1 Single-Agent Markovian Persuasion Models

There is a hidden Markov state $S_t \in \mathcal{S}$ whose transition follows the time-homogeneous kernel $S_{t+1} \sim P(\cdot \mid S_t = s, A_t = a)$. The principal (sender) fully observes S_t and commits to a signaling or recommendation policy π that maps the state to a recommendation $R_t = \pi(S_t)$. The agent (receiver) does not observe S_t.

At each time t, the agent hold a belief $\mu_t \in \Delta(\mathcal{S})$ about S_t, observe the recommendation R_t, and then choose an action A_t. The agent's one-period payoff is given by a function $u(a, s)$ for $a \in \mathcal{A}$, $s \in \mathcal{S}$, and the principal's one-period payoff is $v(A_t, S_t)$. Let H_t denote the history observed by the agent up to time t (e.g., past recommendations and realized outcomes), which induces the belief $\mu_t(\cdot) = \mathbb{P}\left[S_t \in \cdot \,\middle|\, H_t\right]$ under the announced policy. Consider the following three models that differ in how the agent uses information over time:

- The agent at time t is a new, short-lived receiver who cares only about the current period. Their prior is a fixed distribution μ that does not depend on t or on H_t (so effectively $\mu_t \equiv \mu$), and they choose

$$A_t \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_\mu \left[v(a, S_t) \,\middle|\, R_t\right] \tag{1}$$

  that maximizes only their current expected payoff. They do not track information over time and do not care about future rewards. This corresponds to the standard "short-lived receiver" dynamic persuasion setup [Wu et al., 2022].

- The same agent is present over all periods and updates their belief μ_t from history: $\mu_t(\cdot) = \mathbb{P}\left[S_t \in \cdot \,\middle|\, H_t\right]$. However, at each time t they still choose A_t to maximize onluythe current expected payoff, i.e.,

$$A_t \in \arg\max_{a \in \mathcal{A}} \mathbb{E} \left[v(a, S_t) \,\middle|\, H_t, R_t\right], \tag{2}$$

  while how current action affects future information or future payoffs. In this sense the agent may track information over time (their belief μ_t changes with H_t), but their behavior is myopic: they never take actions purely to learn. This captures a long-lived agent who learns passively but does not engage in strategic exploration [Renault et al., 2017, Iyer et al., 2023, Lehrer and Shaiderman, 2021].

- The same agent is present over all periods, forms a belief $\mu_t(\cdot) = \mathbb{P}\left[S_t \in \cdot \,\middle|\, H_t\right]$, and cares about a discounted stream of payoffs

$$\mathbb{E} \left[\sum_{k=t}^{\infty} \beta^{k-t} v(A_k, S_t) \,\middle|\, H_t, R_t\right], \tag{3}$$

  for some $\beta \in (0, 1)$, where the expectation is taken under the probability law induced by the transition kernel P, the principal's signaling policy π, and the agent's strategy. In this case it can be optimal for the agent to take actions that are suboptimal in the current period in order to improve future information and explore deliberately.

2 Networked Markovian Persuasion with Myopic Agents

We now start with the first, simplest model and extend it to a population of agents on a network. The Markov state is now a vector $S_t = (S_{1,t}, \ldots, S_{n,t}) \in \mathcal{S}^n$, where $S_{i,t}$ describes the local state of region $i = 1, \ldots, n$. The regions are connected by a graph $G = (V, E)$; we write $N(i)$ for the neighborhood of i (e.g., i together with its one-hop neighbors on the graph). The state still evolves according to a time-homogeneous transition kernel $S_{t+1} \sim P(\cdot \mid S_t = s, A_t = a)$, where $A_t = (A_{1,t}, \ldots, A_{n,t})$ is the profile of actions taken by the agents in period t. We focus on the case with one representative agent per region and identify agent i with region i. The agent's one-period conditional expected utility in region i is given by a function $v_i : \mathcal{A}_i \times \mathcal{S}^{N(i)} \to \mathbb{R}$, while the principal's one-period conditional expected utility is the average utility across regions,

$$u(A_t, S_t) = \frac{1}{n} \sum_{i=1}^{n} v_i\left(A_{i,t}, S_{N(i),t}\right). \tag{4}$$

We focus on long-run performance in persistent systems such as revenue on ride-sharing platforms or disease control for epidemic mitigation. In such cases, the relevant KPIs are naturally per-period averages (e.g., average waiting time, match rate, or infection incidence). Accordingly, the principal evaluates a (stationary) recommendation policy π by its stationary average utility

$$U(\pi) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi\left[u(A_t, S_t)\right]. \tag{5}$$

Under mild regularity conditions (to be added), the limit exists and there is a stationary distribution d_π over states. In that case, the objective can be written as

$$U(\pi) = \mathbb{E}_{S \sim d_\pi}\left[\mathbb{E}_{A \sim \pi}\left[u(A, S) \,\middle|\, S\right]\right]. \tag{6}$$

This formulation captures the steady performance of the system after transients wash out.

Throughout, the principal fully observes S_t and commits to a (possibly randomized) recommendation policy $\pi : \mathcal{S}^n \to \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_n)$ so that in period t a recommendation profile $R_t = (R_{1,t}, \ldots, R_{n,t}) \sim \pi(\cdot \,|\, S_t)$ is drawn and privately communicated to the agents. We again interpret $R_{i,t}$ as a direct recommendation for the action of agent i. All agents share a common prior μ over S_t and know the policy π, but they do not observe S_t or the recommendations sent to other agents.

At each time t, agent i observes their own recommendation $R_{i,t}$ and then chooses an action $A_{i,t} \in \mathcal{A}_i$ to maximize their current expected payoff:

$$A_{i,t} \,\middle|\, R_{i,t} = r \in \arg\max_{a \in \mathcal{A}_i} \mathbb{E}_\mu\left[u_i(a, S_t) \,\middle|\, R_{i,t} = r\right], \tag{7}$$

where the expectation is taken over the posterior distribution of S_t given $R_{i,t} = r$. By Bayes' rule, this posterior is

$$\mathbb{P}_\mu\left[S_t = s \,\middle|\, R_{i,t} = r\right] = \frac{\mu(s)\mathbb{P}\left[R_{i,t} = r \,\middle|\, S_t = s\right]}{\sum_{s' \in \mathcal{S}^n} \mu(s')\mathbb{P}\left[R_{i,t} = r \,\middle|\, S_t = s'\right]}, \tag{8}$$

where

$$\mathbb{P}\left[R_{i,t} = r \,\middle|\, S_t = s\right] = \sum_{r_{-i} \in \mathcal{A}_{-i}} \pi\left((r, r_{-i}) \,\middle|\, s\right). \tag{9}$$

Following the classic work of Kamenica and Gentzkow [2011], we describe feasible outcomes in terms of the induced distributions over states and recommendations. For a given prior μ on S_t, and for each region i, we define the persuasion set $\mathcal{P}_i(\mu)$ as the set of probability measures $P \in \Delta(\mathcal{S}^n \times \mathcal{A}_i)$ such that the marginal of P on $\mathcal{S}^n$ is μ, and for every $a \in \mathcal{A}_i$ with $P(R_i = a) > 0$,

$$\mathbb{E}_\mu \left[v_i(a, S_t) \,\middle|\, R_{i,t} = a \right] \geq \mathbb{E}_\mu \left[v_i(a', S_t) \,\middle|\, R_{i,t} = a \right] \tag{10}$$

for all $a' \in \mathcal{A}_i$. It is well known that, with Bayesian rational agents, any signaling policy is outcome-equivalent to a direct recommendation policy characterized by some $P \in \mathcal{P}_i(\mu)$. Hence it is without loss of generality to work directly with the persuasion set.

2.1 Learning the Optimal Persuasion Policy

In practice, the principal rarely knows the environment (e.g., transition kernel and utilities) exactly. What the principal does have is a long log of past interactions generated under some baseline recommendation policy π_0. This naturally leads to an off-policy learning viewpoint: given logged data from π_0, can we learn a new stationary recommendation policy π that achieves higher stationary average utility $U(\pi)$ while preserving incentive compatibility?

To understand what the learning algorithm should aim for, it is useful to first ask: if we did know the environment, how would we compute an optimal persuasion policy?

Remark 1. With the off-policy learning formulation, it might be tempting to simply collapse the agent incentive layer and treat the recommendation R_t as the action directly. In that view, the problem reduces to finding a policy $R_t \sim \pi(\cdot \mid S_t)$ that maximizes the long-run value of a reduced MDP with state S_t and action R_t. One could write the reward in reduced form as

$$\mathbb{E}_\pi \left[u'_\pi(r, s) \right] = \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[u(A, S) \,\middle|\, S, R \right] \,\middle|\, S = s, R = r \right] \tag{11}$$

However, this collapsed representation hides an important dependence on the policy π. The conditional distribution of actions, $A_t \,\big|\, (S_t, R_t)$, is generated by agents' best responses to the posterior $\mathbb{P}\left[S_t \,\middle|\, R_t \right]$, and this posterior itself depends on the global recommendation rule π via Bayes' rule. As a consequence, the reduced-form reward u'_π is not a fixed function of (s, r) that is invariant across policies, and the collapsed off-policy learning approach is only valid under a strong invariance assumption that the conditional law $A_t \,\big|\, (S_t, R_t)$ must be the same for all policies π under consideration.

Remark 2. Questions to think about:

- What happens if the prior μ is not observed by the principal? In that case, it is unclear what the persuasion set is. (i) assume μ is common knowledge; (ii) consider a set of μ and discuss robust/ambiguity-set-type extensions; (iii) assume certain best-response model and do structural estimation

- What if the prior μ_t varies over time (as in the second case in Section 1)? The persuasion set can still be defined in a similar way period by period, but finding the optimal policy becomes more complicated. might be straightforward to extend the analysis when μ_t is exogenous, but may require substantially more work when μ_t is endogenous and depends on past signals and actions.

2.2 Local v.s. Global Policies

We impose the following locality assumption on rewards and dynamics. This assumption says that, conditional on the current local environment around a region i, neither the instantaneous payoff nor the law of $S_{i,t+1}$ depends on what happens in distant parts of the network. In particular, externalities and information propagate across the system only through the edges of G. This will potentially allow us to work with local posteriors over $S_{N(i),t}$ and to formulate the persuasion and incentive-compatibility constraints in terms of neighborhoods $N(i)$ rather than the full state S_t. Such locality assumptions are reasonable in many applications where interactions are predominantly spatial or geographic, for example, ride-sharing platforms where waiting times and prices in an area depend mainly on nearby demand and supply, or epidemic and diffusion models where the state of a location next period depends on its own and its neighbors' states (add some references here).

Assumption 1. For each region i there exists a neighborhood $N(i) \subseteq V$ such that:

1. The reward of agents in region i depends only on local state in $N(i)$. In other words, for all $s \in \mathcal{S}^n$,
$$v_i(a_i, s) = v_i(a_i, s_{N(i)}). \tag{12}$$

2. he next-period state of region i depends on (S_t, A_t) only through the local configuration $(S_{N(i),t}, A_{N(i),t})$. Formally, for all measurable $\mathcal{S}' \subseteq \mathcal{S}$,
$$\mathbb{P}\left[S_{i,t+1} \in \mathcal{S}' \,\middle|\, S_t = s, A_t = a\right] = \mathbb{P}\left[S_{i,t+1} \in \mathcal{S}' \,\middle|\, S_t = s', A_t = a'\right] \tag{13}$$
whenever $(s_{N(i)}, a_{N(i)}) = (s'_{N(i)}, a'_{N(i)})$.

References

Krishnamurthy Iyer, Haifeng Xu, and You Zu. Markov persuasion processes with endogenous agent beliefs. *arXiv preprint arXiv:2307.03181*, 2023.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Ehud Lehrer and Dimitry Shaiderman. Markovian persuasion. *arXiv preprint arXiv:2111.14365*, 2021.

Jérôme Renault, Eilon Solan, and Nicolas Vieille. Optimal dynamic information provision. *Games and Economic Behavior*, 104:329–349, 2017.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. *arXiv preprint arXiv:2202.10678*, 2022.