The Econometrics Journal

Econometrics Journal (2018), volume **21**, pp. 298–315.
doi: 10.1111/ectj.12112

Non-parametric Bayesian inference of strategies in repeated games

MAX KLEIMAN-WEINER[†], JOSHUA B. TENENBAUM[†] AND PENGHUI ZHOU[‡]

[†]*Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.*
E-mail: maxkw@mit.edu, jbt@mit.edu

[‡]*Department of Economics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.*
E-mail: phzhou@mit.edu

First version received: March 2017; final version accepted: March 2018

Summary Inferring underlying cooperative and competitive strategies from human behaviour in repeated games is important for accurately characterizing human behaviour and understanding how people reason strategically. Finite automata, a bounded model of computation, have been extensively used to compactly represent strategies for these games and are a standard tool in game theoretic analyses. However, inference over these strategies in repeated games is challenging since the number of possible strategies grows exponentially with the number of repetitions yet behavioural data are often sparse and noisy. As a result, previous approaches start by specifying a finite hypothesis space of automata that does not allow for flexibility. This limitation hinders the discovery of novel strategies that may be used by humans but are not anticipated *a priori* by current theory. Here we present a new probabilistic model for strategy inference in repeated games by exploiting non-parametric Bayesian modelling. With simulated data, we show that the model is effective at inferring the true strategy rapidly and from limited data, which leads to accurate predictions of future behaviour. When applied to experimental data of human behaviour in a repeated prisoner's dilemma, we uncover strategies of varying complexity and diversity.

Keywords: *Bayesian inference, Computational economics, Finite-state automata, Non-parametric inference, Repeated games.*

1. INTRODUCTION

In strategic settings, predicting the actions of other players is essential for both cooperative and competitive intelligent behaviour. Models of how people reason about and infer the strategies of others can give insights into the cognitive systems used by humans in interactive strategics contexts. Repeated games are a key example: players must infer the strategies of others based on their previous interactions so as to achieve their cooperative or competitive goals.

Repeated games offer significantly more strategic possibilities than one-shot games. In this work we will use the repeated two-player prisoner's dilemma as an example to demonstrate our approach, although our model is general in the underlying stage game and in the number of players. We first briefly describe the one-shot prisoner's dilemma: two players simultaneously choose either to cooperate (C) or defect (D). Based on their joint selection of actions, they obtain utility according to the payoff matrix in Figure 1. When the game is played only once, there is

		Player 2	
		C	D
Player 1	C	a, a	$12, 50$
	D	$50, 12$	$25, 25$

Figure 1. Payoff matrix for a two-player prisoner's dilemma. The value of a sets the payoff of joint cooperation and must be $25 < a < 50$.

only a single Nash equilibrium: both players defect. Thus the prisoner's dilemma represents a simplified social dilemma: both players would prefer to receive the Pareto-optimal outcome (C, C), but only (D, D) is the equilibrium outcome.

When the prisoner's dilemma is repeated an indefinite number of times between the same two players, the cooperative outcome can be rationally sustained as dictated by the folk theorem (Fudenberg and Maskin, 1986). Unlike one-shot games where the space of strategies is just a measure over the action space, rational players in repeated games condition their actions on previous outcomes. This results in an exponential growth in the number of strategies as the repeated interaction continues. As a result of this exponential growth, learning these strategies from data seems like an intractable task. Each additional round requires conditioning on greater and greater amounts of data. This complexity is sometimes called the *curse of history* (Pineau et al., 2003).

To succinctly represent strategies with behavioural significance, theorists have turned to a model of bounded computation, the finite state transducer (FST), a type of automaton that compactly represents strategies using only limited memory (Carmel and Markovitch, 1996 and Rubinstein, 1986). In the prisoner's dilemma, a conditional cooperation strategy called tit-for-tat (TFT), which starts off playing C and then copies the previous move of the other player, can be compactly represented using FSTs. These simple strategies have significance for the evolution of cooperation, understanding human behaviour and designing self-regulating cooperative systems (Axelrod and Hamilton, 1981, Kleiman-Weiner et al., 2016 and Littman and Stone, 2005). An example interaction in the two-player repeated prisoner's dilemma where player 1 (P1) is using the TFT strategy is

$$\text{P1:} \quad CCCDCCCDDDC\ldots,$$
$$\text{P2:} \quad CCDCCCDDDCC\ldots.$$

Simple strategies like TFT are often enriched by considering forgiving variants that return to cooperation after a string of mutual defections or a vindictive TFT that defects multiple times after a defection regardless of whether or not the other player returns to cooperation (Nowak and Sigmund, 1992 and Zagorsky et al., 2013). Developing strategies for repeated games in terms of FST enables theorists to capture and study their intuitions about behaviour in a formal model. Using FSTs to represent strategies, one maps the curse of history of strategy inference (where the number of strategies grows exponentially in previous interactions) to a search over the space of possible FSTs. However, there are still *a priori* an infinite number of possible automata one must consider.

So how might people infer strategies from the behaviour of other players? How do theorists generate new candidate FSTs for study from this infinite space? In this work we develop a Bayesian model for strategy inference. The problem of strategy inference can be posed probabilistically as finding $P(\text{strategy}|\text{data})$, i.e. given data from an interaction between players, finding the probability of each strategy (as represented by an FST). Using Bayes' rule we can write the posterior distribution in terms of the data likelihood and a strategy prior:

$$P(\text{strategy}|\text{data}) \propto P(\text{data}|\text{strategy})\,P(\text{strategy}).$$

The core of this work is to formalize the pieces of this relationship and to propose an algorithm for inference. The term $P(\text{data}|\text{strategy})$ is the probability that an FST could have generated a specific sequence of behavioural data. For deterministic FST, this probability distribution is a delta function. However, in reality, behaviour is likely to be 'noisy', i.e. selected play may not coincide exactly with the action prescribed or intended by the strategy. If we assume probabilistic errors, any FST can generate a sequence of data, but with varying probability. The real challenge of inference comes from specifying $P(\text{strategy})$, the prior distribution over possible strategies. Since strategies are represented as FSTs and we want to consider strategies of arbitrary complexity, this is equivalent to specifying a distribution over all possible FSTs.

Solving this inference problem has key implications for learning equilibrium strategies. Under some general assumptions, rational learning leads to Nash equilibria in infinitely repeated games (Kalai and Lehrer, 1993). If each player assigns positive probability to all remaining possible opponent strategies that can occur within future play given past observations, then Bayesian updating will lead in the long run to accurate predictions about future play of the game. Bayesian updating is essentially a process of eliminating 'impossible' strategies and selecting the most probable strategies from the remaining possible choices. The prior plays a key role: for guarantees on rational learning to hold, a learner must correctly assign positive probability over all possible remaining strategies.

One solution common in experimental game theory is to put a uniform distribution over a hand-selected subset of strategies (Bó, 2005, Bó and Fréchette, 2011 and Blonski et al., 2011). However, this approach only allows for one to estimate the relative likelihood of strategies that are specifically hypothesized *a priori*, preventing the discovery of novel strategies for commonly studied games. Furthermore, this method is not robust for games that have not been analytically analysed. For instance, Bó and Fréchette (2011) conducted experiments and found that tit-for-tat and always-defect account for more than 80% of played strategies in repeated prisoner's dilemma games, but later work pointed out that a lesser known strategy of equal complexity called semi-grim can better account for their data (Breitmoser, 2015). Since semi-grim was not in the authors' original prior hypothesis space, it could not be inferred. Likewise, inexperienced players faced with a strategic situation need a robust way to infer the strategies of other players such that, given sufficient evidence, the correct strategy will be inferred.

Besides using a uniform prior over a finite hypothesis space, another approach for predicting behaviour is based on learning methods such as fictitious play (Fudenberg and Levine, 1998). Under this framework, each player best responds to the other player based on the empirical frequency of the other player's actions. While these methods can be powerful at predicting behaviour, they do not infer a model of the other players' strategies. Thus, while reinforcement learning methods can learn the statistical likelihood of certain actions, they do not learn a *causal* model (like a FST) of other players. These methods are less likely to generalize across games or predict the behaviour of others in rare situations.

Here, we present a novel non-parametric Bayesian model for strategy inference in repeated games. We develop a new prior over strategies based on the hierarchical Dirichlet process (HDP) (Teh et al., 2006). The model is non-parametric in the number of states in an FST and implicitly represents the infinite space of possible FST. We derive a Gibbs sampler for efficient inference in this model that successfully infers the actual strategy in simulated interactions. Our model predicts the correct FST even under noisy conditions and can be used to investigate human strategies from behavioural data. Our main contribution is to bring powerful tools from statistical machine learning to the study of strategic behaviour. To our knowledge this is the first application of the HDP in game theory and the analysis of human strategic behaviour in games. This model is a step toward developing computational agents with social intelligence that can predict the behaviour of others in strategic settings.

2. MODEL

We first describe the FST formally and describe its relation to a hidden Markov model (HMM). This relation allows us to leverage a suite of tools from probabilistic graphical models for inferring FSTs. We review the HDP and the HDP-HMM, and extend these models to represent strategies. We call this new model the HDP-FST. The HDP-FST is a generalization of the HDP-HMM, and can be used to represent and infer strategies in repeated games.

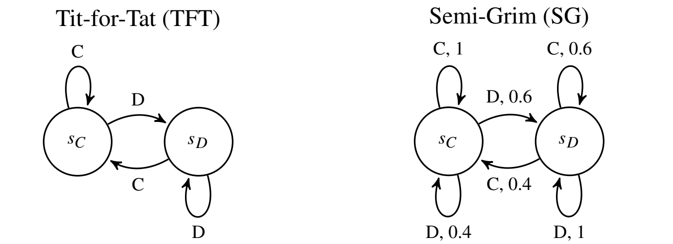
2.1. FST and HMM

An FST is a bounded model of computation that is capable of representing strategies in infinitely repeated games (Rubinstein, 1986). Formally, for player i, an FST is a tuple $\langle S_i, \boldsymbol{O}, \boldsymbol{y}, \boldsymbol{\pi}, \phi, \mathrm{F}\rangle$, namely a finite set of states $S_i = \{s_1, \ldots, s_n\}$, a finite set of input symbols $\boldsymbol{O} = \{o_1, \ldots, o_n\}$, a finite set of emission symbols $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ and a transition relation $\boldsymbol{\pi}$, $\pi_{ij} = \Pr(s_{t+1} = j | s_t = i, o_t \in \boldsymbol{O})$, where o_t is the input observed at time t. The term F characterizes the distribution for emissions at each state $s_i \in S_i$, where ϕ_{s_i} parameterizes the emission y_i such that $y_i | s_i \sim \mathrm{F}(\phi_{s_i})$.

FSTs represent strategies as ‘if–then’ computations. Given two players i and j, *if* player j previously played action o_j, *then* player i transitions to a particular state s and performs action $y_i | s$ at the next iteration of the game. Thus FSTs take in a set of inputs and for each input, update their internal state and produce an emission. In the context of strategies, the emissions of an FST correspond to actions.

For illustration, two FSTs (TFT and semi-grim) are reproduced in Figure 2. The top row shows a FST representation; Arrows show transitions between states given the actions of the other players. The middle row shows Transition matrices between the state at time t (rows) and $t + 1$ (columns), one for each action available to the other player. Each entry gives the probability of transitioning between states. The box above each matrix specifies the other player’s action at time t. The bottom row shows the Emission matrix that probabilistically maps from a state (rows) to an action (columns). In the Prisoner’s Dilemma set up we have $S_i = \{s_C, s_D\}$, $\boldsymbol{O} = \{C, D\}$, $\boldsymbol{y} = \{C, D\}$, $f_{\phi_{s_C}}(C) = 1$ and $f_{\phi_{s_D}}(D) = 1$, where $f_{\phi_{s_C}}$ and $f_{\phi_{s_D}}$ are the probability density functions (pdfs) of $\mathrm{F}(\phi_{s_C})$ and $\mathrm{F}(\phi_{s_D})$, respectively. Since the input symbols in games come from other strategic players, we will use $\boldsymbol{y_{-i}} = \boldsymbol{O}$, where $-i$ are all the players except i. The starting state of an FST is s_0.

Let $s_{i,t} \in S_i$ be the state of player i and let $y_{i,t} \in y_i$ be the action taken by player i at time t. Using TFT as an example, $s_{i,t} = s_C$ and $y_{i,t} = C$. If the action by player j at t is $y_{j,t} = C$, then $s_{i,t+1} = s_C$ and $y_{i,t+1} = C$; otherwise if $y_{j,t} = D$, then $s_{i,t+1} = s_D$ and $y_{i,t+1} = D$ from $\boldsymbol{\pi}$. Thus i plays in round $t + 1$ the action that the j played in round t.



TFT:

$$\boxed{y_j = C} \qquad \boxed{y_j = D}$$

$$\begin{array}{c} \\ s_C \\ s_D \end{array} \begin{array}{c} \begin{array}{cc} s_C & s_D \end{array} \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{cc} s_C & s_D \end{array} \\ \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ s_C \\ s_D \end{array} \begin{array}{c} \begin{array}{cc} y_i = C & y_i = D \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

SG:

$$\boxed{y_j = C} \qquad \boxed{y_j = D}$$

$$\begin{array}{c} \\ s_C \\ s_D \end{array} \begin{array}{c} \begin{array}{cc} s_C & s_D \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{cc} s_C & s_D \end{array} \\ \begin{bmatrix} 0.4 & 0.6 \\ 0 & 1 \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ s_C \\ s_D \end{array} \begin{array}{c} \begin{array}{cc} y_i = C & y_i = D \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

Figure 2. Representations of strategies for the two-player repeated prisoner's dilemma.

The strategy grim-trigger plays cooperate until a defection is played and then defects forever. Semi-grim is a more forgiving version of the grim-trigger that 'forgives' defection and tries to resume cooperation, i.e. even if $\exists k \in \{1, \ldots, T\}$ such that $y_{j,k} = D$, there is a small probability for s_i to return to s_C if $y_{j,t} = C$. Following the example in Figure 2, if $y_{j,t} = D$ and $s_{i,t} = s_C$, then $\Pr(s_{i,t+1} = s_C) = 0.4$. Thus there is some probability that i remains in a cooperative state even if j defects. Similarly, if $s_{i,t} = s_D$ and $y_{j,t} = C$, then $\Pr(s_{i,t+1} = s_C) = 0.4$, i.e. there is some probability of transitioning back to the cooperating state if cooperate was played by j.

So as to infer FST from data, we first describe the relationship between an FST and the HMM, a common probabilistic model for analysis of sequential data such as language or DNA. An HMM is a doubly stochastic Markov chain defined as a tuple $\langle \boldsymbol{s}, \boldsymbol{\pi}, \boldsymbol{y}, \phi, \mathrm{F} \rangle$, where $\boldsymbol{s} = (s_1, s_2, \ldots, s_T)$ represents a sequence of states linked by a transition matrix $\boldsymbol{\pi}$, where $\pi_{ij} = p(s_{t+1} = j | s_t = i)$ with $\pi_{0i} = p(s_1 = i)$. Corresponding to each state in the model is a parallel sequence of observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ with y_t drawn conditionally dependent only on s_t. For each state $s_t \in \{1, \ldots, K\}$ there is a parameter ϕ_{s_t} that reflects the likelihood of the observation at that state: $y_t | s_t \sim F(\phi_{s_t})$.

The difference between an HMM and an FST is that an HMM does not condition on the observed actions made by player j. Conditional on an opponent's action at $t-1$, both models are representationally identical: a set of transition matrices in the FST becomes a single transition matrix like the HMM. This implies that the HMM can be written as a limiting case of the FST. By assuming that player j's action fully specifies the transition probability between s_{t-1} and s_t i.e. each row of the transition matrix is conditionally independent, then the HMM can be augmented into an FST by making the states of the HMM dependent on the other players' actions. The equivalence between these augmented HMMs and FSTs has been formally proven (Kempe, 1997) and has been applied for use in speech recognition (Mohri et al., 2002).

While the FST formalism can represent specific strategies, it does not provide an algorithm or mechanism for enumerating or representing a hypothesis space of strategies. Consider an example sequence of plays in the infinitely repeated prisoner's dilemma where the observed actions are $[(D, D), (D, C), \ldots]$. Player 1's strategy could be always-defect, which is a one-state FST, or could be TFT, a two-state FST, or even a three-state FST where player 1 begins with D, defects until two iterations of C are observed from player 2 and then plays C. This kind of reasoning can generate an infinite space of strategies if the number of states in the FST is not restricted. How do we represent this space formally and apply it tractably for inference?

2.2. *HDP-HMM*

We now develop a new model for strategy inference based on the correspondence between the HMM and FSTs. This model is based on the HDP-HMM, a non-parametric extension of the HMMs (Teh et al., 2006). This is the first time to our knowledge that Bayesian non-parametric models have been applied to a game theoretic context. We first review Bayesian non-parametric models in general, focusing on Dirichlet processes. Then we describe how Dirichlet processes can be generalized with the HDP and how these tools are used for sequence modelling with the HDP-HMM.

We first introduce the Dirichlet process (DP). A DP is a generalized Dirichlet distribution (the conjugate prior of a multinomial distribution), but may contain an infinite number of elements. The DP is commonly used to describe a prior over the distribution of random variables and is parameterized by a base distribution H and a concentration parameter α, where $\alpha > 0$. For instance, consider the DP

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H), \\ H &\sim N(\mu, \sigma^2), \end{aligned} \tag{2.1}$$

where the base distribution (H) is a normal distribution with mean μ and variance σ^2. Draws from the DP, G, would have the same support as H, with one important difference: all draws from the DP are discrete. While H is continuous, implying that the probability that any two samples are equal is 0, this is not the case for G. See Figure 3 for a graphical illustration.

2.2.1. Stick-breaking process. The DP can also be represented by a stick-breaking process. This formalism directly reveals its discrete nature (Sethuraman, 1994). For $k = 1, 2, \ldots,$ let

$$\phi_k \sim H, \qquad \beta'_k \sim \mathrm{Beta}(1, \alpha), \qquad \beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_k). \tag{2.2}$$

Then the random measure defined by $G = \sum_{k=1}^{K} \beta_k \delta_{\phi_k}$ is with probability 1 equal to a sample from $\mathrm{DP}(\alpha, H)$, where δ_ϕ is a probability measure concentrated at ϕ. The construction

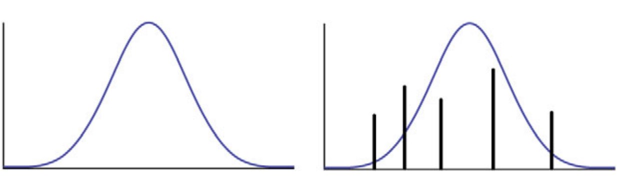


Figure 3. (Left) A normal distribution $H \sim N(\mu, \sigma^2)$ and (right) $G \sim \mathrm{DP}(H, \alpha)$, with H interposed for reference; G is a probability distribution that 'looks like' H, but whose distribution is discrete. [Colour figure can be viewed at wileyonlinelibrary.com]

of β_1, β_2 can be also thought of as starting off with a stick of length 1, breaking it off at $\beta_1 \sim \text{Beta}(1, \alpha)$ and recursively breaking the remaining portion at β_2, β_3 and so on. This process is also called GEM(α) after Griffiths, Engen and McCloskey, where α refers to the same concentration parameter as in the DP, and $\alpha > 0$.

2.2.2. Hierarchical Dirichlet process. The HDP is a set of DPs that are coupled together with a random base measure that is itself a DP. The HDP was developed to apply non-parametric methods to the problem of clustering grouped data. Data can be subdivided into different groups, and within each group there can be clusters that capture latent structure within that group. These models have been used in machine learning to cluster and classify data such as documents and genetic data (Beal et al., 2002, Blei et al., 2003, Gabriel et al., 2002 and Wood et al., 2011).

There are many similarities between these problems and the challenge of inferring strategies. As we have shown, strategies can be represented in terms of states, transition probabilities and emission matrices. The clusters in strategy inference refer to the distinct states in the FST. Given a sequence of observed actions by player i (e.g. C or D) and the corresponding history of actions by player j, we want to identify the underlying and distinct states in i that produced these actions and the transitions between these states. However, to consider all possible strategies, we would have to consider and conduct inference for an infinite number of transition parameters and states, which is not a computationally feasible process.

Next, we need to integrate out the infinite number of transition parameters and represent the process with a finite number of indicator variables. In the HDP there is a natural bias toward using already existing transitions proportional to their prior usage ('rich get richer'). This implies that the latent state sequences ($\boldsymbol{s_i}$) produced by the FST that we observe are *typical trajectories* (Beal et al., 2002), which result in a set of strategies biased toward those of lower complexity (as measured by the number of states in the FST) that most resemble the actual strategy.

We now formally describe the HDP. First, we define a global vector β and local vectors π_k in the method

$$\begin{aligned} \beta &\sim \text{Dirichlet}(\gamma/K, \ldots, \gamma/K), \\ \pi_k|\beta &\sim \text{DP}(\alpha, \beta), \qquad \phi_k \sim H, \end{aligned} \tag{2.3}$$

where π_k represents the transition probabilities out of state k, and ϕ_k parameterizes the distribution of emissions at each state k, drawn from a base distribution H. Since each $\pi_k \sim \text{DP}(\alpha, \beta)$, the states (within each FST) that the transition matrices refer to are *shared* as each DP is drawn from the same β, itself a *discrete* distribution obtained from the global draw parameterized with γ and K. Each atom in β hence represents the prior mean for transition probabilities leading into state k.

The ratio γ/K determines the sparsity of β. For instance, if $\gamma/K \ll 1$, the mass of β will be highly concentrated in just a few components. When $\gamma/K \to \infty$, the mass will gradually be equally dispersed across all the components. As $K \to \infty$, the prior in (2.3) becomes an HDP. Each π_k has concentration parameter α that determines deviation from the mean. This sharing is shown graphically in Figure 4. Since the DP draw of the base measure is necessarily discrete, subsequent draws will be drawn from the same discrete distribution.

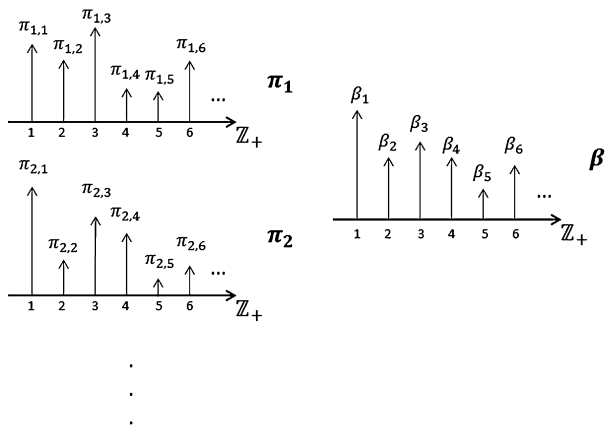


Figure 4. The sparsity of β is shared. (Right) An example β. (Left) Each π_i shares the same atoms as β and $\pi_{i,k}$ has β_k as its expected value.

Similar to the DP, the HDP can also be constructed via a stick-breaking process (Teh et al., 2006 and Van Gael et al., 2008):

$$\begin{aligned} \beta \sim \mathrm{GEM}(\gamma), \qquad \pi_k | \beta \sim \mathrm{DP}(\alpha, \beta), \qquad \phi_k \sim H, \\ s_{t+1} | s_t \sim \mathrm{Multinomial}(\pi_{s_t}), \qquad y_t | s_t \sim \mathrm{F}(\phi_{s_t}). \end{aligned} \tag{2.4}$$

The graphical model for the HDP-HMM is shown in Figure 5.

2.3. HDP-FST

Just as we showed the relation between the FST and the HMM, we will now show that the HDP-HMM is a limiting case of a more general class of models, which we call the HDP-FST. Let $\boldsymbol{y_i} = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})$ be the history of actions for player i where each $y_{i,t}$ is the action player i took at time t. The $\boldsymbol{y_j}$ are the history of actions taken by player j that are observed by

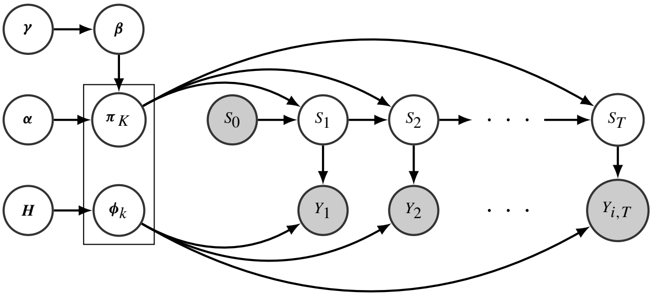


Figure 5. Graphical model for the HDP-HMM: The four 'global' parameters $(\gamma, \alpha, H, \beta)$ generate π_K and ϕ_k, the two parameters that define an FST, and the remainder of the figure follows a typical HMM formalism, with the emission $Y.$ being a function of the state $s.$ and ϕ_k.

player i. For now, we restrict analysis to one additional player, but the model is general for any finite number of players.

In an HMM, the probability distribution of each subsequent state is dependent only on the previous state. However, in an FST, each state is dependent on both the current state and the observed actions of the other player. Adding this additional dependency to the HDP-HMM turns it into the HDP-FST. Let the state sequence of player i be $\boldsymbol{s_i} = (s_{i,1}, \ldots, s_{i,T})$. Player i observes player j's action $y_{j,t}$ at time t and conditions $s_{i,t+1}$ on both the previous state $s_{i,t}$ and the previous action played by player j, $y_{j,t}$. Thus like an FST, conditional on s_t, the HDP-FST only needs to know the other player's previous action $y_{j,t}$ and not his/her full sequence of actions $\boldsymbol{y_j}$.

Heterogeneity between different people could be captured through the hyperparameters, β, ϕ and H, which are now subscripted by i, but retain their interpretation from (2.4). The HDP-FST is formally

$$\begin{aligned} \beta_i &\sim \mathrm{GEM}(\gamma_i), \qquad \pi_{i,k,y_j} | \beta_i \sim \mathrm{DP}(\alpha_i, \beta_i), \\ \phi_{i,k} &\sim H_i, \qquad s_{i,t+1} | s_{i,t}, y_{j,t} \sim \mathrm{Multinomial}(\pi_{i,s_{i,t},y_{j,t}}), \\ y_{i,t} | s_{i,t} &\sim F(\phi_{i,s_{i,t}}). \end{aligned} \tag{2.5}$$

The key difference between (2.5) and (2.4) is that π_i additionally depends on y_j. This is comparable to how an HMM can be augmented into an FST. Intuitively, consider a partition of π_i into $|y_j|$ different $k \times k$ matrices, where $|y_j|$ is the number of unique actions available to player j, and when $|y_j| = 1$, then the HDP-FST is equivalent to the HDP-HMM. The interpretation of the hyperparameters β_i, α_i and γ_i is unchanged. Figure 6 shows the graphical model for the HDP-FST in the case of two players.

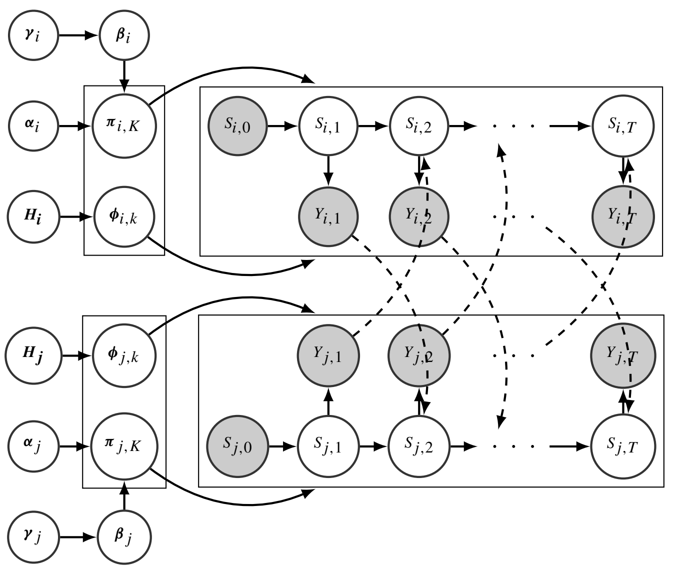


Figure 6. HDP-FST graphical model. The dotted arrows represent the additional conditional dependency added from the HDP-HMM model, where each $s_{i,t}$ is conditioned on $y_{j \neq i,t-1}$, $s_{i,t-1}$ and π_i.

Note that each player is shown with a different set of hyperparameters. These hyperparameters could be different for different players if we believed that the players differed in some capacity. A larger γ for a particular player could correspond to believing that a player is *a priori* more likely to play a smaller FST. Similarly, choosing a smaller γ corresponds to a higher prior probability on strategies with a large number of states allowing for larger FSTs. While the model allows for this variation, in this work we use the same parameters and base distribution for all analyses. Hence for the remainder of the paper,

$$\gamma_i = \gamma_j = \gamma, \qquad \alpha_i = \alpha_j = \alpha, \qquad H_i = H_j = H.$$

2.4. *Inference*

Having described a prior over strategies using HDP-FST, we now describe how to make inference over this hypothesis space tractable using a Gibbs sampler. The Gibbs sampler is a commonly used method for drawing samples from a distribution that cannot be calculated analytically. This method is relevant because the exact Bayesian inference for the model is intractable as the number of states K is infinite, which means we cannot apply a generic forward–backward algorithm as one would commonly do if the K was known in advance. Since the distribution is over strategies, each sample is a specific FST, and by running the Gibbs sampler for many iterations, we can draw enough samples to approximate the posterior distribution.

We now describe the Gibbs sampler for inference in the HDP-FST. Gibbs sampling works by sampling each variable while conditioning on the values of all the other variables in the distribution (Murphy, 2012). Our Gibbs sampler builds on the direct sampling mechanism presented in Teh et al. (2006), and we reference Van Gael et al.'s (2008) description of sampling for the HDP-HMM. Given that the states are exchangeable, we can analytically marginalize out the latent variables π_i, ϕ_i from (2.5). Thus sampling will only involve the latent state sequence $\boldsymbol{s_i}$ and the DP parameter β_i. To resample $s_{i,t}$, we need to calculate the conditional probabilities

$$\begin{aligned} & p(s_{i,t}, s_{j,t} | s_{i,-t}, s_{j,-t}, y_{i,t}, y_{j,t}, y_{i,t-1}, y_{j,t-1}, \beta_i, \beta_j, \alpha, H) \\ & \quad \propto p(y_{i,t} | s_{i,t}, H) \cdot p(y_{j,t} | s_{j,t}, H) \\ & \qquad \cdot p(s_{i,t}, s_{j,t} | s_{i,-t}, s_{j,-t}, y_{i,t-1}, y_{j,t-1}, \beta_i, \beta_j, \alpha), \end{aligned} \tag{2.6}$$

where $s_{i,-t}$ refers to the sequence of states s_i, excluding $s_{i,t}$.

However, the structure of conditional independence in the HDP-FST can simplify this equation and allow for more efficient sampling. The conditional likelihood of $(y_{i,t}, y_{j,t})$ given the states $s_{i,t}$ and $s_{j,t}$, actions $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$, and base distribution H is easy to compute as each $y_{i,t}$ and $y_{j,t}$ is only dependent on its respective states at time t. Further, if the base distribution H and the likelihood F from (2.5) are conjugate, we can analytically update this portion of the likelihood. Given that the state space for these strategies is discrete, the conjugate multinomial-Dirichlet distribution is appropriate and greatly simplifies inference.

Furthermore, we can use the independence structure of the model to avoid having to sample from $\{\boldsymbol{s_i}, \boldsymbol{s_j}\}$ jointly. Because of the Markov property of the model, each $s_{i,t}$ is conditionally independent of all $\boldsymbol{s_j}$ given $s_{i,t-1}$, $s_{i,t+1}$, $y_{j,t-1}$ and $y_{j,t}$, where $y_{j,t-1}$ and $y_{j,t}$ are observed. Therefore each sequence of states $\boldsymbol{s_i}$ can be sampled independently of the hidden states of all other players. This factors the model into independent components (one for each player) that can be treated separately during inference. For this reason, we can sample each player independently when resampling the latent state sequences. We can also further reduce $s_{i,-t}$ to $\{s_{i,t-1}, s_{i,t+1}\}$ given that $s_{i,t}$ is only dependent on player i's state one time period prior and one after. Using this simplification, we can rewrite (2.6) as

$$\begin{aligned} &p(s_{i,t}|s_{i,t-1}, s_{i,t+1}, y_{i,t}, y_{j,t-1}, y_{j,t}, \beta_i, \alpha, H) \\ &\quad \propto p(y_{i,t}|s_{i,t}, H) \cdot p(s_{i,t}|s_{i,t-1}, s_{i,t+1}, y_{j,t-1}, y_{j,t}, \beta_i, \alpha). \end{aligned} \tag{2.7}$$

Now we describe the sampling mechanism for $p(s_{i,t}|s_{i,t-1}, s_{i,t+1}, y_{j,t-1}, y_{j,t}, \beta_i, \alpha)$. First, we remove the subscripts for s_i, β_i and observed j's actions y_j as it is clear which player we are referring to for each variable. At any time t, let $n_{l,m}$ be the total number of transitions from sampled states l to m, excluding time steps t and $t-1$, and let $n_{\cdot,l}$ and $n_{l,\cdot}$ be the total number of transitions into and out of state l, respectively. Let K be the current number of distinct states in $s_1, s_2, \ldots, s_{t-1}$.[1]

Because there is a Dirichlet prior on β, we can define the distribution of s_t generated from a single transition π_l as

$$\begin{aligned} p(s_t|s_{t-1}=l, \beta, \alpha) &= \int_{\pi_l} p(\pi_l, s_t|s_{t-1}=l, \beta, \alpha)\, d\pi_l \\ &= \int_{\pi_l} p(\pi_l|\beta, \alpha)\, p(s_t|s_{t-1}=l, \pi_l)\, d\pi_l \\ &= \int_{\pi_l} \frac{\Gamma(\sum_k \alpha\beta_k)}{\prod_k \Gamma(\alpha\beta_k)} p(s_t|s_{t-1}=l, \pi_l)\, d\pi_l && (2.8) \\ &= \int_{\pi_l} \frac{\Gamma(\sum_k \alpha\beta_k)}{\prod_k \Gamma(\alpha\beta_k)} \prod_{k=1}^{K+1} \pi_{l,k}^{\alpha\beta_k - 1} \prod_{k=1}^{K+1} \pi_{l,k}^{n_{l,k}}\, d\pi_l && (2.9) \\ &= \frac{\Gamma(\sum_k \alpha\beta_k)}{\prod_k \Gamma(\alpha\beta_k)} \frac{\prod_k \Gamma(\alpha\beta_k + n_{l,k})}{\Gamma(\sum_k \alpha\beta_k + n_{l,k})} \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{l,\cdot})} \prod_k \frac{\Gamma(\alpha\beta_k + n_{l,k})}{\Gamma(\alpha\beta_k)}, && (2.10) \end{aligned}$$

where we use the fact that β has a Dirichlet prior in (2.8).[2]

We augment (2.10) with the observed y_{t-1} to find, for a single state $s_t = k$,

$$\begin{aligned} p(s_t = k|s_{t-1}=l, y_{t-1}, \alpha, \beta) &= \frac{\Gamma(\alpha + n_{l,\cdot|y_{t-1}})}{\Gamma(\alpha + n_{l,\cdot|y_{t-1}} + 1)} \frac{\Gamma(\alpha\beta_k + n_{l,k|y_{t-1}} + 1)}{\Gamma(\alpha\beta_k + n_{l,k|y_{t-1}})} \\ &= \frac{\alpha\beta_k + n_{l,k|y_{t-1}}}{\alpha + n_{l,\cdot|y_{t-1}}}, \end{aligned} \tag{2.11}$$

which states that the probability of $s_t = k$ given $s_{t-1} = l$ is proportional to the relative frequency of previous transitions from state l to k ($n_{l,k}/n_{l,\cdot}$) and smoothed by $\alpha\beta_k$, the prior over k.

[1] Since we ignore the ordering of states in β, the K distinct states are labeled $1, \ldots, K$, and $K+1$ refers to a new state.

[2] Recall that π_l refers to the transition probabilities from state l to all other states, and $\pi_{l,k}$ refers to the probability of transitioning from state l to state k

Note that we can sample backward t from $t+1$ as well. Since we observe y_t (i.e. action of player j), we have

$$p(s_t = k|s_{t+1} = m, y_t, \alpha, \beta) = \frac{\alpha\beta_k + n_{k,m|y_t}}{\alpha + n_{k,\cdot|y_t}}. \tag{2.12}$$

From (2.11) and (2.12), we have $p(s_t = k|s_{t-1}, s_{t+1}, \beta, \alpha, y_{t-1}, y_t) \propto$

$$\begin{cases} (\alpha\beta_k + n_{s_{t-1},k|y_{t-1}})\dfrac{\alpha\beta_{s_{t+1}} + n_{k,s_{t+1}|y_t}}{\alpha + n_{k,\cdot|y_t}} & \text{for } k \leq K, k \neq s_{t-1} \\ (\alpha\beta_k + n_{s_{t-1},k|y_{t-1}})\dfrac{\alpha\beta_{s_{t+1}} + n_{k,s_{t+1}|y_t} + 1}{\alpha + n_{k,\cdot|y_t} + 1} & \text{for } k = s_{t-1} = s_{t+1} \\ (\alpha\beta_k + n_{s_{t-1},k|y_{t-1}})\dfrac{\alpha\beta_{s_{t+1}} + n_{k,s_{t+1}|y_t}}{\alpha + n_{k,\cdot|y_t} + 1} & \text{for } k = s_{t-1} \neq s_{t+1} \\ \alpha\beta_k\beta_{s_{t+1}} & \text{for } k = K+1, \end{cases}$$

where y_{t-1} and y_t refers to the observed actions of player j at $t-1$ and t. Using these equations for the conditional updates, we can realize tractable inference in the HDP-FST using a Gibbs sampler.

3. RESULTS

We empirically investigate the effectiveness of this model to infer FSTs under sparse and noisy observations of behaviour. We apply the model to a previously published data set of human behaviour in the infinite discounted prisoner's dilemma (Bó and Fréchette, 2011). To more precisely evaluate the model, we also analyse performance on a simulated data set where the ground truth strategies are known.

3.1. Behavioural results

We used our model to analyse data from a previous human behavioural experiment run by Bó and Fréchette (2011). Subjects were matched into dyads and played a discounted version of the prisoner's dilemma, i.e. the repeated interaction ended with constant probability after each round. In different interactions the value of a in the payoff matrix varied between 32 and 48 (see the payoff matrix in Figure 1). Prior work predicts that as the cooperative outcomes becomes less attractive relative to mutual defection, subjects will be more likely to defect. We analysed all dyadic interactions that lasted at least 10 rounds.

In Figure 7, we show the averaged posterior distribution across the 41 dyads that lasted longer than 10 rounds for three values of a. When a took a low value, the most common strategy inferred was always-defect (AD). As the value of a increased, the always-cooperate (AC) strategy was inferred with higher probability (with an intermediate balance of AC and AD when a took an intermediate value). We found relatively few instances of more complex strategies such as TFT and win–stay–lose–shift (WSLS)[3] . This likely reflects a combination of the simplicity bias

[3] Win–Stay–Lose–Shift describes a strategy where the prior action is repeated if the preferred payoff is obtained, and another action is chosen if not

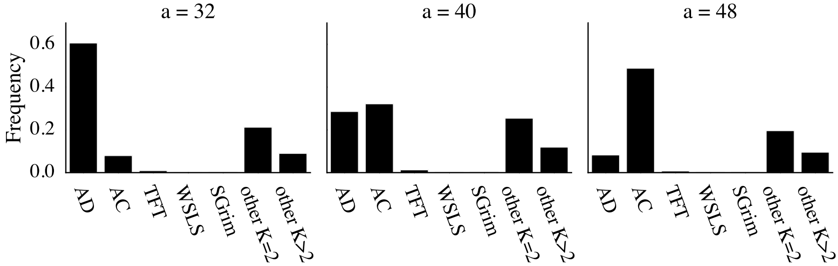


Figure 7. Posterior distribution over strategies averaged across all dyads that played at least 10 rounds of the prisoner's dilemma. Strategies not commonly studied in the literature were clustered by their complexity, where K is the number of states in the strategy.

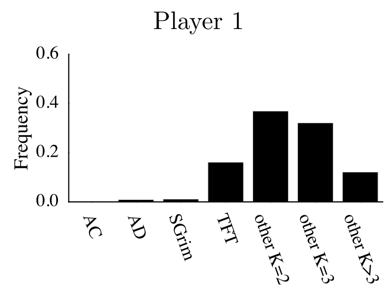
in the non-parametric prior since only a few of 41 dyads were longer than 20 rounds as well as the observation that many of the dyads played (C, C) or (D, D) for the entire interaction, which, while consistent with TFT and other more complex strategies, does not provide for any additional predictive power over simpler strategies. Finally, we note that there was also a significant number of strategies the model discovered that have not been investigated in the literature, particularly those with $K > 2$, where K is the number of states.

We also analysed the few long interactions in the data set that exceeded 20 rounds of repetition. We selected a dyad that seemed to have a fairly complex pattern of interaction and used the model to infer a distribution over strategies for just that dyad. Figure 8 shows the actions taken by the two players and the inferred distribution over strategies for each of the players. The model detected evidence of TFT for player 1 and semi-grim for player 2. As can be seen in the histograms, both the two-state FSTs were insufficient to capture the complexities of the interaction and most of the probability mass was on FSTs with three or more states. This suggests that people's actual strategies are more complex than the simple strategies of reciprocal cooperation (like TFT) predicted by current theory.

3.2. Simulated results

Using the four FSTs listed in Figure 9, TFT, WSLS, tit-for-two-tats and semi-grim, we generated 200 simulated dyadic interactions. Each of these simulations tests features of our model that have been challenging for previous approaches. We chose TFT and WSLS because they are of scientific significance (Nowak and Sigmund, 1993), tit-for-two-tats because it has more than two states and semi-grim because it has probabilistic transitions between states. Furthermore, we tested the algorithm when the observations were perfectly observed and also when 20% of the observations were corrupted by noise.

For all analyses presented here, we fixed $\alpha = \gamma = 0.5$ and used a symmetric prior on $H = [0.3, \ldots, 0.3]$. The first 200 samples of the Gibbs sampler were thrown away as burn-in and the chain was thinned every 2 samples. We ran the sampler until we collected 500 posterior samples. While many of the FSTs described in the literature are those with deterministic transitions and emissions, the model is not restricted in this way and does not represent deterministic strategies any differently from probabilistic ones. Thus so as to compare the probabilistic output of strategies from the model, we round the transition and action matrices to their closest deterministic FST. Since two deterministic FSTs may be identical to each other by merely



P1: $CCCCCDCCDCDDDDDDDDCDCCC$

P2: $CCCCCCDDCDDDDDDDDDDCCCC$

$t\rightarrow$

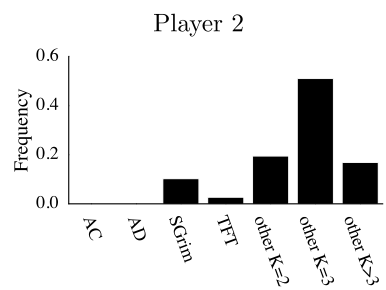


Figure 8. Posterior over strategies for two human players during a long interaction of 23 rounds. Of the strategies discussed in the literature, the model finds some evidence for the play of semi-grim and TFT in both players. However, overall, both players' play is more consistent with larger and more complex FSTs.

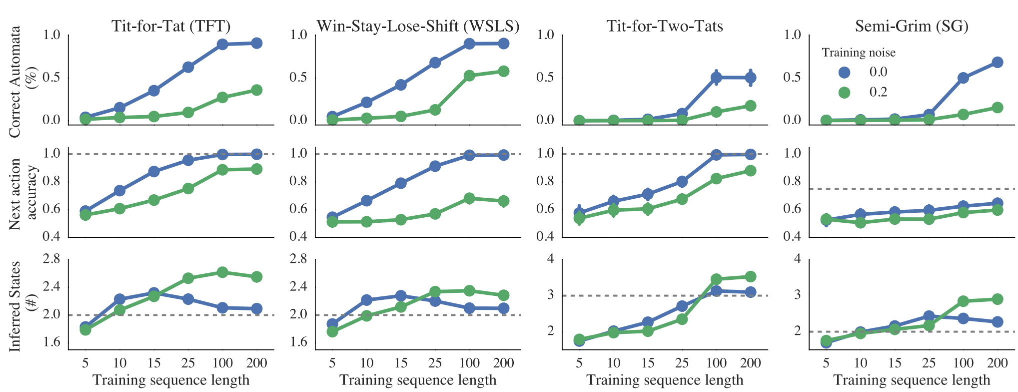


Figure 9. Simulated performance of inference with four different strategies (columns) that generated data while playing a random opponent. Each point is the average model performance of 200 runs. [Colour figure can be viewed at wileyonlinelibrary.com]

relabelling the states (which corresponds to permuting the transition and emission matrix), we clustered FSTs into functionally equivalent strategies by testing for isomorphisms in their graph. Using these two methods, we were able to classify a sample of a probabilistic strategy from the model to a known strategy type (if one was known).

Figure 9 shows the results of this empirical analysis. We evaluated the inferences of the model on three metrics: its ability to infer the correct FST (top row), whether or not it predicted the next action correctly (middle row) and the mean number of states in the distribution of sampled FSAs (bottom row). Since the repeated prisoner's dilemma has only two actions, chance guessing of the next action will result in 50% correct predictions. For semi-grim, the best one can predict is only 75% due to stochasticity in the transitions. In contrast, the chance probability of predicting the correct FST is negligible. In no part of the model did we include any specification of the FSTs commonly studied in the literature, but the model correctly infers them given only sparse and noisy observations.

The top row of Figure 9 shows the model's success in inferring the underlying strategy used to generate the interaction. Due to the simplicity bias inherited from the HDP prior, the simpler strategies (TFT and WSLS) are inferred correctly with less data. When trained on simulated behaviour corrupted by noise (shown in green), predictive performance was impaired but still improved with more training data. Even when the model does not infer the correct FST with high accuracy, it is still effective at predicting the next move. When the training sequence is short, there are many plausible FSTs that are consistent with the data.

Finally, we calculated the average number of states across the FSTs in the posterior sample as an approximate measure of the complexity of the inferred strategy. Given a very small amount of training data, the model mostly infers strategies of low complexity, but as the amount of training data increases, the average number of states in the inferred sample grows. This feature, the ability to increase model complexity as the amount of data grows, comes from the non-parametric prior and balances against overfitting. With a medium amount of training data, the number of FSTs considered grows considerably and even exceeds the actual number of states in the ground truth FST since there are many FSTs consistent with the data. However, as the training data increase, the model places most of its posterior mass on the correct strategy and the average complexity converges to the complexity of the true strategy. When the training signal is corrupted with noise, the complexity of the inferred FSTs exceeds that of the actual sequence, since the model accounts for some of the stochasticity with extra model complexity to account for noisy actions.

With these simulated results we have shown the power of this model to infer the strategies used by players who play strategies described by a single FST out of an infinite space of possible strategies without ever enumerating that space. Our non-parametric model trades off model complexity with data fit and allows for the consideration of more complex models as the amount of data grows, in contrast to previous analyses that use a finite hypothesis space.

4. CONCLUSIONS

In this work we developed the HDP-FST, a new non-parametric Bayesian model for the inference of strategies in repeated games. By extending the HDP, our model inherits many desirable properties of non-parametric models: (a) prior support over a hypothesis space that contains all possible FSTs without actually constructing this infinite space, (b) dynamically trades off the complexity of the inferred model with model fit by biasing the posterior probabilities toward simpler strategies and (c) allows for model complexity to grow with the data. We developed an efficient Gibbs sampler for conditional inference in this model. Using this inference scheme, we showed that from sparse and noisy observations of a dyadic interaction, it both infers the strategies and accurately predicts the expected next action. When applied to human data, the model inferred many strategies that have not been previously examined in the literature on repeated prisoner's dilemma. While we focused on the infinitely repeated prisoner's dilemma, our model applies to any repeated game with a finite number of players and a finite action space.

In future work we would like to develop a beam sampler for these models that would allow for more efficient online inference (Van Gael et al., 2008). It may be possible to adapt these methods to account for non-stationarity in player's strategies by putting a lower weight on earlier actions. Since our approach to inference is probabilistic and causal, it is possible to compose it with other probabilistic models allowing for richer multilevel analyses of human behaviour that can explicitly model individual variation across subjects.

While our model of strategy inference considers all possible FSTs (with a bias toward simple strategies), it does not consider the strategic implications of the FSTs. Consider the example

$$\text{P1:} \quad CCCD,$$
$$\text{P2:} \quad CCCC,$$

where we want to infer the strategy for P2. While both AC and tit-for-two-tat are consistent with P2's play, we intuitively believe that tit-for-two-tat should be more likely, i.e. our intuitions about what strategies are most likely *a priori* may also take into account the strategic nature of those strategies, not just complexity. In future work we will investigate the way the prior over strategies itself might be modulated by payoffs:

$$P(\text{strategy}|\text{data, payoffs}) \propto P(\text{data}|\text{strategy})\,P(\text{strategy}|\text{payoffs}).$$

One possibility is that strategies that are not consistent with a best response could be assigned a lower or even zero probability in the prior. Another possibility is to weight a strategy's prior probability by an estimate of its expected payoff. The modulation of the prior by payoffs might itself be modulated by one's estimate of the strategic sophistication of one's opponent. For instance, if players knew their opponents were very intelligent, they might place a very low prior probability on that opponent using always-cooperate.

Understanding which strategies are 'good' will likely require players who do not just infer strategies, but also plan on using them (Doshi-Velez et al., 2010, Kleiman-Weiner et al., 2016 and Panella and Gmytrasiewicz, 2015). The combination of strategic inference with planning will be essential for developing intelligent agents that flexibly cooperate and compete.

ACKNOWLEDGEMENTS

We thank Victor Chernozukov, Muriel Niederle, Juuso Toikka, Erez Yoeli and the anonymous referee for comments and improvements. MKW was supported by a Hertz Foundation Fellowship and NSF-GRFP. JBT was supported by the Center for Brains, Minds and Machines (CBMM), by NSF STC award CCF-1231216 and by ONR grant N00014-13-1-0333.

REFERENCES

Axelrod, R. and W. D. Hamilton (1981). The evolution of cooperation. *Science 211*, 1390–6.

Beal, M. J., Z. Ghahramani and C. E. Rasmussen (2002). The infinite hidden Markov model. *Advances in Neural Information Processing Systems 1*, 577–84.

Blei, D. M., A. Y. Ng and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Blonski, M., P. Ockenfels and G. Spagnolo (2011). Equilibrium selection in the repeated prisoner's dilemma: axiomatic approach and experimental evidence. *American Economic Journal: Microeconomics 3*, 164–92.

Bó, P. D. (2005). Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American Economic Review 95*, 1591–604.

Bó, P. D. and G. R. Fréchette (2011). The evolution of cooperation in infinitely repeated games: experimental evidence. *The American Economic Review 101*, 411–29.

Breitmoser, Y. (2015). Cooperation, but no reciprocity: individual strategies in the repeated prisoner's dilemma. *American Economic Review 105*, 2882–910.

Carmel, D. and S. Markovitch (1996). Learning models of intelligent agents. In *AAAI/IAAI*, *Volume 1*, 62–7. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Doshi-Velez, F., D. Wingate, N. Roy and J. B. Tenenbaum (2010). Nonparametric Bayesian policy priors for reinforcement learning. In *Advances in Neural Information Processing Systems*, 532–40. Red Hook, NY: Curran Associates, Inc.

Fudenberg, D. and D. K. Levine (1998). *The Theory of Learning in Games*, *Volume 2*, Chapter 2, 29–49. Cambridge, MA: MIT Press.

Fudenberg, D. and E. Maskin (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica 54*, 533–54.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, et al. (2002). The structure of haplotype blocks in the human genome. *Science 296*, 2225–9.

Kalai, E. and E. Lehrer (1993). Rational learning leads to Nash equilibrium. *Econometrica: Journal of the Econometric Society 61*, 1019–45.

Kempe, A. (1997). Finite state transducers approximating hidden Markov models. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 460–7. Cambridge, MA: MIT Press.

Kleiman-Weiner, M., M. K. Ho, J. L. Austerweil, M. L. Littman and J. B. Tenenbaum (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Conference paper. Hoboken, NJ: Wiley.

Littman, M. L. and P. Stone (2005). A polynomial-time Nash equilibrium algorithm for repeated games. *Decision Support Systems 39*, 55–66.

Mohri, M., F. Pereira and M. Riley (2002). Weighted finite-state transducers in speech recognition. *Computer Speech and Language 16*, 69–88.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Nowak, M. and K. Sigmund (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature 364*, 56–8.

Nowak, M. A. and K. Sigmund (1992). Tit for tat in heterogeneous populations. *Nature 355*, 250–3.

Panella, A. and P. J. Gmytrasiewicz (2015). Nonparametric Bayesian learning of other agents? Policies in interactive POMDPs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1875–6. Hingham, MA: Kluwer Academic Publishers.

Pineau, J., G. Gordon and S. Thrun (2003). Point-based value iteration: an anytime algorithm for POMDPs. In *IJCAI*, *Volume 3*, 1025–32. Burlington, MA: Morgan Kaufmann Publishers.

Rubinstein, A. (1986). Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory 39*, 83–96.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–50.

Teh, Y. W., M. I. Jordan, M. J. Beal and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*, 1566–81.

Van Gael, J., Y. Saatci, Y. W. Teh and Z. Ghahramani (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, 1088–95. New York, NY: ACM.

Wood, F., J. Gasthaus, C. Archambeau, L. James and Y. W. Teh (2011). The sequence memoizer. *Communications of the ACM 54*, 91–8.

Zagorsky, B. M., J. G. Reiter, K. Chatterjee and M. A. Nowak (2013). Forgiver triumphs in alternating prisoner's dilemma. *PloS One 8*, e80814.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication files