Bounding reasonable doubt: Implications for plea bargaining*

Yacov Tsur◇

September 19, 2016

Abstract

A bound for reasonable doubt is offered based on the cost of type I and type II errors. The bound increases with the punishment, hence its use as a conviction threshold may leave too many offenders of severe crimes at large. Plea bargaining addresses this limitation but introduces strategic interaction between concerned parties. Considering strategic interaction between defendants and judge/jury, it is shown that to any plea offer there corresponds a *unique* equilibrium. Moreover, all equilibria share the *same* conviction threshold, given by the reasonable doubt bound. The latter property ensures that the plea bargaining procedure is consistent with the 'equality before the law' principle. The former property (that to any plea offer there corresponds a unique equilibrium) bears implications for the design of plea bargain schemes.

Keywords: Reasonable doubt, conviction threshold, plea bargaining, perfect Bayesian equilibrium.

JEL classification: K1, K4.

*Helpful comments by Guni Orshan and three anonymous reviewers are gratefully acknowledged.

◇Department of Environmental Economics and Management, The Hebrew University of Jerusalem, POB 12, Rehovot 7610001, Israel. Email: yacov.tsur@mail.huji.ac.il. Tel: +972-54-7520086. Fax: +972-8-94662676.

1 Introduction

Judicial processes can err in two fundamental ways: convicting the innocent (type I) or acquitting the guilty (type II). Type I error is considered more serious and this asymmetry underlies the principle that guilt must be proven beyond a reasonable doubt. Applying this principle in practice requires evaluating reasonable doubt, which turns out elusive and controversial (Tillers and Gottfried 2006, Weinstein and Dewsbury 2006, Newman 2006, and references they cite). In fact, even agreeing on a definition of reasonable doubt has been challenging (Mulrine 1997, Whitman 2008, Laudan 2011).

The first purpose of this work is to offer a tractable bound for 'reasonable doubt' based on the social cost of type I and type II errors. Consider an offense for which the mandatory punishment (or penalty that fits the crime) is P and let $\varphi(P)$ represent the number of guilty persons that are worth acquitting in order to avoid the conviction of *one* innocent person. Optimizing over the cost of type I and type II errors yields the conviction threshold $b^*(P) = 1/(1 + 1/\varphi(P))$: a defendant is convicted (considered guilty beyond a reasonable doubt) if the probability that he is guilty (based on existing evidence) exceeds $b^*(P)$. A similar bound was offered by Andreoni (1991) based on the calculus of a judge/jury when deciding to convict or acquit.

The notion that the tradeoff between type I and type II errors should be biased in favor of the innocent is age-old.[1] As follows from its definition (see, e.g., Andreoni 1991, Weinstein and Dewsbury 2006), $\varphi(P)$ increases with the punishment P, hence the ensuing bound $b^*(P)$ increases with P as well and

[1]This principle goes back to biblical times and was advocated, inter alia, by the 12'th century Jewish philosopher Maimonides and the 18'th century English jurist Blackstone (Volokh 1997, Tillers and Gottfried 2006). Essentially, $\varphi(P)$ can be interpreted as a penalty-dependent Blackstone ratio (I thank Alon Harel for pointing my attention to this relation.)

approaches one at crimes with draconian sentence. As a result, using $b^*(P)$ as a conviction threshold may leave too many offenders of severe crimes at large – a disturbing situation. Plea bargaining addresses this limitation. The second purpose of this work is to study the role of $b^*(P)$ in the design and implementation of plea bargaining.

Plea bargaining introduces strategic interaction between concerned parties, which, depending on the underlying rules, may involve interaction between defendants and prosecutors (Baker and Mezzetti 2001, Kim 2010) or between defendants and judges/jury (Bjerk 2007). In the plea bargain arrangement considered herein, the prosecution determines the plea offer (the reduced punishment under the plea bargain) and is committed to send to trial those that reject the offer, giving rise to the latter (defendant-judge) interaction. Characterizing the outcome of this interaction, we find that to any plea offer there corresponds a *unique* perfect Bayesian equilibrium (PBE) and all PBEs share the *same* conviction threshold given by the reasonable doubt bound $b^*(P)$.

A meaningful test of the principle "innocent unless proven guilty beyond a reasonable doubt" must have a normative criterion, uniformly applied to all defendants accused of committing a crime of severity P (this, in essence, is a manifestation of 'equality before the law'). The reasonable doubt bound $b^*(P)$, used as conviction threshold, provides such a criterion. The property that all PBEs share the same conviction threshold $b^*(P)$, thus, ensures that the plea bargaining mechanism satisfies the 'equality before the law' principle.

The property that to any plea offer there corresponds a *unique* PBE bears implications regarding the design of plea bargaining schemes. To see this, note that plea offers affect defendants choices and thereby the ensuing defendant-judge/jury interaction and the guilt signals (probabilities) that come out of

this interaction. The *unique* plea offer-PBE correspondence, thus, gives rise to a unique correspondence between plea offers and defendants' guilt probability, which in turn affects the performance of a plea bargaining scheme via its affect on conviction decisions. This enables ranking the performance of different plea offer menus vis-à-vis their effect on (expected) occurrences of type I and type II errors, thereby providing a criterion for the design of optimal plea bargain schemes. In the present framework this task is performed by the prosecution, while accounting for the defendant-judge/jury interaction.[2]

The next section discusses related literature and compares the results found in this work with those established in the literature. Section 3 lays out the ingredients needed to bound reasonable doubt and Section 4 derives the reasonable doubt bound as an outcome of social cost optimization. Section 5 discusses plea bargaining, in which defendants and judges/jury interact while taking the the plea bargain scheme as given, and the prosecution designs the optimal plea-bargain scheme while accounting for this interaction. Section 6 concludes and the appendix contains proofs.

2 Literature context

The role of type I and type II errors in the economics of law enforcement was introduced by Harris (1970), who showed that they can have pronounced effects on Becker's (1968) recommendations regarding levels of law enforcement expenditure and punishment (see also Posner 1973). Harris' work motivated two lines of research. The first deals with reasonable doubt, standards of

[2]The prosecution is assumed to behave in accordance with the social interest (for other modes of prosecution behavior see Landes 1971, Miceli 1990). In addition, the role of plea bargaining in reducing the direct cost of law enforcement (see, e.g., Landes 1971, Adelstein and Miceli 2001, Rakoff 2014) is beside the current scope.

proof and conviction thresholds (Rubinfeld and Sappington 1987, Miceli 1990, Andreoni 1991). The second addresses plea bargaining as a mean to balance tradeoffs between type I and type II errors.

Early economic studies of plea bargaining include Grossman and Katz (1983) and Reinganum (1988), who emphasized its screening role in eliciting defendants' private information (guilty or innocent) via the decision to accept or reject the plea offer. Baker and Mezzetti (2001) showed that this screening role is greatly reduced in the presence of strategic interaction between prosecutors and defendants. In particular, they showed that perfect screening, where only innocent defendants reject the plea offer and go to trial, cannot be maintained in equilibrium, because prosecutors will be reluctant to send innocent defendants to trial, knowing that with some (positive) probability they will be (wrongfully) convicted, thereby motivating guilty defendants to choose the trial option as well, implying that not sending those that reject the plea offer to trial cannot be optimal. Baker and Mezzetti (2001) analyzed plea bargaining as a signaling game played by privately informed defendants and the prosecution, where the latter is not committed to send to trial those that reject the offer but may expand more resources in gathering additional evidence. Kim (2010) extended Baker and Mezzetti's (2001) results to more general situations. The judge/jury behavior in these works was assumed exogenous.

Addressing the limitation of an exogenous jury behavior, Bjerk (2007) analyzed strategic interaction between jury and defendants, where the prosecution designs the plea bargain scheme and is committed to send to trial those that reject the bargain. In Bjerk's (2007) framework, any (arbitrarily chosen) conviction threshold gives rise to a PBE.

The present work draws on both literature veins. We first derive Andreoni's (1991) bound of reasonable doubt as an outcome of optimizing over the social cost of type I and type II errors. In so doing, we formulate this bound in terms of the above-mentioned $\varphi(P)$, which, given its intuitive meaning, clarifies some of the difficulties in measuring this bound and using it as a conviction threshold.

The plea bargaining framework considered in this work is similar to that of Bjerk (2007), in that it involves strategic interaction between defendants and judge/jury, but differs in the following respects: while in Bjerk's (2007) model a (any) conviction threshold gives rise to a PBE, in the present framework a PBE corresponds uniquely to a plea offer and all PBEs share the *same* conviction threshold, given by the reasonable doubt bound $b^*(P)$. As discussed in the introduction, the role of the normative bound $b^*(P)$ as conviction threshold induces an 'equality before the law' property on the plea bargaining scheme.

3 Evidence, errors and cost

Consider a felony for which the mandatory sentence (or 'penalty that fits the crime') is P. Let A be the set containing all possible pieces of evidence that can help identify perpetrators, the number of which is denoted n. An element of A could strengthen innocence (e.g., a convincing alibi) or guilt (e.g., an eye witness, finger print or DNA test). Let $\mathcal{A}$ be the collection of the 2^n subsets $A_i \subseteq A$. A defendant is brought to trial (after collecting evidence and apprehending) with an attached subset of evidences, which is often modified during the trial (on evidence gathering and calculation of guilt probabilities, including cost-benefit considerations, see Posner 1999). At the conclusion of

the trial the defendant is assigned the post-trial subset of evidence A_i.

Let $\tilde{\Pi}_i$ represent the fraction of defendants with evidence set A_i that are guilty, which is a measure of the strength of the guilt signal emitted by A_i. It is useful to rearrange A_i such that the A_i's with the same guilt signal $\tilde{\Pi}_i$ are lumped together, and those with distinct $\tilde{\Pi}_i$, the number of which is $N \leq 2^n$, are ordered according to:

$$\tilde{\Pi}_i > \tilde{\Pi}_j \Rightarrow i > j, \;\; i, j = 1, 2, \ldots, N.$$

Secondly, define

$$t_i = i/N, \; i = 1, 2 \ldots, N.$$

Note that t_i lies between zero and one and the guilt signal $\tilde{\Pi}(t_i)$ increases with t_i. Thirdly, a continuous, increasing and differentiable function $\Pi : [0, 1] \mapsto [0, 1]$, satisfying

$$\Pi(t_i) = \tilde{\Pi}_i, \; i = 1, 2 \ldots N, \;\; \Pi'(\cdot) > 0, \tag{3.1}$$

is constructed. In general, more than one $\Pi(\cdot)$ function satisfying (3.1) exist, but any such function serves equally the purpose of bounding reasonable doubt, addressed in the next section.

As the strength of the guilt signal of suspects with evidence set A_i is $\Pi(t_i)$, we refer to them as *type-t_i* suspects and extend this notation to any $t \in [0, 1]$. It is helpful to define $\theta(t) \in \{G, I\}$ as a Bernoulli variate with success probability $Pr\{\theta(t) = G\} = \Pi(t)$, where G and I stand for 'guilty' and 'innocent', respectively. From the vantage point of the triers-of-fact, a given type-t suspect (with a guilt signal $\Pi(t)$) is a realization (a draw from the distribution) of $\theta(t)$.

4 Bounding reasonable doubt

A verdict can err in two ways: convicting an innocent (type I) or acquitting a guilty (type II). Let $c_1(P)$ and $c_2(P)$ represent, respectively, the social cost (disutility) associated with type I and type II errors, assumed increasing and convex in P. These costs are related to a primordial sense of justice, to the deterrence and retribution roles of punishment, as well as to more tangible costs borne by individuals (fines, prison time, defence expenditures, loss of reputation and opportunities) and the public (see discussion in Andreoni 1991).

While measuring the costs $c_1(P)$ and $c_2(P)$ could be far from trivial, the ratio

$$\varphi(P) \equiv c_1(P)/c_2(P) \tag{4.1}$$

bears an intuitive meaning, as it represents the cost of type I error per unit cost of type II error or, alternatively, the number of guilty defendants that are worth acquitting in order to avoid the conviction of *one* innocent defendant, and is deeply rooted in judicial systems (Volokh 1997). Consistent with its meaning, we assume that $\varphi(P)$ increases with P (see discussions in Andreoni 1991, Weinstein and Dewsbury 2006).[3] This property, we show below, motivates plea bargaining.

Let $b \in [0,1]$ denote the upper bound of 'reasonable doubt,' i.e., a type-t defendant is deemed 'guilty beyond reasonable doubt' if $\Pi(t) > b$ or, alternatively, if $t > \Pi^{-1}(b)$, in which case he is convicted. The expected error cost

[3]The underlying intuition ca be seen by considering a minor felony, for which the penalty is a small fine, and a major crime, for which the penalty is, say, life imprisonment, and comparing the number of guilty suspects that are worth acquitting in order to avoid the conviction of one innocent suspect in both cases. Attaching a larger number to the second (severe crime) amounts to assuming that $\varphi(P)$ increases in P.

(type I and type II) inflicted by type-t under the conviction threshold b is

$$\begin{cases} c_1(P)Pr\{\theta(t) = I\} = c_1(P)[1 - \Pi(t)] & \text{if } t > \Pi^{-1}(b) \text{ (innocent but convicted)} \\ c_2(P)Pr\{\theta(t) = G\} = c_2(P)\Pi(t) & \text{if } t \leq \Pi^{-1}(b) \text{ (guilty but acquitted)} \end{cases}.$$

Summing over all types, the social cost associated with the bound b is

$$C(b) = c_2(P) \int_0^{\Pi^{-1}(b)} \Pi(t)dt + c_1(P) \int_{\Pi^{-1}(b)}^1 (1 - \Pi(t))dt. \tag{4.2}$$

Differentiating with respect to b gives $C'(b) = \Pi^{-1\prime}(b)\left[(c_2(P) + c_1(P))b - c_1(P)\right]$, which, noting (3.1), implies that $C(b)$ attains a global minimum at

$$b^*(P) = \frac{c_1(P)}{c_1(P) + c_2(P)} \equiv \frac{1}{1 + 1/\varphi(P)}\,, \tag{4.3}$$

where $\varphi(P)$ is defined in (4.1). Andreoni (1991) derived this bound from the calculus of a juror (or a judge) contemplating whether to convict a defendant with a certain guilty probability.[4]

Table 1 lists values of $\varphi(P)$ and the corresponding reasonable doubt bounds $b^*(P)$ for felonies of increased severity P. Misdemeanors with small penalties (that do not inflict excessive damage) bear small penalties, thus entail values of $\varphi(P)$ around one (i.e., society is willing to trade one acquitted-guilty for one convicted-innocent). In such cases the burden of proof is about one half ($b^* = 0.5$), in that the plaintiff needs only demonstrate that the probability that he is right (the offender is guilty) exceeds 50 percent. More serious offenses, but still less than severe crimes, entail larger penalties and society's willingness to pay, in terms of the number of acquitted guilty worthy of avoiding one convicted innocent, is correspondingly larger. If, say, $\varphi(P) \approx 3$, the burden of proof increases to 75 percent ($b^* = 0.75$), i.e., it is sufficient that the plaintiff

[4]If the values of correct verdicts (convicting the guilty or acquitting the innocent) are normalized to zero (as in Andreoni 1991), then $-C(b)$ bears a social value interpretation and b^* maximizes this value.

demonstrates that the probability that he is right exceeds 75 percent in order to win the case.[5]

Table 1: Values of $\varphi(P)$ and $b^*(P)$ corresponding to increasing P values.

$\varphi(P)$	$b^*(P)$
1	0.5
3	0.75
10	0.909
100	0.99
1000	0.999

For criminal offenses with harsher penalties, the price society is willing to pay in order to avoid convicting the innocent is larger and the ensuing burden of proof is accordingly stricter, e.g., $\varphi > 10$ implies $b^* > 0.9$. For a draconian sentence, e.g., a capital offense, $\varphi(P)$ is very large, $b^*(P) \approx 1$ and $C(b^*(P)) \approx c_2(P) \int_0^1 \Pi(t)dt$, which can be prohibitively large when too many offenders of severe crime remain at large (because both $c_2(P)$ and $\int_0^1 \Pi(t)dt$ are large).

This observation highlights a potential difficulty in using the reasonable doubt bound $b^*(P)$ as a conviction threshold. Plea bargaining addresses this limitation.

5 Plea bargaining

Three players participate in a plea bargaining situation: prosecution, defendant and judge/jury. They interact in different ways, depending on the underlying judicial structure. Baker and Mezzetti (2001) considered interaction

[5]The standards of proof in the cases where $b^* \approx 0.5$ and $b^* \approx 0.75$ are referred to as 'preponderance of evidence' and 'clear and convincing evidence,' respectively (see Kagehiro 1990).

between prosecutors and defendants, assuming exogenous jury. Bjerk (2007) studied interaction between jury and defendants under exogenous prosecution behavior. We consider a mechanism in which the interaction occurs between defendants and judes/jury (as in Bjerk 2007) and the prosecution designs the plea-bargain scheme while accounting for this interaction.

The discussion in Sections 3-4 pertains to court evidence, available at the conclusion of a trial when conviction decisions are made. Before a suspect is indicted, however, his guilt probability is assessed by the prosecution, which can either dismiss the case, send to trial or offer a plea bargain. The evidence available to the prosecution differs from that available to the judge/jury at the conclusion of a trial, hence the ensuing guilt probability differs as well. Repeating the procedure described in Section 3 with the evidences available to the prosecution gives the guilt probability $\pi(t)$. Thus, $\pi(t)$ is the pre-trial guilt probability of type-t defendants available at the time the prosecution decides to dismiss, send to trial or offer a plea bargain.

The plea bargain scheme, corresponding to an offense of severity (punishment) P, is characterized in terms of a lower bound $a < b^*(P)$ and a plea offer (reduced punishment) menu $Q(t) < P$, satisfying $Q'(t) > 0$, as follows: type-t defendants that fall in the range $a \leq \pi(t) \leq b^*(P)$ qualify to participate in the bargain; a qualifying type-t can either accept the plea offer $Q(t)$ or go to trial; type-t defendants for whom $\pi(t) > b^*(P)$ are sent to trial and those with $\pi(t) < a$ are dismissed. The judge/jury decides to convict or acquit at the conclusion of a trial.

Plea bargain rules vary across jurisdictions (see, e.g., Miceli 1996, Adelstein and Miceli 2001) and the above design addresses the rationale for plea bargaining in this work, namely the near-one conviction threshold $b^*(P)$ of severe

crimes. Because defendants with guilt probability above $b^*(P)$ are not part of the problem, they should not be part of the solution (i.e., should not qualify to participate in the plea bargain). The lower bound for qualification, a, is set endogenously (see subsection 5.2). The $Q(t)$ should increase with t, i.e., with the guilt probability, because smaller type suspects (with weaker guilt signals) will require a smaller punishment in order to give up the chance of acquittal in trial.

The mechanism proceeds along the following steps: in Step 1 the prosecution specifies $a < b^*(P)$ and the plea offer menu $Q(t)$, $t \in [\pi^{-1}(a), \pi^{-1}(b^*(P))]$; in Step 2 the (qualifying) type-t defendant decides to accept the plea offer or go to trial; and in Step 3 the judge/jury decides, at the conclusion of the trial, to convict or acquit. The defendant-judge interaction occurs in Steps 2 and 3, during which both players take $Q(t)$ as given. The optimal a and $Q(t)$ are set by the prosecution in Step 1 by optimizing over the social cost of type I and type II errors while accounting for this interaction.

5.1 Defendant - judge/jury interaction (Steps 2-3)

For brevity, we let $D(t)$ and J denote type-t defendant and judge/jury, respectively. The payoff of J (which represents society's preferences) stems from the costs of type I and type II errors. A type I error occurs if an innocent $D(t)$ pleads guilty and receives the penalty $Q(t)$, inflicting the cost $c_1(Q(t))$, or if the innocent $D(t)$ goes to trial and ends up being convicted, inflicting the cost $c_1(P)$.[6] A type II error occurs when a guilty $D(t)$ pleads guilty and

[6] Rakoff (2014) reports that about 10 percent of guilty pleaders were later proven innocent. The true error is likely higher, since the 10 percent estimate does not include innocent guilty-pleaders that have not been proven innocent. The rate of wrongful convictions in the US has been estimated between 2 percent and 8 percent (see Rakoff 2014, p. 10, Weinstein and Dewsbury 2006).

receives the reduced penalty $Q(t)$, inflicting the cost $c_2(P - Q(t))$[7] or when a guilty $D(t)$ goes to trial and ends up being acquitted, inflicting the cost $c_2(P)$.

The payoff of $D(t)$ follows from the (private) disutility associated with a punishment. This disutility function, denoted $\psi(\cdot)$, satisfies (see Andreoni 1991, Bjerk 2007)[8]

$$\psi(0) = 0,\ \psi' > 0,\ \psi'' \geq 0. \tag{5.1}$$

Thus, if $D(t)$ accepts the plea offer he will experience the disutility $\psi(Q(t))$ and if he chooses to go to trial his disutility is $\psi(P)$ or $\psi(0) = 0$ if convicted or acquitted, respectively. The index

$$q(t) \equiv \psi(Q(t))/\psi(P) \tag{5.2}$$

represents the reduction in disutility associated with the plea offer $Q(t)$ relative to the punishment that fits the crime P. The assumptions that $0 < Q(t) < P$ and $Q'(t) > 0$ imply $q(t) \in (0, 1)$ and $q'(t) > 0$.

During a trial, more evidence is brought up and existing evidence may receive different interpretation, modifying the ensuing guilt probability (or type). It is assumed (following Grossman and Katz 1983, Reinganum 1988, and subsequent literature) that these changes are biased in favor of the innocents, in that the guilt probability of innocent defendants is more likely to decrease during the trial, whereas the guilt probability of guilty defendants is more likely to increase. We account for this by specifying the post-trial type $\tilde{t}$, as

[7]The cost of type II error for plea-bargain qualifiers that accept the plea offer should satisfy the following properties: first, it should decrease with Q (a higher plea offer reduces the discrepancy between the penalty that fits the crime P and the actual penalty Q); second, it should vanish when $Q = P$ (i.e., when a guilty defendant receives the correct penalty); third, it should equal $c_2(P)$ when $Q = 0$ (i.e., when a guilty defendant gets away unpunished). The function $c_2(P - Q)$ satisfies these conditions.

[8]For simplicity it is assumed that guilty and innocent defendants share the same disutility function $\psi(\cdot)$. Allowing different disutility functions for innocent and guilty that satisfy $\psi^I(x) = \gamma\psi^G(x)$, with γ a positive constant, will not affect the results.

perceived by $D(t)$ before the trial, as $\tilde{t} = t + \tilde{\varepsilon}(t)$, where

$$\tilde{\varepsilon}(t) = \begin{cases} \varepsilon^I(t) & \text{if } \theta(t) = I \\ \varepsilon^G(t) & \text{if } \theta(t) = G \end{cases}. \tag{5.3}$$

In (5.3), $\varepsilon^I(t)$ and $\varepsilon^G(t)$ are random errors with (common knowledge) distributions $F^I(\cdot, t)$ and $F^G(\cdot, t)$, respectively, such that the former is more likely to take negative values and the latter is more likely to draw positive values. In particular, $E\{\tilde{\varepsilon}(t)|\theta(t) = G\} > 0$, $E\{\tilde{\varepsilon}(t)|\theta(t) = I\} < 0$ and for any nondecreasing function $\alpha(\cdot)$

$$E\{\alpha(\tilde{\varepsilon}(t))|\theta(t) = I\} < E\{\alpha(\tilde{\varepsilon}(t))|\theta(t) = G\}. \tag{5.4}$$

Because $\tilde{t} \in [0, 1]$, the supports of $\varepsilon^G(t)$ and $\varepsilon^I(t)$ are contained in $[-t, 1-t]$. Figure 1 shows possible densities $f^I(\varepsilon, t)$ and $f^G(\varepsilon, t)$, corresponding to $F^I(\varepsilon, t)$ and $F^G(\varepsilon, t)$, respectively, for a given t.

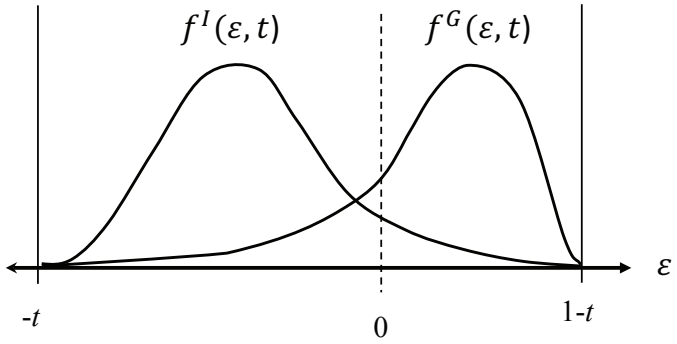

Figure 1: Possible densities $f^G(\cdot, t)$ and $f^I(\cdot, t)$ for a given t.

An impartial J, devoid of any information regarding plea-bargaining, who knows $\pi(t)$ and observes the realization ε at the end of the trial will update the guilt probability of $D(t)$ according to (the Bayesian updating rule)

$$\Pi(\varepsilon, t) \equiv Pr\{\theta(t) = G|\tilde{\varepsilon}(t) = \varepsilon\} = \frac{\pi(t)f^G(\varepsilon, t)}{\pi(t)f^G(\varepsilon, t) + (1 - \pi(t))f^I(\varepsilon, t)}. \tag{5.5}$$

We refer to $\Pi(\varepsilon, t)$ as the naïve post-trial guilt probability and assume that it satisfies

$$\Pi(-t, t) = 0, \;\; \Pi(1-t, t) = 1, \;\; \Pi_\varepsilon(\varepsilon, t) > 0 \;\text{ and }\; \Pi_t(\varepsilon, t) > 0 \tag{5.6}$$

for all $t \in [\pi^{-1}(a), \pi^{-1}(b^*(P))]$ and $\varepsilon \in [-t, 1-t]$, where $\Pi_\varepsilon(\varepsilon, t) \equiv \partial\Pi(\varepsilon, t)/\partial\varepsilon$ and $\Pi_t(\varepsilon, t) \equiv \partial\Pi(\varepsilon, t)/\partial t$.[9]

Judges (jury), however, are not visitors from outer space but live here and now, are aware of the bargaining that takes place outside the court room and will incorporate this information in their assessment of guilt probability. In particular, they will weigh in their belief regarding $D(t)$'s strategy when deciding whether to accept the plea offer or go to trial. This, in turn, depends on $D(t)$'s belief regarding J's belief and so on, giving rise to intricate strategic interaction.

This interaction is analyzed as a signaling game (Harsanyi 1967-1968, Fudenberg and Tirole 1991a,b). Consequently, the post-trial guilt probability, updated by a Bayesian-rational J at the conclusion of the trial upon observing the trial signal ε, depends also on her belief $r = (r^G, r^I)$ that $D(t)$ will (have chosen to) go to trial with probability r^G or r^I if guilty or innocent, respectively. The ensuing post-trial guilt probability, evaluated by J upon observing the trial signal ε, takes the form (see Appendix A)

$$\hat{\Pi}(\varepsilon, r, t) = \frac{r^G \Pi(\varepsilon, t)}{r^G \Pi(\varepsilon, t) + r^I (1 - \Pi(\varepsilon, t))} \; . \tag{5.7}$$

The information available to $D(t)$ and J at the time decisions are made is as follows. The penalty P with the ensuing $\varphi(P)$ and $b^*(P)$, defined in (4.1) and (4.3), respectively, as well as the pre-trial guilt probability $\pi(t)$, the

[9] Requiring $f^G(\varepsilon, t)/f^I(\varepsilon, t) \to 0$ as $\varepsilon \to -t$ and $f^G(\varepsilon, t)/f^I(\varepsilon, t) \to \infty$ as $\varepsilon \to 1-t$ ensure (5.6). Similar properties regarding f^G/f^I are required and justified by Bjerk (2007).

distributions $F^G(\cdot, t)$ and $F^I(\cdot, t)$ of the trial signal $\tilde{\varepsilon}(t)$ and the ensuing naïve guilt probability $\Pi(\varepsilon, t)$, defined in (5.5), are common knowledge. At the time $D(t)$ makes his choice he knows $\theta(t) \in \{G, I\}$ (guilty or innocent) and the private disutility $\psi(\cdot)$. At the time J decides to convict or acquit she observes the realization ε of $\tilde{\varepsilon}(t)$.

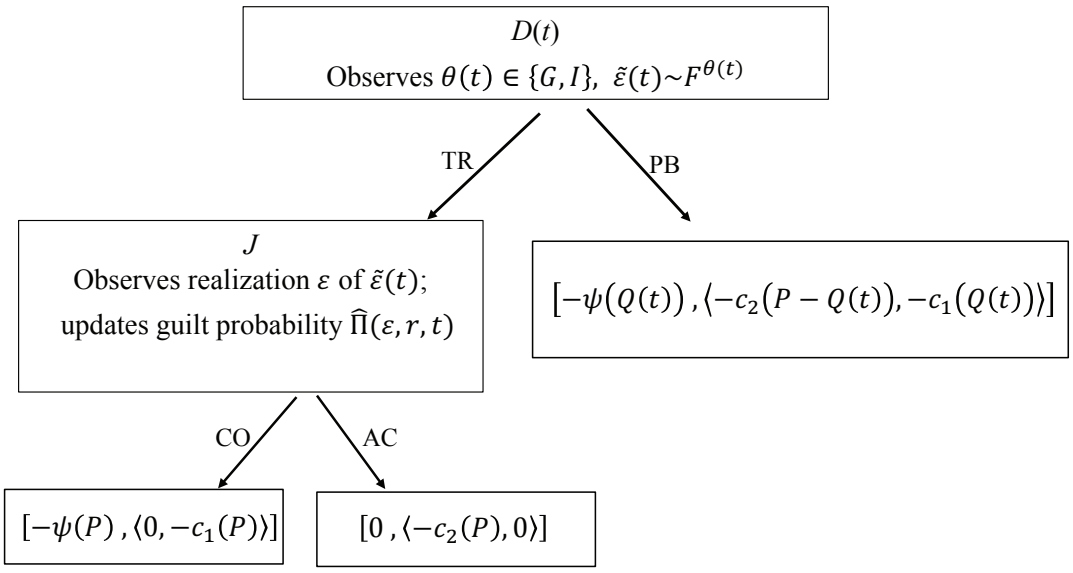

Figure 2: The game in extensive form.

The sequence of moves is presented in Figure 2 and the payoffs are summarized in Table 2. If $D(t)$ accepts the plea offer (PB) his payoff is $-\psi(Q(t))$, whereas if he chooses to go to trial (TR) his payoff is $-\psi(P)$ if convicted and zero if acquitted. J's payoff if $D(t)$ accepts the plea bargain is $-c_2(P - Q(t))$ if $D(t)$ is guilty (see footnote 7) or $-c_1(Q(t))$ if innocent. If $D(t)$ goes to trial, J's payoff is $\langle 0, -c_1(P)\rangle$ if she convicts (CO) and $\langle -c_2(P), 0\rangle$ if she acquits (AC), where $\langle x, y\rangle$ indicates payoff x or y if $D(t)$ is guilty or innocent, respectively.

Considering J's decision (the south-west branch of the game tree), her ex-

Table 2: Payoffs.

		$D(t)$	
		PB (accept the plea offer)	TR (go to trial)
J	CO (convict)	$-\psi(Q(t))\,,\ \langle -c_2(P-Q(t)), -c_1(Q(t))\rangle$	$-\psi(P)\,,\ \langle 0, -c_1(P)\rangle$
	AC (acquit)		$0\,,\ \langle -c_2(P), 0\rangle$

pected payoff if she convicts or acquits is $-c_1(P)(1-\hat{\Pi}(\varepsilon,r,t))$ or $-c_2(P)\hat{\Pi}(\varepsilon,r,t)$, respectively. The decision rule that maximizes her expected payoff is therefore to convict if $-c_1(P)(1-\hat{\Pi}(\varepsilon,r,t)) > -c_2(P)\hat{\Pi}(\varepsilon,r,t)$ or equivalently, if $\hat{\Pi}(\varepsilon,r,t)/[1-\hat{\Pi}(\varepsilon,r,t)] < \varphi(P)$, which in turn can be expressed as: convict if $\hat{\Pi}(\varepsilon,r,t) > 1/[1+1/\varphi(P)] = b^*(P)$, where $b^*(P)$ is the 'reasonable doubt' bound, defined in (4.3). We summarize J's decision rule in:

Property 1. *The decision rule that maximizes J's expected payoff is*

$$\begin{cases} CO\ (convict) & if\ \ \hat{\Pi}(\varepsilon,r,t) > b^*(P) \\ AC\ (acquit) & otherwise \end{cases} . \tag{5.8}$$

It is seen from (5.8) that the plea-bargain mechanism admits the standard of proof given by the reasonable doubt bound $b^*(P)$. As $b^*(P)$ depends neither on the plea offer $Q(t)$ nor on the defendant's type, the standard of proof is uniform, consistent with the 'equality before the law' principle. The plea offer affects $D(t)$'s choice, as well as the post-trial guilt probability (updated by J at the conclusion of the trial), but once the decision to go to trial has been made, the payoffs are independent of the plea offer (see Figure 2) and so is the standard of proof. We summarize this discussion in:

Property 2. *The plea-bargain mechanism admits a uniform standard of proof for conviction, given by the reasonable doubt bound $b^*(P)$.*

Turning to $D(t)$, if he chooses to go to trial his expected payoff is $-\psi(P)$

times the probability of conviction (recalling (5.1), the payoff under acquittal is normalized to zero). In view of (5.8), the probability of conviction in trial, as perceived by $D(t)$ at the time he makes his choice, is $Pr\{\hat{\Pi}(\tilde{\varepsilon}(t), r, t) > b^*(P)|\theta(t)\}$. Thus, $D(t)$'s expected payoff if he goes to trial is $-\psi(P)Pr\{\hat{\Pi}(\tilde{\varepsilon}(t), r, t) > b^*(P)|\theta(t)\}$. If $D(t)$ accepts the plea offer, his payoff is $-\psi(Q(t)) = -\psi(P)q(t)$ (cf. (5.2)). We summarize these considerations in:

Property 3. *If*

$$Pr\{\hat{\Pi}(\tilde{\varepsilon}(t), r, t) > b^*(P)|\theta(t) = G\} = q(t), \tag{5.9}$$

then a guilty $D(t)$ is indifferent between accepting the plea offer or going to trial. If the equality in (5.9) holds as $>$ or $<$, a guilty $D(t)$ prefers the plea offer or the trial, respectively. The same rule applies to an innocent $D(t)$ when the condition $\theta(t) = I$ replaces $\theta(t) = G$ in the conditional probability expression on the left-hand side of (5.9).

A strategy profile consists of J's conviction rule and $D(t)$'s decision rule, characterized in Properties 1 and 3, and is based on $r = (r^G, r^I)$. In a perfect Bayesian equilibrium (PBE), J's beliefs regarding r are consistent with $D(t)$'s decisions. Let $r^* = (r^{G*}, r^{I*})$ denote a PBE.

Notice that if $r^* = (1, 1)$, i.e., J believes that $D(t)$ will go to trial with probability one whether innocent or guilty and $D(t)$, while perfectly aware of J's belief, will indeed go to trial whether guilty or innocent, then, noting (5.7), $\hat{\Pi}(\varepsilon, r^*, t) = \Pi(\varepsilon, t)$. Thus,

$$\Pi^G(t) = Pr\{\Pi(\tilde{\varepsilon}(t), t) > b^*(P)|\theta(t) = G\} \tag{5.10}$$

is the conviction probability perceived by a guilty $D(t)$ at the time he chooses between accepting the plea offer or going to trial in a PBE with $r^* = (1, 1)$.

We can now state the following properties:

Proposition 1. *(i) To any* $q(t) \in (0, \Pi^G(t)]$ *there corresponds a unique PBE* $r^* = (r^{G*}, 1)$ *where* $r^{G*} > 0$ *increases with* $q(t)$ *and approaches* 1 *as* $q(t)$ *approaches* $\Pi^G(t)$*. (ii) If* $q(t) \geq \Pi^G(t)$*, then* $r^{G*} = r^{I*} = 1$*.*

The proof is presented in Appendix B, where it is verified that

Property 4. *If* $q(t) \in (0, \Pi^G(t)]$*, then in a PBE*

$$Pr\left\{\hat{\Pi}(\tilde{\varepsilon}(t), r^*, t) > b^*(P) | \theta(t) = G\right\} = q(t)\,. \tag{5.11}$$

Thus, recalling Property 3, in PBEs where $q(t) \leq \Pi^G(t)$, a guilty $D(t)$ is indifferent between accepting the plea offer and going to trial. Under (5.11), condition (5.4) implies (see Appendix)

$$Pr\left\{\hat{\Pi}(\tilde{\varepsilon}(t), r^*, t) > b^*(P) | \theta(t) = I\right\} < q(t).$$

Thus (Property 3 again), innocent $D(t)$ qualifiers prefer to go to trial rather than plead guilty, i.e., $r^{I*} = 1$. Given $r^{I*} = 1$, equation (5.11) specifies r^{G*} as a function of $q(t)$, such that r^{G*} increases with $q(t)$ and approaches one as $q(t)$ approaches $\Pi^G(t)$, as stated in part (i) of the proposition.

If $q(t) = \Pi^G(t)$, the penalty reduction associated with the plea bargain is not significant enough to justify giving up the chance of acquittal in trial (however small) even for a guilty $D(t)$. If this is the case for $q(t) = \Pi^G(t)$, it must also hold for less favorable plea offers $q(t) > \Pi^G(t)$, implying $r^{G*} = r^{I*} = 1$, as stated in part (ii) of the proposition.

Notice that non-qualifying $D(t)$s, with $\pi(t) > b^*(P)$, must go to trial, so for them $r^{G*} = r^{I*} = 1$ from the outset, implying $\hat{\Pi}(\varepsilon, r^*, t) = \Pi(\varepsilon, t)$. Indeed, when no interaction takes place, the best J can do is to update guilt probabilities according to the naïve rule (5.5).

For $q(t) \in (0, \Pi^G(t))$, all PBEs are semi-separating, with all innocent and some guilty $D(t)$ going to trial, consistent with Bjerk (2007) (and with Baker and Mezzetti 2001, Kim 2010, in the context of strategic interaction between prosecution and defendants). Unlike Bjerk (2007), the different PBEs, while vary uniquely with the plea offer $q(t)$, share the same conviction threshold $b^*(P)$ – the upper bound of reasonable doubt. The latter, thus, retains its role as the conviction threshold in the plea bargaining situation, while the guilt probability, based on which conviction is determined, varies with the strategic interaction and with the information revealed in court.

5.2 Plea-bargain design (Step 1)

The prosecution (denoted R for brevity) specifies the plea bargain scheme in terms of $a < b^*(P)$ and a plea offer menu $Q(t) < P$, or $q(t) = \psi(Q(t))/\psi(P)$, for $t \in [\pi^{-1}(a), \pi^{-1}(b^*(P))]$. In doing so, R optimizes over the social cost of type I and type II errors while accounting for the interaction between $D(t)$ and J, characterized above. The information needed to accomplish this task, in addition to the common knowledge information mentioned above, is the social cost of type I and type II errors, $c_1(P)$ and $c_2(P)$, respectively, as well as $D(t)$'s disutility function ψ.[10]

Denote by $\hat{r}^G(q)$ the equilibrium r^{G*} corresponding to the plea offer $q(t) = q$ and let $\hat{r}(q) = (\hat{r}^G(q), 1)$. Noting Proposition 1($i$), for $q \in (0, \Pi^G(t)]$, $\hat{r}^G(q)$ increases in q and a qualifying $D(t)$ chooses to go to trial with probability $\hat{r}^G(q)$ or one if he is guilty or innocent, respectively. The social cost incurred if a guilty $D(t)$ participates in the plea bargain is $c_2(P - Q)$, where, recalling

[10]Notice that in the $D(t)$-J interaction, knowledge of the cost ratio $\varphi(P)$ was needed but not the individual costs $c_1(P)$ and $c_2(P)$ – see discussion below equation (4.1).

(5.2), $Q = \psi^{-1}(\psi(P)q)$ and the cost is due to the reduced penalty (see footnote 7). The cost incurred if a guilty $D(t)$ goes to trial is zero or $c_2(P)$ if convicted or acquitted, respectively. Now, the (pre-trial) probability of acquittal for a guilty $D(t)$ that goes to trial is $Pr\{\hat{\Pi}(\tilde{\varepsilon}(t), \hat{r}(q), t) \leq b^*(P) | \theta(t) = G\}$, which in a PBE equals $1 - q$ if $q \in (0, \Pi^G(t)]$ (cf. (5.11)). Thus, the social cost inflicted by a guilty $D(t)$ when $q(t) = q \in (0, \Pi^G(t)]$ is

$$C^G(q) = [1 - \hat{r}^G(q)]c_2(P - \psi^{-1}(\psi(P)q)) + \hat{r}^G(q)[1 - q]c_2(P). \tag{5.12}$$

An innocent $D(t)$ goes to trial with probability one and will be convicted if the post-trial guilt probability exceeds the reasonable doubt bound $b^*(P)$, the probability of which (as perceived by the innocent defendant before the trial) is $Pr\{\hat{\Pi}^*(\tilde{\varepsilon}(t), \hat{r}(q), t) > b^*(P) | \theta(t) = I\}$. The innocent $D(t)$ thus inflicts the social cost

$$C^I(q, t) = Pr\{\hat{\Pi}^*(\tilde{\varepsilon}(t), \hat{r}(q), t) > b^*(P) | \theta(t) = I\}c_1(P). \tag{5.13}$$

From R's vantage point at the time the plea bargain scheme is designed, the social cost associated with setting $q(t) = q$ is

$$\pi(t)C^G(q) + (1 - \pi(t))C^I(q, t)$$

and the optimal plea offer for a type-t defendant is

$$q^*(t) = \arg\min_{q \in (0, \Pi^G(t)]} \left\{\pi(t)C^G(q) + (1 - \pi(t))C^I(q, t)\right\}. \tag{5.14}$$

Turning to the choice of the participation criterion $a < b^*(P)$, suppose that without plea bargaining all suspects with $\pi(t) \leq b^*(P)$ are released and those with $\pi(t) > b^*(P)$ are sent to trial.[11] Before the introduction of plea

[11] Changing this assumption will affect the choice of a.

bargaining, the released suspects with $a \leq \pi(t) \leq b^*(P)$ inflict the cost (type II)

$$\int_{\pi^{-1}(a)}^{\pi^{-1}(b^*(P))} \pi(t)c_2(P).$$

The introduction of plea bargaining changes this cost to

$$\int_{\pi^{-1}(a)}^{\pi^{-1}(b^*(P))} \left[\pi(t)C^G(q^*(t)) + (1-\pi(t))C^I(q^*(t),t)\right] dt.$$

Subtracting the former from the latter gives the plea bargain value

$$v(a) = \int_{\pi^{-1}(a)}^{\pi^{-1}(b^*(P))} \left[\pi(t)\left[C^G(q^*(t)) - c_2(P)\right] + (1-\pi(t))C^I(q^*(t),t)\right] dt \tag{5.15}$$

Differentiating $v(a)$ with respect to a and equating to zero, the optimal lower bound for participation, a^*, satisfies

$$a^* = \frac{1}{1+1/\varphi^a(a^*)}\,, \tag{5.16}$$

provided $0 < \varphi^a(a^*) < \varphi(P)$, where

$$\varphi^a(a) = \frac{C^I(q^*(\pi^{-1}(a)), \pi^{-1}(a))}{c_2(P) - C^G(q^*(\pi^{-1}(a)))}\,. \tag{5.17}$$

We summarize the above discussion in:

Property 5. *The optimal plea offer menu and lower bound for participation are $q^*(t)$ and a^* defined in (5.14) and (5.16), respectively.*

6 Concluding comments

The principle 'innocent unless proven guilty beyond reasonable doubt,' underlying judicial processes in democratic societies, requires an upper bound of reasonable doubt to have any meaning at all. Such a bound, corresponding to a felony of severity P, is derived and shown to equal $b^*(P) = 1/(1+1/\varphi(P))$,

where $\varphi(P)$ signifies the number of guilty suspects worth setting free to avoid the conviction of *one* innocent person. As $\varphi(P)$ bears ethical, moral and cultural connotation, it varies across societies that differ in these respects and so is the ensuing burden of proof $b^*(P)$.

Since $b^*(P)$ approaches one as P increases, its use as a conviction threshold may leave too many offenders of major crimes at large and this limitation motivates the use plea bargaining. The latter introduces strategic interaction between concerned parties. Considering interaction between judge/jury and defendants, it is shown that the resulting plea bargain scheme admits a uniform standard of proof, given by the reasonable doubt bound $b^*(P)$, consistent with the 'equality before the law' principle.

In actual practice examples of innocents pleading guilty are not rare (Rakoff 2014), in violation of the model's prediction. One reason might be that the choice to go to trial entails considerable transactions costs (lawyers, time spent in court, anxiety), incurred disregarding the trial's outcome. These costs should be weighed in against the cost of accepting the plea offer. Another explanation is that players' perceptions (or beliefs regarding the other players' perceptions) of the rules of the signaling game may deviate from those giving rise to the decisions underlying the PBE. Yet another explanation could be due to self-interested prosecutors whose behavior deviates from the social goal assumed in this work (see Miceli 1990). The first explanation can be incorporated within the present framework; the other two require extensions.

Appendix

A Defendant-judge/jury interaction

The interaction is analyzed as a signaling game between a type-t defendant, denoted $D(t)$, and a judge/jury, denoted J, as follows (see Figure 2). First, $D(t)$ chooses $a_D \in \{TR, PB\}$, where TR means 'go to trial' and PB means 'accept the plea offer'. If $a_D = PB$, $D(t)$ receives the penalty $Q(t)$ and the game ends. If $a_D = TR$, the defendant goes to trial. At the end of the trial a signal ε is drawn from the distribution of $\tilde{\varepsilon}(t)$ and J (the judge or jury) chooses $a_J \in \{CO, AC\}$ after observing ε and weighing in her belief regarding $D(t)$'s strategy, where CO means 'convict' and AC stands for 'acquit'. The corresponding mix actions are $\alpha_D(\theta(t)) = Pr\{a_D = TR|\theta(t)\}$ and $\alpha_J(\varepsilon, a_D) = Pr\{a_J = CO|\varepsilon, a_D\}$. At the time $D(t)$ chooses a_D, he knows $\theta(t) \in \{G, I\}$ with certainty and $\tilde{\varepsilon}(t)$ up to the distribution $F^{\theta(t)}(\cdot, t)$; at the time J chooses a_J, she observes ε and knows $\theta(t)$ up to a guilt probability updated based on the observed signal ε and her belief regarding a_D.

The expected payoff of $D(t)$, at the time he chooses a_D, under the mixed actions (α_D, α_J) is[12]

$$u_D(\alpha_D, \bar{\alpha}_J(\theta(t)), \theta(t)) = \alpha_D E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t)\}(-\psi(P)) + (1-\alpha_D)(-\psi(P)q(t)),$$

where $\bar{\alpha}_J(\theta(t)) \equiv E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t)\}$ is the pre-trial conviction probability of $D(t)$ given $\theta(t)$, as perceived by $D(t)$ before the trial, and, noting (5.2), $\psi(Q(t)) = \psi(P)q(t)$.

J's payoff at the time she chooses a_J is

$$u_J(\alpha_D, \alpha_J(\varepsilon), \theta(t)) = \begin{cases} \alpha_D\alpha_J(\varepsilon)[-c_1(P)] + (1-\alpha_D)[-c_1(Q(t))] & \text{if } \theta(t) = I \\ (1-\alpha_D)[-c_2(P-Q(t))] & \text{if } \theta(t) = G \end{cases}$$

[12] We follow closely the notation in Fudenberg and Tirole (1991a, pp. 324-326).

For example, in $u_J(\alpha_D, \alpha_J(\varepsilon), G)$, with probability $\alpha_D\alpha_J(\varepsilon)$ the guilty defendant chooses to go to trial and is convicted, inflicting no social cost (the social cost of truthful conviction is normalized to zero) and with probability $1-\alpha_D$ the guilty defendant accepts the plea offer and receives the penalty $Q(t) < P$, inflicting the social cost $c_2(P-Q(t))$ associated with type II error (see footnote 7).

$D(t)$'s strategy $\sigma_D(\cdot|\theta(t))$ sets a probability on $a_D = TR$ given $\theta(t)$; J's strategy $\sigma_J(\cdot|a_D, \varepsilon)$ sets a probability on $a_J = CO$ given $D(t)$'s action a_D and the observed trial signal ε. $D(t)$'s payoff under the strategy profile (σ_D, σ_J) is

$$\begin{aligned} u_D(\sigma_D, \sigma_J, \theta) &= \sum_{a_D}\sum_{a_J} \sigma_D(a_D|\theta)\sigma_J(a_J|a_D, \varepsilon)u_D(a_D, a_J, \theta) \\ &= \alpha_D E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t)\}(-\psi(P)) + (1-\alpha_D)(-\psi(P)q(t)). \end{aligned}$$

J's payoff under the strategy $\sigma_J(\cdot|\varepsilon, a_D)$ is

$$\Pi(\varepsilon, t)\sum_{a_J}\sigma_J(a_J|a_D)u_J(a_D, a_J, G) + (1-\Pi(\varepsilon, t))\sum_{a_J}\sigma_J(a_J|a_D)u_J(a_D, a_J, I),$$

where $\Pi(\varepsilon, t)$, defined in (5.5), is the naïve guilt probability of $D(t)$ at the conclusion of the trial, based on the observed signal ε before J weighs in her belief regarding a_D.

A perfect Bayesian equilibrium (PBE) consists of $\sigma_D^*(\theta(t))$, $\theta(t) \in \{G, I\}$, $\sigma_J^*(\varepsilon)$ and posterior guilt probability $\Pi^*(\varepsilon)$ such that:

$$\sigma_D^*(\theta(t)) \in \arg\max_{\alpha_D}\ u_D(\alpha_D, E\{\alpha_J(\tilde{\varepsilon})|\theta(t)\}, \theta(t)), \tag{A.1}$$

$$\begin{aligned} \sigma_J^*(\varepsilon) \in \arg\max_{\alpha_J(\varepsilon)}\ &\Pi^*(\varepsilon)\left[\alpha_J(\varepsilon)u_J(TR, CO, G) + (1-\alpha_J(\varepsilon))u_J(TR, AC, G)\right] + \\ &(1-\Pi^*(\varepsilon))\left[\alpha_J(\varepsilon)u_J(TR, CO, I) + (1-\alpha_J(\varepsilon))u_J(TR, AC, I)\right] \end{aligned} \tag{A.2}$$

and

$$\Pi^*(\varepsilon) = \sigma_D^*(G)\Pi(\varepsilon, t)/\left[\sigma_D^*(G)\Pi(\varepsilon, t) + \sigma_D^*(I)(1-\Pi(\varepsilon, t))\right]. \tag{A.3}$$

If the denominator on the right side of (A.3) vanishes, $\Pi^*(\varepsilon)$ can be arbitrarily set between zero and 1. The updated guilt probability $\Pi^*(\varepsilon)$ incorporates the observed trial signal ε and J's belief regarding the probability that $D(t)$ chooses to go to trial.

Noting $u_D(\cdot)$, (A.1) gives

$$\sigma_D^*(G) = \begin{cases} 1 & \text{if } E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} < q(t) \\ \in [0,1] & \text{if } E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} = q(t) \\ 0 & \text{if } E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} > q(t) \end{cases} \tag{A.4a}$$

and

$$\sigma_D^*(I) = \begin{cases} 1 & \text{if } E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = I\} < q(t) \\ \in [0,1] & \text{if } E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = I\} = q(t) \\ 0 & \text{if } E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = I\} > q(t) \end{cases} . \tag{A.4b}$$

Noting $u_J(\cdot)$, (A.2) gives

$$\sigma_J^*(\varepsilon) = \begin{cases} 1 & \text{if } \Pi^*(\varepsilon)c_2(P) > (1 - \Pi^*(\varepsilon))c_1(P) \\ 0 & \text{otherwise} \end{cases} . \tag{A.5}$$

From (A.5) it is seen that J's equilibrium strategy is to convict if

$$\Pi^*(\varepsilon)/(1 - \Pi^*(\varepsilon)) > c_1(P)/c_2(P)$$

or, alternatively, if

$$\Pi^*(\varepsilon) > [1 + 1/\varphi(P)]^{-1} \equiv b^*(P), \tag{A.6}$$

where $b^*(P)$ is the reasonable doubt bound defined in (4.3).

Let $r = (r^G, r^I)$, where r^G and r^I represent, respectively, J's belief regarding the probability that $D(t)$ chooses to go to trial if he is guilty or innocent. In a PBE, $r^* = (\sigma_D^*(G), \sigma_D^*(I))$, hence, noting (A.3),

$$\hat{\Pi}(\varepsilon, r^*, t) = \Pi^*(\varepsilon) = \frac{r^{G*}\Pi(\varepsilon, t)}{r^{G*}\Pi(\varepsilon, t) + r^{I*}(1 - \Pi(\varepsilon, t))}, \tag{A.7}$$

verifying (5.7).

B Proof of Proposition 1

(i) We begin by showing:

Lemma B.1. *If* $q(t) \in (0, \Pi^G(t)]$, *then*

$$E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} = q(t) \tag{B.1}$$

must hold in a PBE.

Proof. Suppose $\alpha_J(\varepsilon)$ and $\alpha_D(\theta(t))$ constitute a PBE. From (A.5)-(A.6),

$$\alpha_J(\varepsilon) = \begin{cases} 1 & \text{if } \Pi^*(\varepsilon) > b^*(P) \\ 0 & \text{otherwise} \end{cases}$$

so

$$\alpha_J(\tilde{\varepsilon}(t)) = \begin{cases} 1 & \text{if } \Pi^*(\tilde{\varepsilon}(t)) > b^*(P) \\ 0 & \text{otherwise} \end{cases}. \tag{B.2}$$

Suppose $E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} < q(t)$. Since $\Pi(\varepsilon, t)$ increases in ε (cf. (5.6)), $\Pi^*(\varepsilon)$ increases in ε as well and $\alpha_J(\cdot)$ is nondecreasing. Condition (5.4) then implies that $E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = I\} < q(t)$ and (A.4) gives $\sigma_D(G) = \sigma_D(I) = 1$, which in turn implies, noting (A.3) and (A.7), that $\Pi^*(\varepsilon) = \Pi(\varepsilon, t)$. Thus, (B.2) implies in this case

$$\alpha_J(\tilde{\varepsilon}(t)) = \begin{cases} 1 & \text{if } \Pi(\tilde{\varepsilon}(t), t) > b^*(P) \\ 0 & \text{otherwise} \end{cases}$$

and, noting (5.10),

$$E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} = Pr\{\Pi(\tilde{\varepsilon}(t), t) > b^*(P)|\theta(t) = G\} \equiv \Pi^G(t) \geq q(t),$$

a contradiction (since $q(t) \in (0, \Pi^G(t)]$).

Suppose $E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} > q(t)$ and consider first $E\{\alpha_J(\tilde{\varepsilon})|\theta(t) = I\} < q(t)$. Then, from (A.4), $\sigma_D^*(G) = 0$, $\sigma_D^*(I) = 1$ and (A.3) implies

that $\Pi^*(\varepsilon) = 0$ identically at all ε. The conviction rule (A.6) then implies zero conviction probability, hence $\alpha_J(\tilde{\varepsilon}(t)) = 0$ for all $\tilde{\varepsilon}(t)$ values and $E\{\alpha_J(\tilde{\varepsilon}(t)|\theta(t) = G\} = 0$ – a contradiction.

The case $E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} > q(t)$ and $E\{\alpha_J(\tilde{\varepsilon})|\theta(t) = I\} > q(t)$, under which guilty and innocent qualifiers accept the plea offer, can be ruled out noting that in this case $\sigma_D^*(G) = \sigma_D^*(I) = 0$ and any $\Pi^*(\varepsilon) \in [0, 1]$ constitutes a PBE. Setting $\Pi^*(\varepsilon) = 0$ implies $\alpha_J(\tilde{\varepsilon}) = 0$ and $E\{\alpha_J(\tilde{\varepsilon}(t)|\theta = I\} = 0 < q(t)$ – a contradiction. □

It follows from Lemma B.1 and (5.4), recalling that $\alpha_J(\cdot)$ is nondecreasing, that if $q(t) \in (0, \Pi^G(t)]$, then

$$E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = I\} < E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} = q(t) \tag{B.3}$$

must hold in a PBE. Thus, noting (A.4), (B.3) gives

Lemma B.2. *If $q(t) \in (0, \Pi^G(t)]$, then $r^{G*} \in (0, 1]$ and $r^{I*} = 1$.*

In view of (A.7), in a PBE equation (B.2) can be rendered as

$$\alpha_J(\tilde{\varepsilon}(t)) = \begin{cases} 1 & \text{if } \hat{\Pi}(\tilde{\varepsilon}(t), r^*, t) > b^*(P) \\ 0 & \text{otherwise} \end{cases} \tag{B.4}$$

and condition (B.1) can be expressed as

$$Pr\left\{\hat{\Pi}(\tilde{\varepsilon}(t), r^*, t) > b^*(P)|\theta(t) = G\right\} = q(t), \tag{B.5}$$

verifying Property 4.

Noting (5.6) and (A.7), when $r^{G*} > 0$, $\hat{\Pi}(\varepsilon, r^*, t)$ increases from zero to one as ε increases from its lower support $-t$ to its upper suport $1 - t$. Moreover, $\hat{\Pi}(\varepsilon, r^*, t)$ increases from zero to $\Pi(\varepsilon, t)$ as r^{G*} increases from zero to one. For

a given $q(t) \in (0, \Pi^G(t)]$, these properties ensure the existence of $\hat{\varepsilon} < 1-t$ and $\hat{r}^G \in (0,1]$ satisfying

$$Pr\{\tilde{\varepsilon}(t) > \hat{\varepsilon} | \theta(t) = G\} = q(t) \tag{B.6}$$

and

$$\hat{\Pi}(\hat{\varepsilon}, (\hat{r}^G, 1), t) = b^*(P). \tag{B.7}$$

To see this, use the property that $\hat{\Pi}(\tilde{\varepsilon}(t), (\hat{r}^G, 1), t)$ increases in $\tilde{\varepsilon}(t)$ to express (B.6) as

$$Pr\{\hat{\Pi}(\tilde{\varepsilon}(t), (\hat{r}^G, 1), t) > \hat{\Pi}(\hat{\varepsilon}, (\hat{r}^G, 1), t) | \theta(t) = G\} = q(t),$$

which upon invoking (B.7) can be expressed as

$$Pr\{\hat{\Pi}(\tilde{\varepsilon}(t), (\hat{r}^G, 1), t) > b^*(P) | \theta(t) = G\} = q(t). \tag{B.8}$$

Now, when $q(t) = \Pi^G(t)$, equations (5.10) and (B.8) ensure $\hat{\Pi}(\tilde{\varepsilon}(t), (\hat{r}^G, 1), t) = \Pi(\tilde{\varepsilon}(t), t)$, implying $\hat{r}^G = 1$. Thus, (5.6) and $b^*(P) < 1$ ensure the existence of $\hat{\varepsilon} < 1-t$ that satisfies (B.6) when $q(t) = \Pi^G(t)$. As $q(t)$ decreases from $\Pi^G(t)$ toward zero, equation (B.6) implies that $\hat{\varepsilon}$ must increase toward $1-t$ (the upper support of $\tilde{\varepsilon}(t)$). Equation (B.7) then implies, recalling that $\hat{\Pi}(\varepsilon, (r^G, 1), t)$ increases in both ε and r^G, that $\hat{r}^G$ must decrease toward zero. We thus establish that

Lemma B.3. *r^{G*} decreases from 1 toward 0 as $q(t)$ decreases from $\Pi^G(t)$ toward zero.*

It follows from Lemma B.2 and Lemma B.3 that to any $q(t) \in (0, \Pi^G(t)]$ there corresponds a unique $r^* = (r^{G*}, 1)$ where r^{G*} decreases from 1 toward zero as $q(t)$ decreases from $\Pi^G(t)$ toward zero. This completes the proof of part (i).

(*ii*) Suppose $q(t) > \Pi^G(t)$. Then, repeating the arguments of Lemma B.1 implies that in a PBE

$$E\{\alpha_J(\tilde{\varepsilon}(t))|\theta(t) = G\} < q(t)$$

must hold. Equations (A.4), noting (5.4), then imply $\sigma_D^*(G) = \sigma_D^*(I) = 1$. The intuition is obvious: if qualifying type-t defendants (guilty and innocent) prefer the trial over the plea bargain when $q(t) = \Pi^G(t)$, they will certainly prefer to go to trial in less favorable plea offers $q(t) > \Pi^G(t)$.

References

Adelstein, R. and Miceli, T.: 2001, Toward a comparative economics of plea bargaining, *European Journal of Law and Economics* **11**(1), 47–67.

Andreoni, J.: 1991, Reasonable doubt and the optimal magnitude of fines: Should the penalty fit the crime?, *RAND Journal of Economics* **22**(3), 385 – 395.

Baker, S. and Mezzetti, C.: 2001, Prosecutorial resources, plea bargaining, and the decision to go to trial, *Journal of Law, Economics, & Organization* **17**(1), 149 – 167.

Becker, G.: 1968, Crime and punishment: An economic approach, *Journal of Political Economy* **76**, 169–217.

Bjerk, D.: 2007, Guilt shall not escape or innocence suffer? the limits of plea bargaining when defendant guilt is uncertain, *American Law and Economics Review* **9**(2), 305–329.

Fudenberg, D. and Tirole, J.: 1991a, *Game Theory*, MIT Press, Cambridge, MA.

Fudenberg, D. and Tirole, J.: 1991b, Perfect bayesian equilibrium and sequential equilibrium, *Journal of Economic Theory* **53**(2), 236 – 260.

Grossman, G. M. and Katz, M. L.: 1983, Plea bargaining and social welfare, *The American Economic Review* **73**(4), 749–757.

Harris, J. R.: 1970, On the economics of law and order, *Journal of Political Economy* **78**(1), 165–174.

Harsanyi, J. C.: 1967-1968, Games with incomplete information played by bayesian players, *Management Science* **14**, 159–182, 320–334, 486–502.

Kagehiro, D. K.: 1990, Defining the standard of proof in jury instructions, *Psychological Science* **1**(3), 194–200.

Kim, J.-Y.: 2010, Credible plea bargaining, *European Journal of Law and Economics* **29**(3), 279–293.

Landes, W. M.: 1971, An economic analysis of the courts, *Journal of Law and Economics* **14**(1), pp. 61–107.

Laudan, L.: 2011, Is it finally time to put 'proof beyond a reasonable doubt' out to pasture?, *Research Paper 194*, University of Texas Law.

Miceli, T.: 1996, Plea bargaining and deterrence: An institutional approach, *European Journal of Law and Economics* **3**(3), 249–264.

Miceli, T. J.: 1990, Optimal prosecution of defendants whose guilt is uncertain, *Journal of Law, Economics and Organization* **6**(1), 189–201.

Mulrine, T. V.: 1997, Reasonable doubt: How in the world is it defined?, *American University International Law Review* **12**(1), 195 – 225.

Newman, J. O.: 2006, Quantifying the standard of proof beyond a reasonable doubt: a comment on three comments, *Law, Probability and Risk* **5**, 267 – 269.

Posner, R. A.: 1973, An economic approach to legal procedure and judicial administration, *Journal of Legal Studies* **2**, 399 – 458.

Posner, R. A.: 1999, An economic approach to the law of evidence, *Stanford Law Review* **51**(6), 1477 – 1546.

Rakoff, J. S.: 2014, Why innocent people plead guilty, *The New York Review of Books* pp. 1–12.

Reinganum, J. F.: 1988, Plea bargaining and prosecutorial discretion, *The American Economic Review* **78**(4), 713–728.

Rubinfeld, D. L. and Sappington, D. E. M.: 1987, Efficient awards and standards of proof in judicial proceedings, *The RAND Journal of Economics* **18**(2), 308–315.

Tillers, P. and Gottfried, J.: 2006, Case comment – United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A collateral attack on the legal maxim that proof beyond a reasonable doubt is unquantifiable?, *Law, Probability and Risk* **5**(2), 135–157.

Volokh, A.: 1997, n guilty men, *University of Pennsylvania Law Review* **146**, 173–216.

Weinstein, J. B. and Dewsbury, I.: 2006, Comment on the meaning of proof beyond a reasonable doubt, *Law, Probability and Risk* **5**(2), 167 – 173.

Whitman, J. Q.: 2008, *The Origins of Reasonable Doubt: Theological Roots of the Criminal Trial*, Yale University Press.