Polarizing Persuasion: Notes

Axel Anderson and Nikoloz Pkhakadze

1 Extended Abstract Masquerading as Introduction

This paper considers a Bayesian persuasion game between a single sender and a pair of receivers. As in Kamenica and Gentzkow (2011), the sender chooses an experiment (message service), but cannot directly control the public outcome of the experiment, and the receivers passively update their beliefs using Bayes' rule. The sender's payoff is a monotone function of each receiver's belief on the binary *payoff relevant* state. All agents share a common prior on this payoff relevant state. However, we allow for disparate beliefs on the *payoff irrelevant* state, a binary variable that enters no utility functions, but nonetheless may affect the receivers interpretation of the signal sent by the message service chosen by the sender.

As in the standard persuasion model, if all agents have the same prior beliefs, then the sender chooses a perfectly revealing experiment when her payoff function is convex in receiver beliefs and reveals no information when her payoff function is concave. But this is no longer true once we allow for disparate priors on the payoff irrelevant state. To understand why, first note that when the sender's prior differs from either (or both) receivers' prior, then the sender can design an experiment that reveals information and shifts receiver beliefs up *from the sender's perspective*. We call such upward shifts in beliefs the *misinformation effect*. Lemma 2 establishes that the misinformation effect rules out perfectly concealing messages and provides a lower bound on the receiver's ex post disagreement. Also, the sender's payoff increases in the receivers' ex ante disagreement (Proposition 1).

Since the receivers disagree about the payoff irrelevant state, the sender can design experiments that induce *strict polarization*; namely an experiment in which *every* signal causes the receiver's posterior beliefs in the payoff relevant state to move in opposite directions. Proposition 2 establishes that optimal message services must be strictly polarizing when the sender's payoff is bi-concave and submodular in receiver's beliefs. With sufficient ex ante disagreement between the receivers we can explicitly solve for the optimal message service when the sender's payoff is bi-convex. While this optimal message service does not induce strict polarization, it does imply a weaker form of polarization; namely, that receiver posterior beliefs negatively covary across messages.

Message services need not be polarizing. Receivers *ordinally agree* on a message service when the signal that generates the highest posterior is the same for each receiver, and the

signal that generates the second highest posterior is the same for each receiver, etc. If the sender's payoff function is bi-concave and supermodular, then optimal message services induce ordinal agreement (Proposition 3).

2 Model

This section describes the public communication game between one sender and two receivers.[1] The state is two dimensional $(\theta, \omega) \in \{\theta_0, \theta_1\} \times \{\omega_0, \omega_1\}$, and all player's share a common prior $p \in (0,1)$ that $\theta = \theta_1$. The prior beliefs that $\omega = \omega_1$ are $q_s, q_1, q_2 \in (0,1)$ for the sender, receiver one, and receiver two. We order receivers such that $q_1 < q_2$.

The sender knows the receiver's beliefs and can costlessly commit to any finite *message service*, M; namely, a finite set of signals j and state contingent probabilities $\pi^j(\theta, \omega)$. Let $\mathcal{M}$ be the space of such message services. For any message service M, the sender's subjective belief that signal j is realized:

$$\Pi_s^j \equiv p\left(q\pi^j(\theta_1,\omega_1) + (1-q)\pi^j(\theta_1,\omega_0)\right) + (1-p)\left(q\pi^j(\theta_0,\omega_1) + (1-q)\pi^j(\theta_0,\omega_0)\right) \tag{1}$$

and analogously define receiver i's subjective beliefs as Π_i^j.

One signal is drawn from the chosen message service, and this signal is commonly observed by the sender and each receiver. Let P_s^j (P_i^j) be the posterior belief of the sender (receiver i) after observing signal j. We assume, all agents update their beliefs according to Bayes' rule using their own prior beliefs. Specifically, posterior beliefs obey:

$$P_k^j = \frac{p\left(q_k\pi^j(\theta_1,\omega_1) + (1-q_k)\pi^j(\theta_1,\omega_0)\right)}{\Pi_k^j} \qquad \forall\ k \in \{s,1,2\} \tag{2}$$

We abstract from receiver choices by assuming the sender's C^2 payoff V only depends on the receivers' beliefs on the state variable θ. The sender's maximization problem is thus:

$$V^*(p, q_s, q_1, q_2) = \max_{M\in\mathcal{M}} \sum_j \Pi_s^j V(P_1^j, P_2^j) \qquad \text{s.t. (1) and (2)} \tag{3}$$

While the sender's payoff function does not directly depend on receiver beliefs in state ω, the indirect payoff function V^* does. This owes to the fact that the posterior beliefs P_i^j on θ depend on the prior belief q_i on ω, and the fact that the sender evaluates the probabilities

[1] The model embeds the information structure in Benoit and Dubra (2019) in a Kamenica and Gentzkow (2011) communication game. Critically for our purposes, Alonso and Câmara (2016) allow for heterogeneous prior beliefs.

Π_s^j with her own prior q_s. Altogether, while it would be more precise to refer to ω as *not directly* payoff relevant, we henceforth refer to ω as the *payoff irrelevant state* and θ as the *payoff relevant state*. Throughout we assume that V is strictly increasing in each argument, which trivially implies that V^* is strictly increasing in p.

3 The Sender Benefits from Ex Ante Disagreement

The constraint set in sender maximization (3) only enters the objective function indirectly. The standard approach in the Persuasion literature is to allow the sender to choose the distribution over receiver posteriors subject to a constraint that takes account of receivers' posterior beliefs and Bayesian updating. Toward this end, define $\alpha \in \mathbb{R}$ as the unique solution to $q_s = \alpha q_1 + (1-\alpha)q_2$ (valid by $q_1 \neq q_2$ and all beliefs bounded away from 0 and 1). Then the following Lemma fruitfully reformulates the sender's maximization problem.[2]

Lemma 1 *A message service solves the Sender's maximization* (3)*, if and only if, it induces distributions over beliefs that solves:*

$$V^* = \max_{\Pi_i, P_i} \quad \sum_j \left(\alpha \Pi_1^j + (1-\alpha)\Pi_2^j\right) V(P_1^j, P_2^j) \tag{4}$$

$$s.t. \qquad \sum_j \Pi_i^j P_i^j = p \quad \forall i \tag{5}$$

$$\frac{q_1}{q_2} \leq \frac{P_1^j \Pi_1^j}{P_2^j \Pi_2^j} \leq \frac{1-q_1}{1-q_2} \quad and \quad \frac{q_1}{q_2} \leq \frac{(1-P_1^j)\Pi_1^j}{(1-P_2^j)\Pi_2^j} \leq \frac{1-q_1}{1-q_2} \quad \forall j \tag{6}$$

If we set $q_s = q_1$, and then take the limit $q_1 \to q_2$, the reformulation in Lemma 1 is a special case of the reformulation in Kamenica and Gentzkow (2011).[3] In particular, when $q_1 = q_2$, constraint (6) demands that the distribution over receiver beliefs must be identical. Further, $q_s = q_1 = q_2$ implies $\alpha = 1$, so the sender's expected payoff is evaluated with this common belief distribution. Altogether, in this limit the sender chooses a common belief distribution for all players, subject to the martingale property (5).

Constraint (6) is a consistency requirement on the joint distribution of receiver beliefs and is novel to the current model, constraining the divergence in receiver posterior beliefs on the payoff relevant state θ. Recalling that we have assumed $q_1 < q_2$, we say that *receiver ex ante disagreement increases* when q_1 decreases and/or q_2 rises. By inspection, a decrease in q_1 or

[2] These cursory notes omit all substantial proofs.

[3] The maximization in Lemma 1 is not well defined in the limit $q_1 = q_2 \neq q$, since $\alpha = (q_2 - q)/(q_2 - q_1)$. When $q_1 = q_2 \neq q$ the original maximization (3) is a special case of Alonso and Câmara (2016). They provide an alternative reformulation for this case, which we apply to our model in Section 6.

an increase in q_2, relaxes constraint (6). Thus, we have the following immediate consequence of Lemma 1.

Proposition 1 *The sender's expected value V^* rises in receiver ex ante disagreement.*

4 Polarization vs. Ordinal Agreement

We now turn to our focus: understanding when the sender induces polarization. We say that a message service is *strictly polarizing* if $(P_1^j - p)(P_2^j - p) < 0$ for all signals j that are sent with positive probability. That is, when the posterior beliefs of the receivers move in opposite directions with probability 1. A necessary condition for such polarization is that optimal belief distributions cannot put positive weight on posteriors $P_i^j = p$ for any receiver. The following Lemma implies that this necessary condition is met.

Lemma 2 *At least one of the constraints in* (6) *holds with equality for all signals sent with positive probability in any optimal message service. Consequently, optimal belief distributions cannot put positive probability on $P_i^j = p$ for any receiver.*

In a standard persuasion game with common prior beliefs, concavity of the sender's payoff function V implies that optimal message services reveal no information, i.e. that posterior beliefs equal prior beliefs. Lemma 2 asserts that such *concealing* messages can never be optimal for *any* smooth monotone payoff function given any ex ante disagreement on the payoff irrelevant state. To reconcile these results, assume $q_s = q_2$ and then consider the limit $q_1 \to q_2$. Notice in this limit that constraint (6) implies that $P_1^j = P_2^j$, and polarization must vanish in this limit. The sender will then choose to conceal when $V(P, P)$ is concave in P.

The senders value is *bi-concave* when $V(P_1, P_2)$ concave in P_1 for all P_2 and concave in P_2 for all P_1. The sender's value is *(strictly) submodular* when

$$V(P_1, \hat{P}_2) + V(P_1, \hat{P}_2)(>) \geq V(P_1, P_2) + V(\hat{P}_1, \hat{P}_2) \quad \forall\ (\hat{P}_1, \hat{P}_2) > (P_1, P_2)$$

and (strictly) supermodular when $-V$ is (strictly) submodular. Now notice that V bi-concave and submodular, implies that $V(P, P)$ is concave in P, precisely the assumption needed to induce the sender to conceal all information in the standard common prior model. But in our model with ex ante disagreement about the payoff irrelevant state, these assumptions imply strict polarization.

Proposition 2 (Polarization) *If V is bi-concave and submodular, then optimal messages are strictly polarizing and have binary support (i.e. WLOG we can assume two messages).*

For completeness sake we also consider message services that induce positive covariance in receiver beliefs. Specifically we say that receivers *ordinally agree* on a message service if ever pair of signals (j, j') induces beliefs that obey $(P_1^j - P_1^{j'})(P_2^j - P_2^{j'}) > 0$. Or put another way, the two receivers agree on which signal makes them most confident the state is θ_1, and which signal makes them second most confident the state is θ_1, etc.

Proposition 3 (Ordinal Agreement) *If V is bi-concave and supermodular, then receivers ordinally agree on any optimal message service.*

5 Maximal Ex-Ante Disagreement

In this section, we will show that the sender may induce polarization even when she has a bi-convex value function (i.e. $-V$ bi-concave). For simplicity we focus on the limiting case $q_1 \to 0$ and $q_2 \to 1$.[4] By Proposition 1, this is the highest value case for the sender. In particular, the consistency constraint (6) vanishes in this limit, and the sender's choice of posterior distributions for one receiver is unconstrained by her choice of posterior distributions for the other receiver. That is, this case is formally equivalent to separate *private communication* with each receiver on the payoff relevant state.

In a persuasion model with one receiver, the sender chooses to fully reveal the state when her value is convex in the receiver's belief. Here bi-convexity implies a specific trinomial distribution over extremal posterior beliefs.

Proposition 4 *Assume private communication. If V is bi-convex, then the sender chooses a message service that with three signals inducing posterior beliefs $(P_1, P_2) = ((1,0),(0,1),(1,1))$ with probabilities $(\Pi_1, \Pi_2) = ((0, 1-p),(1-p, 0),(p, p))$.*

Intuitively, bi-convexity of beliefs induces the sender to restrict attention to belief distributions supported on $\{(0,0),(1,0),(0,1),(1,1)\}$. The given distribution first order dominates all other distributions on this support from the point of view of the sender subject to the constraint that $E[P_i|q_i] = p$ (i.e. each receiver does not expect his belief to change).

This belief distribution is not strictly polarizing, since it places positive weight on the perfectly correlated beliefs $(P_1, P_2) = (1, 1)$. However, two of the three signals yield beliefs that are maximally polarized. More specifically, from the sender's ex ante point of view, there is a $1 - p$ chance that posterior beliefs end up being maximally polarized. A cardinal

[4]Similar results will hold when q_1 and q_2 are sufficiently far apart. Results for the general bi-convex case are in progress.

measure of polarization is the covariance between P_1 and P_2, with more negative covariance corresponding to more polarization. Calculating the covariance (from the sender's point of view) for the distribution in Proposition 4 we get $-q_s(1-q_s)(1-p)^2 < 0$. Thus, this measure of polarization is maximized at $q_s = 1/2$ and decreasing in p.

To show how the sender can induce this distribution over beliefs, consider a three signals message service with state contingent probabilities:

$$\begin{array}{ccc} \pi^1: & & \\ & \theta_0 & \theta_1 \\ \omega_0 & 0 & \varepsilon \\ \omega_1 & 1 & 0 \end{array} \qquad \begin{array}{ccc} \pi^2: & & \\ & \theta_0 & \theta_1 \\ \omega_0 & 1 & 0 \\ \omega_1 & 0 & \varepsilon \end{array} \qquad \begin{array}{ccc} \pi^3: & & \\ & \theta_0 & \theta_1 \\ \omega_0 & 0 & 1-\varepsilon \\ \omega_1 & 0 & 1-\varepsilon \end{array} \tag{7}$$

Fixing $q_1 = 0$ and $q_2 = 1$ and taking the limit $\varepsilon \to 0$, yields the desired belief distribution.

6 A Second Reformulation

The reformulation in Lemma 1 is useful for understanding the impact of receiver disagreement on the sender's optimal *payoff*, but requires $q_1 \neq q_2$. In this section we derive a second reformulation that can be applied directly to the case $q_1 = q_2$. Toward this reformulation, let $s = ((1-p)(1-q_s), p(1-q_s), (1-p)q, pq)$ denote the sender's prior belief over the four states $((\theta_0,\omega_0),(\theta_1,\omega_0),(\theta_0,\omega_1),(\theta_1,\omega_1))$, and let $S^j = (S_1^j, S_2^j, S_3^j, S_4^j)$ be the sender's posterior belief vector over these four states given some signal j. Let $\Delta(S)$ be the set of distributions on the 3-simplex: $\{(S_1 + S_2 + S_3) \geq 0 | S_2 + S_3 + S_4 \leq 1\}$ with positive probability on four realizations.

Define receiver i's *relative prior bias* $\varrho_i \equiv \frac{(1-q_i)q_s}{(1-q_s)q_i} - 1$. Thus, ϱ_i is positive when $q_s > q_i$ and negative when $q_s < q_i$. We can then express receiver posterior beliefs on the payoff relevant state θ as a function of the sender's posterior beliefs on the two dimensional state, and reformulate the sender's maximization as a choice over her beliefs, subject to the usual martingale restriction.

Lemma 3 *If a signal generates sender posterior beliefs S then receiver beliefs are:*

$$\mathcal{P}_i(S) \equiv Pr(\theta = \theta_1 | S) = \frac{1 - S_1 - S_3 + S_2\varrho_i}{1 + \varrho_i(S_1 + S_2)}$$

And sender maximization (3) *is equivalent to:*

$$V^* = \max_{\delta \in \Delta(s)} E_\delta[V(\mathcal{P}_1(S), \mathcal{P}_2(S))] \quad s.t.\ E_\delta[S] = s$$

When $\varrho_i = 0$, we have $\mathcal{P}_i(S) = 1 - S_1 - S_2$, that is, the receiver and sender share the same beliefs on the payoff relevant state. Thus, if $\varrho_1 = \varrho_2 = 0$, Lemma 3 reduces to the standard formulation in Kamenica and Gentzkow (2011).

References

Alonso, R., and O. Câmara (2016): "Bayesian persuasion with heterogeneous priors," *Journal of Economic Theory*, 165(C), 672–706.

Benoit, J. P., and J. Dubra (2019): "Apparent Bias: What Does Attitude Polarization Show?," *International Economic Review*, 60, 1675 –1703.

Kamenica, E., and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101(6), 2590–2615.