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 only the 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 Off-Policy Learning with Endogenous Belief

Consider an experiment in which the experimenter (principal) runs a micro-randomized experiment on a single user (agent) and collects a sequence of data $(S_t, R_t, Y_t)_{t\geq 1}$ under an experimental recommendation policy π_0. For example, on a short-video platform the experimenter can randomize how videos are highlighted to the user; in an online advertisement campaign, the experimenter can randomize how promotion emails are sent on each day. The goal is to find a target recommendation policy $\pi : \mathcal{S} \to \Delta(\mathcal{R})$ that optimizes the long-run average outcome for the principal.

A common approach is to model the system as a (time-homogeneous) MDP: conditionally on (S_t, R_t), the next state is drawn as $S_{t+1} \sim P(\cdot \,|\, S_t = s, R_t = r)$. Conventional off-policy evaluation methods then treat R_t as the action in this MDP and use importance weighting or Bellman equations to evaluate the long-run value of a target recommendation policy π.

Consider the case where the user has a (potentially unobserved, misspecified, and endogenous) belief $\mu_t \in \Delta(\mathcal{S})$ about the current state S_t. Rather than always following the recommendation, the user first updates their belief after observing R_t, and then chooses an action that maximizes their own one-period utility $v(a, S_t)$. In particular, if the recommendation policy π is public,[1] the posterior over S_t given $R_t = r$ and prior μ_{t-1} is

$$\mu_t(s; r, \mu_{t-1}) = \mathbb{P}\left[S_t = s \,\middle|\, R_t = r, \mu_{t-1}\right] = \frac{\pi(r \,|\, s)\mu_{t-1}(s)}{\sum_{s'} \pi(r \,|\, s')\, \mu_{t-1}(s')} \tag{4}$$

The user then plays a myopic best response

$$a^*(\mu_t) \in \operatorname{argmax}_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} v(a, s)\, \mu_t(s) \tag{5}$$

Denote the principal's one-period payoff as $u(A_t, S_t)$, which depends on the state S_t and the user's action A_t (but not directly on R_t). As the users always update their belief according to the Bayes rule, the pair (S_t, μ_t) evolves as a time-homogeneous Markov chain: given (S_t, μ_t), the recommendation R_t is drawn from $\pi(\cdot \,|\, S_t)$; the user responds with $A_t \sim a^*(R_t, \mu_t)$; the next state S_{t+1} are drawn from the $P(\cdot \,|\, S_t, A_t)$; and the belief μ_{t+1} is updated deterministically from (μ_t, R_t, A_t) according to the user's learning rule. We denote by d_π the stationary distribution of this Markov chain on $\mathcal{S} \times \Delta(\mathcal{S})$. Under some regularity assumptions (ergodic to be added), the stationary distribution d_π exists and is unique, and the long-run average payoff is well-defined and independent of the initial belief.

Given d_π, the principal's stationary objective under policy π can be written as

$$\mathbb{E}_\pi\left[u(A_t, S_t)\right] = \mathbb{E}_{(S_t,\mu_t)\sim d_\pi}\left[\mathbb{E}_{A_t\sim a^*(\mu_t)}\left[u(A_t, S_t) \,\middle|\, S_t, \mu_t\right]\right], \tag{6}$$

Thus, the principal's value under a policy π depends not only on the transition kernel for S_t but also on the endogenous belief process $\{\mu_t\}$ and the user's best-response strategy. Unless we impose the strong assumption that, for every (s, r), the induced conditional distribution of the user's action $\mathbb{P}\left[A_t = a \,\middle|\, S_t = s, R_t = r\right]$ is invariant across recommendation policies π, all conventional MDP approaches that treat R_t as the action break down.

[1] In practice, users may track only a coarsened state $S'_t \in \mathcal{S}'$ and a belief μ'_t over $\mathcal{S}'$, rather than a full posterior over $\mathcal{S}$. The formula below then applies to (S'_t, μ'_t) with π and μ_{t-1} replaced by their induced versions on $\mathcal{S}'$.

2.1 Identifying Policy Effect in Belief-Dependent Environments

To see how the existence of endogenous belief process makes generic off-policy learning fail even when belief is observed, let's start by considering the identification of off-policy value in such a belief-dependent environment. A key challenge is that the transition of the state $P_\pi(\cdot \mid S_t, \mu_t, R_t)$ now depends on the policy π due to the endogenous belief update:

$$P_\pi(S_{t+1}, \mu_{t+1} \mid S_t, \mu_t, R_t) = P_S(S_{t+1} \mid S_t, a^*(\mu_t)) \cdot I\left\{\mu_{t+1}(s) = \frac{\pi(R_t \mid s)\mu_t(s)}{\sum_{s'} \pi(R_t \mid s')\mu_t(s')}, \forall s\right\}$$

As a result, the usual off-policy identification argument, which assumes a single policy-invariant environment and only reweights using the state–action occupancy, breaks down. In particular, to recover the value of a target policy π from data collected under a logging policy π_0, we would need that every belief transition induced by π can also arise under π_0. Fixing a prior μ_t and a realized recommendation R_t, this requires the posterior under π to match the posterior under π_0:

$$\frac{\pi(R_t \mid s)\mu_t(s)}{\sum_{s'} \pi(R_t \mid s')\mu_t(s')} = \frac{\pi_0(R_t \mid s)\mu_t(s)}{\sum_{s'} \pi_0(R_t \mid s')\mu_t(s')}, \qquad \forall s. \tag{7}$$

When μ_t is a belief over the full state space, this equality forces $\pi(r \mid s) = \pi_0(r \mid s)$. In other words, generic off-policy identification in this belief-dependent environment would require an overlap condition that is much stronger than the usual state–action overlap and is almost always violated.

One way to weaken this requirement is to assume that the agent only tracks a coarsened state $B_t = g(S_t)$ and updates their belief on B_t using aggregated likelihoods (so that the Bayes update depends on $\mathbb{P}\left[R_t = r \mid B_t = b\right]$ rather than on $\mathbb{P}\left[R_t = r \mid S_t = s\right]$). In that case, off-policy identification is still only possible within a restricted class of target policies that are observationally equivalent at the level of the tracked state, in the sense that they induce the same likelihoods $\mathbb{P}\left[R_t = r \mid B_t = b\right]$ as the logging policy. Outside of such restricted classes, the policy dependence of the belief dynamics implies that standard MDP-based OPE methods, which rely on a policy-invariant transition kernel, are misspecified in this setting.

3 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 $S_{t+1} \sim P(\cdot \,|\, S_t = s, A_t = a)$, 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 $v_i : \mathcal{A}_i \times \mathcal{S} \to \mathbb{R}$, while the principal's one-period conditional expected utility is the average utility across regions,

$$u(A_t, S_t) = \frac{1}{n} \sum_{i=1}^{n} v_i \left(A_{i,t}, S_t\right). \tag{8}$$

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

$$U(\pi) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \left[u(A_t, S_t)\right]. \tag{9}$$

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

$$U(\pi) = \mathbb{E}_{S \sim d_\pi} \left[\mathbb{E}_{A \sim \pi} \left[u(A, S) \,\middle|\, S\right]\right]. \tag{10}$$

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:

$$A_{i,t} \,\middle|\, R_{i,t} = r_i \in \arg\max_{a_i \in \mathcal{A}_i} \mathbb{E}_\mu \left[u_i(a_i, S_t) \,\middle|\, R_{i,t} = r_i\right], \tag{11}$$

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

$$\mathbb{P}_\mu \left[S_t = s \,\middle|\, R_{i,t} = r_i\right] = \frac{\mu(s) \mathbb{P}\left[R_{i,t} = r_i \,\middle|\, S_t = s\right]}{\sum_{s' \in \mathcal{S}^n} \mu(s') \mathbb{P}\left[R_{i,t} = r_i \,\middle|\, S_t = s'\right]}, \tag{12}$$

where

$$\mathbb{P}\left[R_{i,t} = r_i \,\middle|\, S_t = s\right] = \sum_{r_{-i} \in \mathcal{A}_{-i}} \pi\left((r_i, r_{-i}) \,\middle|\, s\right). \tag{13}$$

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 $a_i \in \mathcal{A}_i$ with $P(R_i = a_i) > 0$,

$$\mathbb{E}_\mu \left[v_i(a_i, S_t) \,\middle|\, R_{i,t} = a_i \right] \geq \mathbb{E}_\mu \left[v_i(a_i', S_t) \,\middle|\, R_{i,t} = a_i \right] \tag{14}$$

for all $a_i' \in \mathcal{A}_i$. 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.

A natural choice of prior μ is the stationary distribution d_π of S_t induced by the recommendation policy π. This mirrors the classic Bayesian persuasion framework, where the sender observes a realized state drawn from a common prior; the difference in our dynamic setting is that this common prior is itself determined by the policy π. We interpret this formulation as targeting long-run performance: once the system has mixed, agents have learned the steady-state distribution of the state under π, but in each period they don't observe the realized state.[2] They are thus Bayesian with prior d_π and update their beliefs about the current state using the principal's recommendation, exactly as in the static persuasion model. The policy learning problem can then be summarized as the following optimization problem of finding the optimal occupancy measure $\gamma(a, s) = d_\pi(s)\pi(a \,|\, s)$ that maximizes long-term performance subject to incentive constraints:

$$\begin{aligned}
&\max_{\gamma} \sum_{a,s} \left(\frac{1}{n} \sum_i v_i(a_i, s) \right) \gamma(a, s) \\
&\text{s.t.} \sum_s v_i(a_i, s) \sum_{a_{-i}} \gamma((a_i, a_{-i}), s) \geq \sum_s v_i(a_i', s) \sum_{a_{-i}} \gamma((a_i, a_{-i}), s), \\
&\qquad\qquad \forall i, \forall a_i', \forall a_i \text{ with } \sum_{s, a_{-i}} \gamma((a_i, a_{-i}), s) > 0, \\
&\sum_{a,s} \gamma(a, s) = 1, \qquad \gamma(a, s) \geq 0, \qquad \forall a, \forall s \\
&\sum_{a,s} \gamma(a, s) P(s' \,|\, s, a) = \sum_a \gamma(a, s'), \qquad \forall s'.
\end{aligned}$$

3.1 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

[2] Another interesting setting is when agents only observe their own realized local state, but not the states of their neighbors, so the principal's signal nudges their beliefs about the global state of the system.

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 $s \in \mathcal{S}^n$,
$$v_i(a_i, s) = v_i(a_i, s_{N(i)}). \tag{15}$$

2. The 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 $\mathcal{S}' \subseteq \mathcal{S}$,
$$\mathbb{P}\left[S_{i,t+1} \in \mathcal{S}' \,\middle|\, S_t = s, A_t = a\right] = \mathbb{P}\left[S_{i,t+1} \in \mathcal{S}' \,\middle|\, S_t = s', A_t = a'\right] \tag{16}$$
whenever $(s_{N(i)}, a_{N(i)}) = (s'_{N(i)}, a'_{N(i)})$.

However, even if one-step rewards and transitions are local, under a stationary policy the stationary distribution of S_t can have long-range correlations, because local interactions propagate over time. We therefore need a mixing assumption to ensure that an approximate local law is close to the true local law.

Formally, consider the utility function $U(\pi)$. Under Assumption 1, with a global policy π,
$$U(\pi) = \frac{1}{n} \sum_i \sum_{a_i, s_{N(i)}} v_i(a_i, s_{N(i)}) \Gamma_{i,\pi}(a_i, s_{N(i)}), \tag{17}$$
where
$$\Gamma_{i,\pi}(a_i, s_{N(i)}) = \sum_{a_{-i}, s_{-N(i)}} \gamma_\pi(a, s) \tag{18}$$
Now consider a set of local policies $\{\tilde{\pi}_i\}$ that sets
$$\tilde{\pi}_i(a_i \,|\, s_{N(i)}) = \mathbb{E}\left[\pi(a_i \,|\, s_{N(i)}, S_{-N(i)})\right] \tag{19}$$
so that under γ_π the conditional distribution of A_i given $s_{N(i)}$ is the same under π and $\tilde{\pi}$. We need a mixing assumption under which
$$\Gamma_{i,\pi}(a_i, s_{N(i)}) \approx \Gamma_{i,\tilde{\pi}}(a_i, s_{N(i)}) \tag{20}$$
for all i, a_i, $s_{N(i)}$. Under this assumption, the value $U(\tilde{\pi})$ is close to $U(\pi)$, so so restricting attention to local policies incurs only a small regret in the objective. Similarly, using the same notation, we can write the IC constraints as
$$\sum_s v_i(a_i, s_{N(i)}) \Gamma_{i,\pi}(a_i, s_{N(i)}) \geq \sum_s v_i(a'_i, s_{N(i)}) \Gamma_{i,\pi}(a_i, s_{N(i)}). \tag{21}$$
Then, under the mixing assumption, and if v_i is bounded a.s., both the objective and each side of the IC inequalities change only by a small amount when we replace the global policy π by its local counterpart $\tilde{\pi}$. Thus restricting attention to local policies leads to at most a small loss in both feasibility and performance.[3]

[3] We also need to handle the stationarity constraint. This requires that the learned local measures are consistent with the local dynamics induced by π and that, whenever two neighborhoods overlap, their marginals on the overlap coincide.

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.