Pervasive Spurious Normativity or: The Case for Lots of Silly Rules

Gillian K. Hadfield **Dylan Hadfield-Menell**

Abstract

Human societies are awash with rules, with *normative moralization* [Fessler and Navarrete, 2003] attaching to almost everything we do: how we eat, talk, dress, work, uses resource, treat others and so on. Most analysis of the normative world seeks first-order functional explanations for norms. Evolutionary analysis proposes that specific norms survive because they promote fitness, for example. Economic analysis has generally understood norms as value-enhancing solutions to coordination or cooperation games. While there is no doubt that many norms are functional, many are also inherently spurious, with no direct impact on material well-being. The pervasiveness of spurious norms suggests there is a deeper puzzle to solve in explaining the phenomenon of normativity in human societies than only accounting for the content of specific norms. In this paper we identify an important role for the extension of normativity to actions that have little or no direct impact on welfare. We do this by modeling an individual's choice about whether to join a particular community with a known set of rules as a multi-armed bandit problem. Using both analytical and computational methods, we show that a community with pervasive spurious normativity–lots of silly rules–generates higher payoffs for a potential member than a community that restricts normativity to actions with direct benefits for the agent–a few important rules. It does so under the assumption that in the community's equilibrium, someone who punishes any violation of a rule in the community's rule set punishes all violations of rules in the community's rule set. This makes an observation of punishment behavior informative, regardless of the particular violation observed. The community with lots of silly rules then achieves a higher value for an individual by providing the individual with plentiful opportunities to learn about the likelihood that important rules will be punished by other agents in the community. In addition to providing an account of the pervasive spurious normativity that characterizes many non-legal settings, our analytical approach, by providing a framework for evaluating normative systems with arbitrary content, also contributes to the development of the microfoundational account of law introduced by Hadfield and Weingast [2012].

1 Introduction

One of the central attributes that differentiates human from other animal societies and accounts for the enormous gains of human ultra-sociality [Richerson and Boyd, 2008] is the presence of third-party enforced norms [Riedl et al., 2012, Tomasello and Vaish, 2013, Buckholtz and Marois, 2012]. Many of these norms generate direct benefits for individual and group well-being: norms that prescribe reciprocity, fair sharing of rewards or non-interference with property properly claimed by another, for example, can coordinate behavior and sustain incentives for cooperation and investment. These are the norms that are the primary focus of most research into the properties and origins of human normativity (see Tomasello and Vaish [2013] for a review.) Most analyses of norms in the law and economics literature focus on the role that norms play in coordinating 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).

The normative landscape is also, however, populated by many norms that are spurious: "silly" rules about how and what we eat, how we greet each other, what clothes and body decorations we wear, and what rituals we observe. People treat compliance with these norms as important and punish violations, but, except for effects generated by this socially-constructed salience, they have no direct or first-order impact on welfare. 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 naturally right-handed but, particularly in societies with few specialized tools, whether someone is right- or left-handed generally 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 for opprobrium–revealing weakness or evil.

The ubiquity of spurious norms is a puzzle for functionalist accounts of norms. Spurious norms may exist to serve as cheap signals of group membership and thus support group-based reciprocity [McElreath et al., 2003]. Compliance with spurious norms may also help reduce errors in social learning and the transmission within groups of subtle and valuable knowledge via authentic norms [Richerson and Boyd, 2008], in what may be a form of prescriptive over-imitation [Kenward et al., 2012]. But these accounts seem to fall short in explaining the vast quantity of spurious norms we find in most societies. It would seem that a society would do better to minimize costly efforts to punish and conform with norms that produce no material benefits, and so to economize on the number of spurious norms used as markers or retained as a by-product of the cultural transmission of knowledge. The sheer abundance of spurious norms seems to require an account that grants the normative moralization of seemingly irrelevant actions a more significant role in the management of complex environments.

Fessler [1999] suggests one such explanation. He hypothesizes that culture extends the set of actions that are subject to normative moralization as a way of enlarging 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] puts this,

> 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 even spurious norms and is not an account of the emergence and persistence of spurious norms *per se*. But the insight that conformity with spurious norms can be informative for cooperation clearly suggests that retaining spurious norms can improve cooperation.

In this paper we explore a different, but related, account of the value of spurious normativity, focusing on settings in which norms are enforced by third-party collective punishment. (As Boyd and Richerson [1992] and many others have emphasized, 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.) Based on a framework developed to analyze legal order by Hadfield and Weingast [2012] we construct a multi-armed bandit model that analyzes an individual's decision about whether to join a normative community or not. We then experiment computationally with this model and demonstrate that communities that extend normativity to otherwise irrelevant actions 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. The willingness of an individual (whom we will denote Ego) to cooperate in a community–which requires foregoing safe non-cooperative options and exposing oneself to the risk of being exploited–depends on Ego's beliefs about the likelihood that the community (Others) effectively punishes actions that harm 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. Similarly, you will care about whether the community effectively punishes actions that you do not want to see punished. (Hadfield and Weingast [2012] call this *sufficient convergence* between an individual's idiosyncratic normative classification scheme and the common scheme used to coordinate third-party enforcement.) Assume you are a newcomer to a community with a known set of announced norms. This could be a sub-set of a society, such as a club or trading community, as well as a community to which one physically migrates. 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. (See Bicchieri [2006] for a definition of a social norm that does not require that a norm be observed to exist.) Assume also that the only way reliably to learn about the likelihood of punishment is to observe punishment behavior and that an individual who punishes (or not) any norm violation punishes (or not) all norm violations. That is, punishers punish not violation of "a rule" but rather violations of "the rules." You can learn the likelihood that violations will be punished more cheaply if you are given abundant opportunities to observe what happens when there are violations if the violations to which you have to expose yourself 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 result is related to the suggestion of DeScioli and Kurzban [2013] that the phenomenon of moralization, even of arbitrary behaviors, can be explained by the need for individuals to decide which side to take in a dispute. They argue that other mechanisms–siding with a high status individual or with those with whom one has a pre-existing relationship–are costly, either because they empower individuals who can then become despotic or, because individuals have non-overlapping sets of relationships, they fail to achieve coordination and may cause escalation. Impersonal moral norms, on the other hand, apply to individuals regardless of identity (a point also emphasized by the coordination account of law in Hadfield and Weingast [2012]) and so observations of behavior relative to moral norms can serve as a signal that coordinates side-taking and contain the cost of disputes. Importantly, DeScioli and Kurzban [2013]'s account rests, as does ours, on the informative impact of participating in what they call moral condemnation and we call punishment, in a setting where multiple participants in punishment are required (in their model to contain dispute costs, in ours to make punishment effective.) In experimental work, Kurzban et al. [2007] show that individuals are more likely to engage in moralistic punishment when the decision to punish is public. DeScioli and Kurzban [2013] propose that public punishment actions signal to others what side an individual is on; Hadfield and Weingast [2012] propose that public punishment actions signal an individual's ongoing willingness to support a particular normative classification institution (an entity that partitions actions into those that are punishable and those that are not.) Our model here introduces the idea that punishment actions also provide information to potential members of a community (necessary to grow the size of a community) about the extent to which announced norms are in fact enforced by the community.

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 become 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. We also consider in this section the implications of our model for the development of codes, meaning sets of norms defined by the property that drives our results: a person who punishes violation of a particular norm in the set can be expected to punish violation of any norm in the set.

2 Overview of Model

The basic idea of the model is based on a framework developed in Hadfield and Weingast [2012]. 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 that is classified by a classification institution L as "rule violation" and another that 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 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.[1] (That is, on the off-equilibrium path where a violation does occur in the presence of an Other of type t, punishment is

[1] 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.

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. [2] 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, as a tuple: $M = \langle \mathcal{S}, \mathcal{A}, P, R, \delta \rangle$[3]. $\mathcal{S}$ 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.

A solution to M is a *policy*, π, that maps states to actions, $\pi : \mathcal{S} \to \mathcal{A}$. 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.

[2] It is straightforward to generalize our interpretation of c as the cost of complying with spurious rules–self-punishment–to signal Ego's support for the equilibrium rules. Thus the assumption that Ego plays as a punisher is equivalent to an assumption that Ego always complies with spurious rules, which we define shortly.

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

A multi-armed bandit provides an analytically and computationally tractable model of exploration-exploitation tradeoffs 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 $t \to \infty$. 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 $\Omega(\log n)$[4].

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 and 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 $U : G \times T_\theta \to \mathbb{R}$ 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.

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.

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

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.

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.

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

Note that a super-game has high sparsity (s close to 1) when important games are a small fraction of the total number of games.

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 suitable 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[5]:

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

[5] $\mathbb{I}[\psi]$ 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 achieve 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: 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 [Gittins et al., 2011]. 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 the optimal policy maps a distribution over states, a *belief state*, to a 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 for 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 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.[6] This rules out the possibility that Ego can simply choose not to play an important game and then re-enter 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 process is a measurement of the improvement in decision making as a function of information gathering actions [Russell et al., 1995]. 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 sparsity 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 instruct the agent to retire whenever $V(\theta) < 0$. We use $V_+(\theta) = \max\{V(\theta), 0\}$ to 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

[6] 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.

history of observations:

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

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 s. We let n_p be the random number of punishers observed prior to drawing an important game. The distribution over n_p will be a binomial distribution conditioned on τ. Thus, the value of executing this policy can 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 that $\mathbb{E}_{\theta'}[V(\theta')|a + n_p, b + \tau - n_p]$ 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 right-hand 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 left-hand 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, that is, a lot of silly rules.

5 The Cost of Pervasive Normativity: Computational Results

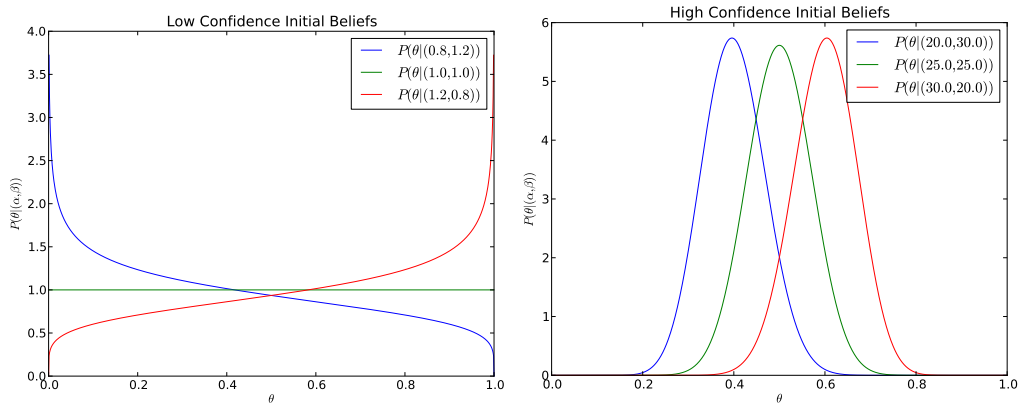
Our results above show that an environment with lots of rules that an agent cares nothing about intrinsically is more valuable for an agent contemplating participation than one in which the only rules are ones that matter on the merits–altering Ego's payoff directly. We assumed, however, that increasing 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 compute 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 reward in an important game is negative, positive and equal to zero. In this work, we chose $\mathbb{E}[\theta] \in \{.4, .5, .6\}$.[7] 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 six initial states.

The optimal policy is invariant to scaling of rewards, so we normalize $R = 1$. Then the cost of participation c can be interpreted generally as $\frac{c}{R}$, the relative cost of participation. $\frac{c}{R}$ is 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.

Figure 2 shows value as a function of $\frac{c}{R}$ for our six selected initial states. We can clearly see the costs of pervasive normativity: value functions at higher sparsity decrease more quickly as participation costs increase. This occurs because the number of rounds per important game increases so more participation costs are paid. Increased participation costs force Ego to pay more for information.

[7] The expected reward is $-.2R$ when $\theta = .4$; 0 when $\theta = .5$ and $.2R$ when $\theta = .6$.

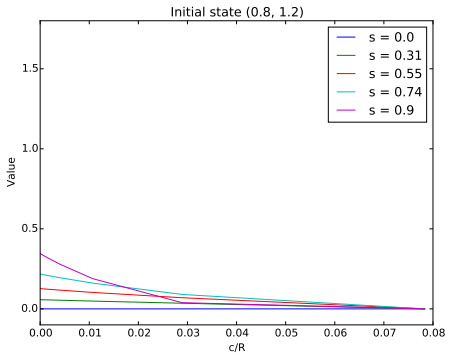


(a) Low Confidence Initial Belief Distributions (b) High Confidence Initial Belief Distributions

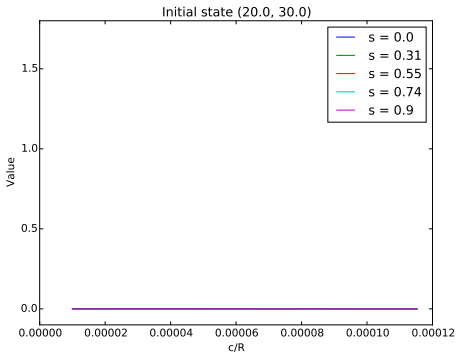
Figure 1: Plots of a selection of initial belief distributions. Each curve corresponds to a different value of $\mathbb{E}[\theta]$. (a) shows low confidence initial beliefs. In these information states, Ego has seen very little data and so her belief is very spread out. In these scenarios, we expect sparsity to be helpful because the gap between full information and partial information is large. Conversely, (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.

Nonetheless, sparsity is valuable in all but two of our cases: when there is high confidence that an important game has negative expected value (case (b) in Figure 2) and when there is high confidence that an important game has positive expected value (case (f) in Figure 2). In all other cases, the information about the proportion of punishers gained from playing games with spurious rules makes some degree of sparsity greater than zero valuable for sufficiently low participation costs. In all these cases (cases (a), (c), (d) and (e)), the value of the super-game is maximal with very high sparsity ($s = .9$) for sufficiently low participation costs. Moreover, in all these cases, a super-game with some positive level of sparsity generates higher value than the super-game that consists only of important only ($s = 0$).

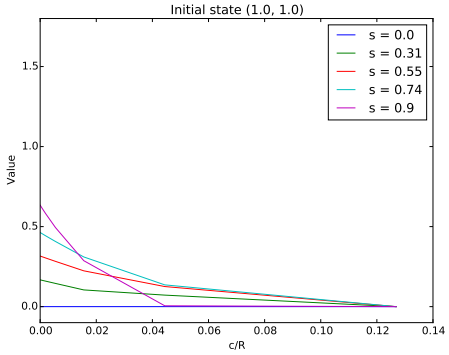
As we might expect, the net gain in value from sparsity is larger in belief states that are very uncertain. For example, compare cases (c) and (d). In case (c) Ego can be thought of as having formed the belief that the proportion of punishers is 50% (and so the expected reward in an important game is 0.) This belief is based, however, on only two observations–one of a punisher and one of a non-punisher. In that case the value of the most sparse environment ($s = .9$) exceeds 0.5. In case (d)–where Ego also holds the belief that the proportion of punishers is 50% but now on the basis of fifty observations–the most sparse environment, while still yielding value, increases the value function by less than half of the amount generated in case (c), and over a much smaller range of participation costs. In both cases, the value of sparsity is that while the initial estimate that an important game is not worth playing when participation costs are any positive amount (retirement is the best policy), there is some chance that this estimate will be revised upwards with additional draws from the distribution of punishers–making a decision to retire sub-optimal. But the chance of this happening in case (d) is much lower than in case (c), so the expected value of continuing to play



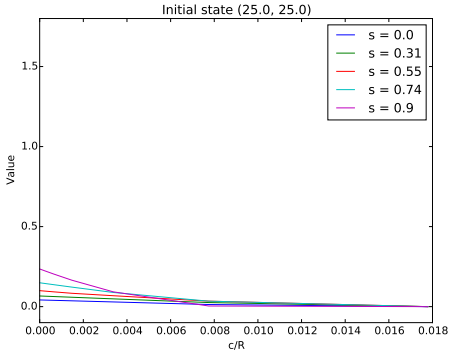
(a) Low Mean, Low Confidence



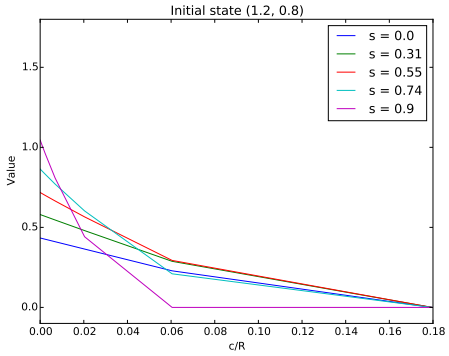
(b) Low Mean, High Confidence



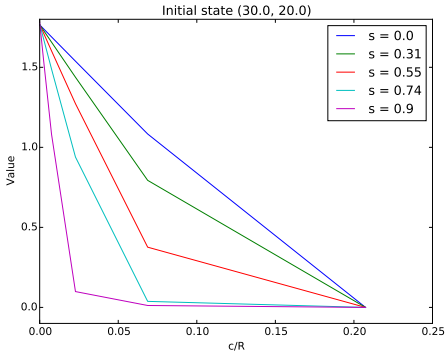
(c) Mid Mean, Low Confidence



(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 ($R = 1$). The increase in value for higher sparsity for low values of $\frac{c}{R}$ in all cases except (f) (high mean, high confidence) shows confirmation of the results in Prop. 3. The least sparse environment ($s = 0.0$) is dominated by a more sparse environment in all cases except (f). For cases with low or mid mean (a-d) participation is suboptimal. unless sparsity is positive. The cost of this sparsity is increased sensitivity to participation costs: value declines more quickly as costs increase for higher levels of sparsity.

is lower. Furthermore, in case (d), participation costs must be extremely low–an order of magnitude lower than in case (c)–for participation to net positive value.

6 Discussion

Our results have surprisingly powerful implications for the structure of normativity in human communities.

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 5 indicate that when expected costs relative to rewards and confidence in initial beliefs 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 by pervasive and sometimes 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 group 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.[8] [9]

The value of silly rules in a community, however, does depend on the extent of uncertainty about the likelihood that norm violations are punished by that community–in our model, the proportion of punishers in the community. We thus expect that silly rules are a feature of new and uncertain environments–as when a group breaks off from an established community and proposes new rules, for example or when an established community is growing due to immigration of "foreigners" who are originally outsiders and hence unsure of how locals behave. Such conditions are, we suggest, implied by a theory of group competition as an explanation for the innovation and survival of functional norms [Boyd and Richerson, 2009]. Environments with "norm entrepreneurs" who are engaged in deliberate efforts to change a norms in order to order to better adapt to changed/changing conditions may also be a setting in which the presence of spurious norms is valuable. Conversely, very stable communities, where the likelihood of punishment is well-established and well-known to all members of the community, are ones in which spurious norms will have little if any value. Persistence of spurious norms in those settings may be a result of over-imitation and the normative moralization of irrelevant actions that may support learning in children [Kenward et al., 2012].

[8] 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.

[9] This prediction distinguishes our account from that of DeScioli and Kurzban [2013]. In their account, normative moralization of potentially arbitrary actions "that are likely to occur in conflicts" (p. 484) serves to reduce losses due to costly conflicts by giving individuals a means of signalling which side they would take in a dispute. This account focuses on settings in which norm violation results in relatively high cost punishments; if norm violation elicits only mild rebuke, there would seem to be little cost-saving pressure to extend normativity.

We motivated our analysis at the outset with the observation that many societies seem to be awash with norms that are apparently meaningless for everyone and which, but for the informative value of moralizing these actions, a society would do just as well or better without. But our analysis has empirical validity beyond the explanation of the persistence of truly silly rules. Our model only assumes that spurious rules are of no value to an individual Ego who is contemplating joining a community. It may well be that a norm that is spurious from Ego's perspective is important to some Others. Our results therefore can also be seen to predict that communities with dense rule sets–lots of rules, period–can generate higher value for Ego, provided the costs incurred to comply with and punish rules that are valued only by Others are not too high.

Finally, 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 282 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 282 together on a stone pillar and named the collection as his Code, he created the possibility for the emergence of a new, and attractively parsimonious, labeling system with 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.

The larger lesson of our model is about the value of analyzing not only the emergence or functionality of specific norms but also the relationships between norms. As our analysis shows, it may

not possible to understand the significance of some norms without understanding the role they play in a normative system. A focus on individual norms in isolation, we suggest, overlooks the need to explain and understand, and hence ultimately our capacity to predict and create, the structure of normativity itself.

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.

Cristina Bicchieri. *The grammar of society: the nature and dynamics of social norms*. Cambridge University Press, New York, 2006.

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 and Peter J. Richerson. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533):3281–3288, 2009.

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

Joshua W. Buckholtz and Ren Marois. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature neuroscience*, 15(5):655, 2012.

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

Peter DeScioli and Robert Kurzban. A solution to the mysteries of morality. *Psychological Bulletin*, 139(2):477, 2013.

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.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

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.

Ben Kenward, Humanistisk samhllsvetenskapliga vetenskapsomrdet, Samhllsvetenskapliga fakulteten, Uppsala universitet, and Institutionen fr psykologi. Over-imitating preschoolers believe unnecessary actions are normative and enforce their performance by a third party. *Journal of experimental child psychology*, 112(2):195–207, 2012.

Robert Kurzban, Peter DeScioli, and Erin O'Brien. Audience effects on moralistic punishment. *Evolution and Human behavior*, 28(2):75–84, 2007.

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.

Richard McElreath, Robert Boyd, and PeterJ Richerson. Shared norms and the evolution of ethnic markers. *Current anthropology*, 44(1):122–130, 2003.

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.

Peter J. Richerson and Robert Boyd. *Not by genes alone: how culture transformed human evolution*. University of Chicago Press, 2008.

Katrin Riedl, Keith Jensen, Josep Call, and Michael Tomasello. No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America*, 109 (37):14824–14829, 2012.

Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25, 1995.

Robert Sugden. The economics of rights, cooperation, and welfare. Palgrave Macmillam, 1986.

Michael Tomasello and Amrisha Vaish. Origins of human cooperation and morality. *Annual Review of Psychology*, 64(1):231–255, 2013.

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('--show', action='store_true')
    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))

    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

    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'$\frac{c}{R}$')
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_valu
    ax.plot(np.r_[max_ret_sparsity_values, [1]], max_c_values + [0], label="({},

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


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


    # The x-axis values, determines the accuracty of the plots. Uses
    # log linear spaces because that gives qualitatively better responses
    max_cost = 0.000001
    for s in sparsity_values:
        max_cost = max(max_cost, largest_possible_c((alpha, beta), s, 0.9))
    costs = np.exp(np.linspace(np.log(0.00001), np.log(max_cost),  args.x_fidelit


    fig = plt.figure()
    ax = fig.add_subplot(1, 1, 1)
    # ax.set_xscale('log')
    ax.set_xlabel('c/R')
```

```
        ax.set_ylabel('Value')

        for s in sparsity_values:
         print "s: {}, ratio: {}, effective samples: {}".format(
                    s, ratio, effective_samples)
         cost_response_curve = compute_val(
                    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([-.1, 1.8])
        plt.legend(loc='best')
        plt.savefig(
            resultfolder+start_values[(ratio, effective_samples)] + ".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 the maximal error is the
    maximal positive reward for all time. With probability (1 - s) Ego
    gets value R with probability \theta. We upper bound by letting
```

```
\theta = 1 and c = 0 then normalize by R to get an upper bound:

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

Error decreases by at least delta_s each step of value iteration so
we need delta_sˆH UB(c,s) <= tol ===> H >= log(tol/UB(c, s)) / log(delta_s)
"""
log_V_ub = np.log(1 - s) - 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 t rounds we will have seen t heads or tails,
    # and we incorporate the priors
    cur_confidence = t + a + b

    # P[i] = i / Nt; i in [0,...,t]
    P = a_vals[:t] / (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_minus1 = 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()
```