Let the Punishment Fit the Crime: Enforcement with Error*

Indranil Chakraborty
Department of Economics
National University of Singapore
indro@nus.edu.sg

and

R. Preston McAfee
Google
preston@mcafee.cc

*We thank Dilip Mookherjee, Harry Paarsch, Larry Samuelson and Ilya Segal for several insights and comments.

1. Motivation

Divergence of social and private interests is a standard feature of many economic situations. Market failures may be minimized or avoided by an appropriate choice of a tax or subsidy schedule that induces the individual to internalize the external costs and benefits of his action. Consider, for instance, the external costs generated as a car travels at different speeds on an expressway. Suppose that as the car drives at a speed x it imposes a net social externality $s(x)$. The external effect includes the danger to other drivers in the event of an accident as well as the possibility of an accident itself, both of which vary with speed. If the driver has a net benefit $B(x)$ from driving at a speed x then he can be induced to drive at the socially optimal speed that maximizes $B(x)-s(x)$ by a *penalty* or *Pigouvian tax* $p(x)$, which, if speed is observed, satisfies $p(x)=s(x)$ regardless of the specific functional form of $B(x)$.

The tax achieves its purpose when the speed x of the car is perfectly measured. If, however, the technology allows only an imperfect measurement of the speed of the car then the driver of the car will not generally choose the socially optimal speed. As an example, suppose $s(x)$ is convex, and Y is the unbiased but imperfect measure of speed. Then $E[s(Y)\,|\,x]>s(x)$. Thus, a penalty function $p(y)=s(y)$ based on the observed speed y will usually result in a choice of speed which is different from the socially optimal speed.

The natural question then is whether a penalty function based on the imperfectly observed speed y can force the agent to internalize the cost $s(x)$ for driving at speed x regardless of his benefit function. We consider the following problem: Suppose a tax or subsidy function $s(x)$ associated with an action level x that is measured perfectly achieves a certain objective.[1] For ease of reference, let us call $s(x)$ a (*social*) *externality function*. The action x is only imperfectly observed as a random variable Y whose distribution depends on the true action x.

[1] The function $s(x)$ could be a Pigouvian tax, or alternately, $s(x)$ could be viewed as a price/penalty function that can help the market to co-ordinate towards the optimum described by Coase (1960). Generally, we will treat $s(x)$ to be a given target tax or transfer function.

We examine how and when a tax or subsidy function $p(y)$ of the observed signal gives rise to the same choice of action by the agent as the tax or subsidy function $s(x)$ regardless of the nature of his benefit function. In this case, there is a Pigouvian solution to the problem of externalities even when the behavior is observed with error.[2]

We will show that in a broad class of circumstances, there is indeed a solution to the problem for risk neutral agents. What is needed is enough information in the signal to separate distributions of behaviors. If two distinct distributions of behaviors produce the identical distribution of signals, then it is not possible to distinguish these distributions with any penalty function. If the function s separates the distributions of behaviors, then no solution can possibly exist, because the signal does not distinguish distributions that the penalty function separates. Therefore separating distributions is a necessary condition for implementation. The remarkable fact is that it is also sufficient. Moreover, the same condition conveniently works for both finite and continuous state spaces, although continuous state spaces may in some cases lead to approximate, rather than exact, implementation.

Interest in the implementability of desired outcomes by penalty functions is not new. Pigou (1952) suggested that forcing an agent to internalize the damage he causes by taxing him the amount of the damage would take the market toward efficiency. Coase (1960) critiqued that the Pigouvian scheme provides the wrong incentives. However, interpreted appropriately (so that the price suggested by Coase is the tax) the Pigouvian principle continues to hold in his examples. In fact, Sandmo (1975) showed that in the absence of government revenue requirements the Pigouvian tax implements the first best, and when there is a revenue requirement the Pigouvian principle extends appropriately. In all these cases, of course, the agent's action is perfectly observable and there is no uncertainty about any element of the model.

[2] Note that the role of y is purely that of a signal on the true action x that actually gives rise to the externality. If the observable y completely determines the level of externality then we are back in the perfect observability case.

Kwerel (1977), Dasgupta, Hammond and Maskin (1980), Duggan and Roberts (2002), among others, assume that there are a finite number of polluting firms with costs not observable to the regulator. They show that the first best outcome (that arises when costs are perfectly observable) is implementable in equilibrium. In these models, the emission level by each firm is perfectly observable. In reality, the total pollution is often not observable. While the rate at which a car pollutes can be observed through some tests, the amount of gasoline burnt or the number of city miles are not easy to observe. Montero (2005) assumes that the emission level is not observable but the emission rate is. The first best outcome cannot be implemented in this framework. In contrast, we consider a situation where the agent cannot perfectly control what the principal observes, *i.e.*, his action gives rise to a stochastic signal.

Similar problems arise in litigation contexts especially in the application of tort laws where advance contracting is not possible. Court rulings are not without errors. A judge may err in favor of the plaintiff just like she may err in favor of the defendant (see Posner, 1973, for a detailed discussion of errors in judicial administration). The presence of such errors has been argued to distort the actions of agents away from socially desirable levels. A part of the literature has taken an economic approach to describe the nature of damages that should be awarded as deterrence for agents to not deviate from the socially optimal behavior. Shavell (2011) shows that liability damages are welfare improving over taxation of actions when, for instance, the optimal level of action for the agent is not publicly known. In fact, if the victims do not bring liability suit or if the court may err in favor of the defendant with a positive probability then awarding only the liability damage is not sufficient to get the agent to take socially optimal decisions. In such cases, it is essential to incorporate a punitive damage and inflating the liability damage by a factor depending on the probability of the error (or the probability with which the suit may not be brought) to restore socially optimal behavior (see Shavell, 2011, and Polinsky and Shavell, 1998).

Png (1986) allows rulings to have both errors in favor of the defendant (type I error) and errors in favor of the plaintiff (type II error) and shows that in

order to make the injurer choose the socially optimal action it is necessary to include both a penalty and a subsidy over the liability. This paper is closest to the work of Png (1986) except that we allow injurer types to differ so that the "penalty" or "transfer" function has to get all types of injurers to act in a socially optimal (or some other predetermined) manner. Thus, while for Png the optimal transfer scheme always exists, the same cannot be guaranteed in our case. The relevant question then is what kind of "evidence" (i.e., the signal of the true action) must the penalty be based upon, to induce the socially optimal behavior by the injurer. We give an intuitively simple necessary and sufficient condition that the evidence must satisfy to implement the socially optimal behavior in such cases. In our case, the associated transfer to the victim may also involve punitive damage as well as subsidy depending on the nature of the evidence.

The principal-agent formulation of our problem demands a few words about its relationship with the agency literature. Our approach is distinct from the standard agency problems in that we require that a single penalty function implement the target externality function for all relevant action levels. In contrast, the agency literature usually employs a menu of contracts, treating the benefit function as private information and extracting it prior to the agent's choice of action, while we seek a single penalty function that works for all benefit functions.[3] The difference in the nature of the problems is most easily seen by observing that in the moral hazard model (cf. Holmstrom, 1979) that also uses a single payment schedule, the first best is always implementable when the agent is risk neutral, which is not true in our setting.[4] Thus, our results sit nicely between the standard agency models and the literature on implementation of tax functions.

Most mechanism design solutions entail complicated mechanisms that are very sensitive to the underlying description of the environment. This sensitivity

[3] In 1990, the Texas legislature considered a proposal to allow drivers to avoid modest speeding tickets, but only if they had purchased speeding coupons in advance. Had the plan passed, Texas would have implemented a menu of speeding fines, but the measure never reached a vote. To our knowledge, no speeding fines involve a menu of contracts.

[4] We will discuss our results in the agency context in greater detail below.

is especially extreme when correlation is exploited to mitigate incentive constraints. In contrast to most related literature, we provide an approximation to the solution that has quite modest informational requirements (means and variances of the error function), which in many applications are knowable. Consequently, our approach is plausibly applicable in real world settings, such as the speeding example discussed above.

2. The Model

Let x denote the action chosen by a *risk neutral* agent and Y be the associated (publicly observable) signal. Y is distributed according to density $f(y \mid x)$ conditional on x, $x \in \mathbb{X}, y \in \mathbb{Y}$. The function f completely describes the relevant (*stochastic*) *environment* of the situation. To keep the exposition simple, we present our results for two cases: the finite-dimensional case where $\mathbb{X}$ and $\mathbb{Y}$ are finite sets of sizes m and n, respectively; and the infinite-dimensional case where x and y take values on the unit interval, *i.e.*, $\mathbb{Y}=\mathbb{X}=[0,1]$, and f is continuous. The results and discussions of section 4 extend straightforwardly to more general measurable sets and distributions in their current forms.[5]

The action x generates a private return $B(x,\theta)$ for the agent with a privately known type θ. The agent has quasi-linear utility

$$U(B(x,\theta),t) = B(x,\theta) - t.$$

from taking action x and making a payment t, if any. We assume that transfers are non-distortionary.

Let us denote by $s(x)$ the externality (or transfer or tax) function that the regulator wants to implement. We are interested in examining when and how a penalty function $p(y)$ makes the agent face the exact same problem for all θ that he would with perfect observation. If the action is measured imperfectly then a

[5] Our results are most interesting when $\mathbb{X}$ is a richer set than $\mathbb{Y}$; this will become clear in later discussions.

function $p(y)$ which we refer to as the *penalty function* of the imperfect observation y will be said to *implement* $s(x)$ if

$$U(B(x,\theta), s(x)) = E[U(B(x,\theta), p(Y)) \mid x]$$

for all $x \in \mathbb{X}$.

Quasi-linearity of utility implies that $p(y)$ implements $s(x)$ if

$$B(x,\theta) - s(x) = B(x,\theta) - E[p(Y) \mid x],$$

i.e.,

$$s(x) = E[p(Y) \mid x].$$

If $s(x)$ is first-best and $p(y)$ implements $s(x)$, then $p(y)$ is first-best in spite of imperfect observability of action. Throughout this paper we denote by A, the (*conditional*) *expectation operator* on $p(\cdot)$; with this notation, p implements s if $s = Ap$.

In the *finite actions* case we assume x takes values $1,2,...,m$ and y takes values $1,2,...,n$ so that $f(y \mid x)$ is the probability that action y is observed when action x is undertaken. In this case where the agent is risk-neutral we have that $p(y)$ implements $s(x)$ if

$$\sum_{y=1}^{n} p(y) f(y \mid x) = s(x)$$

for all x. In matrix notation we write this as $Ap = s$ where A is the $m \times n$ matrix of conditional probabilities and p and s are n- and m- vectors, respectively. With *continuous actions* we have under risk-neutrality that $p(y)$ implements $s(x)$ if

$$\int_{0}^{1} p(y) f(y \mid x) dy = s(x)$$

for all x. The reason for considering both cases is that a finite dimensional approach with finite matrices makes the analysis straightforward. However, the intuition from finite dimensional analyses often does not extend to the infinite dimensional analysis where a continuous set of actions is undertaken. Also, the standard literature on both externality and the basic agency problems in large

parts deal with the continuous models. Thus, it is desirable to treat the continuous actions case separately. While the infinite dimensional analysis does not permit the ease of working with matrices, we are able to verify some of the key findings from the finite actions case.

Throughout this paper we will denote by μ the uniform probability measure on the interval [0,1]. Thus when the actions are continuous A is an integral operator (a continuous linear transformation) from $L_2(\mu)$ to $L_2(\mu)$. In light of the fact that we work with a $L_2(\mu)$ space, the relevant equalities, and other statements and relationships must be interpreted as *almost everywhere* μ and convergences as convergence in $L_2(\mu)$.

3. Can the Crime Fit the Punishment?

When does using the social externality function as a penalty function work even in the presence of error? Our first result shows that in the infinite dimensional space of all possible externality functions only a "negligible" sub-collection of externality functions $s(x)$ can be implemented by $p(y) = s(y)$.

Proposition 1. Consider the continuous action case. In any given environment there are at most a finite number of linearly independent externality functions that can be implemented by $p(y) = s(y)$.

Proof. See Appendix.

In other words, given $f(y \mid x)$ the collection of $L_2(\mu)$ functions that can be implemented by using the externality function as a penalty function belong to a finite dimensional subspace.

Of course, if there is no error, the penalty function $p \equiv s$ implements the externality function s. However, not only is zero error sufficient, it is also necessary for *all s* to implement itself.

Proposition 2. An environment allows all externality functions $s(x)$ to be implemented by itself (*i.e.*, by setting $p(y) = s(y)$) if and only if actions are observable without errors in that environment.

Proof. See Appendix.

We now turn to positive results.

4. Existence

We now examine environments where an externality function can be implemented by *some* penalty function. Existence of a solution p is at the center of problems ranging from extraction of information rent in auctions to implementation of payment functions in agency problems and existence of continuation payoffs in repeated games. The conditions that are generally used to obtain existence for these problems are variations of the *spanning condition* which for the finite dimensional problem requires that the matrix A have full row rank. However, the spanning condition is not necessary for implementation. Also, it is easy to see that if $m > n$, *i.e.*, the size of the action space is larger than the signal space, the full row rank condition cannot be satisfied.

The necessary and sufficient condition for a solution to the finite dimensional $Ap = s$ to exist is in fact that the system of equations be *consistent*. To formalize this idea with an intuitive interpretation in mind, and to generalize to the infinite dimensional framework we need the following definitions:

Definitions. (i) f *separates* distributions G_1 and G_2 over $\mathbb{X}$ if

$$\int_x f(\cdot \mid x) dG_1(x) \neq \int_x f(\cdot \mid x) dG_2(x).$$

(ii) s *separates* distributions G_1 and G_2 over X if

$$E[s(X) \mid X \sim G_1] \neq E[s(X) \mid X \sim G_2].$$

When $\mathbb{X}$ is finite the integrals are substituted by summations.

We can now state the finite dimensional version of the necessary and sufficient condition using the separation terminology.

Proposition 3 (Existence – finite actions). Suppose that the sets of possible actions and signals are finite. Then $s(x)$ is implementable if and only if f separates distributions on $\mathbb{X}$ that are separated by s.

Proof. First observe that the range of a finite dimensional linear transformation is closed. Hence, by the Fredholm alternative theorem (see Appendix) the necessary and sufficient condition for the equation $Ap = s$ to have a solution is that

$$A^T z = 0 \;\Rightarrow\; s^T z = 0$$

$A^T z = 0$ implies $z = q - \tilde{q}$ for some probability distributions q and $\tilde{q}$. Then we can restate the above necessary and sufficient condition as follows: For any pair of distributions q and $\tilde{q}$

$$\text{whenever } s^T q \neq s^T \tilde{q} \text{ we have } A^T q \neq A^T \tilde{q}$$

i.e., f separates distributions on $\mathbb{X}$ that give rise to unequal expectations $E_q[s(X)] \neq E_{\tilde{q}}[s(X)]$. ■

The *only if* portion of proposition 3 is quite straightforward. If f does not separate two distributions of X, say G_1 and G_2, then for any p, $E[p(Y) \mid X \sim G_1] = E[p(Y) \mid X \sim G_2]$ because both G_1 and G_2 give rise to the same distribution of Y. Thus, when $s(x)$ is implementable, *i.e.*, $E[p(Y) \mid x] = s(x)$, we also have $E[s(X) \mid X \sim G_1] = E[s(X) \mid X \sim G_2]$. The remarkable fact is that the condition is sufficient – if f separates any distribution that s separates, then there exists a function p that implements s in the stochastic environment.

The above necessary and sufficient condition is *almost* necessary and sufficient when the signals and actions take a continuum of values. The only

difference is that in the infinite dimensional problem the image of the integral operator is not closed, so we need to also include the possibility that the penalty functions can come arbitrarily close to implementing the externality function $s(x)$. Specifically, we need the following definition:

Definition. Externality function $s(\cdot)$ is *approximately implementable* if there is a sequence of penalty functions $\{p_n(\cdot)\}$ such that $E[p_n(Y) \mid X = x] \to s(x)$.

Clearly, all functions in the range of the operator A are exactly implementable. It is only when a function $s(\cdot)$ is not in the range of A but is in the boundary that it needs to be implemented approximately. Such a function $s(x)$, however, gives rise to an added complication. The existence of a sequence of penalty functions $\{p_n(\cdot)\}$ in itself does not guarantee that the corresponding sequence of choices $\{x_n(\theta)\}$ by the agent with

$$x_n(\theta) \in \arg\max_x \left(B(x,\theta) - E[p_n(Y) \mid x]\right)$$

will also converge to the agent's choice under $s(x)$. If the sequence $\{x_n(\theta)\}$ does not converge to a choice $x(\theta)$ with

$$x(\theta) \in \arg\max_x \left(B(x,\theta) - s(x)\right)$$

then the idea of approximately implementing $s(x)$ becomes meaningless. Technically, for such an $s(\cdot)$ additional conditions must be derived from the incentive compatibility constraint of the specific agency problem to guarantee upper hemi-continuity of the relevant graph. Rather than considering specific cases we sidestep this incentive compatibility issue for the boundary points and consider approximate implementation only in the sense defined above.[6] The infinite dimensional version of our result is as follows.

[6] Once an approximation is constructed it is usually straightforward to check the incentive compatibility in the context of the particular problem.

Proposition 4 (Existence – infinite actions). A penalty function s is at least approximately implementable if and only if f separates distributions on X that are separated by s.

Proof. See Appendix.

Proposition 4 is the infinite dimensional analog of Proposition 3 and has the same interpretation. It is generally straightforward to describe situations in the finite-actions case where an externality function cannot be implemented. The infinite-dimensional case is less straightforward. Functions that cannot be implemented exactly can arise very naturally in the continuous action case. If the conditional distribution $f(y|x)$ is continuous in x and y, discontinuous penalty functions are not exactly implementable. Of course, discontinuous functions may be obtained as limits of continuous functions.

The implementable externality functions are dense in $L_2(\mu)$ only under some stronger conditions on the environment. By the modified Fredholm alternative theorem (see Appendix) the implementable functions are dense if and only if

$$\text{for all } y, \quad \int_0^1 f(y|x)g(x)dx = 0 \;\&\; g(x) \in L_2(\mu) \Rightarrow g(x) = 0.$$

By a construction similar to that in the proof of Proposition 4 (see Appendix) it follows that the implementable function are dense in $L_2(\mu)$ if and only if for two densities $g_1(x)$ and $g_2(x)$

$$\int_0^1 f(y|x)g_1(x)dx = \int_0^1 f(y|x)g_2(x)dx$$

implies $g_1(x) = g_2(x)$, *i.e.*, f separates *every* pair of distinct distributions. This is, in fact, an infinite dimensional counterpart of the finite dimensional condition that matrix A of conditional probabilities have full row rank. We call this the *spanning condition*. The necessary and sufficient condition from propositions 3

and 4 that ties $s(x)$ and $f(y \mid x)$, instead, will be referred to as the *joint separation condition*.

Implementation under Joint Separation and Spanning Conditions

Several variations of spanning conditions, which we collectively refer to as *spanning-type* conditions, have been used in problems ranging from areas of principal-agent problems to repeated games with imperfect monitoring (Cremer and McLean, 1988; McAfee and Reny, 1992; Melumad and Reichelstein, 1989; Fudenberg, Levine and Maskin, 1994).

Cremer and McLean (1988) and McAfee and Reny (1992) examine the problem of implementing a target expected payment function under a spanning-type condition on agent beliefs in models of adverse selection. The two papers also take advantage of a menu of contracts, and hence use selection by agent as an indication of the agent's type. They are then able to fully extract payoffs (implement target transfers, in our language) from all types of agents. In contrast, we do not anticipate using a menu for the externalities problem because that requires contracting in advance. Contracting in advance is, of course, implausible in the case of speeding and absurd in the case of intoxicated drivers.

Models of moral hazard are closer to our model in spirit. Holmstrom (1979) aims to induce appropriate behavior at a single point; in our notation, it would be as if there is only one type of agent. The need to implement a function at multiple points arises when the agent has multiple types. Laffont and Tirole (1986), McAfee and McMillan (1987), among others, consider such models of moral hazard with private types and quasi-linear utilities. As in the rent extraction literature, the agency literature approach the solution with menu of contracts. If the principal in the mixed model of moral hazard and adverse selection of McAfee and McMillan (1987) cannot use a menu of contracts then the implementation problem we consider in this paper arises there, too.

Melumad and Reichelstein (1989) show that under a spanning-type condition on conditional probabilities $f(y \mid x)$, the type-independent transfer $p(y)$ is as good as the type-dependent transfer $p(y,\theta)$. They also demonstrate that the condition is *not* necessary. The relevant program in this problem generally gives the solution in the form of the expected transfer $E_Y[p(Y) \mid x]$ conditional on the agent's action. Whether there is an output-contingent transfer schedule $p(y)$ that can implement the suggested solution involves exactly the same problem that we consider in this paper. Of course, we now know that such implementation is possible under specific conditions obtained by relating $f(y \mid x)$ to $E_Y[p(Y) \mid x]$.

We conclude the discussion in this section by observing that the joint separation condition is quite distinct from the spanning condition with the help of an example.

Example. Consider the standard additive error model where $Y \equiv x + \varepsilon$ with $x \in [0,1]$ and $\varepsilon \sim h(\cdot)$ independently distributed from X, where $h(\cdot)$ is the uniform distribution on [0,1]. The spanning condition holds if for any $g_1(x)$ and $g_2(x)$

$$\int_0^1 h(y-x)g_1(x)dx = \int_0^1 h(y-x)g_2(x)dx \quad \Rightarrow \quad g_1(\cdot) = g_2(\cdot). \tag{1}$$

However,

$$\int_0^1 h(y-x)g_1(x)dx = \int_0^y g_1(x)dx$$

so that (1) is satisfied. Clearly, the spanning condition is satisfied and so the separation condition is automatically satisfied in this case for any externality function. Now suppose that the signal is observed only in its integer part, *i.e.*, $Y \equiv [x+\varepsilon]$ where $[z]$ is the part of the number z that appears before the decimal point, often known as the *floor* of z. For expositional simplicity, we let $[2] \equiv 1$. For two symmetric densities $g_1(x)$ and $g_2(x)$ the distribution of $X + \varepsilon$ is symmetric around 1. Hence, Y takes values 0 and 1 with equal probabilities. It follows that

the spanning condition cannot hold in this case, whereas the joint separation condition can continue to hold. In fact, it is not difficult to check that in any given environment the collection of implementable externality functions is non-empty, thus the joint separation condition is applicable quite generally.

5. Construction of a Penalty Function

Existence theorems are often cold comfort to someone who needs to use a construct. In this section we provide a method of constructing the penalty function in the case of multiplicative errors.

Suppose that $Y = x\varepsilon$, $\varepsilon \geq 0$ and that the externality function s is analytic, so that it can be expressed by a Taylor series:

$$s(x) = \sum_i a_i x^i.$$

Let

$$p(y) = \sum_i \frac{a_i y^i}{E\varepsilon^i}$$

Then

$$E[p(Y) \mid x] = E\sum\left[\frac{a_i (x\varepsilon)^i}{E\varepsilon^i} \mid x\right] = \sum a_i x^i = s(x)$$

whenever the series converges absolutely at each point in the support. This is a full solution for analytic s for the multiplicative case. In addition, for any continuous $s(\cdot)$ on a compact set of $\mathbb{X}$, we can approximate arbitrarily closely by first approximating $s(\cdot)$ with an analytic function and then using the $p(\cdot)$ for the analytic function.

How does the penalty compare to the externality? Suppose that $Y = x\varepsilon$, for $\varepsilon \geq 0$, and the error is either unbiased or upward biased, *i.e.*, $E\varepsilon \geq 1$ and that all the derivatives of $s(x)$ are nonnegative, *i.e.*, $a_i \geq 0$, as arises with the exponential function. Then the penalty is less than the externality, *i.e.*, $p(y) \leq s(y)$, because $E\varepsilon^i \geq (E\varepsilon)^i \geq 1$. Thus $p(y) = \sum \frac{a_i y^i}{E\varepsilon^i} \leq \sum a_i y^i = s(y)$.

The multiplicative error model provides a construction for additive errors for many externality functions. Consider the additive error model $Y = x + \varepsilon$ and assume that $s(\log(z))$ is an analytic function of z, and that $Ee^{j\varepsilon} < \infty$ for each nonnegative integer $j < \infty$. Because $s(\log(\cdot))$ is analytic, we can express $s(\log(z)) = \sum_{j=0}^{\infty} a_j z^j$. Consequently, setting $x = \log(z)$, $s(x) = \sum a_j (e^x)^j = \sum a_j e^{jx}$. It is readily verified that $p(y) = \sum \frac{a_j e^{jy}}{Ee^{j\varepsilon}}$ satisfies $E\left[p(Y)\right] = s(x)$.

6. Small Errors

So far, our main results – existence and construction under multiplicative or additive errors – tell us little about the actual nature of the penalty function. Imagine that a car going at 100 mph generates an externality $s(100)$, would the corresponding penalty function charge more than $s(100)$, or less upon observing a speed of 100 mph? A 10 mph increase in speed at 100 mph could presumably do much more damage than the same increase would do at 70 mph. Would the corresponding penalty function reflect that? Under the hypothesis that errors are small, we can provide a sharper characterization of the penalty functions.

When the error is sufficiently "local" in nature, the Taylor expansion allows close approximation of the penalty function given the externality function. We consider the second order Taylor expansion to derive an approximation. In what follows we let Y be the observation by the regulator and provide a formula for approximating the penalty function that implements an externality function $s(x)$ when the associated observational error is small. Let $\mu(x) = E[Y \mid x]$ and $\sigma^2(x) = E[(Y - \mu(x))^2 \mid x]$.

Proposition 5. Suppose that the family of random variables

$$\tilde{Y}_b \equiv \mu(x) + b(Y - \mu(x))$$

with distribution functions $G_b(\tilde{y} \mid x) = F((\tilde{y} - (1-b)\mu(x))/b \mid x)$ separate, for all small $b > 0$, any pair of distributions on X that are separated by $s(\cdot)$. Assume also that $\mu(x)$ is twice differentiable and monotonic.[7] If $\sigma^2(x)$ is small, the penalty function is approximated by

$$p(z) \approx s(\mu^{-1}(z)) - \frac{1}{2}\sigma^2(\mu^{-1}(z))\frac{d^2}{dz^2}s(\mu^{-1}(z)).$$

Proof. See Appendix.

In particular, when the signal distribution is unbiased, we have

$$p(x) \approx s(x) - \tfrac{1}{2}s''(x)\,\sigma^2(x)$$

It is now much easier to relate the penalty to the externality function. For instance, when the observation is unbiased ($\mu(x) = x$) and the error ($\sigma^2(x)$) is small, the penalty function is smaller or larger than the externality function depending on whether s is convex or concave. The approximation for p is in fact exact when $\sigma^2(x)$ does not depend on x and s is quadratic. To see this observe that if $s(x) = a_0 + a_1 x + a_2 x^2$ then we have

$$s(x) - \tfrac{1}{2}s''(x)\,\sigma^2 = a_0 + a_1 x + a_2 x^2 - a_2\sigma^2$$

And, hence,

$$\begin{aligned} E\left[s(Y) - \tfrac{1}{2}s''(Y)\sigma^2\right] &= E\left[a_0 + a_1 Y + a_2 Y^2 - a_2\sigma^2\right] \\ &= a_0 + a_1 x + a_2\left(x^2 + \sigma^2\right) - a_2\sigma^2 \\ &= a_0 + a_1 x + a_2 x^2 \\ &= s(x) \end{aligned}$$

Thus, the formula fits the quadratic case exactly, and it is not surprising that it represents a second order approximation in general.

[7] The two assumptions are necessary to apply the Taylor approximation result at $\mu(x)$.

7. Conclusion

Given an externality function which implements a social objective, this paper examines the possibility of implementing the social objective when the action is observed with error. Provided that the signal is informative in the sense that it separates certain distributions of actions and agents are risk-neutral, the social objective remains implementable even with observational error. In addition, when errors are small, there is a closed form second-order approximation for the penalty function that depends only on first and second moments and two derivatives of the externality function. The formula is applicable when activity is measured reasonably accurately, which is necessary for an acceptable implementation. This formula is simple enough to lend itself to actual implementation.

There is good reason to suppose that errors are small in most settings in which penalty functions are in use. Courts are generally hostile to punishments that do not fit the crime committed and observational error has been used to disallow evidence, deeming such evidence unreliable. That is, courts are typically hostile to punishments that are not commensurate with the crime. Our approach constructs appropriate *ex ante* incentives in the presence of observational error, but does nothing directly to insure fairness of the realization of the penalty. For this reason courts may be hostile to the approach taken here; while it aligns *ex ante* incentives, it may result in extreme *ex post* punishments. On the other hand, with small errors and a convex externality function, our approach involves reducing the fine below that which would prevail were the error ignored; such a reduction might arguably be viewed as more fair. Thus, fairness considerations are not an absolute bar to implementation, but present thorny issues beyond the scope of the present analysis.

In this paper we have kept the implementation issue simple by assuming that the agent accepts the penalty for what it is and does not contest it. In many legal situations, the signal based on which the penalty is determined is only an evidence of the real act, and as such may be contested in a court. A substantial

punitive damage gives the defendant an incentive to contest the penalty. In this case, if there is a scope for avoiding or lowering the payment then the original payment function will not achieve its objective. On the other hand, if the damage awarded is not high enough there will be a natural incentive for the plaintiff to appeal. The scope for successful appeals then introduces further distortion into the outcome. Excessive punitive damages may also invite frivolous lawsuits (cf. Posner, 1973, for a discussion of the importance of these issues in actual litigations). Thus, extreme payments with an appeal phase make the situation strategic. The implementability issue in that case depends on how the game is played out under different penalty functions. Although more complex, this is clearly an interesting direction to go for future research which we hope to persue.

Our analysis does not apply in the form of first-best implementation when the agent is risk averse. The difficulty arises due to the welfare effect of the redistributive role of tax function under risk aversion. If the agent is risk averse and the penalty is a function of a stochastic signal, the socially optimal penalty function depends on the conditional distribution of the signal. Implementing a function that is first-best when the action is observed perfectly may not be optimal in the stochastic environment due to the *risk cost* to the society.[8] While an externality function may be implementable using the same techniques developed here, it is no longer possible to conclude that the implementation is optimal.

[8] We thank Ilya Segal for pointing out this difficulty in analyzing risk aversion.

Appendix: Proofs

Proof of Proposition 1. First note that the integral operator A is compact. Therefore, if there is a non-zero function $s(x)$ such that $As = s$ then the operator has an eigenvalue 1. The result then follows upon applying Proposition II.4.13 of Conway (1990) and observing that the space of functions s for which $(A-I)s = 0$ is at most finite dimensional. ■

Proof of Proposition 2. The "if" part follows straightforwardly. To show the "only if" part, let us consider I, the identity operator $Is = s \ \forall s \in V$ where V is the relevant (finite or infinite dimensional) space of penalty functions. Now, $As = s$ for all $s \in V$ implies that $(A-I)s = 0$ for all $s \in V$. This implies that $\|A-I\| = 0$ where $\|\cdot\|$ is the norm for the space $B(V)$ of bounded linear operators on V. Thus we have $A = I$ which completes the proof. ■

The proofs of propositions 3 and 4 will use the following result:

Fredholm Alternative Theorem (cf. Keener, 1988). If A is a bounded linear operator in Hilbert space H with a closed range, the equation $Ap = s$ has a solution if and only if $\langle s,u \rangle = 0$ for every u in the null space of the adjoint operator A^*.

Definition. We say that the equation $Ax = b$ has an *approximate solution* if there exists a sequence $\{x_n\}_{n=1}^{\infty}$ in $L_2(\mu)$ with $Ax_n \to b$.

A Modified Fredholm Alternative Theorem. Let A be a compact linear operator. Then $Ax = b$ has either a solution or an arbitrarily close approximate solution if and only if

$$\langle v,b \rangle = 0 \text{ for all } v \text{ satisfying } A^*v = 0.$$

Moreover, all solutions in the equation $Ax = b$ are exact if and only if $\operatorname{ran} A$ is finite dimensional.

Proof. *Only if part*. Suppose that $Ax = b$ is at least approximately solvable. Then there exists a sequence $\{x_n\}_{n=1}^{\infty}$ possibly all identical such that $b_n \equiv Ax_n \to b$.
This implies that for v satisfying $A^*v = 0$

$$\langle v, b\rangle = \lim_n \langle v, b_n\rangle = \lim_n \langle v, Ax_n\rangle = \lim_n \langle A^*v, x_n\rangle = \lim_n 0 = 0$$

If part. Suppose $\langle v, b\rangle = 0$ for all v satisfying $A^*v = 0$, but $Ax = b$ does not have even an approximate solution. Then $b = b^r + b^o$ where b^r is in the closure of the range of A and b^o is in its orthogonal subspace. Therefore, $\langle b^o, Ax\rangle = 0\ \forall x$ so that $\langle A^*b^o, x\rangle = 0\ \forall x$ which implies that $A^*b^o = 0$. Now using the hypothesis of this part we have $\langle b^o, b\rangle = 0$ which implies $\langle b^o, b^o + b^r\rangle = 0$, or, $\langle b^o, b^o\rangle + \langle b^o, b^r\rangle = 0$. Since b^o is orthogonal to b^r this implies that $\langle b^o, b^o\rangle = 0$ or, $b^o = 0$. Hence, $b \in \mathrm{cl}(\mathrm{ran}A)$, *i.e.*, $Ax = b$ either has an exact solution or an approximate solution.

To prove the second part of the result observe that by Problem 7.1.1 of Abraamovich and Aliprantis (2002), a compact operator has a closed range if and only if its range is finite dimensional. Next we show that $Ax = b$ has only exact solutions if and only if $\mathrm{ran}\,A$ is closed. The 'if' part follows from the original Fredholm Alternative Theorem (see above). To see the converse, suppose $\mathrm{ran}\,A$ is not closed. Then there exists a sequence $\{x_n\}_{n=1}^{\infty}$ such that $b \equiv \lim_n Ax_n$ is well defined and $b \notin \mathrm{ran}\,A$. Thus $Ax = b$ is at least approximately solvable. ■

Proof of Proposition 4.
The modified Fredholm alternative theorem implies that a necessary and sufficient condition for a sequence $\{p_n(\cdot)\}$ with the property

$$\int_0^1 p_n(y) f(y \mid x) dy \to s(x)$$

to exist is that

$$\int_0^1 s(x)u(x)dx = 0 \text{ whenever } \int_0^1 f(y \mid x)u(x)dx = 0.$$

Now suppose a $L_2(\mu)$ function $u(x)$ satisfies $\int_0^1 f(y \mid x)u(x)dx = 0$ and does not vanish over some positive μ measure subset. Define

$$v(x) = |u(x)| + u(x) \text{ and } w(x) = |u(x)|.$$

Then v and w are non-negative functions satisfying $u(x) = v(x) - w(x)$. Also, $\mu(x : u(x) = 0) < 1$ implies that $v(x) \neq w(x)$ over a set with positive measure.

Next, $\int_0^1 f(y \mid x)u(x)dx = 0 \Rightarrow$

$$\int_0^1 \int_0^1 f(y \mid x)u(x)dxdy = 0 \Rightarrow \int_0^1 u(x) \int_0^1 f(y \mid x)dydx = 0 \Rightarrow \int_0^1 u(x)dx = 0$$

i.e.,

$$\int_0^1 v(x)dx = \int_0^1 w(x)dx = K \text{ (say)}$$

Since v and w are non-negative functions satisfying $v(x) \neq w(x)$ on a set with positive measure, we have $K > 0$. Thus

$$\tilde{v}(x) = \frac{v(x)}{K} \text{ and } \tilde{w}(x) = \frac{w(x)}{K}$$

are probability density functions satisfying

$$\int_0^1 f(y \mid x)\tilde{v}(x)dx = \int_0^1 f(y \mid x)\tilde{w}(x)dx$$

Thus the necessary and sufficient condition above can be restated as that for two densities $\tilde{v}(x)$ and $\tilde{w}(x)$

$$\int_0^1 s(x)\tilde{v}(x)dx = \int_0^1 s(x)\tilde{w}(x)dx \text{ whenever } \int_0^1 f(y \mid x)\tilde{v}(x)dx = \int_0^1 f(y \mid x)\tilde{w}(x)dx.$$

In other words, $f(y \mid x)$ separates any pair of densities $\tilde{v}(x)$ and $\tilde{w}(x)$, *i.e.*,

$$\int_0^1 f(y \mid x)\tilde{v}(x)dx \neq \int_0^1 f(y \mid x)\tilde{w}(x)dx$$

whenever $\tilde{v}(x)$ and $\tilde{w}(x)$ give rise to separate expectations for s, *i.e.*,

$$\int_0^1 s(x)\tilde{v}(x)dx \neq \int_0^1 s(x)\tilde{w}(x)dx. \blacksquare$$

Proof of Proposition 5

Let Y be the observation with a conditional distribution $F(y \mid x)$. Define $\varepsilon = Y - \mu(x)$, which has a distribution $H(\varepsilon \mid x) = F(\varepsilon + \mu(x) \mid x)$. The corresponding density is $h(\varepsilon \mid x)$. Recall that $\mu(x) = E[Y \mid x]$ and $\sigma^2(x) = E[(Y - \mu(x))^2 \mid x]$ so that

$$E[\varepsilon \mid x] = 0 \quad \text{and} \quad E[\varepsilon^2 \mid x] = \sigma^2(x).$$

For any $b > 0$ but small let $\tilde{p}(\cdot, b)$ solve (suppressing the limits in the integrations)

$$\int \tilde{p}(z + b\varepsilon, b)h(\varepsilon \mid x)d\varepsilon = s(\mu^{-1}(z)) \tag{A1}$$

Our hypothesis guarantees that the functions $\tilde{p}(\cdot, b)$ exist for all small $b > 0$.[9] Existence in the case of $b = 0$ is, of course, immediate from the monotonicity of $\mu(x)$. Our target is the solution at $b = 1$. Note that $p(z, 0) = s(\mu^{-1}(z))$, so that $p_1(z, 0) = \frac{d}{dz} s(\mu^{-1}(z))$ and $p_{11}(z, 0) = \frac{d^2}{dz^2} s(\mu^{-1}(z))$.

Taking the derivative with respect to b of both sides of equation (A1) above we have

$$\int \left[\tilde{p}_1(z + b\varepsilon, b)\varepsilon + \tilde{p}_2(z + b\varepsilon, b) \right] h(\varepsilon \mid x)d\varepsilon = 0,$$

which at $b = 0$ gives

$$\tilde{p}_1(z, 0)\int \varepsilon h(\varepsilon \mid x)d\varepsilon + \tilde{p}_2(z, 0)\int h(\varepsilon \mid x)d\varepsilon = 0$$

or, $\tilde{p}_2(z, 0) = 0$. This also implies $p_{12}(z, 0) = 0$.

---

[9] Note that if $s(\cdot)$ is on the boundary of the range of operator A and can only be approximately implementable, then the function $s(\cdot)$ must be replaced by its exactly implementable approximation, say $\tilde{s}(\cdot)$, at this and the next few steps. The result is still unaltered since $\tilde{s}(\cdot)$ approximates $s(\cdot)$.

Taking the second derivative with respect to b of both sides of equation (A1), we have

$$\int \left[\tilde{p}_{11}(z+b\varepsilon,b)\varepsilon^2 + 2\tilde{p}_{12}(z+b\varepsilon,b)\varepsilon + \tilde{p}_{22}(z+b\varepsilon,b) \right] h(\varepsilon \mid x) d\varepsilon = 0 .$$

Setting $b=0$,

$$\int \left[\tilde{p}_{11}(z,0)\varepsilon^2 + 2\tilde{p}_{12}(z,0)\varepsilon + \tilde{p}_{22}(z,0) \right] h(\varepsilon \mid x) d\varepsilon = 0$$

or,

$$0 = \tilde{p}_{11}(z,0)E[\varepsilon^2 \mid x] + \tilde{p}_{22}(z,0)$$

Hence

$$\begin{aligned} \tilde{p}_{22}(z,0) &= -p_{11}(z,0)E[\varepsilon^2 \mid x] \\ &= -\sigma^2 \frac{d^2}{dz^2} s(\mu^{-1}(z)) \end{aligned}$$

Now we use the second order approximation on the first argument of $p(x,b)$:

$$\begin{aligned} \tilde{p}(z,b) &\approx s(\mu^{-1}(z)) + b\tilde{p}_2(z,0) + \frac{1}{2}b^2\tilde{p}_{22}(z,0) \\ &= s(\mu^{-1}(z)) - \frac{1}{2}b^2\sigma^2 \frac{d^2}{dz^2} s(\mu^{-1}(z)) \end{aligned}$$

At $b=1$,

$$\tilde{p}(z,1) \approx s(\mu^{-1}(z)) - \frac{1}{2}\sigma^2 \frac{d^2}{dz^2} s(\mu^{-1}(z))$$

It is straightforward at this point to see that scaling the error down and scaling b up in the same amount keeps the entire calculation the same. Hence, $b=1$ is without loss of generality and we have the result. ■

References

Abramovich, Y.A. and C.D. Aliprantis, *Problems in Operator Theory*, American Mathematical Society, Graduate Studies in Mathematics, vol. 51, 2002.

Coase, R.H., "The problem of social cost," *The Journal of Law and Economics*, vol. 3, 1960, 1-44.

Conway, J.B., *A Course in Functional Analysis*, second edition, Springer-Verlag New York, Inc., 1990.

Cremer, K., and R. McLean, "Full extraction of the surplus in Bayesian and dominant strategy auctions," *Econometrica*, vol. 56, 1988, 1247-1257.

Dasgupta, P., P. Hammond, and E. Maskin, "On imperfect information and optimal pollution control," *Review of Economic Studies*, vol. XLVII, 1980, 857-860.

Duggan, J. and J. Roberts, "Implementing the efficient allocation of pollution," *The American Economic Review*, vol. 92, 2002, 1070-1078.

Fudenberg, D., D. Levine, and E. Maskin, "The Folk Theorem with imperfect public information," *Econometrica*, vol. 62, 1994, 997-1039.

Holmstrom, B., "Moral hazard and observability," *Bell Journal of Economics*, vol. 10(1), 1979, 74-91.

Keener, J.P., *Principles of Applied Mathematics*, Addison-Wesley Publishing Company, 1988.

Kwerel, E, "To tell the truth: Imperfect information and optimal pollution control," *Review of Economic Studies*, vol. 44(3), 1977, 595-601.

Laffont, J-J., and J. Tirole, "Using cost observation to regulate firms," *Journal of Political Economy*, vol. 94(3), 1986, 614-641.

McAfee, R.P., and P. Reny, "Correlated information and mechanism design," *Econometrica*, vol. 60(2), 1992, 395-421.

McAfee, R.P., and J. McMillan, "Competition for agency contracts," *RAND Journal of Economics*, vol. 18(2), 1987, 296-307.

Melumad, N.D., and S. Reichelstein, "Value of communication in agencies," *Journal of Economic Theory*, vol. 47, 1989, 334-368.

Montero, J-P, "Pollution markets with imperfectly observed emissions," *RAND Journal of Economics*, vol. 36(3), 2005, 645-660.

Pigou, A.C., *The Economics of Welfare*, Macmillan, London, 1952.

Png, I.P.L., "Optimal subsidies and damages in the present of judicial error," *International Review of Law and Economics*, vol 6(1), 1986, 101-105.

Polinsky, A.M. and S. Shavell, "Punitive damages: An economic analysis," Harvard Law Review, vol. 111(4), 1998, 869-962.

Posner, R.A., "An economic approach to legal procedure and judicial administration," *The Journal of Legal Studies*, vol. 2(2), 1973, 399-459.

Sandmo, A., "Optimal taxation in the presence of externalities," *Swedish Journal of Economics*, vol. 77(1), 1975, 86-98.

Shavell, S., "Corrective taxation versus liability," *American Economic Review*, vol. 101(3), 2011, 273-76.