Pervasive Spurious Normativity

Gillian K. Hadfield Dylan Hadfield-Menell

Abstract

This paper proposes a mathematical model for a simplified version of the game defined in Hadfield and Weingast [2012] which proposes that legal order can be described as an equilibrium in third-party decentralized enforcement coordinated by a centralized classification institution. We explore the attractiveness of joining a new group (which is assumed to have settled on an enforcement equilibrium already) where groups differ in terms of the frequency of interactions in which norm violation is possible (normative interactions) and thus punishment is called for. We show that groups in which normative interactions are frequent but involve relatively unimportant rules may achieve higher value for participants.

1 Introduction

One of the attributes that differentiates human from other animal societies is the presence of norms and in particular norms over behaviors that do not appear to have a material impact on well-being apart from the fact that people will punish the failure to adhere to a norm. In all known human societies there is a rich normative landscape, which attaches normative valence to actions including many that have no immediate material implications for those who are expected to help enforce them through social sanctions. Most societies for example have rules about what it is appropriate to eat when, what tone of voice to use in what settings, how close to stand to someone else, how to behave when waiting with others to get into a venue or access to a resource, what information should be shared with whom and when, who can participate in particular trades or economic activities and so on. In many cases, particular norms are arbitrary, even though violation of them is treated as worthy of punishment by others. We call this spurious normativity. Hadfield [1999], for example, reviews the anthropological literature on the sexual division of labor across pre-industrial societies. While almost all societies categorize work as either women's work or men's work, the particular classifications vary substantially cross-culturally: what is women's work in one society is men's work in another. The classification is arbitrary.[1]

Fessler and Navarrete [2003] call the process by which patterns of behavior are imbued with moral sentiments that motivate sanctioning of violations of the pattern *normative moralization*. They use as an example the normative moralization of handedness. Most people are right-handed but particularly in societies with few specialized tools, whether someone is right- or left-handed often has no material consequences for others. Nonetheless, many cultures treat using one's right hand as a morally approved category–denoting purity or politeness–and one's left hand as cause

[1] Hadfield [1999] shows that there is functionality to an arbitrary classification scheme of enforced norms: it coordinates investments in specialized learning by gender which raises the value of heterosexual matching in the marriage market.

for opprobrium–revealing weakness or evil. Fessler [1999] hypothesizes that, driven forward by the emotions of shame and pride that are triggered by violation or compliance with cooperative norms, culture extends the set of actions that are subject to normative moralization as a way of extending the set of actions that can be used as information about cooperation beyond those actions that directly involve cooperation. A norm that says "it is wrong to fail to look where you are going" generates direct cooperative benefits, helping people to avoid crashing into one another. A norm that says "it is wrong for a man to walk down the street wearing shorts" does not generate cooperative benefits–*unless* people in this society treat conformity to this norm as informative about a person's likelihood of behaving in conformity with norms that do generate cooperative benefits. As Fessler [1999] (who identifies the the "watch where you are going" and "men don't wear shorts out in public" norms in an Indonesian village) puts this:

> Ego's inclusion in cooperative activity is dependent on her ability to meet Others' expectations, and those expectations are, in turn related to shared standards for behaviors which are relevant to cooperation. As a consequence, the significant adaptive advantage offered by participation in cooperative activities generated selective pressure for an increase in the attention paid to these standards. Late [evolving] second order emotions [emotions that arise in response to others' first order emotions] were the vehicle through which this increase in attention was achieved. Moreover, because late second order emotions entail a sensitivity to the reactions of all individuals, Ego must be concerned with her performance vis--vis shared standards when interacting with any other member of her group. *It is only a small step from this situation to one in which the shared standards with which Ego is concerned are not limited to the question of cooperative activity–once Ego is concerned with how all Others evaluate her, it is not difficult for shared standards governing other types of behavior to become salient as well.* This is because an Other may extrapolate from situations that do not involve cooperation to those that do–an Other may think "if that individual does not follow shared standards in this context, how can I be confident that he will do so if I invite her to engage in cooperative activity?"(pre-pub p. 34, emphasis added)

Fessler's account focuses on the evolution of emotions in response to norm violation (in particular shame) to motivate voluntary conformance with norms. But as Boyd and Richerson [1992] and many others have emphasized, and Fessler's own account of shame as a second order response to Ego's actual or imagined experience of hostility or criticism from Other triggered by norm violation implies, effective third-party punishment plays a significant role in supporting norm compliance. Indeed, Mathew et al. [2012] argue that even small-scale cooperation among kin and close associates may require third-party punishment to achieve evolutionary stability.

In this paper we analyze mathematically the potential benefits of extending normative moralization to behaviors that are from a material perspective irrelevant or at least of small consequence to most people, that is, the value of pervasive and spurious normativity. We ask: does a community generate higher payoffs for participants if it punishes violation of apparently meaningless rules or if it focuses more narrowly only on norms that are functional in the sense of generating direct benefits? Intuitively, one might expect that only functional norms that govern behaviors that matter for payoffs would emerge and stabilize in equilibrium: punishment is costly and why would a society expend punishment resources on ensuring conformity with rules that have no impact on material well-being? Most analyses of norms in the law and economics literature assume that norms coordinate outcomes that improve welfare. Sugden [1986], McAdams [2005] and Myerson [2004] propose, for example, that property rules emerge because they solve the coordination problems that arise in costly contests over resources (Hawk-Dove games).

We consider the impact of extending normativity to apparently arbitrary actions on the prospect for sustaining welfare-improving behavior in an equilibrium social order based on collective third-party punishment. We demonstrate that communities that extend normativity in this way can generate higher value for participants than those that restrict the range of normativity.

The intuition of our result is as follows. Communities with pervasive spurious normativity provide agents with plentiful and cheap opportunities to observe punishment behavior by others. An individual's willingness to cooperate in a community–which requires foregoing safe non-cooperative options and exposing oneself to the risk of being exploited–depends on that individual's beliefs about the likelihood that the community effectively punishes conduct that harms the individual. If you are going to risk exposing yourself to harm, you want to know if your community contains enough people who will punish the perpetrator to give you confidence that harm is reasonably deterred. Assume you are a newcomer to a community or that the community was recently handed a new set of norms. (Think here of rule-of-law building efforts in developing countries.) Assume that the community is in an equilibrium in terms of punishment behavior but you do not know the likelihood of effective punishment of norm violations. Assume also that the only way to learn about the likelihood of punishment is to observe punishment behavior. You can learn this information more cheaply if you are given abundant opportunities to observe what happens when there are violations if the violations that you have to expose yourself to don't really matter very much. You don't really care whether men walk down the street in shorts but by taking a walk yourself you can see how others react to shorts-wearing men and thus gain information about how they would react to violations you do care about–careless driving for example. Thus, if you could participate solely as an observer except when your own interests were directly at stake, you would prefer to live in a world with pervasive even if spurious normativity–abundant opportunities to observe reactions to norm violations–than one that was narrowly focused on punishing just the stuff you care about. This will still be true even if you are required to participate in the community–complying with and punishing spurious norms–so long as those costs, which increase with the pervasiveness of norms, are not too great.

Our results have implications for the evolution and microfoundations of law.

Hadfield and Weingast [2012] present a model that derives characteristic features of law–such as generality, stability, uniqueness and universality–as attributes necessary to support an equilibrium in which behavior is patterned on the classifications of behavior articulated by a centralized institution. Enforcement is assumed to come exclusively from decentralized collective punishment of conduct classified as punishable by the centralized institution; there is no centralized enforcement apparatus such as the state. This theory of the microfoundations of law proposes, contrary to most economic and positive political theories of law, which define law as a set of rules enforced by a centralized enforcement apparatus (see Hadfield and Weingast [2014]), that law is an innovation in the mechanism used to coordinate the same enforcement mechanism that supports other normative social orders–decentralized collective punishment. This has important implications for our understanding of how law developed and how it can be built in environments where it is currently lacking.

The equilibrium legal order in Hadfield and Weingast [2012] is supported by a particular specification of beliefs. Their model posits that an agent (call the agent Ego) treats other agents' (Others') failure to punish behavior classified by the institution as punishable, including behavior that has no impact on Ego's (or perhaps any agent's) payoff, as informative of the likelihood that Others will also fail to punish wrongful behavior that does have an impact on Ego's payoff. This belief structure creates an incentive for individuals to participate in collective punishment, potentially mitigating the free-rider problem in collective punishment. Ego has no incentive to participate in collective punishment

(the model assumes standard preferences) except to maintain an equilibrium in which behavior that reduces Ego's long-term payoff is deterred by the threat of collective punishment. Effectively, an agent's participation in punishment is treated as information about that agent's continued assessment that an equilibrium in which collective punishment is coordinated by the classifications articulated by the central institution will benefit that agent and hence that the agent will be willing to incur costs to support the equilibrium.

The classification institution in this model is serving as what Hadfield and Weingast [2012] call an authoritative steward of a very simple binary partition of behaviors into those that are punishable and those that are not. This can be interpreted as the construction of a sparse labeling system: lawful and not lawful.[2] We can interpret the beliefs underlying legal order as beliefs about whether people are punishers or not of behavior to which the constructed (that is, not natural) label "unlawful" is attached.

Our model here provides a basis for understanding how a simple binary classification scheme that is comprehensive–covering a wide variety of conduct–can emerge in a setting in which equilibrium depends on 1) voluntary participation in punishment and 2) a belief structure in which punishment of an action labeled punishable is considered informative about the likelihood of punishment of other actions labeled punishable, even when the assignment of the label is potentially arbitrary. We suggest that understanding how such a classification scheme and belief structure can emerge is critical for understanding the emergence of law.

The strategy of our paper is as follows. We first give an overview of the model and basic notation in Section 2, together with some technical background from the analysis of multi-armed bandit games and partially observed Markov decision processes. To build intuition, we then present in Section 3 analytical results for the limiting case in which Ego bears no cost of complying with spurious norms or punishing their violation. Because the games we analyze quickly becomes analytically complex but relatively easy to compute once we introduce a positive cost of complying with norms and punishing their violation, we turn to computational results in Section 4. Section 5 relates our results to conjectures about the likely growth and stability of communities in which norms are more or less pervasive and in which a legal institution that reduces ambiguity about norm violations and increases the informativeness of punishing behavior (by linking disparate rules into a code such that punishment of one rule is informative about the likelihood of punishment of another) exists.

2 Overview of Model

The basic idea of the model is as follows. Consider an infinitely repeated game setting in which an agent Ego is faced with the choice in each period of participating or sitting out. If choosing to sit out, Ego receives a payoff normalized to 0. If Ego chooses to participate, she plays a randomly selected game g with two randomly selected agents drawn from a population (Others). We model these games in reduced form. In each game, one of the Others is randomly selected and presented with an opportunity to choose between two actions, one which is classified by a classification institution L as "rule violation" and another which is classified as "not rule violation." If Other chooses "rule violation" the remaining Other and Ego each independently choose either to punish or not punish. Rule violations are deterred by collective punishment, that is, punishment that requires more than one agent to punish. For example, in Hadfield

[2]Cooter [1998] also suggests that the effectiveness of law can be understood as deriving from the classification of behavior as lawful or not. Cooter however presumes the existence of preferences based on this simple binary classification, that is, that at least some people inherently prefer to avoid actions labeled unlawful. This presumes that a category "law", which extends to potentially arbitrary actions, exists.

and Weingast [2012] two buyers and a seller engage in repeated contract and performance games. Actions for the seller are drawn from a set of possible contract performances, some of which are classified as breach and others which are not breach. A decision by a buyer not to purchase from a third-party seller in one period constitutes punishment, specifically a boycott. Breach is deterred when the seller expects both buyers to boycott in response to breach. Games are distinguished by the rule that may be violated. For example, there could be a game in which rule "watch where you are going" may be violated and one in which "men should not wear shorts in the street" may be violated.

We assume, and Ego believes, that the community of Others is playing an equilibrium in the super game that consists of the sequence of repeated games. Others are of two types, t: punishers ($t = 1$), who punish anyone who chooses an action classified by L as a rule violation, and non-punishers ($t = 0$), who never punish anyone. Let θ be the true proportion of punishers in the equilibrium. We assume that an Other's type is observable by the other participants in any particular game, that is, only in the context of the opportunity for rule violation. We focus on sub-game perfect equilibria in which the knowledge that two punishers are present deters rule violation.[3] (That is, on the off-equilibrium path where a violation does occur in the presence of an Other of type t, punishment is carried out that imposes costs on the violator that exceed the present value of benefits from violation.) We do not model how this equilibrium is generated or supported but we observe that the equilibrium is not destabilized by the presence of non-punishers. We assume, however, that Ego plays as a punisher, bearing an expected cost c in each round. c can be thought of as the cost to Ego of signaling that she is a punisher. For simplicity we assume that Ego is never presented with an opportunity for rule violation.[4] Ego's participation in the game is assumed to be on the margin, with no impact on the equilibrium played by the Others. Ego is able to observe rule violations, the types of Others and punishments in games in which she participates.

3 Formal Model Specification

Before providing a formalization of our model we provide a brief overview of the theory of Markov decision processes and multi-armed bandits.

3.1 Technical Background and Notation

We define a *Markov decision process* (MDP), M, is a tuple: $M = \langle \mathcal{S}, \mathcal{A}, P, R, \delta \rangle$[5]. $\mathcal{S}$ is a set of states. $\mathcal{A}$ is a set of actions. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a function that assigns probability to state transitions for each state-action pair. If Ego is in state, s, and selects action a the probability of transitioning to s' is given by $P(s, a, s')$. R is a (bounded) reward function that maps states, to an interval of $\mathbb{R}$, w.l.o.g., $R : \mathcal{S} \to [0, 1]$. $\delta \in [0, 1)$ is a discount factor that expresses Ego's preference for current versus future rewards.

[3] For an example of such a game, see Boyd et al. [2010]. They present an evolutionary game model in which punishment is a heritable strategy and deterrence requires multiple punishers. A population with a fraction of punishers can be stable in equilibrium when punishers can signal that they are punishers at low cost and thus avoid the costs of punishment if there are too few punishers present.

[4] It is straightforward to generalize our interpretation of c as the cost of complying with spurious rules to signal Ego's support for the equilibrium rules.

[5] In standard treatments, T is typically used for the transition distribution, we use P here to avoid confusion with our model specification.

A solution to M is a *policy*, π, that maps states to actions, $\pi : \mathcal{S} \rightarrow \mathcal{A}$. The *value* of a state, s, under π is the sum of expected discounted rewards received by starting in s and selecting actions according to π:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \delta^t R(s_t) | s_0 = s, \pi\right].$$

The optimal policy, π^*, maximizes this value and we will write $V = V^{\pi^*}$ for the optimal value function. Standard results show that a unique optimal value function exists[Puterman, 1994].

In a *partially observed Markov decision process* (POMDP), we additionally define a distribution over *observations* ($\mathcal{O}$) for a each state. A policy now maps a history of observations to an action, as the agent does not know the true state of the world. A POMDP can always be converted into a (continuous state) MDP where each state is a distribution over states of the world and the transition distribution is defined by Bayesian inference.

An interesting class of POMDPs are *multi-armed bandits* (MAB). In a multi-armed bandit, an agent is given access to several distributions. At each time step, that agent must selects a distribution to sample from, and receives a reward that is equal to the value of the sample.

A multi-armed bandit provides an analytically and computationally tractable model of exploration-exploitation trade-offs that occur with practical agents. In particular, success in this class of problems requires explicit reasoning about the impact of information on future decision making quality. Recent applicability to online banner advertising has led to rapid progress in both theoretical and computational methods for MABs [Auer et al., 2002]. The optimal full information policy (which knows the distributions of the arms) will always select the arm with the highest mean.

A key result from Lai and Robbins [1985] lower bounds the number of times an optimal (partial information) policy selects a suboptimal arm in expectation. This result holds for the class of consistent policies: policies where the probability that the optimal arm is chosen at time t approaches 1 as $t \rightarrow \infty$. We additionally require a smoothness constraint on the MAB distributions.

Proposition 1. *[Lai and Robbins, 1985] Let Θ be a class of MAB arms (i.e., a class of distributions) with parameter θ. Let $\mu(\theta)$ be the corresponding mean. If*

$$\forall \theta \in \Theta, \forall \delta > 0, \exists \theta' \neq \theta \text{ such that } \mu(\theta) \leq \mu(\theta') \leq \mu(\theta) + \delta$$

then, for any consistent policy the expected number of times a suboptimal arm is selected in the first n rounds is $\Omega(\log n)$[6].

3.2 Model Description

We define our super game as a tuple: $\langle G, T_\theta, \Pi, U, \delta, c \rangle$ where G is a distribution over games and T_θ is a distribution over punishment types t in the population of Others. We will abuse notation somewhat as use T and G to refer to the support of the corresponding distributions where the meaning is obvious. Π is Ego's prior distribution over the parameters of T_θ, and $U : G \times T_\theta \rightarrow \mathbb{R}$ is a mapping from types and games to immediate payoffs for Ego. The understanding is that this mapping represents the results of the Others playing their role in the equilibrium. δ is Ego's discount parameter for future rewards. c expresses a participation cost. This can be understood as the expected cost of to Ego of signaling to an Other that she is a punisher.

[6] If the function $g(n)$ is $\Omega(f(n))$, then there is a positive constant c such that $g(n) \geq cf(n) \forall n$

Ego begins in period 1 with perfect knowledge of how actions are classified by L, all payoffs and the distribution of games. Ego does not know the distribution of types of Others but holds a prior which we will specify shortly. Ego updates her beliefs about the distribution of types using Bayes' rule. The super game is defined as follows:

An initial game $g_0 \sim G$ is drawn. Then, for each period j:

1. Ego chooses whether to participate or not. If she opts out, then she collects 0 payoff and the next round starts.
2. If Ego opts in, she incurs the cost of signaling that she is a punisher c and a type, $t_j \sim T_\theta$ is for the remaining Other is drawn. If a non-punisher is drawn, a rule violation occurs; if a punisher is drawn, no rule violation occurs.
3. The agent observes whether a punisher is present and whether a rule violation occurs and collects payoff $U(g_j, t_j)$.
4. A game $g_{j+1} \sim G$ is drawn for the next round.

We assume that in equilibrium and given the ruleset created by L, there are two types of games from Ego's perspective: those to which Ego is indifferent and those that Ego cares about. Games to which Ego is indifferent always generate a reward of 0 for Ego. Suppose, for example, that a game involves a rule requiring genuflecting by an Other. We assume Ego realizes no costs or benefits from the Other's choice about whether to genuflect or not, other than the cost of signaling that she is a punisher. We will thus use represent the type of Other with an indicator variable that idicates if she is a punisher for this game.

Games Ego cares about are ones in which Ego receives a positive reward, R if there is another punisher present in the game and a negative reward $-R$ if there is not. We call these important games. We formalize the set of important games as follows:

$$G' = \{g \in G | U(g, \cdot) \neq 0\}$$
$$U(g, t | g \in G') = (2t - 1)R - c.$$

We will use $\mathbb{E}U = \mathbb{E}_{g,t}[U(g,t)|g \in G']$ to denote the expected utility of an important game. We let s denote the *sparsity* of the process generating games: the probability that a game is unimportant.

$$s = 1 - P(g \in G'); g \sim G.$$

Critically, we assume that the sparsity of games does not alter the (expected) rate at which important games are presented to Ego. To be concrete, we assume the expected discounted reward obtained from important games is independent of s. This condition can be attained through a suitably modification of δ as a function of s:

Proposition 2. *Setting*

$$\delta_s = 1 - (1 - s)(1 - \delta)$$

ensures that the expected sum of discounted rewards from important games is independent of s:

$$\forall s, \in [0, 1) \;\; \mathbb{E}_{g_j, t_j}\left[\sum_{j=0}^{\infty} \delta^j U(g_j, t_j) \middle| g_j \in G'\right] = \mathbb{E}_{g_j, t_j}\left[\sum_{j=0}^{\infty} \mathbb{I}[g_j \in G'] \delta_s^j U(g_j, t_j) \middle| s\right] \text{ [7]}.$$

[7] $\mathbb{I}[\psi]$ is the indicator function for the condition ψ.

Proof. We first show that it is sufficient to ensure that the expected value of δ_s^j is the same given that j is a round with an important game:

$$\begin{aligned}\mathbb{E}_{g_j,t_j}\left[\sum_{j=0}^{\infty}\mathbb{I}[g_j \in G']\delta_s^j U(g_j,t_j)\middle| s\right] &= \sum_{j=0}^{\infty}\mathbb{E}_{g_j,t_j}\left[\mathbb{I}[g_j \in G']\delta_s^j U(g_j,t_j)\middle| s\right] \\ &= \sum_{j=0}^{\infty}\delta_s^j \mathbb{E}_{g_j,t_j}[U(g_j,t_j)|s, g_j \in G']\mathbb{E}_{g_j}\left[\mathbb{I}[g_j \in G']|s\right] \\ &= (1-s)\mathbb{E}U\sum_{j=0}^{\infty}\delta_s^j\end{aligned}$$

Where the first line holds by the linearity of expectation, and the fact that g_j, t_j are independent iid draws from a stationary distribution. Substituting the form of the infinite geometric series, we see that

$$\frac{\mathbb{E}U}{1-\delta} = \frac{(1-s)\mathbb{E}U}{1-\delta_s} \tag{1}$$

is sufficient to acheive our goal. Substituting the form for δ_s in the theorem statement and reducing shows that this condition is satisfied. □

It can be easily shows that this model describes a class of MABs. If the parameters that describe equilibrium were known, the decision problem would be trivial. The safe option corresponds to a constant arm, which is a degenerate distribution. Optimal policies for bandits with constant arms exhibit clear structure: is that if it is optimal to choose a known option in round j, it will be optimal to choose the known option in round $j+1$ as well (Bradt et al., 1956). The argument is straighforward: if Ego reaches a point at which her estimate of the tradeoff between risking a negative payoff and learning so as to improve future decisions leads her optimally to choose not to participate, then her information state can never change and so her optimal choice can never be any different than the current opt-out decision. Thus, in our game, if Ego ever retires in a round, then she will never participate again. We refer to the decision not to participate at any point, then, as a decision to retire.

Futhermore, an MAB is an instance of a POMDP, so it the optimal policy maps a distribution over states, a *belief state*, to decision between retirement and particpation. We give our agent a Beta prior over this parameter so that the belief space for our agent is a two dimensional lattice equivalent to $\mathbb{Z}_+^2$. Initially, the belief state is (α_0, β_0) and can be understood as the state an agent would be in if she had seen α_0 punishers and β_0 non-punishers. The conditional probability that a punisher is present in the first game is

$$p_{\alpha\beta} = \frac{\alpha}{\alpha+\beta}.$$

Once the games begin, Ego updates the prior beliefs using Bayes' rule, which in the Beta distribution means adding the counts of punishers and non-punishers observed to the prior values. In the following, we will use $\alpha_i(\beta_i)$ to represent the number of observed punishers (non-punishers) prior to round i.

A second useful result from the theory of multi-armed bandits is that the optimal *policy* is a function that maps a sequence of observations to a decision about retirement. If we restrict to two types, punishers and non-punishers, this problem is a partially observed version of a Markov decision process where the state is the probability, p, of drawing a punisher. From the theory of partially observable Markov decision processes (POMDPs), this optimal policy can also be represented as a mapping from a distribution over p to an action about retirement [Puterman, 1994]. This reduces a partially observed process to a fully observed deterministic process in *belief space*.

4 The Value of Sparsity

Consider first the case in which the participation cost, c, is zero. In this case, Ego only faces a risky choice in periods in which she is presented with an important game. In all other periods, the per-period the expected payoff of playing the risky arm is a constant 0. Thus we can also conclude that if Ego retires, she will retire in a period in which she is playing an important game. In order to maintain the structure of a multi-armed bandit problem, we specify that if Ego chooses not to participate in an important game then the next game in the sequence is also an important game. We can think of this as a suspension of the game.[8] This rules out the possibility that Ego can simply choose not to play an important game and then reenter to play unimportant games in the hope of learning more before the next important game comes along. A decision not to participate is a decision to retire assuming a rational agent.

We let $i = 1$ represent the state in which there is a punisher present in the game. The value of a state is then characterized by the following recursion. To simplify notation we let p_{α_i,β_i} be the probability that a punisher will be present in the game in the belief state $(\alpha_i, \beta_i)$and we abuse notation somewhat by letting $V((\alpha_i, \beta_i); s, \delta)$ represent the discounted expected value of the super-game with sparsity s and discount factor δ_s in the state (α_i, β_i).

We now show a property of the *value of perfect information* (VPI) in our super-game. The VPI for a state in a decision processis a measurement of the improvement in decision making as a function of information gathering actions [**?**]. It is defined as the amount a rational agent is willing to pay to remove all uncertainty associated with a particular random variable.

Proposition 3. *If the participation cost, c, is 0, then, for any belief state, (α_i, β_i), and discount rate δ, the corresponding VPI goes to zero as the sparsity ratio goes to 1. That is*

$$\lim_{s\to 1} VPI((\alpha_i, \beta_i); s, \delta) = 0 \tag{2}$$

Proof. Given θ, it is easy to compute the value of participation:

$$V(\theta) = \mathbb{E}U \sum_{t=0}^{\infty} \delta^t = (2\theta - 1) \sum_{t=0}^{\infty} \delta^t = \frac{2\theta - 1}{1 - \delta}. \tag{3}$$

The optimal full information policy π_0 will retire whenever $V(\theta) < 0$. We use $V_+(\theta) = \max\{V(\theta), 0\}$ denote the value of π_0 as a function of θ. The VPI is computed as the difference between the expected value of V_+ and the value of the optimal policy that only uses the history of observations:

$$VPI((\alpha_i, \beta_i); s, \delta) = \mathbb{E}_\theta \left[V_+(\theta) | \, (\alpha_i, \beta_i) \right] - V^*((\alpha_i, \beta_i); s, \delta) \tag{4}$$

We proceed by lower bounding V. V is the value of the optimal policy so it is weakly lower bounded by any arbitrary policy. A useful candidate is one-step greedy policy, π_g, that always participates for unimportant games and retires in important games if the expected value of participation is negative (disregarding the benefit of new information). We let τ be the random number of games played before an important game is drawn. τ is geometrically distributed with success parameter $1 - s$. We let n_p be the random number of punishment actions observed prior to drawing an important game. The distribution over n_p will be a binomial distribution conditioned on τ. Thus, the value of

[8] In any multi-armed bandit game, the decision to stop suspends the game: deciding to return to the game implies making the risky pull that was rejected previously.

executing this policy ca be written as a joint expectation under τ and n_p:

$$V^{\pi_g}((\alpha_i, \beta_i); s, \delta) = \mathbb{E}_{\tau, n_p}\left[\max\left\{\mathbb{E}_{\theta'}\left[V(\theta')|(\alpha_i + n_p, \beta_i + \tau - n_p)\right], 0\right\}|(\alpha_i, \beta_i), s\right]. \tag{5}$$

We will be interested in the limit of this value, as $s \to 1$. Before proceeding with that, we note that, from the law of large numbers, we have $\mathbb{E}_{\theta'}[V(\theta')|a + n_p, b + \tau - n_p]$ will concentrate about $V(\theta)$. Thus,

$$\mathbb{E}_{\theta'}\left[V_+(\theta')|\left(\alpha_i, \beta_i\right)\right] = \lim_{\tau \to \infty} \mathbb{E}_{n_p}\left[\max\left\{\mathbb{E}_{\theta'}\left[V(\theta')|\left(\alpha_i + n, \beta_i + \tau - n_p\right)\right], 0\right\}|\tau, (\alpha_i, \beta_i)\right]. \tag{6}$$

Note that $\mathbb{E}[V(\theta)|(\alpha_i, \beta_i)]$ only depends on the ratio of (α_i, β_i), so the difference between the left and righthand sides of 6 is caused by the fact that the maximum must be taken at finitely many ratios (for finite τ). Furthermore these ratios are evenly spaced out, so the lefthand side can only increase as τ increases. We can use this to lower bound the limit of V^{π_g}:

$$\lim_{s \to 1} V^{\pi_g}((\alpha_i, \beta_i); s, \delta) = \lim_{s \to 1} \mathbb{E}_{\tau, n_p}\left[\max\left\{\mathbb{E}_{\theta'}\left[V(\theta')|(\alpha_i + n_p, \beta_i + \tau - n_p)\right], 0\right\}|s\right] \tag{7}$$

$$\geq \lim_{s \to 1} P(\tau \geq c(s)) \min_{\tau' \geq c(s)} \mathbb{E}_{n_p}\left[\max\left\{\mathbb{E}_{\theta'}\left[V(\theta')|(\alpha_i + n_p, \beta_i + \tau' - n_p)\right], 0\right\}|s, (\alpha_i, \beta_i), \tau'\right]$$

$$+ P(\tau \leq c(s)) \min_{\tau' < c(s)} \mathbb{E}_{n_p}\left[\max\left\{\mathbb{E}_{\theta'}\left[V(\theta')|(\alpha_i + n_p, \beta_i + \tau' - n_p)\right], 0\right\}|s, (\alpha_i, \beta_i), \tau'\right] \tag{8}$$

$$\geq \lim_{s \to 1} P(\tau \geq c(s)) \mathbb{E}_{n_p}\left[\max\left\{\mathbb{E}_{\theta'}\left[V(\theta')|\,\alpha_i + n_p, \beta_i + c(s) - n_p\right], 0\right\}|c(s), (\alpha_i, \beta_i)\right] \tag{9}$$

Using the form of the cumulative distribution of a geometric variable, $P(\tau \geq c(s)) = 1 - P(\tau < c(s)) = s^{c(s)}$. We set

$$c(s) = -\log 1 - s$$

so that $\lim_{s \to 1} c(s) = \infty$, and $\lim_{s \to 1} s^{c(s)} = 1$. Combining 6 and 9 with these facts allows us to deduce the following:

$$\lim_{s \to 1} V^{\pi_{gi}}((\alpha_i, \beta_i); s, \delta) \geq \mathbb{E}_\theta\left[V_+(\theta)|(\alpha_i, \beta_i)\right] \tag{10}$$

Thus, $\lim_{s \to 1} VPI((\alpha_i, \beta_i); s, \delta) \leq 0$. However, we have that, for any s, $VPI((\alpha_i, \beta_i); s, \delta) \geq 0$ by standard properties of VPI. This shows the result. □

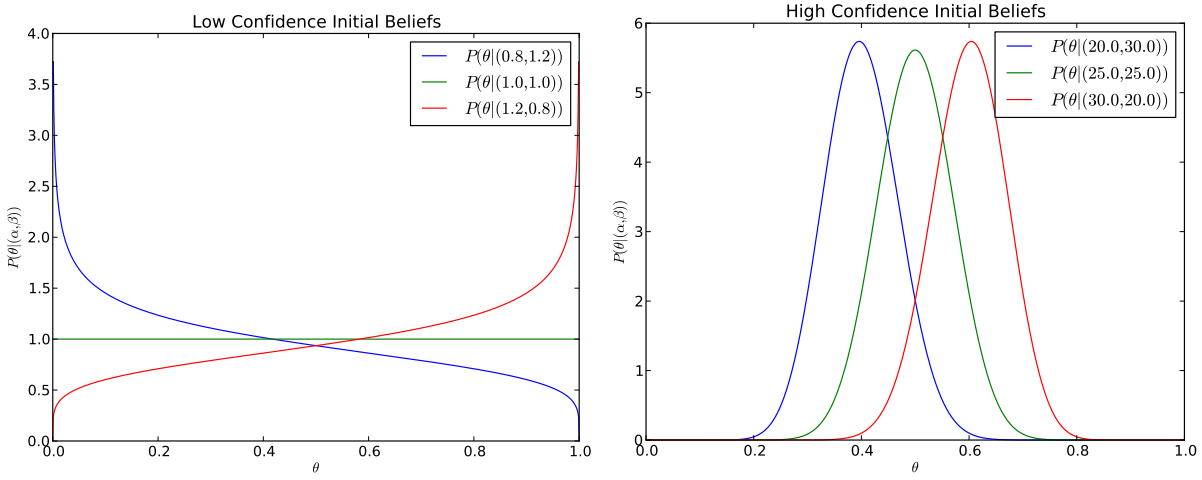
Proposition 4. *If the participation cost, c, is zero, then for any (α_i, β_i) such that $V_+(\frac{\alpha_i}{\alpha_i + \beta_i}) > 0$, VPI is strictly positive for $s = 0$.*

Proof. From Lai and Robbins [1985], we have that, for consistent and assymptotically efficient policies, the expected number of pulls of a suboptimal arm after n rounds is lower bounded by $c \log n$, where c is a positive constant that measures the similarity of the reward distributions for the arms. This class contains the optimal policy. Thus, we have that for any finite s,

$$VPI((\alpha_i, \beta_i); s, \delta) > 0. \tag{11}$$

□

The combination of these two propositions shows that for any (α_i, β_i), the corresponding value will eventually increase as s goes to 1. Thus, in the case where participation costs can be neglected, Ego will prefer an equilibrium with a higher s.



(a) Low Confidence Initial Belief Distributions

(b) High Confidence Initial Belief Distributions

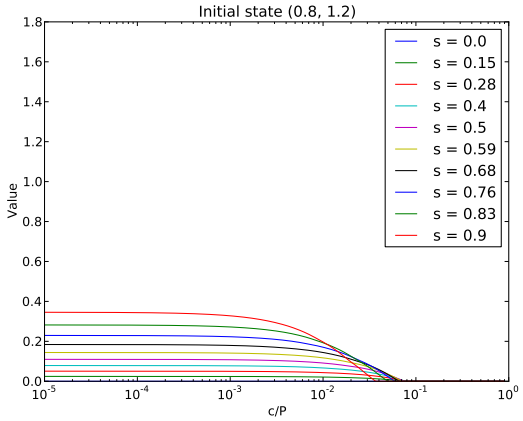
Figure 1: Plots of a selection of initial belief distributions. Each curve corresponds to a different value of $\mathbb{E}[\theta]$. (a) shows low confidence initial beliefs. In these information states, Ego has seen very little data and so her belief is very spread out. In these scenarios, we expect sparsity to be helpful because the gap between full information and partial information is large. Conversly, (b) shows belief distributions after Ego has seen 50 effective samples. The corresponding distribution is more concentrated, so we expect less positive impact from sparsity.

5 The Cost of Pervasive Normativity: Computational Results

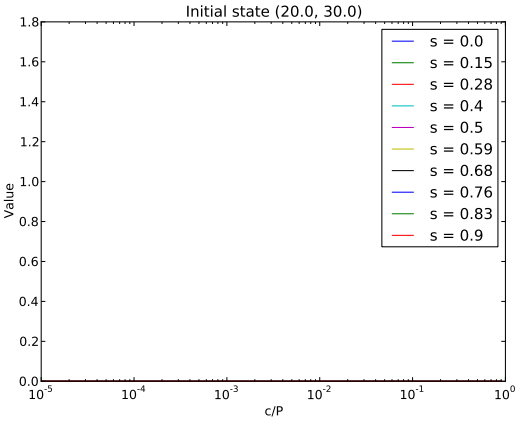
Our results above show that an environment with lots of rules that an agent cares nothing about intrinsically is more valuable for an agent contemplating participation than one in which the only rules are ones that matter on the merits–altering Ego's payoff directly. We assumed, however, that increasing the number of spurious rules is costless to Ego and of course this is not generally likely to be true. If Ego is going to participate in a community with lots of spurious rules, Ego is also likely to bear costs, specifically the cost of participating in collective punishment and the cost of complying with spurious rules. In this section, we relax the assumption that $c = 0$. Doing so, however, increases the analytical complexity. We therefore turn to computational methods to explore environments in which Ego enjoys both costs and benefits from an increase in the number of spurious rules.

To illustrate the effect of participation costs, we select six initial belief states and computed values as a function of c. Our initial beliefs vary the expected value of θ and the variance of the belief about its mean. We selected initial states to cover scenarios where the expected values of an important game is negative, positive and equal to zero. In this work, we chose $\mathbb{E}[\theta] \in \{.4, .5, .6\}$. We varied Ego's confidence in her current estimate by varying the effective number of samples $(\alpha + \beta)$ in the initial belief. Figure 1 shows the corresponding beta distributions for our selection of initial states.

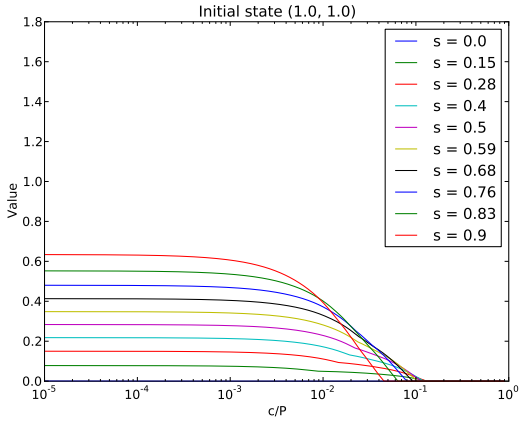
The optimal policy is invariant to scaling of rewards, so we fix $P = 1$ and let $\frac{c}{P}$ be the independent variable in our computations. We compute these values with a variant of value iteration that takes advantage of the structure of the state space. A python script to generate these plots is included as appendix A. We set the parameters of our computation to allow for at most 10^{-8} of error in the computation.



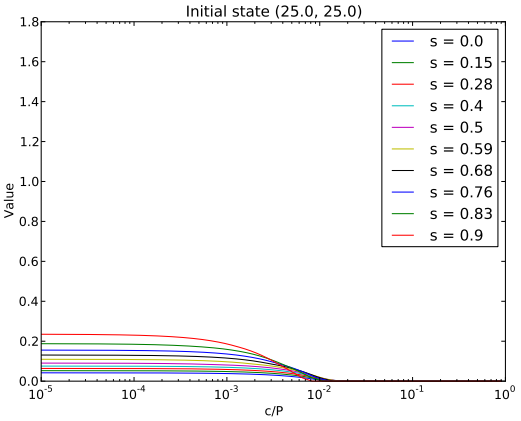
(a) Low Mean, Low Confidence



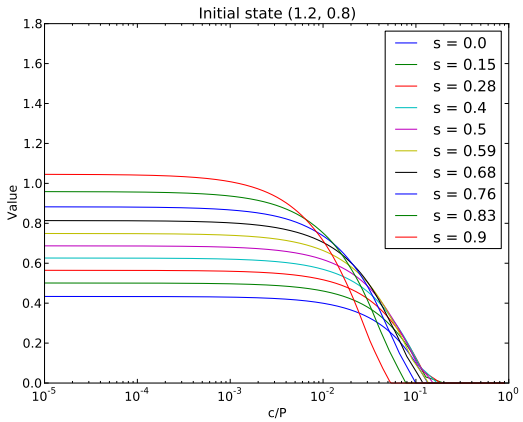
(b) Low Mean, High Confidence



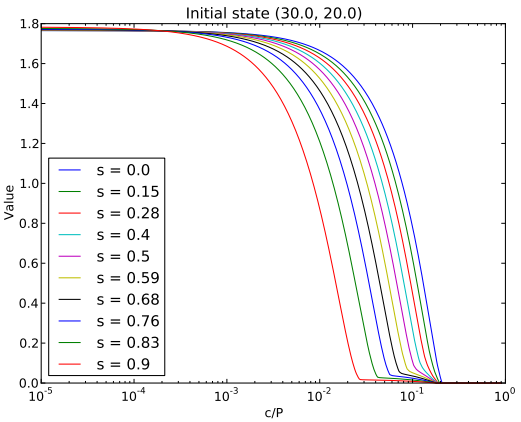
(c) Mid Mean, Low Confidence



(d) Mid Mean, High Confidence



(e) High Mean, Low Confidence



(f) High Mean, High Confidence

Figure 2: Plots of $V((\alpha_i, \beta_i); s, .9)$ for select initial states. The probability density functions that correspond to these beliefs are shown in Fig. 1. The one step value of participating in an equilibrium with enforcement is 1 ($P = 1$).The increase in value for larer sparsity for low values of $\frac{c}{P}$ shows confirmation of the results in Prop. 3. In comparing vertically, we can see that sparsity has a larger effects on states with low mean and low confidence: (f) is essentially unimpacted by sparsity for low $\frac{c}{P}$ while for (a) and (c) participation is suboptimal unless sparsity is positive. The cost of this sparsity is increased sensitivity to participation costs.

Figure 2 shows value as a function of $\frac{c}{P}$ for several pairs of sparsity and initial states. We can clearly see the costs of pervasive normativity: value functions at higher sparsity decreases more quickly and for lower setting of the participation costs. This occurs because the number of rounds per important game increases so more participation costs are paid. In essence, increased participation costs force Ego to pay more for information and so she may prefer an equilibrium with less sparsity.

As we might expect, the net gain in value from sparsity is larger in belief states with that are very uncertain. For example, the value function for initial state (30, 20) is essentially invariant to sparsity for low $\frac{c}{P}$. Similarly, gains from increased sparsity are more pronouced with low means. Essentially all value in those states is due to improved decision making abilities from information.

6 Implications for the Microfoundations of Law

Our results have surprisingly powerful implications for the structure of normativity in human communities. [LETheory readers: we are only sketching these implications for this draft. We anticipate that at least some of them will be capable of more formal analytic or computational demonstration.]

6.1 Gossip, Silly Rules and Durability

The value of participating in a super-game depends on the expected cost of punishing rule violations relative to the reward Ego expects if violations of rules she cares about are deterred. The computations in Section 4 indicate that when expected costs relative to rewards are sufficiently low, *ceteris paribus*, Ego enjoys higher value in environments with more rules that she does not care about. This provides an interesting explanation for the observation from ethnographic studies that simple societies are characterized both by pervasive and apparently spurious rules that are effectively enforced by low-cost collective punishments. Wiessner [2005], for example documents the use of gossip, group criticism and mocking as the principal means by which norms are enforced among the Ju/'hoansi Bushmen of northwest Botswana. In her observations, violations of norms were punished by escalating criticism and rarely got to the point of physical violence. Assuming that the rewards generated for individuals aggregate to raise group well-being, our model thus can be read to predict that communities that succeed in securing an equilibrium with many rules that impose low compliance costs and which are enforced by low-cost means will outperform communities with fewer rules and more costly forms of punishment.[9]

6.2 Birds of a Feather

The rewards Ego enjoys when joining a community depend on the rules of that community. In comparing across communities with comparable forms of punishment, and comparable numbers of rules, Ego will prefer a community with rules for important games that, when effectively enforced, generate higher rewards, R. We have not modeled the source of rules in a community but if we suppose that rules emerge that reflect the interests of the members of a community, this suggests that Ego is more likely to find rules that achieve higher rewards in communities with a number of agents with similar preferences, that is in more homogeneous communities.

[9] Our model takes into account Ego's willingness to bear the higher cost of punishment in more sparse environments. We assume, but have not shown, that the willingness of Others to punish is not reduced with sparsity–that is, that Others have incentives comparable to Ego's.

6.3 The Emergence of Law

As communities grow more heterogeneous–which they naturally do as they generate value and support greater specialization through the division of labor–we expect ambiguity about rule violations to increase. Increasing ambiguity increases Ego's expected cost of participating in a punishment regime if errors in punishment–failing to punish when a violation is perceived by an Other and punishing when the Other does not perceive a violation–are themselves treated as rule violations which produce punishment. This is a feature of the punishment schemes of many communities: members of the Flemish cloth merchants guild in the 13th Century, for example, included in their rules the provision that any member who failed to observe a boycott of a buyer who had cheated another guild member was barred from trading and no guild member was to "house the goods" or "keep company" with the non-punisher. Increasing ambiguity also reduces the expected rewards Ego enjoys in important games because with some probability a violation perceived by Ego is not perceived as such by Others and hence the probability of a violation increases. Thus the cost of punishing relative to rewards decreases with decreasing ambiguity. Hadfield and Weingast [2012] argue that the central function of law is to serve as a unique classification institution, capable of resolving ambiguity about what counts as a rule violation. Moreover, they show that for a classification institution to effectively secure an equilibrium of legal order around a given set of rules enforced only by decentralized collective punishment, the institution must possess legal attributes such as neutrality, openness, clarity, consistency and stability. Our analysis here predicts that communities that introduce law in the form of a classification institution with legal attributes that reduces ambiguity will enjoy higher value and greater durability.

6.4 Hammurabi's Code

One of the key assumptions of our model is that people who punish any rule violation are expected to punish all rule violations. This is a distinctive feature of the labeling system generated by a legal regime: people are "law breakers" or not; they are "law-abiding" or not. Cooter [1998] proposes that "law" is a meaningful category and that people have preferences over behaviors solely on the basis of whether they are labeled "lawful" or not. Our model captures this idea by treating observation of punishment behavior in the context of any rule as informative of the probability of punishment in important games. The model can be interpreted as representing, for example, a community in which legal order is coordinated around a single legal code. Hammurabi's Code from ancient Babylon, for example, consisted of 247 individual rules, such as "If any one hire an ox or an ass, and a lion kill it in the field, the loss is upon its owner" (Rule 244) and "If any one open his ditches to water his crop, but is careless, and the water flood the field of his neighbor, then he shall pay his neighbor corn for his loss" (Rule 55). These rules likely emerged individually over time. We can imagine that knowing whether someone punished Rule 244 may or may not have helped to predict whether they would also punish Rule 55. But when Hammurabi placed all 247 together on a stone pillar and named the collection as his Code, he created the possibility for the emergence of two types of people: those who punished violations of the Code and those that did not. Our analysis suggests that the creation of collections of rules, rather than disparate rules, can generate value. Suppose, for example, that Ego cares about five rules, enjoying rewards when violations of each of them is deterred. Our model treats these five rules as integrated into a single super game in which the observation of punishment behavior in any game is informative, and equally so, of the likelihood of deterrence of violations in any of the five games Ego cares about. But suppose instead that these rules are not connected in this way. Suppose that punishment behavior in each game Ego cares about is only predicted by punishment behavior in non-overlapping subsets of unimportant games. We could then decompose our single super-game into five distinct super-games, each one of which would be considerably less sparse than our original game. Our results predict that

Ego's value in a community with distinct super-games–with disconnected rules and a belief structure about punishment that limits the informativeness of observing punishment of any individual rule–will be lower than the value enjoyed in a community with a comprehensive code.

6.5 Legal Pluralism and the Nation State

Finally, we can combine some of the above observations to shed light on the phenomena of legal pluralism and the emergence of the nation state. Modern advanced legal systems are characterized by comprehensive systems of rules, with the label 'law' attributed to any rule generated by a government organized within the constitutional framework of a state. Indeed, law is frequently equated with the rules generated by a government, as distinguished from those generated from other entities such as corporations or schools or that emerge organically from social interaction [Ellickson, 1991]. But as Hadfield and Weingast [2013] emphasize, prior to the emergence of the nation state, there were many institutions that coordinated legal order in different spheres, often in competition. Medieval Europe, for example, was characterized by multiple legal orders, with rules generated by merchant guilds, towns, churches, local rulers and more. Many societies, particularly those seen as struggling to establish the rule of law, are characterized by multiple legal orders–some governing family relations, others governing commercial dealings for example. Our model suggests a way of thinking about the tradeoffs Ego will face between participating in multiple legal communities, each of which coordinates punishment over some subset of rules and participating in a comprehensive legal community. On the one hand, rewards may be higher when a legal community is comprised of a relatively homogeneous group with shared interests–such as a community of traders–who can select rules that serve Ego's interests. But such a community will also have lower sparsity. On the other hand a system with a single system of comprehensive rules may achieve less alignment with Ego's interests but may also, because of its higher sparsity, provide Ego with more information about the value of continuing to participate.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Robert Boyd and Peter J Richerson. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology*, 13(3):171–195, 1992.

Robert Boyd, Herbert Gintis, and Samuel Bowles. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, 328(5978):617–620, 2010.

Robert Cooter. Expressive law and economics. *The Journal of Legal Studies*, 27(S2):585–607, 1998.

RC Ellickson. Order without law: how neighbors settle disputes. 1991.

Daniel MT Fessler. Toward an understanding of the universality of second order emotions. *Beyond nature or nurture: Biocultural approaches to the emotions*, pages 75–116, 1999.

Daniel MT Fessler and Carlos David Navarrete. Meat is good to taboo. *Journal of Cognition and Culture*, 3(1):1–40, 2003.

Gillian K Hadfield. A coordination model of the sexual division of labor. *Journal of Economic Behavior & Organization*, 40(2):125–153, 1999.

Gillian K Hadfield and Barry R Weingast. What is law? a coordination model of the characteristics of legal order. *Journal of Legal Analysis*, 4(2):471–514, 2012.

Gillian K Hadfield and Barry R Weingast. Law without the state: legal attributes and the coordination of decentralized collective punishment. *Journal of Law and Courts*, 1:3–34, 2013.

Gillian K Hadfield and Barry R Weingast. Constitutions as coordinating devices. *Institutions, Property Rights, and Economic Growth: The Legacy of Douglass North*, page 121, 2014.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

S Mathew, R Boyd, and M van Veelen. Human cooperation among kin and close associates may require enforcement of norms by third parties. In *PJ Richerson and M. Christiansen. Strüngmann Forum Report*, volume 12, 2012.

Richard H McAdams. Expressive power of adjudication, the. *U. Ill. L. Rev.*, page 1043, 2005.

Roger B Myerson. Justice, institutions, and multiple equilibria. *Chi. J. Int'l L.*, 5:91, 2004.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.

R Sugden. The economics of righs, cooperation, and welfare. Palgrave Macmillam, 1986.

Polly Wiessner. Norm enforcement among the ju/hoansi bushmen. *Human Nature*, 16(2):115–145, 2005.

Appendix A: Computing $V((\alpha_i, \beta_i); s, c, \delta)$

```
"""
File: sparse_bernoulli.py

Usage:
$ python sparse_bernoulli.py n_sparsity_values x_fidelity
                             [--resultfolder path/to/results]

Ex: to compute with 10 sparsity values with 500 samples along the x axis and
store the results in folder path/to/figures/ do
$ python sparse_bernoulli.py 10 500 --resultfolder path/to/figures

Not specifying result_folder defaults to the current folder
"""
from __future__ import division

import numpy as np
import matplotlib.pyplot as plt
import argparse
import sys

eps = 1e-8
```

```
def main():
    parser = argparse.ArgumentParser()
    parser.add_argument('n_sparsity_values', type=int)
    parser.add_argument('x_fidelity', type=int)
    parser.add_argument('--resultfolder', type=str, default="")
    args = parser.parse_args()

    # These are the values we'll use to compute cost response curves
    sparsity_values = np.log(np.linspace(np.exp(0),
                                         np.exp(.9),
                                         args.n_sparsity_values))

    # The x-axis values, determines the accuracty of the plots. Uses
    # log linear spaces because that seems qualitatively better responses
    costs                   = np.exp(np.linspace(np.log(0.00001),
                                                 np.log(.9),
                                                 args.x_fidelity))
    max_ret_sparsity_values = np.exp(np.linspace(np.log(0.001),
                                                 np.log(.95),
                                                 args.x_fidelity))

    start_values = {(.4, 2) : 'low-mu-high-sigma',
                    (.4, 50): 'low-mu-low-sigma',
                    (.5, 2) : 'mid-mu-high-sigma',
                    (.5, 50): 'mid-mu-low-sigma',
                    (.6, 2) : 'high-mu-high-sigma',
                    (.6, 50): 'high-mu-low-sigma'}

    resultfolder = args.resultfolder

    for ratio, effective_samples in start_values:
        # compute the costs for different sparsity values
        alpha = ratio * effective_samples
        beta = (1-ratio) * effective_samples

        fig = plt.figure()
        ax = fig.add_subplot(1, 1, 1)
        ax.set_xscale('log')
        ax.set_xlabel('c/P')
        ax.set_ylabel('Value')

        for s in sparsity_values:
```

```
            print "s: {}, ratio: {}, effective samples: {}".format(
                        s, ratio, effective_samples)
            cost_response_curve = compute_cost_response_curve(
                        alpha, beta, s, costs)
            ax.plot(costs, cost_response_curve,
             label="s = {:.2}".format(s))

        plt.title('Initial state ({}, {})'.format(alpha, beta))
        ax.set_ylim([0, 1.8])
        plt.legend(loc='best')
        plt.savefig(
            resultfolder+start_values[(ratio, effective_samples)] + ".pdf")

    fig = plt.figure()
    ax = fig.add_subplot(1, 1, 1)
    ax.set_xlabel('s')
    # ax.set_xscale('log')
    ax.set_xlim([.001, 1])
    ax.set_ylabel(r'$\frac{c}{P}$')
    plt.title("Maximal Participation Cost vs Sparsity")

    for ratio, effective_samples in start_values:
        print "ratio: {}, effective samples: {}".format(ratio, effective_samples)
        # compute the costs for different sparsity values
        alpha = ratio * effective_samples
        beta = (1-ratio) * effective_samples

        #compute the largest c such that participation is optimal
        max_c_values = []
        for s in max_ret_sparsity_values:
            max_c_values.append(largest_possible_c((alpha, beta), s, 0.9))
        print "s = 0, max_c = {}".format(max_c_values[0])
        best_s = np.argmax(max_c_values)
        print "s = {}, max_c = {}".format(
            max_ret_sparsity_values[best_s], max_c_values[best_s])
        ax.plot(
            np.r_[max_ret_sparsity_values, [1]],
            max_c_values + [0], label="({}, {})".format(alpha, beta))

    plt.legend(loc='best')
    plt.savefig(resultfolder + "retirement_points.pdf")

def largest_possible_c(s0, s, delta):
```

```
    c_min, c_max = (0, 1)
    while np.abs(c_max - c_min) > eps:
        c = (c_min + c_max)/2.0
        V = sparse_bernoulli_value_iteration(s0, s, c, delta)
        if V > 0:
            c_min = c
        else:
            c_max = c
    return c_min

def sparse_bernoulli_value_iteration((a, b), s, c, delta,
                                     tol=eps, verbose=true):
    """
    Takes a belief state (a, b) and computes the V((a, b); s, c, delta)

    Computation is done with value iteration so that the error is less
    than tol
    """
    delta_s = 1 - (1-s)*(1-delta) # As is Prop 2

    """
    Values are initialized to 0, so maximal the maximal error is the
    maximal positive reward for all time. With probability (1 - s) Ego
    gets value P with probability \theta. We upper bound by letting
    \theta = 1 and then normalize by P to get an upper bound:

         UB(c, s) = (1-s -c/P)/(1-delta_s)

    This ammount decreases by delta_s each step of value iteration so
    we need delta_s^H UB(c,s) <= tol ===> H >= log(tol/UB(c, s)) / log(delta_s)
    """
    log_V_ub = np.log(1 - s - c) - np.log(1 - delta_s)
    H_lb = ( np.log(tol) - log_V_ub ) / np.log(delta_s)
    H = int(np.ceil(H_lb))

    if verbose:
        sys.stdout.write('\r s: {} c: {} H: {}                        '.format(
                s, c, H))
        sys.stdout.flush()

    # Vector of a counts
    a_vals = np.linspace(0, H-1, H) + a
```

```
    # Allocate vectors to store values
    Vt = np.zeros(H)
    Vt_minus1 = np.zeros(H-1)

    # Take the horizion from H-1 to 0
    for t in range(H-1, 0, -1):
        # after cur_h rounds we will have seen cur_h heads or tails,
        # and we incorporate the priors
        cur_confidence = cur_h + a + b

        # P[i] = i / Nt; i in [0,...,cur_h]
        P = a_vals[:cur_h] / (cur_confidence)

        # do a value iteration backup
        Vt_minus1 = backup(Vt, Vt_minus1, P, s, delta_s, c)

        # set up for the next round, reuse the preallocated memory
        # to avoid unnnecessary realloc calls
        tmp = Vt
        Vt = Vt_minus1
        Vt = tmp[:-1] # decrease size by 1
    return Vt[0]

def backup(Vt, Vt_minus1, P, s, delta, c):
    """
    Computes a value iteration back for the super game

    V: vector of values at time t, in increasing order of the number of heads
    Vt_minus1: vector to return values for time t-1 (avoids reallocating memory)
    P: vector of transition proabilities encoding probability of heads at time t-1
    s: sparsity level
    delta: discount factor
    c: participation costs
    """
    # First compute expected value of important game
    # Vt_minus1[i] = delta * (P(tails) * Vt[i] + P(heads) * Vt[i+1])
    Vt_minus1 = delta * (Vt[:-1]*(1-P) + Vt[1:]*(P))
    # Expected reward at next step is 2*\theta - 1 - c
    Vt_minus1 += 2*P - 1 - c
    # Ego decides whether or not to retire
    # After this line Vt_minus1 = P(important game) * E[Rt + Vt| important game]
    Vt_minus1 = (1-s)*np.maximum(Vt_minus1, 0)
    # Same as before with different rewards but its repeated
```

```
    # and we don't want to allocate extra space
    Vt_minus1 += s*np.maximum(delta*(Vt[:-1]*(1-P) + Vt[1:]*P) - c, 0)
    return Vt_minus1

def compute_val(alpha, beta, s, c_vals, delta=0.9):
    """
    computes [V((alpha, beta); s, c, delta) for c in c_vals]
    """
    vals = []
    for c in c_vals:
    vals.append(sparse_bernoulli_value_iteration((alpha, beta), s, c, delta))
    return np.asarray(vals)

if __name__=='__main__':
    main()
```