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 $(a_t, a_{t+1}, \ldots, a_{t'})$ 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 $S$, $\mathcal{P}(S)$ represents space of probability measures on $S$ and $|S|$ 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.
for the users to align their objective with team objective. We conclude in section IV.

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).

\[
P(x_t) = Q(x_t) \quad \text{(1a)}
\]

\[
P(x_t| x_{t-1}) = Q_x(x_t|x_{t-1}), \quad \text{(1b)}
\]

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 \( t \), \( v_t \in \mathcal{V} \subseteq \{0, 1\} \), 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.

\[
P(v_t|x_t,v_{t-1}) = Q_v(v_t|x_t) = p \text{ if } v_t \neq x_t.
\]

Based on actions of previous buyers and its private observation buyer \( t \) takes two actions: \( a_t \in \mathcal{A} \subseteq \{0, 1\} \), which correspond to either buying or not buying the good, and \( b_t \in \mathcal{B} \subseteq \{*\} \) 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

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

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 \( a_t, b_t \) through a (deterministic) policy \( g_t: \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \times \mathcal{V} \to \mathcal{A} \times \mathcal{B} \) as

\[
(a_t, b_t) = g_t(a_{1:t-1}, b_{1:t-1}, v_t). \quad (3)
\]

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.

\[
J \triangleq \sup_g \limsup_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}^g \{ R(X_t, A_t, B_t) \}. \quad (4)
\]

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 \( \psi = (\psi_t)_t \) as \( \gamma_t = \psi_t(a_{1:t-1}, b_{1:t-1}) \) where \( \psi_t: \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \to (\mathcal{V} \to \mathcal{A} \times \mathcal{B}) \). The corresponding common agent’s problem is

\[
J_c \triangleq \sup_{\psi} \limsup_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}^\psi \{ R(X_t, A_t, B_t) \}. \quad (5)
\]

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 \( \pi_t \) at time \( t \) as a probability measure on current state of the system given the common information i.e. \( \pi_t(x_t) = P^\psi(x_t|a_{1:t-1}b_{1:t-1}\gamma_{1:t}) \).

The following lemma shows that the common agent faces a Markov decision problem (MDP).

\textbf{Lemma 1:} \( (\Pi_t, \Gamma_t) \) is a controlled Markov process with state \( \Pi_t \) and action \( \Gamma_t \) such that

\[
P^\psi(\pi_{t+1}|\pi_t, \gamma_t) = P(\pi_{t+1}|\pi_t \gamma_t) \quad (6a)
\]

\[
\mathbb{E}^\psi\{ R(X_t, A_t, B_t)|a_{1:t-1}b_{1:t-1}\gamma_{1:t} \} = \mathbb{E}\{ R(X_t, A_t, B_t)|\pi_t \gamma_t \} \quad (6b)
\]

\[
\hat{R}(\pi_t, \gamma_t) \quad (6c)
\]

where \( \gamma = \theta[\pi] \) is the maximizer in the following equation

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

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

\[
(a_t, b_t) = m_t(\pi_t, v_t) = \theta_t[\pi_t](v_t). \quad (8)
\]

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

\[
\begin{array}{c|c|c}
\hline
\gamma(v_t) & 0 & 1 \\
\hline
\gamma(v_t = 0) & & \\
\hline
\gamma(v_t = 1) & & \\
\hline
\end{array}
\]
Finally the beliefs $\mu_t^*$ can be easily derived from $\pi_t$ 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).

### 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

$$J_t = \max_{\pi_t} \mathbb{E}^p \{ R(X_t, A_t, B_t) \}. \quad (9)$$

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 $(g_t^*, \mu_t^*)$ of strategy and belief profile where $g_t^*$ is the strategy of buyer $t$, $g_t^* : A^{t-1} \times B^{t-1} \times V \rightarrow \mathcal{P}(A \times B)$, and $\mu_t^*$ is a belief as a function of player $t$'s history on the random variables not observed by it till time $t$ i.e. $\mu_t^* : A^{t-1} \times B^{t-1} \times V \rightarrow \mathcal{P}(X^t \times V^t)$. 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 $X_t$, $\mu_t^*(x) \equiv \mathbb{P}^p(X_t = x | a^{t-1}, b^{t-1}, v_t), x \in \{0, 1\}$ 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 $\pi_t$ as

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

With $\phi[\cdot]$ defined through (10), for every history $(a_{1:t-1}, b_{1:t-1}, v_t)$, $\pi_t$ 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 $g_t^*(a_{1:t-1}, b_{1:t-1}, v_t) = \phi[\pi_t](v_t)$. 

### Table I

<table>
<thead>
<tr>
<th>$\gamma^L$</th>
<th>0, *</th>
<th>1, *</th>
<th>1, 1</th>
<th>1, *</th>
<th>0, 1</th>
<th>0, 1</th>
<th>0, 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 0</td>
<td>1, 0</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
</tr>
<tr>
<td>1, 0</td>
<td>1, 0</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
</tr>
<tr>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
</tr>
<tr>
<td>1, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
<td>0, 1</td>
</tr>
</tbody>
</table>

\[ a_t, b_t(v_t = 0) \]. Since the common belief is updated as $\pi_{t+1} = F(\pi_t, \gamma_t(v_t))$ and $v_t$ is binary valued, there exist two types of $\gamma$ functions: learning ($\gamma^L$) and non-learning ($\gamma^NL$). $\gamma^L$ leads to update of belief through $F(\cdot)$ in (9) that is informative of the private observation $v_t$, whereas $\gamma^NL$ 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.

### Fig. 1: Decentralized team optimal policy

\[ \hat{R}(\pi_t, \gamma_t) = R(\pi_t, \gamma_t) + E[V(\Pi')|\pi_t] \quad (11) \]

### Fig. 2: Strategic optimal policy

\[ t(\pi, a, b) = -E\{V(\Pi')|\pi_t\} \quad (12) \]

for $V(\cdot)$ defined in (7) and the distribution of $\pi'$ is given through the kernel $F(\cdot|\pi_t)$ in (9a). 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 \((\pi, a, b)\), the second term in the RHS of (11) depends on \( \gamma \) 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 \( S \) be the set consisting of \( \pi(1) \) where the team optimal policy coincides with the strategic optimal policy and \( S^c \) 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 \( S \) and user reports its observation,

\[
t(\pi, a_t, b_t) = -c \cdot I(\pi(1) \in S)I(b_t = 1).
\] (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 expected average payment made by the designer, which is the gap between the mechanism (including incentives) and the public belief \( \pi(1) \) as well as the objective of the game. The value of the product is modeled as a Markov chain statistics. This would also characterize the gain from approximations or bounds on the induced Markov chain statistics. This also characterizes 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 \( \psi \) such that \( \pi_{t+1} = F(\pi_t, \gamma_t, a_t, b_t) \).

**Proof:** Fix \( \psi \)

\[
\pi_{t+1}(x_{t+1}) = P^\psi(x_{t+1} | a_{t+1}; b_t, \gamma_{t+1}; x_t)
\] (14a)

\[
= \sum_{x_t} P^\psi(x_{t+1}, x_t | a_{t+1}; b_t, \gamma_{t+1}; x_t)
\] (14b)

\[
= \sum_{x_t} P^\psi(x_t | a_{t+1}; b_t, \gamma_{t+1}; x_t)Q(x_{t+1} | x_t)
\] (14c)
Now,

\[
P^\psi(x_t|a_{t-1}b_{t-1}\gamma_{t-1}) = \frac{P^\psi(x_t, a_t, b_t|a_{t-1}b_{t-1}\gamma_{t-1})}{\sum_{\hat{x}_t} P^\psi(\hat{x}_t, a_t, b_t|a_{t-1}b_{t-1}\gamma_{t-1})} \tag{15a}
\]

\[
P^\psi(x_t|a_{t-1}b_{t-1}\gamma_{t-1}) = \sum_{\hat{x}_t} P^\psi(\hat{x}_t, a_t, b_t|a_{t-1}b_{t-1}\gamma_{t-1}) \times \frac{\sum_{v_t} P^\psi(v_t|a_{t-1}b_{t-1}\gamma_{t-1}, x_t)}{\sum_{\hat{x}_t} \sum_{v_t} P^\psi(\hat{x}_t, a_t, b_t, v_t|a_{t-1}b_{t-1}\gamma_{t-1})} \tag{15b}
\]

\[
P^\psi(x_t|a_{t-1}b_{t-1}\gamma_{t-1}) = \frac{\sum_{v_t} \sum_{\hat{x}_t} P^\psi(\hat{x}_t, a_t, b_t, v_t|a_{t-1}b_{t-1}\gamma_{t-1}) I(\gamma_{t}(v_t)) (a_t, b_t) Q_v(v_t|x_t)}{\sum_{\hat{x}_t} \sum_{v_t} P^\psi(\hat{x}_t, a_t, b_t, v_t|a_{t-1}b_{t-1}\gamma_{t-1}) I(\gamma_{t}(v_t)) (a_t, b_t) Q_v(v_t|x_t)} \tag{15c}
\]

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

We conclude that

\[
P(\hat{x}_t|a_{t-1}\gamma_{t-1}) = \pi_t(x_t) \sum_{v_t} I(\gamma_{t}(v_t)) (a_t, b_t) Q_v(v_t|x_t) \tag{16}
\]

thus,

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

where \(F\) is independent of policy \(\psi\).

\section*{Claim 2: (\Pi_t, \Gamma_t)} is a controlled Markov process with state \(\Pi_t\) and action \(\Gamma_t\) such that

\[
P^\psi(\pi_{t+1}|\pi_{t:t-1}, \gamma_{t:t-1}) = P(\pi_{t+1}|\pi_{t:t-1}) \tag{18}
\]

\[
\mathbb{E}^\psi \{ R(X_t, A_t, B_t)|\gamma_{t:t-1}\} = \mathbb{E} \{ R(X_t, A_t, B_t)|\gamma_{t:t-1}\} =: \hat{R}(\pi_t, \gamma_t) \tag{19}
\]

\section*{Proof:}

\[
P^\psi(\pi_{t+1}|\pi_{t:t-1}, \gamma_{t:t-1}) = \sum_{a_{t:1}, b_{t:1}} P^\psi(\pi_{t+1}, a_t, b_t|\pi_{t:t-1}, \gamma_{t:t-1}) \tag{21a}
\]

\[
= \sum_{a_{t:1}, b_{t:1}} \sum_{v_t} I(\gamma_{t}(v_t)) (a_t, b_t) Q_v(v_t|x_t) \sum_{a_t, b_t} P^\psi(\pi_{t+1}, a_t, b_t|\pi_{t:t-1}, \gamma_{t:t-1}) \tag{21b}
\]

\[
= \sum_{a_{t:1}, b_{t:1}} \sum_{v_t} I(\gamma_{t}(v_t)) (a_t, b_t) Q_v(v_t|x_t) \sum_{a_t, b_t} \sum_{v_t} P^\psi(x_t|a_{t:t-1}, \gamma_{t:t-1}) \tag{21c}
\]

\[
= \sum_{a_{t:1}, b_{t:1}} \sum_{v_t} I(\gamma_{t}(v_t)) (a_t, b_t) Q_v(v_t|x_t) \sum_{a_t, b_t} \sum_{v_t} P(\pi_{t+1}|\pi_t, \gamma_t) \tag{21d}
\]

\[
= P(\pi_{t+1}|\pi_t, \gamma_t) \tag{21e}
\]

\section*{REFERENCES}


