Incentive design for learning in user-recommendation systems with time-varying states

Deepanshu Vasal, Vijay Subramanian and Achilleas Anastasopoulos

Abstract— **We consider the problem of how strategic users with asymmetric information can learn an underlying time-varying state in a user-recommendation system. Users who observe private signals about the state, sequentially make a decision about buying a product whose value varies with time in an ergodic manner. We formulate the team problem as an instance of decentralized stochastic control problem and characterize its optimal policies. With strategic users, we design incentives such that users reveal their true private signals, so that the gap between the strategic and team objective is small and the overall expected incentive payments are also small.**

I. Introduction

In a classical Bayesian learning problem, there is a *single decision maker* who makes noisy observations of the state of nature and based on these observations eventually learns the true state. It is well known that through the likelihood ratio test, the probability of error converges exponentially to zero as the number of observations increases and the true state is learnt asymptotically. With the advent of the internet, in today's world, there are many scenarios where strategic agents with different observations (i.e. information sets) interact with each other to learn the state of the system that in turn affects the spread of information in the system. One such scenario was studied by the authors in their seminal paper [1] where they studied the occurrence of fads in a social network, which was later generalized by authors in [2]. The authors in [1] and [2] study the problem of learning over a social network where observations are made sequentially by *different decision makers* (users) who act *strategically* based on their own private information and actions of previous users. It is shown that herding (information cascade) can occur in such a case where a user discards its own private information and follows the majority action of its predecessors (fads in social networks). As a result, all future users repeat this behavior and a cascade occurs. While a good cascade is desirable, there's a positive probability of a bad cascade that hurts all the users in the community. Thus from a social (i.e. team) perspective, it is highly desirable to avoid such situations. Avoiding such bad cascades is an active area of research, for example [3] and [4] propose alternative learning models that aim at avoiding such bad cascades. In this paper, our goal is to analyze this model and design incentives to avoid bad cascades.

Most of the literature for this problem assumes time-invariant state of the nature. However, there are situations where the state of the nature, for e.g. popularity of a product, could change over time, as a consequence of endogenous or exogenous factors (for e.g., owing to the entering of a new competitor product or improvement/drop in quality of the product). In this paper we consider a simple scenario where users want to buy a product online. The product is either good or bad (popular or unpopular) and the value of the product (state of the system) is represented by X_t, which is changing exogenously via a Markov chain. The state is not directly observed by the users but each user receives a private noisy observation of the current state. Each user makes a decision to either buy or not buy the product, based on its private observation and action profile of all the users before its.

The strategic user wants to maximize its expected value of the product. But its optimal action could be misaligned with the team objective of maximizing the expected average reward of the users. Thus the question we seek to address is whether it is possible to incentivize the users to align them with the team objective. To incentivize users to contribute in the learning, we assume that users can also send reports (at some cost) about their private observations after deciding to buy or to not buy the product. The idea is similar to leaving a review of the product. Thus users could be paid to report their observations to enrich the information of the future participants. Our objective is to use principles of mechanism design to construct the appropriate payment transfers (taxes/subsidies). Although, our approach deviates from general principles of mechanism design for solution of the game problem to *exactly* coincide with the team problem. However, this analysis could provide the bounds on the gap and an acceptable practical design.

We use uppercase letters for random variables and lowercase for their realizations. We use notation $a_{t:t'}$ to represent vector $(a_t, a_{t+1}, \dots a_{t'})$ when $t' \geq t$ or an empty vector if $t' < t$. We denote the indicator function of any set A by $I_A(\cdot)$. For any finite set $\mathcal{S}$, $\mathcal{P}(\mathcal{S})$ represents space of probability measures on $\mathcal{S}$ and $|\mathcal{S}|$ represents its cardinality. We represent the set of real numbers by $\mathbb{R}$. We denote by P^g (or E^g) the probability measure generated by (or expectation with respect to) strategy profile g. All equalities and inequalities involving random variables are to be interpreted in *a.s.* sense. We use the terms users and buyers interchangeably.

The paper is structured as follows. In section II, we present the model. In section III, we formulate the team problem as an instance of decentralized stochastic control and characterize its optimal policies. In section IV, we consider the case with strategic users and design incentives

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48105 USA e-mail: {dvasal, vgsubram, anastas} at umich.edu

for the users to align their objective with team objective. We conclude in section V.

II. Model

We consider a discrete-time dynamical system over infinite horizon. There is a product whose value varies over time as (a slowly varying) discrete time Markov process $(X_t)_t$, where X_t takes value in the set $\{0,1\}$; 0 represents that product was bad (has low intrinsic value) and 1 represents and product is good (has high intrinsic value).

$$P(x_1) = \hat{Q}(x_1) \tag{1a}$$

$$P(x_t|x_{1:t-1}) = Q_x(x_t|x_{t-1}), \tag{1b}$$

such that $Q_x(x_t|x_{t-1}) = \epsilon$ if $x_t \neq x_{t-1}$, for $0 < \epsilon < 1$.

There are countably infinite number of exogenously selected, selfish buyers that act sequentially and exactly once in the process. Buyer t makes a noisy observation of the value of the product at time t, $v_t \in \mathcal{V} \overset{\triangle}{=} \{0,1\}$, through a binary symmetric channel with crossover probability p such that these observations are conditionally independent across users given the system state (i.e. noise is i.i.d.) i.e. $P(v_t|x_{1:t}v_{1:t-1}) = Q_v(v_t|x_t) = p$ if $v_t \neq x_t$. Based on actions of previous buyers and its private observation buyer t takes two actions: $a_t \in \mathcal{A} \overset{\triangle}{=} \{0,1\}$, which correspond to either buying or not buying the good, and $b_t \in \mathcal{B} \overset{\triangle}{=} \{*,1\}$ where * represents not reporting its observation and 1 represent reporting truthfully. Based on these actions and the state of the system, the buyer gets reward $R(x_t,a_t,b_t)$ where

$$\begin{aligned} &R(x_t,a_t,b_t) \\ &= -c\cdot I(b_t=1) + \begin{cases} 1/2, & x_t=1, a_t=1 \\ -1/2, & x_t=0, a_t=1 \\ 0, & a_t=0 \end{cases}, \end{aligned} \tag{2}$$

where c is cost of reporting its observation truthfully. The actions are publicly observed by future buyers whereas the observations $(v_t)_t$ are private information of the buyers.

III. Team problem

In this section we study the team problem where the buyers are cooperative and want to maximize the expected average reward per unit time for the team. At time t, buyer t's information consists of its private information v_t and publicly available information $a_{1:t-1}, b_{1:t-1}$. It takes action a_t, b_t though a (deterministic) policy $g_t : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \times \mathcal{V} \to \mathcal{A} \times \mathcal{B}$ as

$$(a_t,b_t) = g_t(a_{1:t-1}, b_{1:t-1}, v_t). \tag{3}$$

The objective as a team (or for a social planner) is to maximize the expected average reward per unit time for all the users i.e.

$$J \overset{\triangle}{=} \sup_g \limsup_{\tau\to\infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}^g\{R(X_t,A_t,B_t)\}. \tag{4}$$

Since the decision makers (i.e. the buyers) have different information sets, this is an instance of a decentralized stochastic control problem. We use techniques developed in [5] to find structural properties of the optimal policies. Specifically, we equivalently view the system through the perspective of a common agent that observes at time t, the common information $a_{1:t-1}, b_{1:t-1}$ and takes action $\gamma_t : \mathcal{V} \to \mathcal{A}\times\mathcal{B}$, which is a partial function that, when acted upon buyer's private information v_t, generates its action (a_t, b_t). The common agent's actions $(\gamma_t)_t$ are taken through common agent's strategy $\psi = (\psi)_t$ as $\gamma_t = \psi_t[a_{1:t-1}, b_{1:t-1}]$ where $\psi_t : \mathcal{A}^{t-1} \times \times \mathcal{B}^{t-1} \to (\mathcal{V} \to \mathcal{A}\times\mathcal{B})$. The corresponding common agent's problem is

$$J^c \overset{\triangle}{=} \sup_\psi \limsup_{\tau\to\infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}^\psi\{R(X_t,A_t,B_t)\}. \tag{5}$$

This procedure transforms the original decentralized stochastic control problem of buyers to a centralized stochastic control problem of the common agent. Thus an optimal policy of common agent can be translated to optimal policy for the buyers. In order to characterize common agent's optimal policies, we find an information state for the common agent's problem. We define a belief state π_t at time t as a probability measure on current state of the system given the common information i.e. $\pi_t(x_t) \overset{\triangle}{=} P^\psi(x_t|a_{1:t-1}b_{1:t-1}\gamma_{1:t})$. The following lemma shows that the common agent faces a Markov decision problem (MDP).

Lemma 1: $(\Pi_t, \Gamma_t)_t$ is a controlled Markov process with state Π_t and action Γ_t such that

$$P^\psi(\pi_{t+1}|\pi_{1:t}\gamma_{1:t}) = P(\pi_{t+1}|\pi_t\gamma_t) \tag{6a}$$

$$\begin{aligned} &\mathbb{E}^\psi\{R(X_t,A_t,B_t)|a_{1:t-1}b_{1:t-1}\gamma_{1:t}\} \\ =&\mathbb{E}\{R(X_t,A_t,B_t)|\pi_t\gamma_t\} \qquad (6b)\\ =&:\hat{R}(\pi_t,\gamma_t) \qquad (6c) \end{aligned}$$

and there exists an update function F, independent of ψ such that $\pi_{t+1} = F(\pi_t, \gamma_t, a_t, b_t)$.

Proof: See Appendinx. ■

Lemma 1 implies that for common agent's problem, it can summarize the common information $a_{1:t-1}, b_{1:t-1}$ in the belief state π_t. Furthermore there exists an optimal policy for the common agent of the form $\theta_t : \mathcal{P}(\mathcal{X}) \to (\mathcal{V} \to \mathcal{A}\times\mathcal{B})$ that can be found as solution of the following dynamic programming equation in the space of public beliefs π_t as, $\forall \pi, \gamma^* = \theta[\pi]$ is the maximizer in the following equation

$$\rho + V(\pi) = \max_\gamma \ \hat{R}(\pi,\gamma) + \mathbb{E}\{V(\Pi')|\pi\gamma\}, \tag{7}$$

where the distribution of π' is given through the kernel $P(\cdot|\pi\gamma)$ in (6a) and $\rho \in \mathbb{R}, V : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ are solution of the above fixed point equation. Based on this public belief π_t and its private information x_t, each user t takes actions as

$$(a_t,b_t) = m_t(\pi_t,v_t) = \theta_t[\pi_t](v_t). \tag{8}$$

We note that since states, actions and observations belong to a binary set, there are sixteen partial functions γ possible that are shown in Table I below where $\gamma = \begin{bmatrix}\gamma(v_t=0)\\ \gamma(v_t=1)\end{bmatrix} =$

$\begin{bmatrix} a_t, b_t(v_t = 0) \\ a_t, b_t(v_t = 1) \end{bmatrix}$. Since the common belief is updated as $\pi_{t+1} = F(\pi_t, \gamma, \gamma(v_t))$ and v_t is binary valued, there exist two types of γ functions: learning (γ^L) and non-learning (γ^{NL}). γ^L leads to update of belief through $F(\cdot)$ in (6a) that is informative of the private observation v_t, whereas γ^{NL} leads to uninformative update of belief. Eight of them are dominated in reward for example v_t need not be reported if it is revealed through a_t, or if it can be revealed indirectly by absence of reporting.

γ^L	$\begin{bmatrix}0,*\\1,*\end{bmatrix}$	$\begin{bmatrix}1,*\\0,*\end{bmatrix}$	$\begin{bmatrix}1,1\\1,*\end{bmatrix}$	$\begin{bmatrix}1,*\\1,1\end{bmatrix}$	$\begin{bmatrix}0,1\\0,*\end{bmatrix}$	$\begin{bmatrix}0,*\\0,1\end{bmatrix}$
	$\xcancel{\begin{bmatrix}0,1\\1,1\end{bmatrix}}$	$\xcancel{\begin{bmatrix}1,1\\0,1\end{bmatrix}}$	$\xcancel{\begin{bmatrix}0,1\\1,*\end{bmatrix}}$	$\xcancel{\begin{bmatrix}1,1\\0,*\end{bmatrix}}$	$\xcancel{\begin{bmatrix}0,*\\1,1\end{bmatrix}}$	$\xcancel{\begin{bmatrix}1,*\\0,1\end{bmatrix}}$
	$\xcancel{\begin{bmatrix}0,1\\0,1\end{bmatrix}}$	$\xcancel{\begin{bmatrix}1,1\\1,1\end{bmatrix}}$				
γ^{NL}	$\begin{bmatrix}0,*\\0,*\end{bmatrix}$	$\begin{bmatrix}1,*\\1,*\end{bmatrix}$				

TABLE I

IV. Game problem

We now consider the case when the buyers are strategic. As before, buyer t observes public history $a_{1:t-1}, b_{1:t-1}$ and its private observation v_t and thus takes its actions as $(a_t, b_t) = g_t(a_{1:t-1}, b_{1:t-1}, v_t)$. Its objective is to maximize its expected reward

$$J_t = \max_{g_t} \mathbb{E}^g\{R(X_t, A_t, B_t)\}. \tag{9}$$

Since all buyers have different information, this defines a dynamic game with asymmetric information. An appropriate solution concept is Perfect Bayesian Equilibrium (PBE) [6] that requires specification of an assessment $(g_t^*, \mu_t^*)_t$ of strategy and belief profile where g_t^* is the strategy of buyer t, $g_t^* : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \times \mathcal{V} \to \mathcal{P}(\mathcal{A} \times \mathcal{B})$, and μ_t^* is a belief as a function of buyer t's history on the random variables not observed by it till time t i.e. $\mu_t^* : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \times \mathcal{V} \to \mathcal{P}(\mathcal{X}^t \times \mathcal{V}^t)$. In general, finding a PBE is hard [6] since it involves solving a fixed point equation in strategies and beliefs that are function of histories although there are few cases where there exists an algorithm to find them [7], [8]. For this problem, since users act exactly once in the game and are thus myopic, it can be found easily in a forward inductive way, as in [1], [2]. Moreover, a belief on X_t, $\mu_t^*(x) \triangleq P^{g^*}(X_t = x | a^{t-1}, b^{t-1}, v_t), x \in \{0, 1\}$ is sufficient and any joint belief consistent with $\mu_t^*(x)$ along with equilibrium strategy profile g^* constitute a PBE. For any history, users compute a belief equilibrium strategy depending on v_t and π_t as

$$\gamma_t^* = \phi[\pi_t] = \arg\max_{\gamma_t} \hat{R}(\pi_t, \gamma_t) \tag{10}$$

With $\phi[\cdot]$ defined through (10), for every history $(a_{1:t-1}, b_{1:t-1}, v_t)$, π_t is updated using forward recursion through $\pi_{t+1} = F(\pi_t, \phi(\pi_t), a_t, b_t)$ and equilibrium strategies are generated as $g_t^*(a_{1:t-1}, b_{1:t-1}, v_t) = \phi[\pi_t](v_t)$. Finally the beliefs μ_t^* can be easily derived from π_t and private information v_t through Bayes rule.

We numerically solve (7) using value iteration to find team optimal policy, shown in Figure 1, for parameters $p = 0.2, \epsilon = 0.001$ and $c = 0.05$. For the same parameters, Figure 2 shows optimal policy for a strategic user that solves (10).

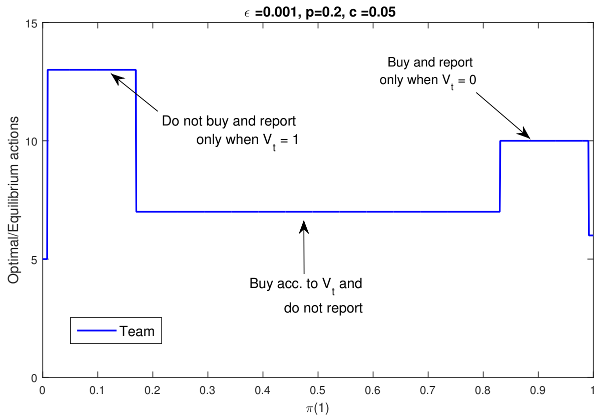

Fig. 1: Decentralized team optimal policy

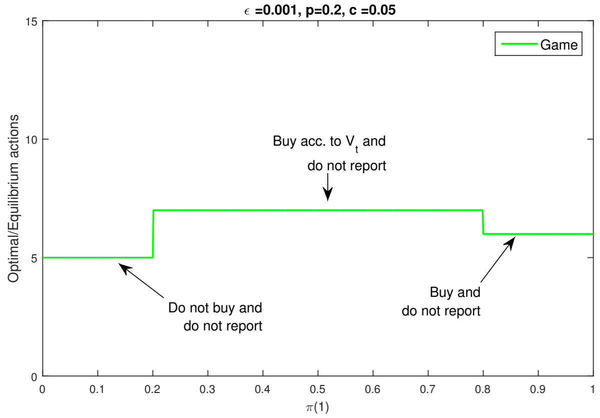

Fig. 2: Strategic optimal policy

A. *Incentive design for strategic users*

Our goal is to align each buyers' objective with the team objective. In order to do so, we introduce incentives (tax or subsidy) for user t, $t : \mathcal{P}(\mathcal{X}) \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ such that its effective reward is given by $\hat{R}(\pi_t, \gamma_t) - t(\pi_t, a_t, b_t)$.

We first note that a user can not internalize social reward through incentives as is done in a pivot mechanism [9]–[12], i.e. there does not exist an incentive mechanism such that the following equation could be true

$$\hat{R}(\pi, \gamma) - t(\pi, a, b) = \hat{R}(\pi, \gamma) + \mathbb{E}\{V(\Pi') | \pi\gamma\} \tag{11}$$

$$\text{i.e.} \qquad t(\pi, a, b) = -\mathbb{E}\{V(\Pi') | \pi\gamma\} \tag{12}$$

for $V(\cdot)$ defined in (7) and the distribution of π' is given through the kernel $P(\cdot | \pi\gamma)$ in (6a). The left side of (11) is buyers' effective reward and right side is the objective of the team problem as in (7). Such a design is not feasible because while $t(\cdot)$ can depend only on public observations (π, a, b), the second term in the RHS of (11) depends on γ as well which is not observed by the designer.

We observe in Figures 1, 2 that team optimal policy coincides with the strategic optimal policy for a significant range of $\pi(1)$. Let $\mathcal{S}$ be the set consisting of $\pi(1)$ where the team optimal policy coincides with the strategic optimal policy and $\mathcal{S}^c$ be the complement set. In order to align the two policies, we consider the following incentive design such that a user is paid c units by the system planner whenever the public belief $\pi(1)$ belongs to the set $\mathcal{S}^c$ and user reports its observation,

$$t(\pi, a_t, b_t) = -c \cdot I(\pi(1) \in \mathcal{S}) I(b_t = 1). \tag{13}$$

These payments are made after any report for enforcement purposes. This is agreed upon, i.e., system planner commits to this. With these incentives, the optimal policy of the strategic user is shown in Figure 3. Figure 4 compares the time average reward achieved through these policies, found through numerical results. This shows that the gap between the team objective and the one with incentives is small. Intuitively, this occurs because the buyers learn the true state of the system relatively quickly (exponentially fast) compared to the expected time spent by the Markov process X_t in any state. Equivalently, the time spent by the process $(\Pi_t(1))_t$ in the set $\mathcal{S}^c$ is small. Yet it is crucial for the social objective that learning occurs in this region. Also in Figure 4, the gap between the mechanism (including incentives) and the mechanism where incentives are subtracted signifies the expected average payment made by the designer, which is relatively small.

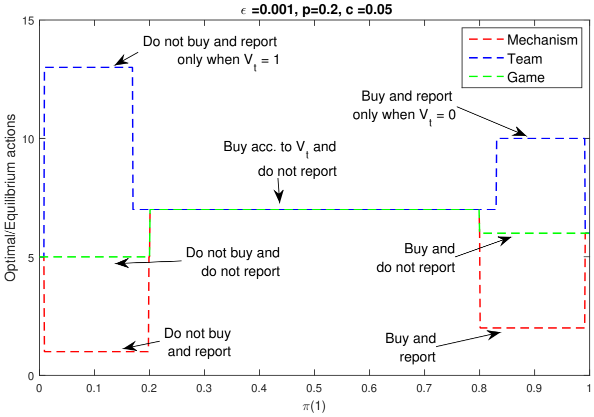

Fig. 3: Strategic optimal policy with incentives

V. Conclusion

We considered a sequential buyers game where a countable number of strategic buyers buy a product exactly once in the game. The value of the product is modeled as a Markov process and buyers privately make noisy observation of the value. We model the team problem as an instance of decentralized stochastic control problem and characterize structure of optimum policies. When users are strategic, it is modeled as a dynamic game with asymmetric information. We show that for some set $\pi_t \in \mathcal{S}$ that occurs with high probability, the strategic optimal policy coincides with the team optimal policy. Thus only outside this set, i.e., when $\pi_t \in \mathcal{S}^c$, buyers need to be incentivized to report their observations so that higher average rewards can be achieved for the whole team. Since numerically $\mathcal{S}^c$ occurs with low probability, the expected incentive payments are low. However, even though infrequent, these incentives help in the learning for the team as a whole, specifically for the future users. This suggests that using such a mechanism for the more general case could be a useful way to bridge the gap between strategic and team objectives.

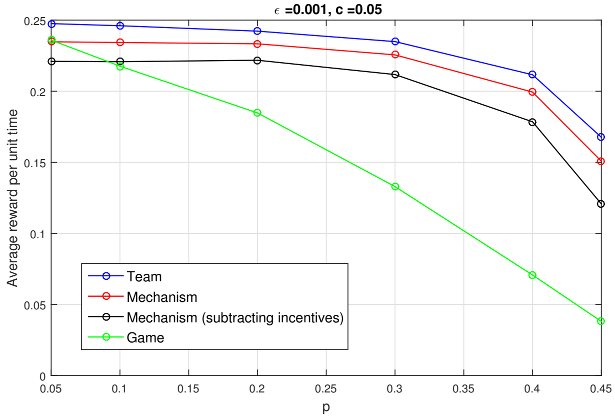

Fig. 4: Expected time average cost comparison for different policies

Future work involves characterizing team-optimum policies analytically and studying the resulting social utility through approximations or bounds on the induced Markov chain statistics. This would also characterize the gain from introducing "structured" incentives. Finally, incentives designs could be studied that minimize total expected incentives and guarantee voluntary participation.

Appendix

Claim 1: There exists an update function F, independent of ψ such that $\pi_{t+1} = F(\pi_t, \gamma_t, a_t, b_t)$.

Proof: Fix ψ

$$\pi_{t+1}(x_{t+1}) = P^{\psi}(x_{t+1} | a_{1:t} b_{1:t} \gamma_{1:t}) \tag{14a}$$

$$= \sum_{x_t} P^{\psi}(x_{t+1}, x_t | a_{1:t} b_{1:t} \gamma_{1:t}) \tag{14b}$$

$$= \sum_{x_t} P^{\psi}(x_t | a_{1:t} b_{1:t} \gamma_{1:t}) \hat{Q}(x_{t+1} | x_t) \tag{14c}$$

Now,

$$\begin{aligned}
&P^{\psi}(x_t|a_{1:t}b_{1:t}\gamma_{1:t}) \\
&= \frac{P^{\psi}(x_t, a_t, b_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t})}{\sum_{\hat{x}_t} P^{\psi}(\hat{x}_t, a_t, b_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t})} && \text{(15a)} \\
&= P^{\psi}(x_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t})\times \\
&\quad \frac{\sum_{v_t} P^{\psi}(a_t, b_t v_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t}, x_t)}{\sum_{\hat{x}_t} P(\hat{x}_t, a_t, b_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t})} && \text{(15b)} \\
&= \frac{P^{\psi}(x_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t-1})\sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t)}{\begin{array}{c}\sum_{\hat{x}_t} P^{\psi}(\hat{x}_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t-1}) \\ \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|\hat{x}_t)\end{array}} && \text{(15c)}
\end{aligned}$$

where first part in numerator in (15c) is true since given policy ψ, γ_t can be computed as $\gamma_t = \psi_t(a_{1:t-1}b_{1:t-1})$.

We conclude that

$$\begin{aligned}
&P(x_t|a_{1:t}, \gamma_{1:t}) \\
&= \frac{\pi_t(x_t) \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t)}{\sum_{\hat{x}_t} \pi_t(\hat{x}_t) \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|\hat{x}_t)}, && \text{(16)}
\end{aligned}$$

thus,

$$\pi_{t+1} = F(\pi_t, \gamma_t, a_t, b_t) \tag{17}$$

where F is independent of policy ψ. ■

Claim 2: $(\Pi_t, \Gamma_t)_t$ is a controlled Markov process with state Π_t and action Γ_t such that

$$P^{\psi}(\pi_{t+1}|\pi_{1:t}\gamma_{1:t}) = P(\pi_{t+1}|\pi_t\gamma_t) \tag{18}$$

$$\begin{aligned}
&\mathbb{E}^{\psi}\{R(X_t, A_t, B_t)|\gamma_{1:t}a_{1:t-1}b_{1:t-1}\} \\
&= \mathbb{E}\{R(X_t, A_t, B_t)|\gamma_t\pi_t\} && \text{(19)} \\
&=: \hat{R}(\pi_t, \gamma_t) && \text{(20)}
\end{aligned}$$

Proof:

$$\begin{aligned}
&P^{\psi}(\pi_{t+1}|\pi_{1:t}, \gamma_{1:t}) \\
&= \sum_{a_t, b_t} P^{\psi}(\pi_{t+1}, a_t, b_t|\pi_{1:t}, \gamma_{1:t}) && \text{(21a)} \\
&= \sum_{a_t, b_t} \mathbf{1}_{\{F(\pi_t, \gamma_t, a_t, b_t)\}}(\pi_{t+1}) \sum_{v_t} P^{\psi}(a_t, b_t v_t|\pi_{1:t}, \gamma_{1:t}) && \text{(21b)} \\
&= \sum_{a_t, b_t, x_t} \mathbf{1}_{\{F(\pi_t, \gamma_t, a_t, b_t)\}}(\pi_{t+1}) P^{\psi}(x_t|\pi_{1:t}, \gamma_{1:t}) \\
&\qquad \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t) && \text{(21c)} \\
&= \sum_{a_t, b_t, x_t} \pi_t(x_t)\mathbf{1}_{\{F(\pi_t, \gamma_t, a_t, b_t)\}}(\pi_{t+1}) \\
&\qquad \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t) && \text{(21d)} \\
&= P(\pi_{t+1}|\pi_t, \gamma_t) && \text{(21e)}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}(R(X_t, A_t, B_t)|\pi_{1:t}, \gamma_{1:t}) \\
&= \sum_{x_t, a_t, b_t v_t} R(x_t, a_t, b_t)P(x_t, a_t, b_t, v_t|\pi_{1:t}, \gamma_{1:t}) && \text{(22a)} \\
&= \sum_{x_t, a_t, b_t} R(x_t, a_t, b_t)P(x_t|\pi_{1:t}, \gamma_{1:t}) \\
&\qquad \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t) && \text{(22b)} \\
&= \sum_{x_t, a_t, b_t} R(x_t, a_t, b_t)\pi_t(x_t) \\
&\qquad \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t) && \text{(22c)} \\
&= \hat{R}(\pi_t, \gamma_t) && \text{(22d)}
\end{aligned}$$

■

References

[1] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of Political Economy*, vol. 100, no. 5, pp. pp. 992–1026, 1992. [Online]. Available: http://www.jstor.org/stable/2138632

[2] L. Smith and P. Sörensen, "Pathological outcomes of observational learning," *Econometrica*, vol. 68, no. 2, pp. 371–398, 2000. [Online]. Available: http://dx.doi.org/10.1111/1468-0262.00113

[3] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011. [Online]. Available: http://restud.oxfordjournals.org/content/78/4/1201.abstract

[4] T. N. Le, V. Subramanian, and R. Berry, "The impact of observation and action errors on informational cascades," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, Dec 2014, pp. 1917–1922.

[5] A. Nayyar, A. Mahajan, and D. Teneketzis, "Optimal control strategies in delayed sharing information structures," *IEEE Trans. Automatic Control*, vol. 56, no. 7, pp. 1606–1620, July 2011.

[6] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*, ser. MIT Press Books. The MIT Press, 1994, vol. 1.

[7] A. Nayyar, A. Gupta, C. Langbort, and T. Başar, "Common information based markov perfect equilibria for stochastic games with asymmetric information: Finite games," *IEEE Trans. Automatic Control*, vol. 59, no. 3, pp. 555–570, March 2014.

[8] D. Vasal, V. Subramanian, and A. Anastasopoulos, "A systematic process for evaluating structured perfect Bayesian equilibria in dynamic games with asymmetric information," Tech. Rep., Aug. 2015. [Online]. Available: http://arxiv.org/abs/1508.06269

[9] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *The Journal of finance*, vol. 16, no. 1, pp. 8–37, 1961.

[10] E. H. Clarke, "Multipart pricing of public goods," *Public choice*, vol. 11, no. 1, pp. 17–33, 1971.

[11] T. Groves, "Incentives in teams," *Econometrica: Journal of the Econometric Society*, pp. 617–631, 1973.

[12] D. Bergemann and J. Valimaki, "The dynamic pivot mechanism," *Econometrica*, vol. 78, no. 2, pp. 771–789, mar 2010.