/2) > 0$ by convexity of d_i). One way to solve this problem is to repay only for the damage, thereby allowing for some budget surplus. In this case,

$$t_i^{PP}(e) = d_i(p_i^*) - \sum_{j \in Ri} \left\{ d_j \left(\sum_{l \in S^j \setminus i} a_{lj}e_l^* + a_{ij}e_i \right) - d_j(p_j^* - a_{ij}e_i^*) \right\},$$

Agent i is enterally compensated for the damage and pays the cost of the last units of pollution it is responsible (including to itself). It implements the efficient emission plan and the derived welfare distribution z^{PP} satisfies non-negativity

$$z_i^{PP} = b_i(e_i^*) - \sum_{j \in R^i} \{d_j(p_j^*) - d_j(p_j^* - a_{ij}e_i^*)\},$$

However, in general both regulation mechanisms based on the damage due to the last units of pollution fail to satisfy RPI. It is because with convex damage the equilibrium and efficient emission for other agents j changes if i 's pollution impacts a_{ij} change. Is it possible to satisfies efficiency and RPI? Could we find an impossible result for convex damage? We know that for the general case we can define a welfare distribution that satisfies RPI and non-negativity. Is there a regulation mechanism that implements efficiency and lead to this welfare distribution?

Remark: The polluter-pays principle can be implemented under asymmetric information with convex damages as well by replacing reports on damages by reports on ambient pollution \hat{p} . It requires to know the d_i functions for every $i \in N$ in addition to b and a .

References

- Ambec, S., and Y. Sprumont (2002): “Sharing a River,” *Journal of Economic Theory* 107:453–462.
- Barret, S. (1994): “Conflict and Cooperation in Managing International Water Resources,” Policy Research Working Paper #1303, The World Bank, Washington.
- Carraro, C and D. Siniscalco (1993) “Strategies for the International Protection of the Environment,” *Journal of Public Economics* 52:309–328.
- Chander P. and H. Tulkens (1997): “The Core of an Economy with Multilateral Environmental Externalities,” *International Journal of Game Theory* 26:379–401.
- Dugan, J. and J. Robert (2002): “Implementing the Efficient Allocation of Pollution,” *American Economic Review* 92(4): 1070–1078.
- Finus, M. and S. Tjøtta (2003): “The Oslo Protocol on sulfur reduction: the great leap forward?”, *Journal of Public Economics* 87:2031–2048.
- Fleurbaey, M. (2008): *Fairness, responsibility and welfare*, Oxford University Press, Oxford.
- Mäler, K.-G. and A. de Zeeuw: (1998) “The acid differential game,” *Environmental and Resource Economics* 12:167-184.
- Montero, J.-P. (2008): “A Simple Auction Mechanism for the Optimal Allocation of the Commons” *American Economic Review* 98(1):496-518.
- Montgomery, W.D. (1972): “Markets in Licenses and Efficient Pollution Control Programs,” *Journal of Economic Theory* 5:395–418.