Controlled Information Fusion with Risk-Averse CVaR Social Sensors

Sujay Bhatt and Vikram Krishnamurthy, *Fellow, IEEE*

***Abstract*— Consider a multi-agent network comprised of risk-averse social sensors and a controller that jointly seek to estimate the state of a Markov chain, given noisy measurements.**

The network of social sensors perform *Bayesian social learning* - each sensor fuses the information revealed by previous social sensors along with its private valuation using Bayes' rule - to optimize a local cost function. The controller sequentially modifies the cost function of the sensors by discriminatory pricing (control inputs) to realize long term global objectives.

We formulate the stochastic control problem faced by the controller as a Partially Observed Markov Decision Process (POMDP) and derive structural results for the optimal control policy as a function of the risk-aversion factor in the Conditional Value-at-Risk (CVaR) cost function of the sensors. We show that the optimal price sequence when the sensors are risk-averse is a *super-martingale*; i.e, it decreases on average over time.

***Index Terms*— Social Learning, Social Sensors, Monopoly Pricing, Structural Results, POMDP, Controlled Fusion, Risk-averse, CVaR**

I. INTRODUCTION

Social learning is the process by which social sensors are influenced by the behaviour of other sensors in a multi-agent network. A social sensor is an information processing system having the following attributes:

i.) It affects the behaviour of other sensors.
ii.) It shares quantized information (decisions/actions) and has its own dynamics.
iii.) It has limited processing capabilities - boundedness.
iv.) It is rational - fuses all available information using Bayes' rule to take optimal action.

Social learning shares similarities with distributed detection [1], and distributed hypothesis testing [2]. We present a model of Bayesian social learning with focus on controlled information fusion using exogenous inputs to the multi-agent network.

The contribution of this paper is two fold:

1) We study the interaction of a controller and a network of risk-averse social sensors, using a Partially Observed Markov Decision Process (POMDP) framework. The nature of the interaction is such that the controller can influence the information fusion in social sensors.
2) We obtain structural results for the optimal policy of the controller and characterize the properties of the optimal (price) sequence, when the social sensors are risk-averse.

[3] considers the interaction of a global controller and a network of social sensors, where the risk-neutral sensors perform social learning to estimate an underlying parameter and to optimize a local utility function. The objective of the controller is to detect a change in the parameter as soon as possible by observing the actions of the sensors. Well known inefficiencies of the standard social learning model [4], [5] like herding (sensors choose the same action irrespective of their private valuation) and informational cascades (information fusion results in no improvement in uncertainty) are shown to be the consequence of the belief state (probability distribution on the parameter) belonging to suitably defined regions in the belief space. Using this characterization, structural results on the optimal policy of the controller are derived. [6] extends the analysis to the case where the sensors display an aversion to risk; i.e, social sensors having risk aversion as an additional attribute. It is shown that when the sensors are risk-averse (modeled using CVaR), the herding behaviour is more pronounced - social learning is absent when the sensors are sufficiently risk-averse.

[7], [8] consider monopoly pricing in the presence of social learning and establish various properties of the value function and the optimal policy for the monopoly. It is shown that using discriminatory pricing the monopoly is able to delay the process of herding in risk-neutral social sensors, to suit its needs. The optimal price (control) sequence is shown to be a super-martingale. We consider monopoly pricing and social learning under CVaR risk-measure[1]; see [9] for an overview of risk measures.

CVaR is an extension of VaR that gives the total loss given a loss event, and is a coherent risk measure; see [10]. The value at risk (VaR) is the percentile loss namely, $\text{VaR}_\alpha(x) = \min\{z : F_x(z) \geq \alpha\}$ for cdf F_x, and $\text{CVaR}_\alpha(x) = \mathbb{E}\{X | X > \text{VaR}_\alpha(x)\}$. CVaR is one of the 'big' developments in risk modelling in the last 15 years. In this paper, we choose CVaR risk measure as it exhibits the following properties: (i) It associates higher risk with higher cost. (ii) It ensures that risk arises only from the services. (iii) It is convex.

S. Bhatt (sh2376@cornell.edu) and V. Krishnamurthy (vikramk@cornell.edu) are with the Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850.

This material is based upon work supported, in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under grant 12346080, National Science Foundation under grant 1714180, and Schmidt Sciences.

[1]A risk measure $\varrho : \mathcal{L} \to \mathbb{R}$ is a mapping from the space of measurable functions to the real line which satisfies the following properties: (i) $\varrho(0) = 0$. (ii) If $S_1, S_2 \in \mathcal{L}$ and $S_1 \leq S_2$ a.s then $\varrho(S_1) \leq \varrho(S_2)$. (iii) if $a \in \mathbb{R}$ and $S \in \mathcal{L}$, then $\varrho(S + a) = \varrho(S) + a$. The risk measure is coherent if in addition ϱ satisfies: (iv) If $S_1, S_2 \in \mathcal{L}$, then $\varrho(S_1 + S_2) \leq \varrho(S_1) + \varrho(S_2)$. (v) If $a \geq 0$ and $S \in \mathcal{L}$, then $\varrho(aS) = a\varrho(S)$. The expectation operator is a special case where subadditivity is replaced by additivity.

Organization and Main Results

Sec. II details the Bayesian social learning model for the process of information fusion by CVaR social sensors and the pricing protocol employed by the controller. The controller's long term objective is to maximize a discounted reward function.
Sec. III formulates the stochastic control problem faced by the controller as a Partially Observed Markov Decision Process (POMDP) and is solved using dynamic programming. The structure of the value function and optimal policy is completely characterized.
Sec. IV describes the nature of the price sequence that is input to the multi-agent network. It is shown that the controller prices high initially and subsequently lowers the price, i.e the price input sequence over time is a super-martingale.

II. CVaR Social Learning Model and Controller Objective

We consider the classical sequential social learning framework [3], [6], [11]. The social sensors and the controller jointly seek to estimate the underlying state of a Markov chain to meet the desired objectives. The controller sequentially chooses price inputs to the multi-agent network in exchange for services and the sensors decide to utilize the services depending on its quality. The decision (action) of each sensor depends on the cost, risk factor, private valuation and the decisions of the other sensors in the network. Amazon Web Services (AWS), for example, provides an on-demand cloud platform with a wide range of services like storage, developer tools, analytics etc for client-side applications at different prices. AWS is only partially aware of the *quality* of its services, while the clients learn about it from the experience of other clients and self-valuation.

Let $x \in \mathcal{X} = \{1(\text{Low}), 2(\text{High})\}$ denote the state and P denote the transition matrix of the Markov chain having an initial distribution (on the quality) as $\pi_0 = (\pi_0(i), i \in \mathcal{X})$, where $\pi_0(i) = \mathbb{P}(x_0 = i)$.

Each sensor acts once in a predetermined sequential order indexed by $k = 1, 2, \cdots$. The index k can also be viewed as the discrete time instant when sensor k acts. Each sensor k obtains noisy private valuations, $y_k \in \mathcal{Y} = \{1(\text{Low}), 2(\text{High})\}$, of the quality x_k and considers this in addition to the actions of its predecessors. The controller does not have any information about $x_k \in \mathcal{X}$ but infers it from the information revealed by the actions of the individual sensors, $a_k \in \mathcal{A} = \{1(\text{Don't Utilize}), 2(\text{Utilize})\}$, and chooses the price inputs[2] $u_k \in [0, 1]$ at each time k (or at each sensor k).

A. CVaR Social Learning Model and Pricing Protocol

The social learning model and the pricing protocol of the controller is as follows:

1. *Sensor's Private Observation*: Social sensor k's private observation denoted by $y_k \in \mathcal{Y} = \{1, 2\}$ is a noisy measurement of the true quality. It is obtained from the observation likelihood distribution as,
$$B_{ij} = \mathbb{P}(y_k = j | x_k = i) \tag{1}$$
The discreteness of the observation distribution captures the *boundedness* or the limited processing capabilities of the sensor.
2. *Social Learning and Private Belief update*: Social sensor k updates its private belief by fusion of the observation y_k and the prior public belief $\pi_{k-1}(i) = \mathbb{P}(x_k = i | a_1, \ldots, a_{k-1})$ as the following Hidden Markov Model (HMM) update
$$\eta_k^{y_k} = \frac{B_{y_k} P' \pi_{k-1}}{\mathbf{1}' B_{y_k} P' \pi_{k-1}} \tag{2}$$
where B_{y_k} denotes the diagonal matrix having $[\mathbb{P}(y_k|x_k = 1)\ \mathbb{P}(y_k|x_k = 2))]$ along the diagonal and $\mathbf{1}$ denotes the 2-dimensional vector of ones. HMM update is a consequence of Bayes' rule, information on the state conditioned on the new observation.
3. *Sensor's Action*: Social sensor k executes an action $a_k \in \mathcal{A} = \{1, 2\}$ to myopically minimize its cost. Let $c(x_k, a_k)$ denote the cost incurred if the sensor takes action a_k when the underlying state is x_k.
The form of the state-action dependent cost is taken as $c(x_k, a_k) = u_k - v(x_k)$ (see [11] for a justification), where v is the valuation of the services by each sensor and u_k is the price chosen by the controller at time k. It is assumed without loss of generality that
$$v(x_k) = \begin{cases} 0 & \text{if } x_k = 1; \\ 1 & \text{if } x_k = 2. \end{cases}$$
The state-action dependent costs for $x \in \mathcal{X}$ are thus given as:
$$c(x_k, a_k) = \begin{cases} \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \text{if } a_k = 1; \\ \begin{bmatrix} u_k \\ u_k - 1 \end{bmatrix} & \text{if } a_k = 2. \end{cases}$$
The sensor chooses an action a_k to minimize the CVaR measure as
$$\begin{aligned} a_k &= \underset{a \in \mathcal{A}}{\text{argmin}} \{\text{CVaR}_\alpha(c(x_k, a))\} \qquad (3) \\ &= \underset{a \in \mathcal{A}}{\text{argmin}} \{\min_{z \in \mathbb{R}} \{z + \frac{1}{\alpha} \mathbb{E}_{y_k}[\max\{(c(x_k, a) - z), 0\}]\}\} \end{aligned}$$
Here $\alpha \in (0, 1]$ reflects the degree of risk-aversion for the sensor (the smaller α is, the more risk-averse the sensor is). Note that when $\alpha = 1$, the cost function is the risk-neutral cost function as in [7], [8], [11]. Define
$$\mathcal{G}_k := \sigma\text{- algebra generated by } (u_1, a_1, u_2, a_2, \ldots, u_k, y_k) \tag{4}$$
$\mathbb{E}_{y_k}$ denotes the expectation with respect to private belief, i.e, $\mathbb{E}_{y_k} = \mathbb{E}[.|\mathcal{G}_k]$ when the private belief is updated after observation y_k.
4. *Controller Reward*: We consider two possible reward functions for the controller. The controller chooses one

[2]The range of prices chosen by the controller is normalized to $[0, 1]$ for convenience.

of the following reward functions at $k = 0$ and accrues the corresponding reward at each time k as

Case 1. Self-Interested: The controller accrues a reward when the sensors utilize the services,

$$r_{u_k} = (u_k - \beta)\mathcal{I}(a_k = 2|\pi_k). \tag{5}$$

Case 2. Altruistic: The controller accrues a reward when the sensors act according to their valuations,

$$r_{u_k} = (u_k - \beta)\mathcal{I}(a_k = y_k|\pi_k). \tag{6}$$

Here $\mathcal{I}$ denotes the indicator function and $\beta \in (0, 1)$ is a fixed[3] cost incurred by the controller. It could denote the cost of service. The controller being self-interested can be seen as profit maximizing, and being altruistic can be seen as social welfare maximizing[4].

5. *Public Belief update*: Sensor k's action is shared by the controller with the multi-agent network and the public belief on the quality is updated according to the social learning Bayesian filter (see [3], [6]) as follows

$$\pi_k = T^{\pi}(\pi_{k-1}, a_k) = \frac{R_{a_k}^{\pi_{k-1}} P' \pi_{k-1}}{\mathbf{1}' R_{a_k}^{\pi_{k-1}} P' \pi_{k-1}}. \tag{7}$$

Here, $\pi_k(i) = \mathbb{P}(x_k = i|a_1, \ldots, a_k)$, $R_{a_k}^{\pi_{k-1}} = \text{diag}(\mathbb{P}(a_k|x = i, \pi_{k-1}), i \in \mathcal{X})$, where $\mathbb{P}(a_k|x = i, \pi_{k-1}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(a_k|y, \pi_{k-1})\mathbb{P}(y|x_k = i)$ and

$$\mathbb{P}(a_k|y, \pi_{k-1}) = \begin{cases} 1 & \text{if } a_k = \underset{a \in \mathcal{A}}{\text{argmin}}\{\text{CVaR}_\alpha(c(x_k, a))\}; \\ 0 & \text{otherwise.} \end{cases}$$

Note that π_k belongs to the unit simplex

$$\Pi(2) \triangleq \{\pi \in \mathbb{R}^2 : \pi(1) + \pi(2) = 1, 0 \leq \pi(i) \leq 1 \text{ for } i \in \{1, 2\}\}$$

Here the expectation in CVaR measure is with respect to the sigma-algebra $\mathcal{G}_k$ with $y_k = y$. Social learning filter update is a consequence of Bayes' rule, information on the state conditioned on the new action.

6. *Fusion Price*: Let the history recorded by the controller and the multi-agent network be denoted as $\mathcal{H}_k = \{\pi_0, u_1, a_1, \cdots, u_k, a_k\}$. The controller chooses $u_{k+1} = \mu_{k+1}(\mathcal{H}_k) \in [0, 1]$ for the sensor $k+1$ and the protocol is repeated for all the sensors in the system. Here μ_{k+1} denotes the pricing policy at time $k + 1$.

Fig. 1 shows the CVaR social learning model.

B. Controlled Fusion Objective

The controller chooses the price inputs to the social sensors sequentially as

$$u_k = \mu_k(\mathcal{H}_{k-1}) \in [0, 1] \tag{8}$$

[3]Note that β could be made state dependent without affecting the nature of the results in the paper. Here it is assumed to be independent of the state for simplicity.

[4]Social welfare is maximized when the controller and the sensors in the multi-agent network take decisions considering network externalities; see (Chapter 4, [11]). We shall see in Sec. IV that $\mathcal{I}(a = y)$ improves the value of information fused by the successive sensors, thereby promoting welfare.

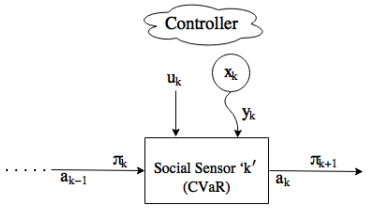


Fig. 1: CVaR Social Learning model. The social sensor k receives the public belief π_{k-1} from all its predecessors. y_k denotes the private valuation of the quality and u_k denotes the price charged by the controller for the services. The decision a_k is shared by the controller and the updated public belief π_{k+1} is received by the successive sensors.

where $\mathcal{H}_k = \{\pi_0, u_1, a_1, \cdots, u_{k-1}, a_{k-1}\}$. Since $\mathcal{H}_k$ is increasing with time k, to implement a controller, it is useful to obtain a sufficient statistic that does not grow in dimension. The public belief π_{k-1} computed via the social learning filter (7) forms a sufficient statistic for $\mathcal{H}_k$ and[5] (8) can be written as

$$u_k = \mu_k(\pi_{k-1}). \tag{9}$$

The optimal pricing policy μ^* is the stationary policy that maximizes the cumulative discounted reward

$$J_\mu(\pi) = \mathbb{E}_\mu\{\sum_{k=1}^{\infty} \rho^k r_{u_k}|\pi_0 = \pi\}. \tag{10}$$

Here $u_k = \mu(\pi_{k-1})$ and $\rho \in [0, 1)$ denotes the economic discount factor indicating the degree of impatience of the controller. In (10), the controller seeks to find the optimal policy μ^* such that

$$J_{\mu^*}(\pi_0) = \sup_{\mu \in \boldsymbol{\mu}} J_\mu(\pi_0). \tag{11}$$

The stochastic control problem faced by the controller was formulated as a partially observed Markov decision process (POMDP) with dynamics (7) and objective (11), and is solved using dynamic programming.

III. Structure of Optimal Pricing Policy

In this section, we characterize the nature of the optimal pricing policy for (10). It is shown that due to the structure of the social learning filter in (7), the choice of price inputs reduces from a continuum to a finite number at every belief.

Assumptions:

(A1) The observation distribution $B_{xy} = \mathbb{P}(y|x)$ is TP2(total positive of order 2), i.e, the determinant of the matrix B is non-negative; see [12].

(A2) The transition probability matrix P is TP2.

Lemma 1 ([12]). *Let η^y denote the private belief update (2) with a public prior belief π. Under (A1), η^y is increasing[6] in y, i.e, $\eta^{y=1}(1) > \eta^{y=2}(1)$.* □

[5]The rewards are a function of the price and the state (see Lemma 8 and Lemma 9), and hence restriction to Markov policies is without loss of generality.

[6]$\pi_2 \geq \pi_1$ if the determinant

$$\begin{vmatrix} \pi_1(1) & \pi_1(2) \\ \pi_2(1) & \pi_2(2) \end{vmatrix} \geq 0$$

Theorem 2 ([12]). *Let the instantaneous rewards be non-decreasing in π. Under (A1) and (A2), the value function $V(\pi)$ with finite number of actions at every belief, is monotone and convex.*

The proof of monotonicity involves using the assumption that the instantaneous reward is increasing in π to establish that the function $Q(\pi, u)$ is increasing in π by inductive arguments. The proof of convexity follows by establishing the convexity of the instantaneous rewards with respect to π.

The optimal policy μ^* and the value function $V(\pi)$ for the POMDP satisfy the Bellman's dynamic programming equation

$$\begin{aligned} Q(\pi, u) &= r_u + \rho \sum_{a \in \mathcal{A}} V(T^{\pi}(\pi, a))\sigma(\pi, a), \\ \mu^*(\pi) &= \underset{u \in [0,1]}{\operatorname{argmax}}\, Q(\pi, u), \\ V(\pi) &= \max_{u \in [0,1]} Q(\pi, u), \quad J_{\mu^*}(\pi_0) = V(\pi_0). \end{aligned} \tag{12}$$

Theorem 3. *Given a risk-aversion factor $\alpha \in (0, 1]$, let $u^H(\pi) = 1 - \frac{\eta^{y=2}(1)}{\alpha}$ and $u^L(\pi) = 1 - \frac{\eta^{y=1}(1)}{\alpha}$ denote two possible prices at the belief π. Under (A1) and (A2), the Q function (12) can be simplified for the rewards (5) and (6) as*
Case 1.) Self-Interested:

$$Q(\pi, u) = \begin{cases} (u - \beta) + \rho V(P'\pi) & \text{if } u \in [0, u^L(\pi)]; \\ (u - \beta) \times \mathbf{1}' B_{y=2}\pi & \\ +\rho \mathbb{E}V(\pi) & \text{if } u \in (u^L(\pi), u^H(\pi)]; \\ 0 & \text{otherwise}. \end{cases}$$

and $V(\pi) = \max Q(\pi, u)$, where $V(\pi) \geq 0$.
Case 2.) Altruistic:

$$Q(\pi, u) = \begin{cases} (u - \beta) + \rho \mathbb{E}V(\pi) & \text{if } u \in (u^L(\pi), u^H(\pi)]; \\ 0 & \text{otherwise}. \end{cases}$$

and $V(\pi) = \max Q(\pi, u)$, where $V(\pi) \geq 0$.

Here,

$$\mathbb{E}V(\pi) = \mathbf{1}' B^{\pi}_{y=1} P'\pi \times V(\eta^{y=1}) + \mathbf{1}' B^{\pi}_{y=2} P'\pi \times V(\eta^{y=2}).$$

The prices $u^H(\pi)$ and $u^L(\pi)$ are such that sensors utilize the services when $y = 2$ and $y = \{1, 2\}$ respectively. The proof of Theorem 3 will be given in the appendix. Theorem 3 represents the Q function (12) over a price input range $[0, 1]$ for rewards (5) and (6) respectively, in *three* and *two* regions. The following corollaries highlight why such partitions are useful.

Corollary 4. *Let the controller reward be given by (5). At every public belief $\pi \in \Pi(2)$, it is sufficient to choose one of the three prices $\{u^L(\pi), u^H(\pi), u^H(\pi) + \epsilon\}$ for any $\epsilon > 0$.* □

Corollary 5. *Let the controller reward be given by (6). At every public belief $\pi \in \Pi(2)$, it is sufficient to choose one of the two prices $\{u^H(\pi), u^H(\pi) + \epsilon\}$ for any $\epsilon > 0$.* □

The Q function and hence the value function $V(\pi)$ is monotone from Theorem 2. The following theorem completely characterizes the optimal pricing policy when the controller aims to maximize the reward. The proof is given in the appendix.

Theorem 6. *Let $P = I$, i.e, the quality is a random variable. For every public belief $\pi \in \Pi(2)$ and an $\epsilon > 0$, the optimal policy $\mu^*(\pi) = \operatorname{argmax}_u Q(\pi, u)$ is given as*
Case 1.) Self-Interested:

$$\mu^*(\pi) = \begin{cases} u^H(\pi) + \epsilon & \text{if } \pi(2) \in [0, \pi^*(2)); \\ u^H(\pi) & \text{if } \pi(2) \in [\pi^*(2), \pi^{**}(2)); \\ u^L(\pi) & \pi(2) \in [\pi^{**}(2), 1]. \end{cases} \tag{13}$$

for $\pi^*(2),\ \pi^{**}(2) \in [0, 1]$.
Case 2.) Altruistic:

$$\mu^*(\pi) = \begin{cases} u^H(\pi) + \epsilon & \text{if } \pi(2) \in [0, \hat{\pi}^*(2)); \\ u^H(\pi) & \text{if } \pi(2) \in [\hat{\pi}^*(2), 1]. \end{cases} \tag{14}$$

for $\hat{\pi}^*(2) \in (0, 1)$.

Discussion: From Theorem 3, Corollary 4 and Corollary 5, the value function in (12) can be represented as

(Self-Interested)

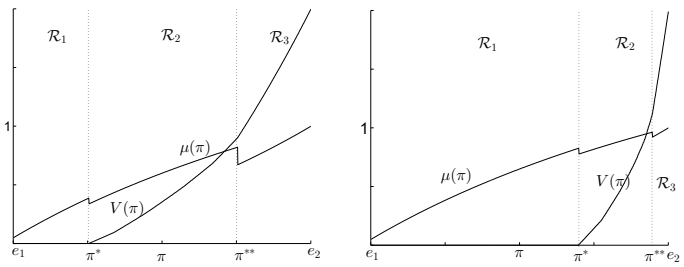
$$\begin{aligned} V(\pi) = \max\{&0, (u^L(\pi) - \beta) + \rho V(P'\pi), \\ &(u^H(\pi) - \beta) \times \mathbf{1}' B_{y=2}\pi + \rho \mathbb{E}V(\pi)\} \end{aligned}$$

(Altruistic)

$$V(\pi) = \max\{0, (u^H(\pi) - \beta) + \rho \mathbb{E}V(\pi)\}$$

The key takeaway is that due to the structure of the social learning filter, the choice of price inputs at every belief is reduced to a finite number of values instead of the range $[0, 1]$. Characterizing the optimal policy amounts to selecting among the these price inputs as a function of the public belief. Theorem 6 completely determines the regions in the belief space $\Pi(2)$ where it is optimal to choose a particular price input, when $P = I$. Fig. 2a and Fig. 2b show the value function and the optimal policy for two different risk-aversion factors (α) in a simple numerical example.

Let $\mathcal{R}_1,\ \mathcal{R}_2,\ \mathcal{R}_3$ denote the three regions determined by Theorem 6 where $u^H(\pi)+\epsilon,\ u^H(\pi),\ u^L(\pi)$ respectively are optimal. $\mathcal{R}_1$ is the *cut-off* region - the controller terminates the services to the multi-agent network. In the *Self-Interested* case, the price inputs are such that no sensor has an incentive to utilize the services. $\mathcal{R}_2$ is the *social learning* region - the sensors act according to their private valuations. The price inputs are such that the sensor having a high valuation $y = 2$ will utilize the services, while the sensor having low valuation $y = 1$ finds it prohibitive. Since the sensors act according to their valuation, sensor deciding at a future instant can successfully infer the private valuation of its predecessors; in other words, the information fusion reduces uncertainty about the quality of the service. $\mathcal{R}_3$ is the *herding* region - every sensor utilizes the services. The controller chooses a low input $u^L(\pi)(< u^H(\pi))$, which prompts the sensor with even a low valuation $y = 1$ to utilize the service.



(a) Risk-aversion factor $\alpha = 0.9$ (b) Risk-aversion factor $\alpha = 0.3$

Fig. 2: Value function and optimal pricing policy of the controller in the *self-interested* case. $P = I$, $B = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$, the discount factor $\rho = 0.7$, and $\mathcal{R}_1 - [0, \pi^*)$, $\mathcal{R}_2 - [\pi^*, \pi^{**})$, and $\mathcal{R}_3 - [\pi^{**}, 1]$ are the cut-off, social learning and herding regions respectively. It can be seen that the width of $\mathcal{R}_1$ increases with increased aversion to risk. This is equivalent to saying that risk-averse sensors that show an increased aversion to risk, choose to utilize the services only when they are reasonably certain about the quality. So it is profitable to the controller if it offers services only when it believes that the quality is high.

Notice that when the controller chooses $u = u^L(\pi)$ (when $\pi(2) \in [\pi^{**}(2), 1]$), the value function is $V(\pi) = \frac{(u^L(\pi) - \beta)}{(1-\alpha)}$ - a fixed payoff. This means the controller induces a herd (sensors choose the same action irrespective of their private valuation) that leads to an information cascade (information fusion results in no improvement in uncertainty) - public belief is frozen.

In the *Altruistic* case, the price inputs (two at every belief) are chosen so as to encourage the sensors to act according to their valuations. This implies that the controller chooses inputs to maximize the width of the *social learning* region $\mathcal{R}_2$. The *herding* region $\mathcal{R}_3$ is absent as $u = u^L(\pi)$ is not chosen by the controller. The *cut-off* region indicates the flexibility to terminate the services when the expected valuation is less than the cost of service.

IV. Properties of the Optimal Price Sequence

In this section, we describe the relation between the optimal policy (13) and (14), and the price sequence $u_k = \mu^*(\pi_k)$.

Theorem 7. *Let $\mathcal{F}_k$ be the σ-algebra generated by $(u_1, a_1, u_2, a_2, \ldots, u_{k-1}, a_{k-1}, u_k, a_k)$, where π_0 is the initial belief. The optimal price sequence $u_k = \mu^*(\pi_{k-1})$ is a super-martingale when the quality is a random variable, i.e, $P = I$ for any $\alpha \in (0, 1]$.*

Proof. Let $P = I$. The public belief π_k is a martingale for $P = I$, i.e, $\mathbb{E}[\pi_{k+1}|\mathcal{F}_k] = \pi_k$; see [8], [11].
It can easily be verified[7] that $u^H(\pi)$ is a concave function and $u^L(\pi)$ is a convex function of π for $\alpha \in (0, 1]$.

i.) *Self-Interested*: For $\epsilon \to 0$, we have for $\pi_k(2), \ \pi_{k+1}(2) \in [0, \pi^{**}(2))$, $u_k = u^H(\pi_k)$ and it satisfies $\mathbb{E}[u^H(\pi_{k+1})|\mathcal{F}_k] \leq u_k$ by Jensen's inequality.
We know that $u^L(\pi) \leq u^H(\pi)$ from Lemma 1. For the case of $\pi_k(2) \in [\pi^*(2), \pi^{**}(2))$ and $\pi_{k+1}(2) \in [\pi^{**}(2), 1]$, we have

$$\mathbb{E}[u_{k+1}|\mathcal{F}_k] = \mathbb{E}[u^L(\pi_{k+1})|\mathcal{F}_k] \leq \mathbb{E}[u^H(\pi_{k+1})|\mathcal{F}_k] \leq u_k.$$

Note that the belief is frozen in $[\pi^{**}(2), 1]$, so $\pi_{k+1}(2) \in [\pi^*(2), \pi^{**}(2))$ and $\pi_k(2) \in [\pi^{**}(2), 1]$ is irrelevant.

ii.) *Altruistic*: Here $\pi^{**}(2) = 1$. For $\epsilon \to 0$, we have for $\pi_k(2), \ \pi_{k+1}(2) \in [0, 1]$, $u_k = u^H(\pi_k)$ and it satisfies $\mathbb{E}[u^H(\pi_{k+1})|\mathcal{F}_k] \leq u_k$ by Jensen's inequality.

□

When the controller is profit maximizing or self-interested, it initially chooses higher price inputs to encourage sensors with higher valuation to utilize the services. Decisions at higher prices are more informative[8], which in turn results in higher public beliefs when $a = 2$. Due to the concavity of the pricing policy, higher belief causes the future price inputs to increase. Once sufficient information about the quality is accumulated, the controller either chooses low price inputs to allow every sensor to utilize the services or terminates its services to the multi-agent network.

When the controller is altruistic, it always chooses high price inputs to encourage the sensors to act according to their private valuations.

Appendix

Lemma 8. *The instantaneous reward $(u - \beta)\mathcal{I}(a = 2|\pi)$ is given as*

$$\sum_{j \in \mathcal{Y}} \sum_{i \in \mathcal{X}} (u - \beta)\mathcal{I}(u \leq 1 - \frac{\eta^{y=j}(1)}{\alpha}) B_{ij} \pi(i). \tag{15}$$

Lemma 9. *The instantaneous reward $(u - \beta)\mathcal{I}(a = y|\pi)$ is given as*

$$(u - \beta)\ \mathcal{I}(u^L(\pi) < u \leq u^H(\pi)). \tag{16}$$

The proofs follow from the structure of the social learning filter (see Theorem 2, [6]), property of the CVaR measure (see Lemma 6, [6]), and Bayes' rule. It is omitted due to lack of space.
We will prove Theorem 3 and Theorem 6 for the *Self-Interested* case. The proof for the *Altruistic* case follows similarly.
Proof of Theorem 3:
Consider $Q(\pi, u)$ as in (10) for $u \in [0, 1]$.

i.) Let $u \in [0, u^L(\pi)]$. Recall that $u^L(\pi) = 1 - \frac{\eta^{y=1}(1)}{\alpha}$. The instantaneous reward in (15) is $(u - \beta)$. The

[7]Note that the matrix B is TP2. It can be seen that derivative of $u^H(\pi)$ is strictly decreasing and the derivative of $u^L(\pi)$ is strictly increasing with respect to $\pi(2)$ for any $\alpha \in (0, 1]$.

[8]Informativeness is in the sense of Blackwell; see [12]. For any two observation matrices B_1 and B_2, B_1 is more informative than B_2 in the Blackwell sense ($B_1 \succ_B B_2$) if $B_2 = B_1 Q$, for any stochastic matrix Q. Note here that when $u = u^H(\pi)$, the action likelihood matrix in (7) $R^H = B$; and when $u = u^L(\pi)$, the action likelihood matrix $R^L = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. We have for $Q = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$, $R^L = R^H Q \Rightarrow R^H \succ_B R^L$.

continuation payoff $\sum_{a\in\mathcal{A}} V(T^{\pi}(\pi,a))\sigma(\pi,a)$ is given as follows. From (7), $R_a^{\pi} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$.

$$(\Rightarrow) \sum_{a\in\mathcal{A}} V(T^{\pi}(\pi,a))\sigma(\pi,a) = V(P'\pi).$$
$$\therefore Q(\pi,u) = (u-\beta) + \rho V(P'\pi). \quad (17)$$

ii.) Let $u \in (u^L(\pi), u^H(\pi)]$. The instantaneous reward in (15) is $(u-\beta)\times \mathbf{1}'B_{y=2}\pi$. From (7), $R_a^{\pi} = B$.

$$(\Rightarrow) \sum_{a\in\mathcal{A}} V(T^{\pi}(\pi,a))\sigma(\pi,a) = \mathbb{E}V(\pi).$$
$$\therefore Q(\pi,u) = (u-\beta)\times \mathbf{1}'B_{y=2}\pi + \rho\mathbb{E}V(\pi). \quad (18)$$

iii.) Let $u \in (u^H(\pi), 1]$. This implies that $u > 1 - \frac{\eta^{y=2}(1)}{\alpha}$. The instantaneous reward in (15) is 0. From (7), $R_a^{\pi} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. Since $\mathbb{P}(a=1)=1$, the controller doesn't accrue any profit by offering services. Therefore the instantaneous and continuation payoff is 0.

$$(\Rightarrow) Q(\pi,u) = 0. \quad (19)$$

The result follows from (17), (18) and (19). □

Proof of Theorem 6:
Define the following:

$$\delta^* = \min\{\pi(2)|\ \eta^{y=1}(2) \geq 1-\alpha\},$$
$$\gamma^* = \{\pi | (u^H(\pi)-\beta)\times \mathbf{1}'B_{y=2}\pi + \rho\mathbb{E}V(\pi) = 0\},$$
$$\pi^*(2) = \max\{\delta^*, \gamma^*\},$$
$$\pi^{**}(2) = \{\pi(2)|(u^L(\pi)-\beta) + \rho V(\pi) = (u^H(\pi)-\beta)\times \mathbf{1}'B_{y=2}\pi + \rho\mathbb{E}V(\pi)\}.$$

i.) Consider $\pi(2) \in [0,\pi^*(2))$. We will show that $\{\max\ Q(\pi,u) = 0\}$.
Let $V(0)$ denote the value at $\pi = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. As $\pi(2)\to 0$, we have $\mathbb{E}V(0) \to V(0)$ and $u^H(0) = u^L(0) \to 1-\frac{1}{\alpha}$. Assume on the contrary $V(\pi) = (u^L(\pi)-\beta) + \rho V(\pi)$. As $\pi(2)\to 0$, $V(0) = \frac{(1-\frac{1}{\alpha}-\beta)}{(1-\rho)}$. Since $\alpha\in(0,1]$, $\frac{1}{\alpha} \geq 1$ and $V(0) < 0$. From Theorem 3, $V(\pi)\geq 0$. Contradiction.
Similarly if $V(\pi) = (u^H(\pi)-\beta)\times\mathbf{1}'B_{y=2}\pi + \rho\mathbb{E}V(\pi)$, we have $V(0) < 0$. Therefore, $V(0) = 0$ and we have $Q(0,u^H(0)) < 0$ and $Q(0,u^L(0)) < 0$.
From the convexity of the value function, $\mathbb{E}V(\pi) \geq V(\pi)$. Since $Q(\pi,u^H(\pi)) < 0$ for $\pi(2) = [0,\pi^*(2))$, by definition of $\pi^*(2)$, we have

$$(u^H(\pi)-\beta)\times\mathbf{1}'B_{y=2}\pi + \rho\mathbb{E}V(\pi) < 0$$
$$V(\pi)\geq 0 \to \mathbb{E}V(\pi)\geq 0 \text{ by Jensen's Inequality.}$$
$$\therefore (u^H(\pi)-\beta) < 0.$$
$$(u^H(\pi)-\beta) < 0 \to (u^L(\pi)-\beta) < 0 \text{ from Lemma 1.}$$

If on the contrary $V(\pi) = (u^L(\pi)-\beta)+\rho V(\pi)$, then $V(\pi) < 0$; a contradiction.

$$\therefore Q(\pi,u^L(\pi)) < 0 \text{ for all } \pi(2) = [0,\pi^*(2)].$$
$$\Rightarrow V(\pi) = 0 \text{ for all } \pi(2) = [0,\pi^*(2)].$$

ii.) $\pi^{**}(2) = \{\pi(2)|Q(\pi,u^H(\pi)) = Q(\pi,u^L(\pi)))\}$. We will show that for $\pi(2)\in(\pi^{**}(2),1]$, $Q(\pi,u^L(\pi))) > Q(\pi,u^H(\pi))) > 0$.
Assume $Q(\pi,u^H(\pi)) > Q(\pi,u^L(\pi))$ on the contrary. Consider $\pi(2)\to 1$. Let $V(1)$ and $b(=\mathbf{1}'B_{y=2}\pi)\in[0,1]$ denote the values at $\pi = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. We have

$$u^H(1) = u^L(1) \to 1 \text{ and } \mathbb{E}V(1)\to V(1)$$
$$\Rightarrow (1-\beta)\times b + \rho\mathbb{E}V(1) > (1-\beta) + \rho V(1)$$
$$\Rightarrow b > 1, \text{ a contradiction as } \beta > 0.$$
$$\Rightarrow Q(\pi,u^L(\pi)) > Q(\pi,u^H(\pi)).$$

From Theorem 3, $V(\pi)\geq 0$ and therefore, $\mathbb{E}V(\pi)\geq 0$.

For $\pi(2)\in[\pi^{**}(2),1]$, $(u^H(\pi)-\beta)\times\mathbf{1}'B_{y=2}P'\pi > 0$
$$(\Rightarrow) Q(\pi,u^H(\pi)) > 0.$$

iii.) Since $\pi^{**}(2) = \{\pi(2)|Q(\pi,u^H(\pi) = Q(\pi,u^L(\pi)))\}$ and $Q(\pi,u^L(\pi)) < 0$ for all $\pi(2) = [0,\pi^*(2)]$, from part (ii) we have

$$Q(\pi,u^H(\pi))) > Q(\pi,u^L(\pi))) \text{ for all } \pi(2)\in[\pi^*(2),\pi^{**}(2)).$$

Note that $Q(\pi,u^H(\pi))) > 0$ for all $\pi(2)\in[\pi^*(2),\pi^{**}(2))$ by definition of $\pi^*(2)$ and the fact that $Q(\pi,u)\uparrow\pi$ (Theorem 2). □

REFERENCES

[1] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors Part I. Fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.

[2] A. Lalitha, A. Sarwate, and T. Javidi, "Social learning and distributed hypothesis testing," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 551–555, IEEE, 2014.

[3] V. Krishnamurthy, "Quickest detection POMDPs with social learning: Interaction of local and global decision makers," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5563–5587, 2012.

[4] 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. 992–1026, Oct., 1992.

[5] A. V. Banerjee, "A simple model of herd behavior," *The Quarterly Journal of Economics*, vol. 107, no. 3, pp. 797–817, Aug., 1992.

[6] V. Krishnamurthy and S. Bhatt, "Sequential Detection of Market Shocks With Risk-Averse CVaR Social Sensors," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1061–1072, 2016.

[7] S. Bose, G. Orosel, M. Ottaviani, and L. Vesterlund, "Dynamic monopoly pricing and herding," *The RAND Journal of Economics*, vol. 37, no. 4, pp. 910–928, 2006.

[8] S. Bose, G. Orosel, M. Ottaviani, and L. Vesterlund, "Monopoly pricing in the binary herding model," *Economic Theory*, vol. 37, no. 2, pp. 203–241, 2008.

[9] S. Mitra and T. Ji, "Risk measures in quantitative finance," *International Journal of Business Continuity and Risk Management*, vol. 1, no. 2, pp. 125–135, 2010.

[10] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, pp. 21–41, 2000.

[11] C. Chamley, *Rational herds: Economic models of social learning*. Cambridge University Press, 2004.

[12] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.