When is a majority sufficient to persuade a population to adopt a convention?

Miharu Naono

Independent - work carried out at Kyoto Univ.

m.naono2020@gmail.com

Jonathan Newton

Institute of Economic Research, Kyoto Univ.

newton@kier.kyoto-u.ac.jp

Arina Nikandrova

City St George's, University of London

arina.nikandrova@city.ac.uk

September 23, 2026

We study a model of social coordination in which agents repeatedly decide whether to adopt a behavior. Agents form beliefs about population adoption rates through random sampling, update these beliefs in response to a planner's signals, and best respond to their posteriors. The planner seeks to promote adoption via a threshold-based disclosure rule. We show that optimal disclosure initiates a dynamic feedback: successful persuasion increases the fraction of agents with persuadable beliefs, which in turn amplifies the effect of future persuasion. We characterize conditions under which, starting from any state where adoption already holds a simple majority, the planner can induce convergence to the full-adoption convention.

Keywords: conventions, coordination, evolutionary dynamics, sampling best response, information design, Bayesian persuasion, majority persuasion.

JEL Classification: C73, D83.

Websites: jonathannewton.net and sites.google.com/view/anikandrova/home

We thank seminar audiences at Keio University, Nagoya University, National Taiwan University, Econometric Society Meetings in Abu Dhabi 2026, Hanoi 2026. Newton's work was supported by JSPS KAKENHI 22H00826.

During the preparation of this manuscript, the authors used OpenAI's ChatGPT to improve language, readability and presentation. All substantive arguments, analysis, results and conclusions were developed, reviewed and verified by the authors, who take full responsibility for the content.

1 Introduction

Nothing attracts a crowd like a crowd.

– Soul Asylum - "Black Gold" (1993)

Planners often seek to turn partial adoption of a behavior into a widely adopted convention. Public-health authorities may encourage prophylactic measures such as mask wearing and hand washing.[1] Employers may encourage employees to attend the office on the same days or seek to normalize fathers' take-up of parental leave.[2] Industry bodies may promote adoption of a common technical standard.[3] Parent groups may encourage families to delay giving children smartphones.[4] In these environments, individuals may prefer to adopt the target behavior only if enough others do the same.[5]

A natural intuition is that an initial majority should be sufficient for adoption to become self-reinforcing: once the target behavior is more common than the alternative, one might expect it to spread. However, if non-adoption is safer, more familiar, or more robust to strategic uncertainty, individuals may require more than a simple majority before adopting becomes individually optimal. In this paper, we ask when a planner can use credible disclosure about others' behavior to overcome this resistance to adoption and steer the population from any initial strict majority of adopters towards full adoption of the target convention.

We study this question in a model of dynamic social coordination in which agents choose whether to adopt a behavior and occasionally revise their choices.

[1] The UK Department of Health and Social Care's "Hands. Face. Space" campaign (2020) promoted hand washing, face coverings, and physical distancing during the COVID-19 pandemic.

[2] Aviva publicized fathers' take-up of its equal parental leave policy and emphasized that taking leave was compatible with a successful career. See Aviva (2018), "Equal parental leave shows men are eager to share childcare."

[3] The Internet Society's World IPv6 Launch brought together Internet service providers, equipment manufacturers, and web companies to coordinate permanent deployment of IPv6, while publicizing participating organizations. See Internet Society, "World IPv6 Launch: FAQ."

[4] Smartphone Free Childhood organizes a collective pledge to delay children's smartphone ownership until age 14 and makes participation visible through information about how many families are joining at each school and in the surrounding area. See Smartphone Free Childhood, "Parent Pact."

[5] Heiman et al. (2023) find that perceived mask wearing by others predicts subsequent increases in individuals' own mask wearing. Microsoft's 2022 Work Trend Index reports that 73% of surveyed hybrid employees would attend the office more frequently if they knew their direct teammates would be present (see Microsoft (2022), "Hybrid Work Is Just Work. Are We Doing It Wrong?".) Dahl et al. (2014) present evidence that fathers' leave-taking influences their coworkers' subsequent parental leave decisions.

Agents' incentives are summarized by a best-response threshold, denoted ξ. This is the share of adopters in the population at or above which adoption is a best response when the current adoption rate is known to agents. We assume that $\xi > 1/2$, so that an agent who expects the two behaviors to be equally common strictly prefers not to adopt — that is, non-adoption is risk dominant. Larger values of ξ correspond to stronger strategic resistance to adoption.

Agents do not directly observe the population adoption rate. Instead, each agent observes the behavior of a random sample of k peers: a citizen sees whether acquaintances wear masks, a worker observes how many colleagues come to the office on the same day as she, or a member of an organization sees whether peers follow a new norm. On the basis of their samples, agents form beliefs about the current population adoption rate. Different samples generate different beliefs: an agent who observes many adopters is more optimistic about adoption than one who observes few. In the baseline model, heterogeneous samples generate beta-distributed beliefs whose means equal the observed sample adoption shares.

After sampling, agents observe a public signal from a planner who knows the current adoption rate and seeks to promote full adoption. The planner seeks to steer the population towards the target convention, irrespective of whether this objective aligns with agents' welfare.[6] We model the planner's persuasion technology as a threshold-based signal. At the outset, the planner chooses a threshold and commits to truthfully reporting whether the current adoption rate is at least this threshold. The chosen disclosure threshold remains fixed throughout. Such disclosure can be interpreted as an adoption report indicating whether a publicly announced benchmark has been reached.

Agents update their sample-based beliefs using the planner's report and adopt whenever their posterior expectation of the adoption rate is at least ξ — that is, agents myopically best respond to their posterior beliefs. Figure 1 illustrates the full cycle of sampling, persuasion and strategy updating.

The heterogeneity generated by sampling creates both an opportunity for persuasion and a trade-off. Some agents observe enough adopters to adopt without further information, while others are too pessimistic to be persuaded at the current adoption rate. Between these groups are marginal agents whose

[6] In some applications, adoption may generate external benefits that agents do not fully internalize, such as public health benefits, organizational spillovers, or compatibility gains from a shared standard. In others, the planner may be an interested actor whose objective is not socially aligned. Our results characterize when credible information disclosure can implement the target convention, irrespective of its welfare implications.

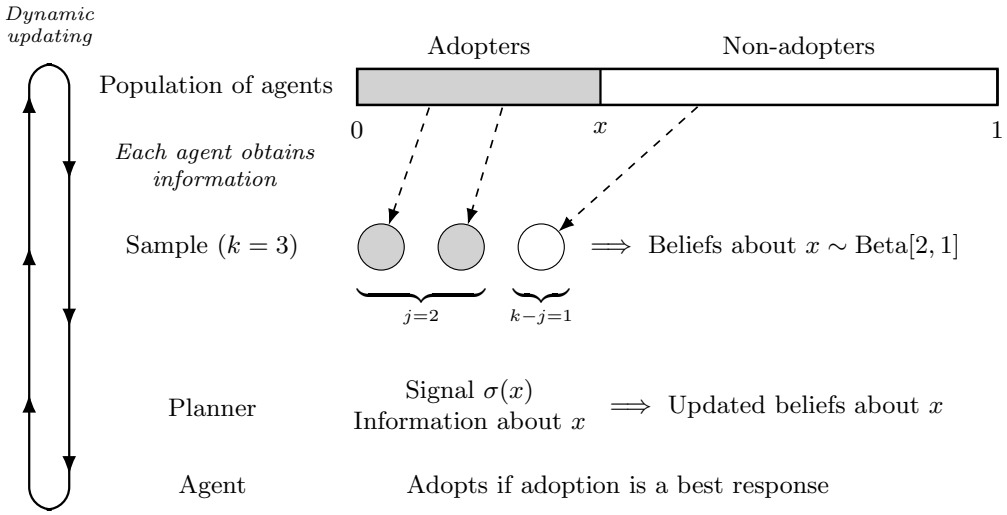

Figure 1: Dynamic updating in a population. Each agent in a population of unit mass obtains information about the current adoption rate x by sampling, before receiving a signal from a planner who knows x and wishes to encourage agents to adopt. The agent then adopts if his posterior expectation of x is large enough to make adoption a best response.

choices can be influenced by the planner's disclosure. Those whose samples contain more adopters are easier to persuade, but persuading only these relatively optimistic agents may be insufficient to make adoption grow. Persuading more doubtful marginal agents requires a higher disclosure cutoff, which can be truthfully met only at higher adoption rates. The planner must therefore balance the difficulty of persuasion against the need to persuade enough agents to sustain adoption growth.

The key mechanism is a dynamic feedback between adoption and persuadability. Adoption grows when the share of revising agents who choose it exceeds the current adoption rate. As adoption rises, future random samples from the population contain more adopters, which increases the share of agents whose sample-based beliefs are sufficiently favorable for the planner's report to induce adoption. A suitable disclosure policy can thus initiate and sustain a self-reinforcing adoption cycle, eventually leading to full adoption of the target convention.

Under our baseline belief specification and within the class of threshold disclosure rules, our core result characterizes the combinations of sample size k and best-response threshold ξ for which the planner can induce convergence to population-wide adoption from every initial strict majority of adopters.

Finding 1. If the agents' sample size k is even, then the planner can induce convergence to full adoption from every initial strict majority of adopters if and only if ξ does not exceed the probability of at least $k/2$ heads in k fair coin tosses. A similar characterization applies when k is odd.

In the benchmark without disclosure, for even sample sizes, an initial strict majority guarantees convergence to full adoption only when $\xi \leq 1/2$ — that is, only when adoption is weakly risk dominant.[7] Threshold disclosure raises this bound on ξ to $3/4$ for $k = 2$ and $11/16$ for $k = 4$. Persuasion can therefore sustain the spread of a convention even when non-adoption is risk dominant.

The case $k = 2$ makes the mechanism transparent. An agent who observes one adopter and one non-adopter forms uniform beliefs on $[0, 1]$, with mean $1/2$. Without disclosure, this agent does not adopt when $\xi > 1/2$. Suppose instead that the planner reports whether adoption is at least $1/2$. A positive report rules out adoption rates below $1/2$, which results in uniform beliefs on $[1/2, 1]$ with mean $3/4$. The agent therefore adopts whenever $\xi \leq 3/4$. The report persuades by ruling out low adoption rates while leaving open the possibility that adoption is considerably more widespread.

To interpret the result in payoff terms, consider a symmetric two-player coordination game in which miscoordination yields payoff of zero, illustrated in Figure 2.[8] In this game, the bound $\xi \leq 3/4$ allows the payoff from coordination on non-adoption to be up to three times the payoff from coordination on adoption. Despite this threefold payoff disadvantage, when $k = 2$, the planner can use threshold disclosure to gradually persuade the entire population to converge to adoption from every initial strict majority.

For even and odd sample sizes considered separately, the largest ξ compatible with majority persuasion decreases with sample size k. Larger samples generate more concentrated beliefs, which are harder to move through disclosure. However, when even and odd sample sizes are considered together, integer effects make the largest persuadable ξ locally non-monotone in k.

The comparison between $k = 2$ and $k = 3$ illustrates the role of odd sample sizes. With $k = 2$, growth from every initial strict majority requires persuading agents who observe one adopter in their samples; with $k = 3$, persuading those who observe two is needed. Therefore, the marginal agent whose adoption is

[7] For odd sample sizes, the benchmark bound is no longer exactly $1/2$ due to integer effects, specifically the impossibility of evenly split samples (see Lemma 7).

[8] Setting miscoordination payoffs equal to zero in Figure 2 helps to illustrate these individual incentives, but results hold across equivalence classes of games that have the same best response correspondence and thus the same ξ.

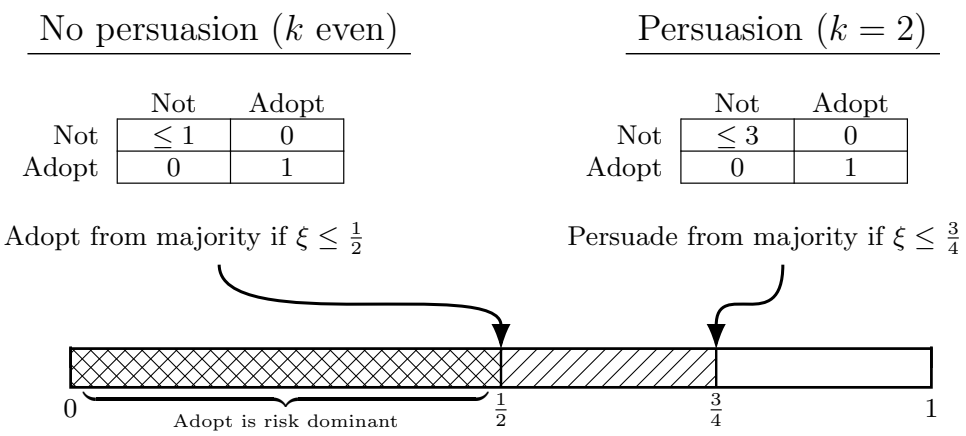

Figure 2: When is an initial majority of adopters sufficient to induce full adoption? Correspondence between coordination game payoffs and the perfect information threshold ξ. Entries in the game matrices denote payoff of the row player. Without persuasion, for even sample sizes, an initial strict majority guarantees convergence to full adoption only when (Adopt, Adopt) is a weakly risk dominant Nash equilibrium. In contrast, persuasion can induce convergence even when payoff differences are large.

required to sustain growth begins with a higher sample adoption share when $k = 3$: $2/3$ rather than $1/2$. Following a report that population adoption is at least $1/2$, this agent's posterior expected adoption rate is $7/9$, compared with $3/4$ for a marginal agent with an evenly split sample when $k = 2$. This more favorable posterior belief explains why the largest persuadable ξ rises from $3/4$ when agents sample two peers to $7/9$ when they sample three.

Finding 2. Across all sample sizes $k \geq 2$, the largest value of ξ for which the planner can induce convergence to full adoption from every initial strict majority is $7/9$. This bound is attained at $k = 3$.

Modeling contribution. Our modeling contribution is to show how finite-sample belief distributions create a channel for information design inside a best-response dynamic. Without signals, or if agents treat sample means as point estimates that are unresponsive to the planner's signal, our model reduces to the sampling best response dynamics of Oyama et al. (2015). In our model, however, agents infer beliefs as in Salant and Cherry (2020); Sawa and Wu (2023), which allows the planner to influence them through Bayesian persuasion (Kamenica and Gentzkow, 2011).

While our baseline model assumes Bayesian updating of beliefs, substantial

evidence in behavioral economics demonstrates that individuals systematically deviate from this benchmark (see, for example Benjamin, 2019). To account for non-Bayesian updating, we extend our model to allow for under- and overinference. Under our specification, underinference makes sample-based beliefs more diffuse without changing their mean, thereby making them easier to move. Underinference does not, however, affect how the planner's threshold disclosure excludes all incompatible adoption rates. The effect of underinference therefore operates entirely through the beliefs formed from sampling.

Finding 3. Underinference facilitates persuasion, while overinference hinders persuasion.

The paper proceeds as follows. First, we briefly discuss related literature. Section 2 describes the model. Section 3 studies sustainable persuasion in a population, asymptotic stability and when a majority suffices to achieve full adoption of the target convention. Section 4 studies under- and overinference. Section 5 concludes. Proofs are relegated to the appendix, as are extensions of the model to environments with assortative interaction, non-standard preferences, general priors, and non-Bayesian inference from sampling.

Related literature

A large literature examines the emergence of conventions across a range of strategic settings. Theoretical contributions include Young (1998); Binmore (2005); Carvalho (2013); Naidu et al. (2017); Newton (2021). Empirical work includes Young and Burke (2001); Gulesci et al. (2025). A recurrent theme is the resilience of risk dominant conventions under behavioral rules, including best response, in which choice depends on payoff differences (Young, 1993; Blume, 2003; Peski, 2010; Nax and Newton, 2022). With this regularity as a reference point, starting with Bergin and Lipman (1996), much work has examined when alternative behavioral rules select non-risk dominant conventions (see Newton, 2018, for a survey). The current paper stays within the framework of best response, but relies on information design to mitigate, and potentially overturn, the inherent bias towards risk dominance.

Under sampling best response dynamics (Oyama et al., 2015), agents play best responses to their sample means. Sawa and Wu (2023) instead allow agents to infer beliefs from samples (see Salant and Cherry, 2020). The resulting dynamics can differ from sampling best response if either

(a) Inference is biased, in the sense that the expected adoption rate (using our terminology) inferred from a sample does not equal the sample mean; or

(b) payoffs are non-linear in strategy shares.

In our model, inference is unbiased and payoffs are linear, hence without persuasion, dynamics would be identical to both of the above dynamics. Thus, our paper isolates and illustrates a third channel by which inferring belief distributions rather than point estimates can affect sample-based choice: persuasion.

The current paper's evolutionary information design parallels the evolutionary mechanism design of Sandholm (2007, 2005), who considers a planner who alters payoffs through Pigouvian pricing, thus ensuring convergence to efficiency. In contrast, our model keeps payoffs fixed and instead manipulates beliefs through information disclosure. Relatedly, Murakami (2021) also studies a planner who uses signals to target conventions through drift:

(a) Agents' strategies are contingent on the signal. Moreover, the planner observes and can condition the signal on these contingent strategies.

(b) Repeatedly sending one signal, the planner allows responses to the unused signal to drift until they enter the basin of attraction of the target convention, at which point switching signals secures convergence.

In our model, there is no unused component of strategies, therefore no drift.

2 Model

2.1 Game of social coordination

Consider the symmetric two-player **game** with strategies A and B given in Figure 3a. Coordination on A yields payoff $1+\beta$ and coordination on B yields payoff 1. Parameter β can be positive or negative or zero, and it represents the payoff advantage or disadvantage of coordination on A rather than B. Miscoordination results in zero payoff for the A player and $\beta - \alpha$ for the B player.

Assume that $\alpha > 0$, so that (A, A) and (B, B) are strict Nash equilibria, with (A, A) the unique risk dominant Nash equilibrium. That is, if an opponent randomizes uniformly across A and B, then a player's expected payoff is maximized by playing A. This bias towards A typically gives A an advantage over B in models of equilibrium selection based on individual best response.

	A	B
A	$1+\beta$	0
B	$\beta-\alpha$	1

(a) General case

	A	B
A	$1+\alpha$	0
B	0	1

(b) $\beta = \alpha$

Figure 3: Symmetric two-player coordination games. Entries denote the payoff of the row player for each strategy combination.

(a) The general case. Parameter β is the payoff advantage or disadvantage of coordination on A relative to coordination on B. Parameter $\alpha > 0$ gives the advantage of strategy A over B in terms of maximizing a potential function. This parameterization suffices to represent all 2-player, 2-strategy symmetric coordination games up to affine transformation of payoffs.

(b) The special case of $\beta = \alpha$. This choice of parameters sets both miscoordination payoffs equal to zero. In addition, this matrix gives an exact potential function (Monderer and Shapley, 1996) for the game in (a).

Although A is the risk dominant strategy, B is the strategy that the planner wants players to adopt. However, before we consider the planner's means of persuasion, we describe how players in a population interact with one another and how they form beliefs about the strategies of other players.

2.2 Population of agents

Consider a **population** of unit mass. At time $t \in \mathbb{R}_+$, each member of the population is playing either A or B. Denote the share of the population playing B at time t by $x(t) \in [0,1]$. The time argument will often be omitted for brevity — that is, we often write x alone. As is standard in the literature on best response dynamics (see, for example Weibull, 1995; Sandholm, 2010), agents in the population would like to best respond to the mixed strategy given by $x(t)$. However, they do not know $x(t)$, so instead respond to their beliefs about $x(t)$. The trajectory of $x(t)$ will be determined by these responses.

2.3 Best response when x is known to agents

Given that a share x of the population is playing B, and a share $1-x$ is playing A, the true expectation of the payoffs that an agent would obtain from

playing A and B against a randomly selected opponent are

$$\pi(A,x) := (1-x)(1+\beta) + x \cdot 0, \qquad \pi(B,x) := (1-x)(\beta-\alpha) + x \cdot 1.$$

Therefore, B is a best response to x under full information if and only if $\pi(B,x) \geq \pi(A,x)$. Rearranging,

$$x \geq \frac{1+\alpha}{2+\alpha} =: \xi. \qquad \text{Best response threshold}$$

That is, B is a best response to x if and only if $x \geq \xi$. This best response threshold ξ is the symmetric mixed strategy Nash equilibrium probability of B for the game in Figure 3. Observe that ξ is independent of β as the β terms cancel when we consider the difference between $\pi(A,x)$ and $\pi(B,x)$.

The assumption that $\alpha > 0$, equivalent to risk dominance of the equilibrium (A,A), implies that $\xi > 1/2$. That is, if $x = 1/2$, so that half of the agents in the population are playing A and half are playing B, then A is the unique best response. This gives a natural bias of best response and similar dynamics towards strategy A. In the literature, this bias is known as asymmetry (Peski, 2010; Newton, 2021). The goal of this paper is to study conditions under which a planner can manipulate information to reverse this bias.

2.4 Beliefs when x is unknown to agents

The agents in the population do not know x, the true proportion of the population playing B. Instead, each agent forms a belief about x. A **belief** is a probability measure ν on $[0,1]$. It represents an agent's subjective probability distribution over x. For a given belief ν, denote the expectation of x by

$$\mathbb{E}_\nu := \int_{[0,1]} x \, \nu(dx). \qquad \text{Expectation of } x \text{ under beliefs } \nu$$

For each $x \in [0,1]$, let μ_x denote the distribution of beliefs in the population. That is, μ_x is a probability measure on the space of all probability measures on [0,1]. Each μ_x represents the population's distribution of beliefs when the actual proportion of agents choosing strategy B is x. The mapping $x \mapsto \mu_x$ describes beliefs that agents might hold about x. These beliefs might be highly concentrated (e.g. Dirac delta functions) or widely spread (e.g. uniform distributions).

2.5 Sampling to form beliefs

Our baseline model of how agents form their beliefs is based on sampling (see Salant and Cherry, 2020). To generate μ_x assume that each agent in the population follows the following two step procedure to obtain their beliefs ν:

Step 1: Sample $k \geq 2$ agents uniformly at random from the population. With probability $\binom{k}{j} x^j (1-x)^{k-j}$, the number of B players in the sample equals j.

Step 2: For a homogeneous sample $j = 0$ or $j = k$, the agent forms point beliefs on $x = 0, 1$ respectively. For a heterogeneous sample $0 < j < k$, beliefs ν follow a Beta$[j, k-j]$ distribution.

This way of forming beliefs has some appealing properties.

Remark 1. For all j, we have $\mathbb{E}_\nu = j/k$. This is consistent with the literature on sampling best response dynamics (see, e.g. Oyama et al., 2015). In that literature, decisions are based on the expected value of x given a sample, and thus it is unnecessary to model beliefs beyond the expected value. In this paper, we wish to manipulate the beliefs, therefore inferring only the expectation from the sample is insufficient. Instead we infer a set of beliefs (Salant and Cherry, 2020). ◁

Remark 2. The beta distribution is flexible. It can be symmetric or skewed and is able to accommodate both precise and diffuse beliefs. Furthermore, the unnormalized density $x^{j-1}(1-x)^{k-j-1}$ is a Bayesian update from the improper prior $x^{-1}(1-x)^{-1}$. Indeed, the Bayesian update of any beta distribution after sampling is a beta distribution. Thus, it fits well in a Bayesian framework. ◁

Remark 3. *Wisdom of the crowd*. Aggregate beliefs are correct in that, for all x,

$$\int_{\mathcal{P}} \mathbb{E}_\nu \, \mu_x(d\nu) = x$$

where $\mathcal{P}$ is the space of all probability measures on $[0, 1]$. ◁

Remark 4. Larger true values of x are associated with belief in larger x. Formally, if $x' < x''$, then $\mu_{x'}$ is first order stochastically dominated by $\mu_{x''}$, which means that for any continuous, increasing utility function u,

$$\int_{\mathcal{P}} \left(\int_0^1 u(x)\nu(dx) \right) \mu_{x'}(d\nu) \leq \int_{\mathcal{P}} \left(\int_0^1 u(x)\nu(dx) \right) \mu_{x''}(d\nu). \qquad \text{(SD)}$$

◁

Remark 5. The distribution of beliefs μ_x is setwise convergent. That is, for all $x^* \in [0,1]$, as $x \to x^*$, we have $\mu_x(A) \to \mu_{x^*}(A)$ for every measurable set A. ◁

Remark 6. Although the assumption of point beliefs for homogeneous samples $j = 0$ and $j = k$ is consistent with sampling best response dynamics (Remark 1), it might be considered extreme. Fortunately, this assumption is unnecessary for our results. Specifically, we can choose different beliefs for $j = 0$ and $j = k$ as long as they do not violate that beliefs are strictly increasing in j in the sense of the monotone likelihood ratio property. As long as this holds, it turns out that persuading agents with $j = 0$ is never necessary, and persuading agents with $j = k$ is never sufficient, thus relevant thresholds later in the paper are independent of the fine details of the beliefs of such agents. ◁

2.6 Planner sends a signal about x to influence agents' beliefs

We restrict the planner to binary threshold disclosures.[9] At the outset, the planner chooses a cutoff $\tau \in [0,1)$ and commits to the **signal** $\sigma(x) \in \{0,1\}$:

$$\sigma(x) = \begin{cases} 0 & \text{if } x < \tau \\ 1 & \text{if } x \geq \tau \end{cases} \qquad \text{Threshold signal}$$

Thus, at each time t, the planner reports whether current adoption is at least τ.[10] The corresponding likelihood of signal $\sigma(x) = \bar{\sigma} \in \{0,1\}$ as

$$\Pr(\bar{\sigma}|x) = \bar{\sigma} \cdot \mathbf{1}_{\{x \geq \tau\}} + (1 - \bar{\sigma}) \cdot \mathbf{1}_{\{x < \tau\}}. \qquad \text{Likelihood}$$

Agents know the planner's threshold τ. Assume that if an observed signal is impossible under the prior then an agent does not update her beliefs. Otherwise, assume that belief updating is Bayesian. Let $\nu(\cdot|\bar{\sigma})$ be the posterior beliefs of an agent with prior ν after observing signal $\sigma(x) = \bar{\sigma}$. Specifically, $\nu(\cdot|0)$ is ν conditioned on $x < \tau$ and $\nu(\cdot|1)$ is ν conditioned on $x \geq \tau$.

Consider our belief formation procedure from Section 2.5. Clearly, if ν is formed from a homogeneous sample and is thus a point belief on $x = 0$ or

[9] For a receiver with a fixed prior and two actions, binary messages are sufficient: all messages inducing A can be pooled, as can all messages inducing B. Each pooled posterior is a convex combination of posteriors inducing the same action and therefore continues to induce that action.

[10] The case of $\tau = 1$ is identical to $\tau = 0$ in that, for no value of x is $\sigma(x)$ informative (i.e. leads to belief updating) for beliefs derived from any sample that occurs with positive probability. Therefore, to keep discussion and proofs concise, we omit $\tau = 1$ from the outset.

$x = 1$, then there is no update in response to a signal. In contrast, if ν is formed from a heterogeneous sample ($0 < j < k$) it admits a probability density that is updated in proportion to likelihoods in the standard Bayesian way in response to a signal. In this case, we abuse notation and write $\nu(x)$ and $\nu(x|\bar{\sigma})$ as the densities of ν and $\nu(\cdot|\bar{\sigma})$ at x. Therefore,

$$\nu(x|\bar{\sigma}) = \frac{\Pr(\bar{\sigma}|x)\nu(x)}{\int_0^1 \Pr(\bar{\sigma}|y)\nu(y)\,dy}, \qquad \boxed{\text{Bayesian update}}$$

which, for $\bar{\sigma} = 0$ and $\bar{\sigma} = 1$ respectively, simplifies to

$$\nu(x|0) = \frac{\mathbf{1}_{\{x<\tau\}}\nu(x)}{\int_0^\tau \nu(x)\,dx} \qquad \text{and} \qquad \nu(x|1) = \frac{\mathbf{1}_{\{x\geq\tau\}}\nu(x)}{\int_\tau^1 \nu(x)\,dx}.$$

Write the expectation of x of an agent with beliefs ν conditional on signal $\bar{\sigma}$ as

$$\mathbb{E}_{\nu|\bar{\sigma}} := \int_0^1 x\,\nu(x \mid \bar{\sigma})\,dx. \qquad \boxed{\text{Conditional expectation}}$$

Note that, if $\tau = 0$, then the planner will always send signal $\bar{\sigma} = 1$, beliefs will not be updated, and expectations will remain unchanged, $\mathbb{E}_{\nu|\bar{\sigma}} = \mathbb{E}_\nu$.

2.7 Population dynamics with sampling and signals

A strategy is a **best response** if it maximizes expected payoff. Given beliefs ν and signal $\bar{\sigma}$, denote the set of best responses by $BR(\nu, \bar{\sigma})$. Observe that, for given x, payoff $\pi(A, x)$ and $\pi(B, x)$ are affine functions of x, therefore the expectations can be taken inside the argument of the function. That is, expected payoffs given beliefs ν and signal $\bar{\sigma}$ equal $\pi(A, \mathbb{E}_{\nu|\bar{\sigma}})$ and $\pi(B, \mathbb{E}_{\nu|\bar{\sigma}})$.

It then follows from the definition of ξ that A is a best response if and only if the conditional expectation of x is less than or equal to ξ, and B is a best response if and only if the conditional expectation of x is greater than or equal to ξ.

$$A \in BR(\nu, \bar{\sigma}) \iff \mathbb{E}_{\nu|\bar{\sigma}} \leq \xi \quad \text{and} \quad B \in BR(\nu, \bar{\sigma}) \iff \mathbb{E}_{\nu|\bar{\sigma}} \geq \xi$$

Given the distribution of beliefs μ_x and the planner's signal $\sigma(x)$, the share of agents for whom B is a best response is given by

$$BR(x) := \int_{\mathcal{P}} \mathbf{1}_{B \in BR(\nu, \sigma(x))}\,\mu_x(d\nu). \qquad \boxed{\text{Share of agents with } B \text{ as a best response}}$$

The distribution over the number j of B players in a sample is non-decreasing in x in the sense of first order stochastic dominance. Furthermore, beliefs satisfy the monotone likelihood ratio property in j. Additionally, from its threshold structure, $\sigma(x)$ is also non-decreasing in x. Monotone likelihood ratio ordering of beliefs is preserved by conditioning on either signal. Thus, our characterization of $BR(\nu, \bar{\sigma})$ tells us that the indicator function of whether a given agent has B as a best response is non-decreasing in j and observed signal $\bar{\sigma}$. Therefore, the aggregate share of agents for whom B is a best response is non-decreasing in x.

Lemma 1. *$BR(x)$ is non-decreasing in x.*

With $BR(x)$ defined, we can define the **best response dynamics**. At any given point in time $t \in \mathbb{R}_+$, some share of agents observe samples of agents, form beliefs, observe the current signal $\sigma(x(t))$ and best respond. The share of updating agents for whom B is a best response is $BR(x(t))$. To avoid expositional clutter, assume that agents who are indifferent between playing A and B will play B. Consequently, the share of updating agents who choose B is exactly $BR(x(t))$. Assume that each agent updates, on average, once per unit of time. This is a normalization and is without loss of generality. We obtain the dynamic

$$\dot{x}(t) = BR(x(t)) - x(t). \qquad \text{Best response dynamic}$$

Remark 7. Observe that $\dot{x} = 0$ if and only if $x = BR(x)$. That is, stationary points of the dynamic are **equilibria** when agents best respond to their beliefs formed from sampling and observing the signal. ◁

From a dynamic perspective, the stability of equilibria also matters. An equilibrium x^* is asymptotically stable if, for small enough departures away from x^*, the dynamics return the population to x^*.

Definition 1. An equilibrium x^* is **asymptotically stable** if there exists a neighborhood U of x^* such that, if $x(0) \in U$, then $\lim_{t\to\infty} x(t) = x^*$.

Setwise convergence (Remark 5) of belief distributions μ_x implies piecewise continuity of $BR(x)$ and existence of an asymptotically stable equilibrium.

Lemma 2 (Continuity and existence)**.**

(i) For fixed signal $\sigma(x) \equiv \bar{\sigma}$, $BR(x)$ is continuous in x.

(ii) Considering signal $\sigma(x)$, there is only one possible discontinuity, where $BR(x)$ may jump upwards at $x = \tau$.

(iii) There exists at least one asymptotically stable equilibrium. Indeed, $x = 0$ is such an equilibrium.

This completes the description of the model and preliminary results. We now characterize when threshold disclosure can induce convergence to B.

3 Analysis

This section develops the threshold characterization underlying our main results. In Section 3.1 we study the problem of persuading a single agent who has sampled j plays of B in a sample of size k. We derive a **persuasion threshold**, the minimum x required to persuade such an agent to also play B. In Section 3.2 we show that, conditional on persuading agents who sample at least j plays of B, there exists a **sustainable growth threshold**, the minimum x above which the population dynamics lead to growth in the share of persuadable agents. In Section 3.3 we combine these thresholds, showing that the planner targets a specific value of j to obtain an **overall threshold** above which the population can be persuaded to converge to strategy B. In Section 3.4 we ask under what conditions this overall threshold is less than or equal to half. That is, we answer the question in the paper's title of **when is a majority sufficient** to persuade a population to adopt a convention. In Section 3.5 we show that, under optimal threshold persuasion, the target convention is asymptotically stable.

3.1 Persuading a single belief

Consider an agent who samples j draws of B from a sample size of k, and consequently forms the belief ν. In this section, we derive a **persuasion threshold** $\tau^*(k, j, \xi)$ such that, whenever $\tau \geq \tau^*(k, j, \xi)$, a signal $\bar{\sigma} = 1$ indicating that $x \geq \tau$ suffices to persuade the agent to play B. That is, $\tau^*(k, j, \xi)$ is the lowest value of x at which such an agent is persuadable.

To derive persuasion threshold $\tau^*(k, j, \xi)$, we first derive $\mathbb{E}_{\nu|1}$, the expectation of x conditional on receiving a signal $\bar{\sigma} = 1$. For $j = 0$ and $j = k$, beliefs ν are point beliefs on $x = 0$ and $x = 1$ respectively and are not updated in response to a signal. For $0 < j < k$, beliefs are updated, and the expectation of x under beliefs ν conditional on $x \geq \tau$ is

$$\mathbb{E}_{\nu|1} = \int_0^1 x\, \nu(x \mid \bar{\sigma} = 1)\, dx = \frac{\int_\tau^1 x\, \nu(x)\, dx}{\int_\tau^1 \nu(x)\, dx} = \frac{\int_\tau^1 x^j (1-x)^{k-j-1}\, dx}{\int_\tau^1 x^{j-1} (1-x)^{k-j-1}\, dx}.$$

This is a ratio of upper incomplete beta functions. It can be shown that this reduces to a ratio of cumulative distribution functions of the binomial distribution. Writing $BinCDF(j;k,\tau)$ as the binomial CDF, the probability of up to j successes from k draws with success probability τ, we have the following lemma.

Lemma 3 (Conditional expectations)**.**

$$\mathbb{E}_{\nu|1} = \begin{cases} 0 & \text{if} \quad j = 0, \\ \frac{j}{k} \cdot \frac{BinCDF(j;k,\tau)}{BinCDF(j-1;k-1,\tau)} & \text{if} \quad 0 < j < k, \\ 1 & \text{if} \quad j = k. \end{cases}$$

These expectations of x conditional on signal $\bar{\sigma} = 1$ are what we require to derive the persuasion threshold $\tau^*(k, j, \xi)$, which we do in three cases.

First, if $j = 0$, then ν is a point belief on $x = 0$ and is not updated in response to any signal: $\mathbb{E}_{\nu|1} = \mathbb{E}_\nu = 0$. It is impossible to induce a posterior expectation above ξ and the agent is unpersuadable. That is, $\tau^*(k, j, \xi)$ is undefined.

Second, if $\mathbb{E}_\nu \geq \xi$, then B is already a best response to ν, so a signal is not necessary to persuade the agent. This will always be the case for $j = k$, where ν is a point belief on $x = 1$, but may also be true for other values of j. Setting $\tau = 0$ so that the signal is always equal to 1 and thus uninformative, we have

$$\mathbb{E}_{\nu|1} = \mathbb{E}_\nu \geq \xi,$$

so that B remains a best response at any value of x. That is, $\tau^*(k, j, \xi) = 0$.

Third, consider $0 < \mathbb{E}_\nu < \xi$, so that B is not a best response to ν, but $j > 0$ and thus ν is continuous with full support and a density on $[0, 1]$. This is the most interesting case, as it is possible to persuade an agent with such beliefs, but only with an informative signal ($\tau > 0$) and thus only at some values of x. Note that

$$\mathbb{E}_{\nu|0} \leq \mathbb{E}_\nu \leq \mathbb{E}_{\nu|1}.$$

Therefore, $\mathbb{E}_{\nu|0} < \xi$, implying $B \notin BR(\nu, 0)$. That is, for any value of τ chosen by the planner, the agent with belief ν is unpersuadable at $x < \tau$ (implying signal $\bar{\sigma} = 0$). In contrast, when $x \geq \tau$ (implying signal $\bar{\sigma} = 1$), we have

$$\mathbb{E}_{\nu|1} = \mathbb{E}_\nu[x \mid x \geq \tau] \geq \tau,$$

therefore, if the planner sets $\tau = \xi$, then $\mathbb{E}_{\nu|1} \geq \xi$ and $B \in BR(\nu, 1)$. However, in general, τ can be set lower than this. Indeed, it can be reduced until the

expectation conditional on $\bar{\sigma} = 1$ is exactly equal to ξ.

$$\tau^*(k, j, \xi) = \min \left\{ \tau \in [0, 1) : \ \mathbb{E}_{\nu|1} = \mathbb{E}_\nu \left[x \mid x \geq \tau \right] \geq \xi \right\}.$$

Combining the above cases with Lemma 3, we can summarize our persuasion thresholds $\tau^*(k, j, \xi)$ in the following lemma.

Lemma 4 (Persuasion thresholds)**.**

$$\tau^*(k, j, \xi) \quad \begin{cases} \textit{undefined} & \textit{if} \quad \mathbb{E}_\nu = 0, \\ \textit{solves} \quad \frac{j}{k} \cdot \frac{BinCDF(j;k,\tau^*)}{BinCDF(j-1;k-1,\tau^*)} = \xi & \textit{if} \quad 0 < \mathbb{E}_\nu < \xi, \\ = 0 & \textit{if} \quad \xi \leq \mathbb{E}_\nu. \end{cases}$$

$\tau^*(k, j, \xi)$ is weakly increasing in ξ and weakly decreasing in j for $j = 1, \ldots, k$.

The fact that $\tau^*(k, j, \xi)$ weakly decreases in j implies that if a positive signal persuades agents who have sampled j' draws of B, then it will also persuade agents who have sampled $j'' > j'$ draws of B. That is, if $x \geq \tau^*(k, j, \xi)$, then all agents sampling at least j agents who play B can be persuaded to play B.

Remark 8 (Optimality of threshold signals)**.** In a static problem, where the planner maximizes the ex ante probability that a single binary-action agent with prior ν adopts B, an upper-tail threshold signal structure is optimal. This threshold disclosure is consistent with the characterization in Kolotilin (2015) and the monotone-persuasion result of Mensch (2021).

In our dynamic model, however, restricting binary disclosure to a threshold form entails loss of generality. The first example in Appendix D shows that a non-threshold signal can persuade an agent to adopt B at a low state where no threshold signal can do so. This construction paves the way to using binary messages as a relay in our dynamic setting. By pooling selected low states with sufficiently favorable high states, one message can induce relatively pessimistic types to play B and initiate B's growth at low adoption levels, while the other message can sustain growth once adoption is higher. Proposition 3 shows that this relay mechanism can strictly enlarge the basin of attraction of full adoption.

We nevertheless focus on deterministic monotone binary disclosure. With two messages, any such disclosure takes a threshold form (up to its treatment of the cutoff point) and has a natural interpretation as a credible report of whether adoption has crossed a publicly announced benchmark. Our results therefore characterize optimal disclosure within this economically natural class rather the

unrestricted information-design optimum. For $k = 2$, however, the restriction is without loss within the class of deterministic binary disclosures: Appendix D shows that threshold disclosure is optimal for every $\xi > 1/2$. ◁

3.2 Ensuring persuasion leads to sustainable growth

In the previous subsection we showed that the planner can persuade all agents who have sampled at least j agents who play B if and only if $x \geq \tau^*(k, j, \xi)$. However, persuading such a set of agents is only part of the story. The planner wants the share of agents playing B to increase. For this to be the case, persuading agents who satisfy a given criterion (in this case those who sample at least j draws of B) to play B must lead to an increase in the share of agents satisfying the criterion. That is, persuading a set of persuadable agents must lead to the share of persuadable agents to grow. Then, a virtuous cycle is achieved and strategy B spreads across the population.

In this subsection we derive another threshold, the **sustainable growth threshold** $x^*(k, j)$. Let the set of agents for whom B is a best response equal the set of agents who sample at least j players who play B. Given this, the number of B players will grow if x is above the threshold[11]

$$x^*(k,j) := \min \left\{ x' \in [0,1] : \ \forall x \geq x', \ \underbrace{1 - BinCDF(j-1;k,x)}_{\substack{\text{share of agents who sample} \\ \text{at least } j \text{ } B \text{ agents}}} \geq x \right\}.$$

Let us verify this definition for different values of j. Again, let the set of players for whom B is a best response equal the set of players who sample at least j players who play B. The number of B players will grow if

$$\dot{x} = BR(x) - x = 1 - BinCDF(j-1;k,x) - x > 0. \tag{Gr}$$

For $j = 0$, every sample has B as a best response, so (Gr) holds for every $x < 1$. Consequently, $x^*(k, 0) = 0$. Of course, under our belief formation process, a homogeneous sample of A agents never leads to B as a best response, so this case never applies. However, we include it for completeness.

For $j = 1$, (Gr) reduces to $1 - (1 - x)^k - x > 0$, which holds for all $x \in (0, 1)$. Consequently, $x^*(k, 1) = 0$. This has a qualitative implication: persuading all agents with heterogeneous samples suffices to induce convergence of the entire

[11] For the case $j = 0$, note that $BinCDF(-1;k,x^*) = 0$.

population to B.

For $2 \leq j < k$, the shape of the Binomial distribution ensures that there exists a unique interior minimal $x^* = x^*(k, j)$ such that (Gr) holds for all $x \in (x^*, 1)$.

For $j = k$, inequality (Gr) is never satisfied. Indeed, equality is reached at $x = 0$ and $x = 1$, so we have $x^*(k, k) = 1$. This has a qualitative implication: from any interior state, no matter how close it is to $x = 1$, to induce convergence of the entire population to B, at least some agents with heterogeneous samples must be persuaded.

Lemma 5 (Sustainable growth threshold)**.**

$$x^*(k,j) \begin{cases} = 0 & \text{if} \quad j = 0, 1, \\ \text{is the unique interior solution to } 1 - BinCDF(j-1; k, x) = x & \text{if} \quad 2 \leq j < k, \\ = 1 & \text{if} \quad j = k. \end{cases}$$

$x^*(k, j)$ is weakly increasing in j for $j = 0, \ldots, k$.

Thus, persuading all agents whose samples contain at least j plays of B increases the share of B-players only when the current adoption rate satisfies $x > x^*(k, j)$. The more samples that can be persuaded (i.e. the lower is j), the lower our sustainable growth threshold. However, recall that lower values of j give a higher persuasion threshold. There exists a trade off between these thresholds, which we address next.

3.3 Combining thresholds to obtain an overall threshold

Two thresholds have been considered. For the planner to persuade players who sample $j \geq 1$ players or more to play B, we require x to be greater than or equal to the persuasion threshold $\tau^*(k, j, \xi)$. Conditional on persuading only those players, for the share of B players to then grow, we require x to be above the sustainable growth threshold $x^*(k, j)$. Combining the two, we see that for the population to be persuaded to play B in such a manner, we require a share $\max\{\tau^*(k, j, \xi), x^*(k, j)\}$, plus possibly some arbitrarily small amount, of players to play B.

$$x \geq \max\left\{\tau^*(k, j, \xi), x^*(k, j)\right\}.$$

This combined threshold may vary across j and the planner can target the j that makes persuasion possible at the lowest values of x. That is, we take the

$k\backslash\xi$	5/9	4/7	3/5	5/8	2/3	3/4	7/9	4/5
2	0.11	0.14	0.20	0.25	0.33	0.50	0.56	0.60
3	0.33	0.36	0.40	0.44	0.50	0.50	0.50	0.56
4	0.23	0.27	0.33	0.38	0.46	0.61	0.65	0.69
5	0.37	0.40	0.44	0.48	0.50	0.58	0.64	0.68
9	0.41	0.44	0.48	0.50	0.53	0.67	0.68	0.70
19	0.46	0.48	0.50	0.54	0.58	0.69	0.71	0.74
31	0.48	0.50	0.54	0.56	0.61	0.70	0.73	0.76
51	0.50	0.52	0.55	0.57	0.62	0.71	0.74	0.77

Table 1: **Overall thresholds.** Values of $\tau^+(k,\xi)$ for various values of k and ξ. Note that $\tau^+(k,\xi)$ increases in ξ and typically increases in k, except for integer effects. That is, greater natural bias towards strategy A (i.e. higher ξ) makes it harder to persuade players to choose B. More concentrated priors after sampling but before receiving a signal (i.e. higher k) also make it harder to persuade players to choose B. For $k = 2, 3$, values $\xi = 3/4, 7/9$ are the highest ξ for which a strict majority of the population playing B is sufficient to persuade the remainder of the population. It is not difficult to calculate these exact thresholds for higher k, but the fractions become somewhat more cumbersome. In the $k = 51$ case, both numerator and denominator have 16 digits.

minimum over j to obtain an **overall threshold**

$$\tau^+(k,\xi) = \min_{j:\,0<j\leq k} \max\Big\{\tau^*(k,j,\xi),\ \ x^*(k,j)\Big\}.$$

As persuasion thresholds are continuous and weakly increasing in ξ and sustainable growth thresholds are independent of ξ, the overall threshold is continuous and weakly increasing in ξ. It typically increases in k, as larger sample sizes lead to more concentrated beliefs ν after sampling, which are harder to persuade. However, there is sometimes local non-monotonicity in k due to integer effects. Example values of $\tau^+(k,\xi)$ for different values of k and ξ are given in Table 1. Plots of $\tau^+(k,\xi)$ are given in Figure 4.

Remark 9. Consider a planner who adopts optimal persuasion to maximize the range of x over which the population converges to $x = 1$. That is, the planner sets $\tau = \tau^+(k,\xi)$. From the definition of our thresholds, it follows that

(i) If $x(0) < \tau^+(k,\xi)$, then $\dot{x}(t) \leq 0$ for all $t \geq 0$ and $\lim_{t\to\infty} x(t) = 0$;

(ii) If $x(0) > \tau^+(k,\xi)$, then $\dot{x}(t) \geq 0$ for all $t \geq 0$ and $\lim_{t\to\infty} x(t) = 1$;

(iii) If $x(0) = \tau^+(k,\xi)$, then $\dot{x}(0) \geq 0$. If sustainable growth binds for the marginal persuaded sample, then $\dot{x}(0) = 0$ and $x = \tau^+(k,\xi)$ is an unstable

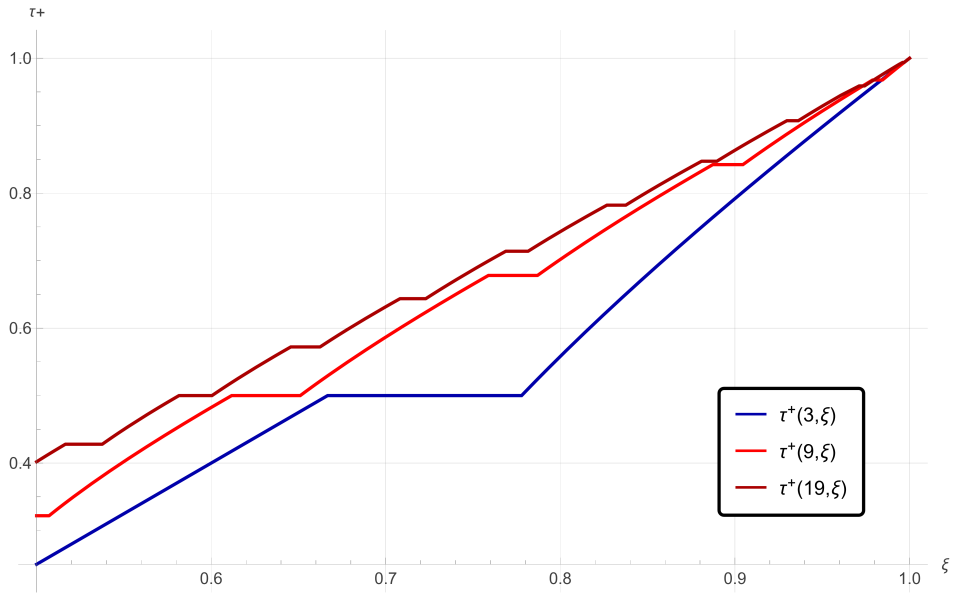

Figure 4: Overall thresholds. Plots of overall threshold $\tau^+(k,\xi)$ for $k = 3, 9, 19$ and $\xi \in (1/2, 1)$. Observe that the overall threshold is weakly increasing in ξ. For these k, it also increases in k. Higher values of ξ and larger sample sizes k make it more difficult for the planner to persuade the population.

equilibrium. Otherwise, $\dot{x}(0) > 0$, so the trajectory immediately enters $(\tau^+(k,\xi), 1)$ and (ii) applies.

Note that, from any initial condition $x(0)$, dynamics under optimal persuasion never cross $x = \tau^+(k,\xi)$, thus the signal $\sigma(x)$ never changes. A consequence is that there is no possibility of agents inferring x from a change in the signal. ◁

3.4 When is a majority sufficient?

Consider the model without persuasion. In our setup, this is equivalent to setting $\tau = 0$ so that the signal is uninformative. Without signals, agents best respond to their beliefs following sampling. Consequently, beliefs only matter via the expected value of x, which equals j/k for an agent who has sampled j draws of B from k. The model reduces to the sampling best response dynamics (Oyama et al., 2015), under which risk dominance of A (i.e. $\alpha > 0$) implies that a majority of players playing A is sufficient for A to take over the population.

Lemma 6 (Without a signal, risk dominance with a majority wins).
Consider $\tau = 0$. If $x(0) < \frac{1}{2}$, then $x(t) \to 0$ as $t \to \infty$.

Despite the advantage afforded to A due to risk dominance, it can sometimes be true that a majority for either strategy suffices for it to take over the popu-

lation. This happens if and only if information from sampling is too coarse to distinguish between A and B at $x = 1/2$, specifically when sample size k is odd and the best response threshold ξ is low enough that a sample majority of one in favor of B is enough to make B a best response.

Lemma 7 (Without a signal, starting from $x(0) = 1/2$). *Consider $\tau = 0$.*

(i) If k is even or $\xi > \frac{k+1}{2k}$, $x(0) = \frac{1}{2}$, then $x(t) \to 0$ as $t \to \infty$.

(ii) If k is odd and $\xi \leq \frac{k+1}{2k}$, then $x = \frac{1}{2}$ is an equilibrium which is asymptotically unstable. Furthermore, if $x(0) > \frac{1}{2}$, then $x(t) \to 1$ as $t \to \infty$.

Lemma 7(ii) gives sufficient conditions under which a majority of agents playing B causes the entire population to converge to B under sampling best response dynamics with no signals thus no persuasion. Lemma 7(i) together with continuity shows that these conditions are also necessary. It is natural to ask how persuasion changes these conditions. In general, we determine conditions under which persuasion can negate the incentive-driven bias towards A. That is, we give conditions under which a majority of agents playing B is sufficient to persuade the entire population to converge to B under dynamic updating.

Stated formally, we give conditions under which it is possible to choose a threshold τ such that $x(0) > 1/2$ implies $x(t) \to 1$ as $t \to \infty$. Given our previous discussion and definitions, observe that the ability to choose such a threshold is equivalent to our overall threshold τ^+ being less than or equal to half,

$$\tau^+(k, \xi) \leq \frac{1}{2}. \qquad \boxed{\text{Majority persuasion is possible}}$$

Given that $\tau^+(k, \xi)$ increases in ξ and tends to increase in k, there is a trade off between k and ξ when it comes to satisfying this inequality. Moreover, for any value of k, we can find low enough $\xi \in (1/2, 1)$ such that the inequality is satisfied and a majority of B players suffices to persuade the entire population.

Lemma 8 (Possibility of a majority being sufficient).

(i) For all $k \geq 2$, there exists $\xi \in (1/2, 1)$ such that $\tau^+(k, \xi) \leq 1/2$.

(ii) There exists a maximum such ξ, which we denote $\xi_{\frac{1}{2}}(k)$.

Our next result provides a precise characterization of $\xi_{\frac{1}{2}}(k)$. As a majority suffices if and only if $\xi \leq \xi_{\frac{1}{2}}(k)$, this formally answers the question of when a majority is sufficient to persuade a population to adopt a convention.

In general, for even k, it turns out that the largest persuadable ξ equals the probability that when a fair coin is tossed k times, at least half of the tosses come up heads. For odd k, the expression is similar but does not simplify to the same extent.

Theorem 1 (When is a majority sufficient?)**.**
Let $J_l \sim Bin(l, \frac{1}{2})$. Then,

$$\xi_{\frac{1}{2}}(k) = \begin{cases} \Pr\left[J_k \leq \frac{k}{2}\right] = \frac{1}{2} + \frac{1}{2}\Pr\left[J_k = \frac{k}{2}\right], & \text{if } k \text{ is even,} \\ \frac{k+1}{2k}\frac{\Pr[J_k \leq \frac{k+1}{2}]}{\Pr[J_{k-1} \leq \frac{k-1}{2}]} = \frac{1}{2} + \frac{1}{2}\frac{\frac{1}{k+1} + \Pr\left[J_k = \frac{k+1}{2}\right]}{\frac{k}{k+1} + \Pr\left[J_k = \frac{k+1}{2}\right]}, & \text{if } k \text{ is odd.} \end{cases}$$

In general, persuasion is harder at higher values of k. Indeed, both expressions in the theorem are decreasing in k. As $k \to \infty$, we have that $\xi_{\frac{1}{2}}(k) \to 1/2$ and the ability to persuade disappears. However, due to integer effects, specifically the impossibility of evenly split samples when k is odd, the maximum value of $\xi_{\frac{1}{2}}(k)$ across k is obtained at $k = 3$ and not $k = 2$.

Corollary 1 (Maximum persuadable ξ)**.**
Across $k \geq 2$, $\xi_{\frac{1}{2}}(k)$ is maximized at $k = 3$, giving $\xi_{\frac{1}{2}}(3) = 7/9$.

For $k = 3$ and $\xi \leq 7/9$, a simple majority of the population playing B is sufficient to persuade the entire population to converge to B. This bound is surprisingly high. To see this, we convert back from ξ to payoffs. Recalling that $\xi = (1+\alpha)/(2+\alpha)$, we have that the bounds $\xi_{\frac{1}{2}}(k)$ correspond to upper bounds $\alpha_{\frac{1}{2}}(k)$ on the incentive parameter α. Consider the case of zero payoff for miscoordination (i.e. $\alpha = \beta$, Figure 3b). The payoff for coordination on A is then $1 + \alpha$, so the maximum such payoff for which a simple majority suffices for convergence to B under optimal persuasion is given by

$$1 + \alpha_{\frac{1}{2}}(k) = \frac{\xi_{\frac{1}{2}}(k)}{1 - \xi_{\frac{1}{2}}(k)}.$$

These payoffs are shown in Figure 5 for $k = 2, 3, 20$.

Remark 10. Compare the examples in Figure 5 to the model without persuasion. For $k = 2$ and $k = 20$, in the absence of persuasion, Lemma 7(i) tells us that any incentive advantage for A (i.e. $\alpha > 0$) makes a simple majority insufficient for convergence to B. Further, for $k = 2$, the process will converge to $x = 0$ from

	A	B
A	**3**	0
B	0	1

(a) $k = 2$

	A	B
A	**3.5**	0
B	0	1

(b) $k = 3$

	A	B
A	$\approx$ **1.43**	0
B	0	1

(c) $k = 20$

Figure 5: Payoff asymmetries overcome by persuasion. Entries denote the payoff of the row player. Consider the case $\beta = \alpha$ in Figure 3, so that pairwise miscoordination payoffs are zero. Payoff matrices illustrate the largest incentive bias in favor of A for which an initial majority of B players suffices for convergence to full adoption of B under optimal persuasion.

any $x(0) < 1$. This starkly contrasts with the maximum incentive advantages $\alpha_{\frac{1}{2}}(k)$ under which persuasion from a majority is possible under both

- small sample sizes, $\quad k = 2 \implies \alpha_{\frac{1}{2}}(k) = 2,$
- large sample sizes, $\quad k = 20 \implies \alpha_{\frac{1}{2}}(k) \approx 0.43.$

For $k = 3$, in the absence of persuasion, Lemma 7(ii) tells us that a simple majority suffices for convergence to B if and only if $\xi \leq 2/3$. This corresponds to $\alpha \leq 1$. This compares to our bound with persuasion of $\alpha_{\frac{1}{2}}(3) = 5/2$. The difference corresponds to the marginal effect of persuasion when added to the existing effect of coarse information from a small, odd sample size. ◁

Remark 11. Using Stirling's expansion of the Gamma function, and hence factorials, the expressions in Theorem 1 can be approximated.[12]

$$\xi_{1/2}(k) = \begin{cases} \dfrac{1}{2} + \dfrac{1}{\sqrt{2\pi k}} - \dfrac{1}{4\sqrt{2\pi}}\, k^{-3/2} + O(k^{-5/2}), & k \text{ even}, \\ \dfrac{1}{2} + \dfrac{1}{\sqrt{2\pi k}} + \left(\dfrac{1}{2} - \dfrac{1}{\pi}\right) k^{-1} + O(k^{-3/2}), & k \text{ odd}. \end{cases}$$

Thus, $\xi_{1/2}(k)$ converges to $1/2$ at rate $k^{-1/2}$. Relative to the common leading approximation $1/2 + (2\pi k)^{-1/2}$, odd k has a positive correction of order k^{-1}, whereas even k has a negative correction of order $k^{-3/2}$. Furthermore, the change in the leading approximation $1/2 + (2\pi k)^{-1/2}$ across consecutive values of k is $O(k^{-3/2})$. Consequently, the value of $\xi_{1/2}(k)$ for odd k exceeds the value at either neighboring even k by $(1/2 - 1/\pi)\, k^{-1} + O(k^{-3/2})$. ◁

[12] See, for example, the NIST Digital Library of Mathematical Functions, Equation (5.11.3): https://dlmf.nist.gov/5.11#E3.

3.5 Asymptotic stability

The asymptotic stability of $x = 0$ (all A) and $x = 1$ (all B) depends on whether a single contrarian element in a sample is enough to incentivize the alternative strategy. The reason for this is simple: if a strategy is present in some small quantity ε, then the probability of it appearing at least once in a sample of size $k \geq 2$ is greater than ε, whereas the probability of it appearing at least twice is less than ε. Consequently, $x = 1$ is asymptotically stable if and only if players who sample $j = k - 1$ draws of B are persuaded to play B.

Proposition 1 (Stability of $x = 1$)**.**

(i) If $\tau = 0$, then $x = 1$ is asymptotically stable if and only if $\xi \leq \frac{k-1}{k}$.

(ii) If $\tau \in (0, 1)$, then $\tau \geq \tau^*(k, k-1, \xi)$ if and only if $x = 1$ is asymptotically stable.

By definition of $\tau^+(k, \xi)$, we have that if $\tau = \tau^+(k, \xi)$, then $\dot{x} > 0$ on $(\tau^+(k, \xi), 1)$, which implies asymptotic stability of $x = 1$. Therefore, $\tau^+(k, \xi) \geq \tau^*(k, k-1, \xi)$. It follows that setting a threshold $\tau = \tau^+(k, \xi)$ to maximize the range of persuadable x has the side effect of making $x = 1$ asymptotically stable.

Regarding $x = 0$, we know from Lemma 2 that it is asymptotically stable for all thresholds τ. That is, a signal cannot break asymptotic stability of $x = 0$. Further, independent of τ, we can show a minimum size of the basin of attraction of $x = 0$. Specifically, if $x(0)$ is low enough that (i) players sampling a single B are not persuaded; and (ii) persuading players sampling at least two B is not enough for sustainable growth, then the process converges to $x = 0$.

Proposition 2 (Stability of $x = 0$)**.**

For all $\tau \in [0, 1)$, if

$$x(0) \in \Big[0, \min\left\{\tau^*(k, 1, \xi),\, x^*(k, 2)\right\}\Big),$$

then $x(t) \to 0$ as $t \to \infty$.

4 Under- and overinference

Consider agents who either underinfer or overinfer, relative to a Bayesian, from the data they observe. Would the planner in our model prefer that agents underinfer or overinfer? In general, there may be competing incentives. On

the one hand, the planner may prefer agents to draw weak inferences from their samples, keeping their beliefs diffuse so that they remain amenable to persuasion. On the other hand, the planner wishes agents to draw strong inferences from the signal.

Underinference and overinference are represented by weighting the likelihood by a power of $c > 0$, where $c < 1$ represents underinference and $c > 1$ represents overinference. This specification is due to Grether (1980). A thorough exposition and comparison to other biases can be found in Benjamin (2019).

Consider under- and overinference with respect to samples. Recall that a beta distribution with parameters j and $k - j$ can be understood as an update of an improper prior density $x^{-1}(1-x)^{-1}$ after observing j successes from k draws from a binomial distribution. If we weight the likelihood of such draws by a power of c, we instead obtain a beta distribution with parameters cj and $c(k-j)$. Using ν^c to represent this density,

$$\nu^c(x) = \frac{x^{cj-1}(1-x)^{c(k-j)-1}}{B(cj,\, c(k-j))}. \qquad \boxed{\text{Beliefs from sample with under/overinf}}$$

Note that we still have $\mathbb{E}_{\nu^c} = j/k$. Under or overinference does not change an agent's expected beliefs after observing a sample, though it does make these beliefs more or less diffuse.

Next, consider under- and overinference with respect to the observed signal. Consider the posterior after the signal $\bar{\sigma} = 1$. Recall that the likelihood for the signal is given by an indicator function:

$$\Pr(\bar{\sigma} = 1 \mid x) = \mathbf{1}_{\{x \geq \tau\}},$$

so that when we raise it to any power c,

$$\left[\mathbf{1}_{\{x \geq \tau\}}\right]^c = \mathbf{1}_{\{x \geq \tau\}}. \qquad \boxed{\text{Likelihood from signal with under/overinf}}$$

Therefore, updating from the signal is unaffected by over or underinference.

$$\nu^c(x \mid x \geq \tau) = \frac{\mathbf{1}_{\{x \geq \tau\}}\, \nu^c(x)}{\int_\tau^1 \nu^c(y)\, dy}, \qquad \mathbb{E}_{\nu^c|1} = \frac{\int_\tau^1 x^{cj}(1-x)^{c(k-j)-1}\, dx}{\int_\tau^1 x^{cj-1}(1-x)^{c(k-j)-1}\, dx}.$$

As a consequence, the only effect of under/overinference is on inference from the sample. Changing the parameter c is like changing j and $k - j$ by

the same proportions. That is, it is similar to changing sample size while keeping the sample mean the same. We saw in our previous section that low sample size is associated with easy persuasion and high sample size is associated with harder persuasion. Thus, underinference leads to easier persuasion and overinference leads to harder persuasion. In Figure 6 we see how τ^+ changes with ξ for different values of c, with higher c leading to a higher τ^+ curve. In Figure 7 we see how τ^+ changes with c for different values of ξ. Notably, underinference allows our previous bound on persuasion by simple majority to be outperformed. Our bound was $\xi_{\frac{1}{2}}(3) = 7/9$, but a sample size of 5 together with $c = 1/3$ allows persuasion by simple majority for ξ as high as $4/5$.

Definition 2. Let quantities defined prior to our introduction of c be similarly defined using the amended posterior expectations. To examine comparative statics according to c, add c to arguments. For example, persuasion thresholds $\tau^*(k, j, \xi, c)$ are defined as $\tau^*(k, j, \xi)$ for varying c. Similarly define $\tau^+(k, \xi, c)$ and $\xi_{\frac{1}{2}}(k, c)$. Sustainable growth thresholds $x^*(k, j)$ are independent of c.

Theorem 2 (Comparative statics of under- and overinference)**.**

(i) If $\tau^*(k, j, \xi, c') = 0$ for some c', then $\tau^*(k, j, \xi, c) = 0$ for all c.

(ii) Otherwise, $\tau^*(k, j, \xi, c)$ strictly increases in c.

Recall that the overall threshold τ^+ depends on both persuasion threshold τ^* and sustainable growth threshold x^*. Consequently, τ^+ increases in c on intervals on which τ^* makes the marginal difference, and is constant in c on intervals on which x^* makes the marginal difference (see Figure 7). In summary, τ^+ is weakly increasing in c. Persuading the population is easier when c is lower.

An implication of weakly increasing τ^+ is weakly decreasing $\xi_{\frac{1}{2}}(k, c)$. That is, underinference increases and overinference decreases the range of ξ such that a majority suffices to persuade the population to adopt B. The following corollary summarizes these implications of Theorem 2. In words, the planner would like agents to underinfer. Underinference allows the planner to (i) use lower thresholds (for fixed ξ), and (ii) increase the range of ξ for which a majority is sufficient.

Corollary 2 (Further comparative statics)**.**

(i) $\tau^+(k, \xi, c)$ is weakly increasing in c.

(ii) $\xi_{\frac{1}{2}}(k, c)$ is weakly decreasing in c.

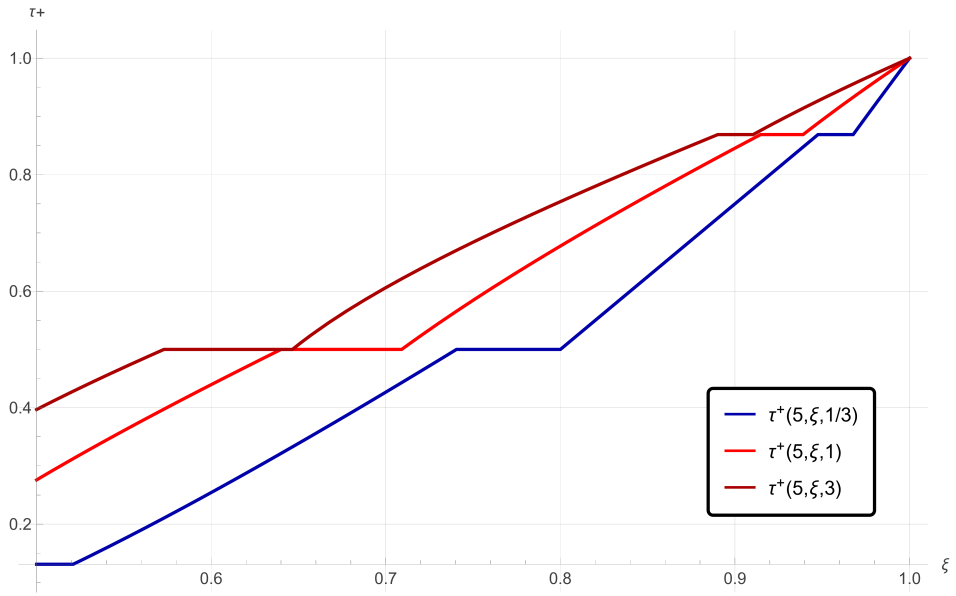

Figure 6: Overall thresholds with under/overinference. Plots of overall threshold $\tau^+(k, \xi, c)$ for $k = 5$, $\xi \in (1/2, 1)$, $c = 1/3, 1, 3$. The overall threshold is weakly increasing in ξ and weakly increasing in c. Underinference makes it easier and overinference makes it more difficult for the planner to persuade the population.

5 Conclusion

This paper derives conditions under which a planner can gradually persuade a population to adopt a target convention. Effective persuasion leverages a feedback loop: if a sufficient share of the population is persuaded to adopt the target convention today, then the share of persuadable agents grows endogenously, which in turn makes subsequent persuasion more effective.

A central lesson is that successful persuasion by the planner requires intermediate levels of private information. Private information must be noisy for beliefs to be movable by the planner's signal, but also informative so that higher adoption generates a larger future pool of persuadable agents. When this is satisfied, a simple majority can persuade the population to converge to the target convention even when private incentives strongly favor the alternative.

Our model is readily extendable to non-Bayesian and biased forms of belief updating. Non-Bayesian analysis further reinforces the message that the precision of inference matters. Underinference broadens the range of environments in which persuasion is effective; overinference has the opposite effect.

Several limitations qualify the analysis. First, the planner commits to a threshold disclosure rule that agents understand and trust. This captures many

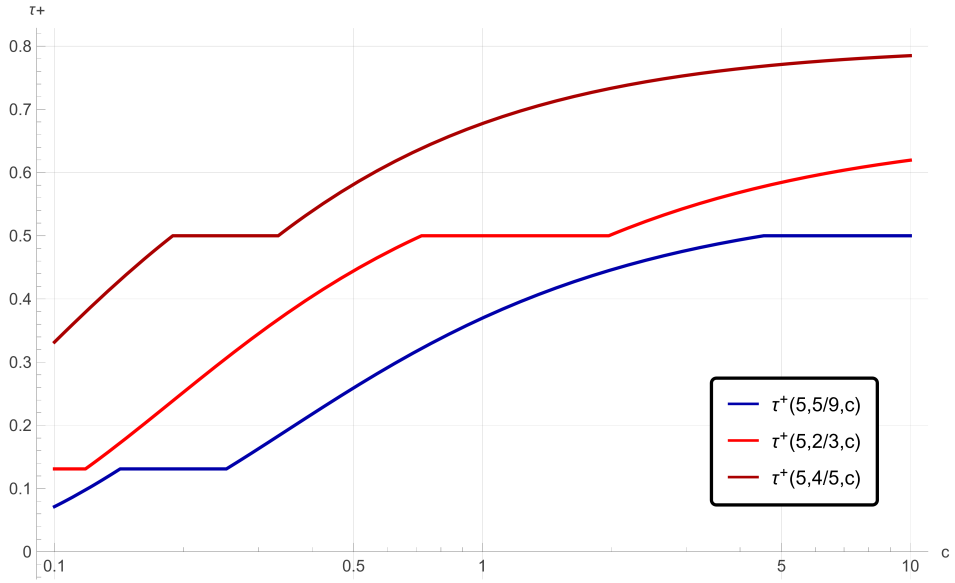

Figure 7: Overall thresholds with under/overinference. Plots of overall threshold $\tau^+(k, \xi, c)$ for $k = 5$, $\xi = 5/9, 2/3, 4/5$, $c \in [1/10, 10]$. The overall threshold is weakly increasing in c. Underinference makes it easier and overinference makes it more difficult for the planner to persuade the population.

forms of verifiable public information, but it abstracts from cheap talk, misinformation, partial credibility, and non-monotone information structures. Accordingly, the thresholds derived here should be read as a useful benchmark: weakening credibility may reduce the planner's power, while allowing non-monotone signals may increase it.

Second, agents update myopically: a revising agent best responds to beliefs formed using only his current random sample and the planner's current signal. Neither past information nor strategic anticipation of future adoption inform the agent's decision. This keeps the state one-dimensional and isolates the feedback from current adoption to future samples. Furthermore, under optimal threshold disclosure, not keeping track of signal histories is largely innocuous because trajectories never cross the threshold and therefore the signal never changes. This robustness is noteworthy as it is not shared by non-monotone disclosure. In Appendix D we shows how non-monotone binary signals can sometimes outperform threshold signals by pooling low and high states and using the two messages as a relay along the adoption path. However, in contrast to the case of monotone disclosure, if agents were to retain and interpret the history of such a non-monotone signal, it could cause the unraveling of the pooling that makes the relay persuasive.

Finally, the baseline model assumes well-mixed samples of a common size,

but in real populations agents differ in how much they observe and also in whom they observe. In richer environments, persuasion may be amplified when influential or well-connected agents become persuaded early, but weakened when homophily or clustering prevents adoption from generating broader evidence. Allowing heterogeneous sample sizes and sampling on a network would clarify when the effect of persuasion spreads broadly across a population and when it remains trapped within local neighborhoods.

A Proofs of results in Section 2

Proof of Lemma 1.
Write ν_j for the beliefs induced after sampling j agents who play B. Let

$$r_j(\bar{\sigma}) := \mathbf{1}\Big\{ \mathbb{E}_{\nu_j|\bar{\sigma}} \geq \xi \Big\}, \qquad \bar{\sigma} \in \{0,1\}.$$

Then, by definition

$$BR(x) = \sum_{j=0}^{k} r_j(\sigma(x)) \binom{k}{j} x^j (1-x)^{k-j} \;=\; \mathbb{E}[\, r_J(\sigma(x))\,], \quad J \sim \text{Bin}(k,x).$$

Step 1: Monotonicity of $r_j(\bar{\sigma})$ in j.

Clearly ν_0 is FOSD-dominated by every other sample belief. Furthermore, as there is no update in response to a signal, this is also true for $\nu_0(\cdot|\sigma(x))$.

Similarly, ν_k FOSD any beliefs. Furthermore, as there is no update in response to a signal, this is also true for $\nu_k(\cdot|\sigma(x))$.

If $0 < j' < j'' < k$ then $\nu_{j'}$, $\nu_{j''}$ exhibit a monotone likelihood ratio (MLR). That is $\nu_{j''}(x)/\nu_{j'}(x)$ is increasing in x. This implies that $\nu_{j''}$ FOSD $\nu_{j'}$.

Conditioning on $x \geq \tau$ or $x < \tau$ preserves MLR and hence FOSD. Thus, $\mathbb{E}_{\nu_{j''}|\bar{\sigma}} \geq \mathbb{E}_{\nu_{j'}|\bar{\sigma}}$ for each $\bar{\sigma} \in \{0,1\}$.

Therefore $j \mapsto r_j(\bar{\sigma})$ is nondecreasing for fixed $\bar{\sigma}$.

Step 2: Monotonicity in the signal.

For any belief ν, $\mathbb{E}_{\nu|0} \leq \mathbb{E}_{\nu} \leq \mathbb{E}_{\nu|1}$, so $r_j(1) \geq r_j(0)$ for all j.

Step 3: Monotonicity in x.

If $x' < x''$ then $J' \sim \text{Bin}(k,x')$ is FOSD by $J'' \sim \text{Bin}(k,x'')$.

It follows that

(i) If $x', x'' < \tau$, then $\sigma(x') = \sigma(x'') = 0$ and

$$BR(x') = \mathbb{E}[r_{J'}(0)] \underset{\text{by Step 1}}{\leq} \mathbb{E}[r_{J''}(0)] = BR(x'').$$

(ii) If $x', x'' \geq \tau$, then $\sigma(x') = \sigma(x'') = 1$ and

$$BR(x') = \mathbb{E}[r_{J'}(1)] \underset{\text{by Step 1}}{\leq} \mathbb{E}[r_{J''}(1)] = BR(x'').$$

(iii) If $x' < \tau \leq x''$, then by Step 2

$$BR(x') = \mathbb{E}[r_{J'}(0)] \underset{\text{by Step 2}}{\leq} \mathbb{E}[r_{J'}(1)] \underset{\text{by Step 1}}{\leq} \mathbb{E}[r_{J''}(1)] = BR(x'').$$

In all cases $x' < x''$ implies $BR(x') \leq BR(x'')$. □

Proof of Lemma 2.
For $\bar{\sigma} \in \{0, 1\}$ write

$$A_{\bar{\sigma}} := \left\{ \nu \in \mathcal{P} : \mathbb{E}_{\nu|\bar{\sigma}} \geq \xi \right\} \quad \text{and} \quad BR_{\bar{\sigma}}(x) := \int_{\mathcal{P}} \mathbf{1}_{A_{\bar{\sigma}}}(\nu)\, \mu_x(d\nu).$$

Let $r_j(\bar{\sigma})$ continue to be defined as in Lemma 1.

Part (i): Continuity under a fixed signal.

Fix $\bar{\sigma}$. The integrand $\mathbf{1}_{A_{\bar{\sigma}}}$ is bounded and measurable, and $x \mapsto \mu_x$ is setwise continuous (Remark 5). Setwise continuity implies

$$\lim_{x \to x^*} BR_{\bar{\sigma}}(x) = \int_{\mathcal{P}} \mathbf{1}_{A_{\bar{\sigma}}}\, \mu_{x^*}(d\nu) = BR_{\bar{\sigma}}(x^*),$$

so $BR_{\bar{\sigma}}$ is continuous.

Part (ii): Single possible jump and it is upward at $x = \tau$.

With the threshold signal $\sigma(x) = \mathbf{1}_{\{x \geq \tau\}}$, we have

$$BR(x) = \begin{cases} BR_0(x), & x < \tau, \\ BR_1(x), & x \geq \tau, \end{cases}$$

hence BR is continuous on $[0, \tau)$ and $(\tau, 1]$ by Part (i). At $x = \tau$,

$$\lim_{x \uparrow \tau} BR(x) = BR_0(\tau), \qquad \lim_{x \downarrow \tau} BR(x) = BR_1(\tau).$$

For any ν, $\mathbb{E}_{\nu|0} \leq \mathbb{E}_{\nu|1}$, so $A_0 \subseteq A_1$ and therefore $BR_0(\tau) \leq BR_1(\tau)$. Thus the only possible discontinuity is an upward jump at $x = \tau$.

Part (iii): $x = 0$ is an asymptotically stable equilibrium.

$k \geq 2$ with $\xi > 1/2$ implies that $\xi > 1/k$.

$E_{\nu_0} = E_{\nu_0|\bar{\sigma}} = 0$ for $\bar{\sigma} \in \{0, 1\}$. Therefore, $r_0(\bar{\sigma}) = 0$.

If $\tau = 0$, then $\sigma(x) \equiv 1$. We have $E_{\nu_1} = E_{\nu_1|1} = 1/k$. Therefore, $r_1(1) = 0$.

If $\tau > 0$, consider $x < \tau$. $E_{\nu_1|0} \leq E_{\nu_1} = 1/k$. Therefore, $r_1(0) = 0$.

Hence, for either $x \in (0, \tau)$ when $\tau > 0$, or arbitrary x when $\tau = 0$, we have

$$BR(x) \leq 1 - (1-x)^k - kx(1-x)^{k-1} = O(x^2),$$

therefore $BR(x) - x < 0$ for sufficiently small $x > 0$.

Moreover, $BR(0) = 0$, so $x = 0$ is an asymptotically stable equilibrium. □

B Proofs of results in Section 3

B.1 Results on expectations and thresholds

Proof of Lemma 3.

If $j = 0$ (resp. $j = k$), ν is a point mass at 0 (resp. 1), so $\mathbb{E}_{\nu|1} = 0$ (resp. 1).

Now take $0 < j < k$. The conditional mean given $\bar{\sigma} = 1$ (i.e. $x \geq \tau$) is

$$\mathbb{E}_{\nu|1} = \frac{\int_\tau^1 x^j (1-x)^{k-j-1}\, dx}{\int_\tau^1 x^{j-1}(1-x)^{k-j-1}\, dx}.$$

Using the beta $B(\cdot,\cdot)$ and regularized incomplete beta $I_\tau(\cdot,\cdot)$ functions,

$$\int_\tau^1 x^j (1-x)^{k-j-1}\, dx = B(j+1, k-j)\left[1 - I_\tau(j+1, k-j)\right],$$

$$\int_\tau^1 x^{j-1} (1-x)^{k-j-1}\, dx = B(j, k-j)\left[1 - I_\tau(j, k-j)\right].$$

Hence,

$$\mathbb{E}_{\nu|1} = \underbrace{\frac{B(j+1,k-j)}{B(j,k-j)}}_{\substack{=\frac{j}{k} \\ \text{from } B(\cdot,\cdot) \text{ definition}}} \cdot \underbrace{\frac{1 - I_\tau(j+1,k-j)}{1 - I_\tau(j,k-j)}}_{\substack{=\frac{\text{BinCDF}(j;k,\tau)}{\text{BinCDF}(j-1;k-1,\tau)} \\ \text{by binomial-beta identities}}}.$$

□

Proof of Lemma 4.

Cases $j = 0$ and $j = k$.

If $j = 0$, ν_0 is a point mass at 0, so $\mathbb{E}_{\nu_0|1} = 0$ for all τ. No τ can make $\mathbb{E}_{\nu_0|1} \geq \xi$, so $\tau^*(k, 0, \xi)$ is undefined.

If $j = k$, ν_k is a point mass at 1, so $\mathbb{E}_{\nu_k|1} = 1 \geq \xi$ for all τ. Therefore, $\tau^*(k, k, \xi) = 0$.

Case $0 < j < k$.

As ν_j is continuous with full support on $(0, 1)$, the expectation $\mathbb{E}_{\nu_j|1}$ is continuous and strictly increasing in τ on $[0, 1)$. Moreover,

$$\phi_j(0) = \mathbb{E}_{\nu_j} = \frac{j}{k} \quad \text{and} \quad \lim_{\tau \uparrow 1} \mathbb{E}_{\nu_j|1} = 1.$$

Therefore,

(I) If $\xi \leq \frac{j}{k}$, then $\mathbb{E}_{\nu_j|1} \geq \xi$ for all $\tau \geq 0$. Therefore, $\tau^*(k, j, \xi) = 0$.

(II) If $\frac{j}{k} < \xi < 1$, by continuity and strict monotonicity there exists a unique $\tau^* = \tau^*(k, j, \xi) \in (0, 1)$ such that $\mathbb{E}_{\nu_j|1} = \xi$.

Monotonicities.

$\mathbb{E}_{\nu_j|1}$ is strictly increasing in τ, so the minimal $\tau \in [0, 1]$ solving $\mathbb{E}_{\nu_j|1} \geq \xi$ is weakly increasing in ξ.

For fixed τ, if $0 < j' < j'' < k$, then by MLR (hence FOSD) in j and preservation of MLR under truncation, the conditional distribution $\nu_{j''}(\cdot \mid 1)$ FOSD $\nu_{j'}(\cdot \mid 1)$, so $\mathbb{E}_{\nu_{j''}|1} \geq \mathbb{E}_{\nu_{j'}|1}$. Therefore, $\tau^*(k, j, \xi)$ is weakly decreasing in j. The boundary case $j = k$ is consistent with this inequality. □

Proof of Lemma 5.

Recall the growth condition (Gr):

$$f(x) := 1 - \text{BinCDF}(j - 1;\, k,\, x) - x = I_x\big(j,\, k - j + 1\big) - x \geq\ 0,$$

where $I_x(\cdot, \cdot)$ is the regularized incomplete beta.

Cases $j = 0$ and $j = k$.

If $j = 0$, then $f(x) = 1 - x \geq 0$ for all x. Thus, $x^*(k, 0) = 0$.

If $j = k$, then $f(x) = x^k - x$, which is ≥ 0 only at $x = 0, 1$. Thus $x^*(k, k) = 1$.

Case $j = 1$.

If $j = 1$, then $f(x) = 1 - \text{BinCDF}(0; k, x) - x = 1 - (1-x)^k - x \geq 0$ for all x. Thus $x^*(k, 1) = 0$.

Case $2 \leq j \leq k-1$.

If $2 \leq j \leq k-1$, then $f(x) = I_x(j, k-j+1) - x$ is C^1 on $(0,1)$ with first derivative

$$f'(x) = \frac{x^{j-1}(1-x)^{k-j}}{B(j, k-j+1)} - 1 =: p_j(x) - 1.$$

$f'(0) = -1 < 0$ and $f'(1) = -1 < 0$, which implies that there exists $\varepsilon > 0$ such that $f(x) < 0$ on $(0, \varepsilon)$ and $f(x) > 0$ on $(1-\varepsilon, 1)$. Continuity of $f(x)$ then implies that there exists $x^* \in (0,1)$ such that $f(x^*) = 0$.

Regarding uniqueness, the second derivative is

$$f''(x) = p_j(x) \left(\frac{j-1}{x} - \frac{k-j}{1-x} \right).$$

$f''(x) > 0$ on $\left(0, \frac{j-1}{k-1}\right)$ and $f''(x) < 0$ on $\left(\frac{j-1}{k-1}, 1\right)$. That is, f is strictly convex then strictly concave. Together with $f(x)$ being < 0 at x arbitrarily close to 0 and > 0 at x arbitrarily close to 1, this implies that $f(x)$ crosses zero exactly once in $(0,1)$ and stays above zero at higher values of x. There is a unique $x^*(k, j) \in (0,1)$ satisfying

$$f(x) = 1 - \text{BinCDF}(j-1;\, k,\, x) - x \geq 0 \quad \iff \quad x \in \{0\} \cup [x^*(k,j),\, 1].$$

Monotonicity in j.

If $2 \leq j' < j'' \leq k-1$, then for every $x \in [0,1]$,

$$1 - \text{BinCDF}(j''-1;\, k,\, x) \leq 1 - \text{BinCDF}(j'-1;\, k,\, x).$$

Hence the set $\{x' : \forall\, x \geq x',\ 1 - \text{BinCDF}(j-1; k, x) \geq x\}$ shrinks as j increases, so its minimum element $x^*(k, j)$ is weakly increasing in j.

Adding $x^*(k, 0) = x^*(k, 1) = 0$ and $x^*(k, k) = 1$ to our consideration, we see that $x^*(k, j)$ is weakly increasing in j for all $j = 0, 1, \ldots, k$.

□

B.2 Results on when a majority is sufficient

Proof of Lemmas 6 and 7.

Set $\tau = 0$. Let $m := \lceil k\xi \rceil$, so

$$BR(x) = \Pr[\mathrm{Bin}(k, x) \geq m], \qquad x \in [0, 1].$$

Define $g(x) := BR(x) - x$, so $\dot{x} = g(x)$. Note that $g(0) = 0$ and, as $m > k/2$, symmetry of the binomial distribution implies $g(1/2) \leq 0$. Differentiating,

$$g'(x) = k \Pr[\mathrm{Bin}(k-1, x) = m-1] - 1.$$

As $\xi > \frac{1}{2}$, we have $m - 1 \geq \frac{k-1}{2}$, so that $\Pr[\mathrm{Bin}(k-1, x) = m-1]$ is strictly increasing in x on $(0, \frac{1}{2}]$. Hence g' is strictly increasing on $(0, \frac{1}{2}]$ with $g'(0) = -1 < 0$.

Lemma 6: If $x(0) < \frac{1}{2}$, then $x(t) \to 0$.

$g(0) = 0 \geq g(1/2)$, $g'(0) = -1$ and g' increasing on $(0, \frac{1}{2}]$ together imply that $g(x) < 0$ on $(0, \frac{1}{2})$. Therefore, from $x(0) < \frac{1}{2}$, the trajectory is strictly decreasing. By continuity and being bounded below by 0, $x(t)$ must decrease to some x' such that $g(x) = 0$. The only such point is $x' = 0$.

Lemma 7(i): If $x(0) = \frac{1}{2}$ and either k is even or k is odd with $\xi > \frac{k+1}{2k}$, then $x(t) \to 0$.

At $x = \frac{1}{2}$, for both subcases we have the strict inequality

$$BR\left(\tfrac{1}{2}\right) = \Pr\left[\mathrm{Bin}(k, \tfrac{1}{2}) \geq m\right] < \tfrac{1}{2}$$

Thus $g(\frac{1}{2}) < 0$, the trajectory immediately enters $(0, \frac{1}{2})$, and part (i) gives $x(t) \to 0$.

Lemma 7(ii): If k is odd and $\xi \leq \frac{k+1}{2k}$, then $x = \frac{1}{2}$ is an unstable equilibrium.

In this case $m = \frac{k+1}{2}$, so by symmetry $BR(\frac{1}{2}) = \frac{1}{2}$ and $g(\frac{1}{2}) = 0$. Moreover,

$$g'\left(\tfrac{1}{2}\right) = k \Pr\left[\mathrm{Bin}\left(k-1, \tfrac{1}{2}\right) = \tfrac{k-1}{2}\right] - 1 = k \frac{\binom{k-1}{k-1/2}}{2^{k-1}} - 1 \; > \; 0$$

which holds because the central binomial coefficient exceeds the average coefficient. That is, $\binom{k-1}{k-1/2} > \frac{1}{k} \sum_{j=0}^{k-1} \binom{k-1}{j} = \frac{2^{k-1}}{k}$ for $k \geq 3$. Therefore, $x = \frac{1}{2}$ is asymptotically unstable.

It remains to show that trajectories starting to the right of $1/2$ converge to 1.

Substituting for m

$$g'(x) = k \Pr[\mathrm{Bin}(k-1, x) = \tfrac{k-1}{2}] - 1,$$

which is strictly decreasing on $(1/2, 1)$, so g is strictly concave on this interval. Since

$$g\left(\frac{1}{2}\right) = 0 \qquad \text{and} \qquad g(1) = 0,$$

strict concavity implies

$$g(x) > 0 \qquad \text{for all } x \in \left(\frac{1}{2}, 1\right).$$

Thus, from $x(0) \in (\frac{1}{2}, 1)$, the trajectory is strictly increasing and bounded above by 1. By continuity, $x(t)$ must increase to some x' such that $g(x) = 0$. The only such point is $x' = 1$. □

Proof of Lemma 8.

Let

$$\mathcal{S}_k := \left\{\xi \in (\tfrac{1}{2}, 1) : \ \tau^+(k, \xi) \leq \tfrac{1}{2}\right\}.$$

We show $\mathcal{S}_k \neq \varnothing$ and that $\xi_{\frac{1}{2}}(k) := \max \mathcal{S}_k$ exists.

Nonemptiness.

Take $j = 1$. By Lemma 5, $x^*(k, 1) = 0 \leq \frac{1}{2}$. By Lemma 3,

$$\mathbb{E}_{\nu_1|1} = \frac{1}{k} \frac{\text{BinCDF}(1;\, k,\, \tau)}{\text{BinCDF}(0;\, k-1,\, \tau)} = \frac{1}{k} \frac{(1-\tau)^k + k\tau(1-\tau)^{k-1}}{(1-\tau)^{k-1}} = \frac{1 + (k-1)\tau}{k}.$$

At $\tau = \frac{1}{2}$, this equals $\frac{k+1}{2k} > \frac{1}{2}$, so for $\xi \in (\frac{1}{2}, \ \frac{k+1}{2k}]$ we have $\tau^*(k, 1, \xi) \leq \frac{1}{2}$ and

$$\tau^+(k, \xi) \ \leq \ \max\{\tau^*(k, 1, \xi),\, x^*(k, 1)\} \ \leq \ \tfrac{1}{2},$$

so $\xi \in \mathcal{S}_k$. Hence $\mathcal{S}_k \neq \varnothing$.

Show that $\mathcal{S}_k \neq (\frac{1}{2}, 1)$.

For $\tau = \frac{1}{2}$,

$$\max_{j<k} \mathbb{E}_{\nu_j|1} < 1.$$

Thus, for ξ close enough to 1, we have $\tau^*(k, j, \xi) > \frac{1}{2}$ for every $j < k$. The remaining case $j = k$ has $x^*(k, k) = 1$. Hence, by definition, $\tau^+(k, \xi) > \frac{1}{2}$ for ξ close enough to 1.

Existence of a maximum.

As persuasion thresholds are weakly increasing and continuous in ξ (Lemma 4), x^* independent of ξ (Lemma 5), it follows from the definition of τ^+ that

$\tau^+(k,\xi)$ is weakly increasing and continuous in ξ. Therefore, $\mathcal{S}_k$ is an interval of the form $(\frac{1}{2}, \xi_{\frac{1}{2}}(k)]$ for some $\xi_{\frac{1}{2}}(k) \leq 1$. □

Proof of Theorem 1.

By definition of τ^+, majority persuasion is possible if and only if there exists $j \in \{1,\dots,k\}$ such that

$$x^*(k,j) \leq \frac{1}{2} \qquad \text{and} \qquad \tau^*(k,j,\xi) \leq \frac{1}{2}.$$

Sustainable growth.

$x^*(k,j) \leq \frac{1}{2}$ implies that

$$1 - \text{BinCDF}(j-1;\, k,\, \tfrac{1}{2}) \;\geq\; \tfrac{1}{2}.$$

The largest j that satisfies this is

$$j^* = \begin{cases} \frac{k+1}{2}, & k \text{ odd}, \\ \frac{k}{2}, & k \text{ even}. \end{cases}$$

Persuasion.

As $\tau^*(k,j,\xi)$ is weakly decreasing in j and weakly increasing in ξ (Lemma 4), the range of ξ such that $\tau^*(k,j,\xi) \leq \frac{1}{2}$ is larger at larger values of j.

Together with the fact that we desire $x^*(k,j) \leq \frac{1}{2}$, this implies that it suffices to consider $\tau^*(k,j^*,\xi)$. The largest ξ that satisfies $\tau^*(k,j^*,\xi) \leq \frac{1}{2}$ then solves

$$\mathbb{E}_{\nu_{j^*}|1} = \frac{j^*}{k} \frac{\text{BinCDF}(j^*;\, k,\, \frac{1}{2})}{\text{BinCDF}(j^*-1;\, k-1,\, \frac{1}{2})} = \xi \qquad \text{(Lemma 3)}$$

Substituting and rearranging, we obtain the expressions from the theorem statement. □

Proof of Corollary 1.

Using Theorem 1, $\xi_{\frac{1}{2}}(k)$ is decreasing in even k and also decreasing in odd k. Therefore, to find the maximum over $k \geq 2$, it suffices to compare $k = 2, 3$.

$$\xi_{\frac{1}{2}}(2) = \frac{3}{4} < \frac{7}{9} = \xi_{\frac{1}{2}}(3).$$

□

B.3 Results on asymptotic stability

Proof of Proposition 1.

(i) No signal ($\tau = 0$).

With $\tau = 0$, agents play B iff $j/k \geq \xi$.

If $\xi > \frac{k-1}{k}$, only sample $j = k$ gives B as a best response, so for all $x \in (0,1)$,

$$BR(x) - x = x^k - x,$$

which is < 0 for all $x \in (0,1)$, so $x = 1$ is not asymptotically stable.

If $\xi \leq \frac{k-1}{k}$, then samples $j = k$ and $j = k-1$ have B as a best response, so

$$BR(x) - x \;\geq\; x^k + k\,x^{k-1}(1-x) - x,$$

which is > 0 on $(1-\varepsilon, 1)$ for some $\varepsilon > 0$, so $x = 1$ is asymptotically stable.

(ii) With signal.

If $\tau \geq \tau^*(k, k-1, \xi)$, then for x in some neighborhood of 1 we have $\sigma(x) = 1$ and hence $BR(x) - x \geq x^k + k\,x^{k-1}(1-x) - x$. As in part (i), this is strictly positive on a neighborhood of 1, so $x = 1$ is asymptotically stable.

Conversely, if $\tau < \tau^*(k, k-1, \xi)$, then by the monotonicity of τ^* in j (Lemma 4) no sample $j \leq k-1$ has B as a best response after any signal. Thus, for x near 1, $BR(x) - x = x^k - x < 0$ on $(0,1)$, so $x = 1$ is not asymptotically stable.

□

Proof of Proposition 2.

Take $0 < x < \min\{\tau^*(k,1,\xi), x^*(k,2)\}$. Agents with $j = 0$ are never persuaded. Agents with $j = 1$ are not persuaded either. If $x < \tau$, then the signal is 0 and $\mathbb{E}_{\nu_1|0} \leq \mathbb{E}_{\nu_1} = 1/k \leq 1/2 < \xi$. If $x \geq \tau$, then $\tau \leq x < \tau^*(k,1,\xi)$, so $\mathbb{E}_{\nu_1|1} < \xi$. Hence only agents with $j \geq 2$ can play B. Therefore

$$BR(x) \leq 1 - (1-x)^k - kx(1-x)^{k-1} \underbrace{<}_{\text{by } 0<x<x^*(k,2)} x.$$

Thus $\dot{x} < 0$ and any trajectory starting from such an x converges to 0, which is an equilibrium by Lemma 2. □

C Proofs of results in Section 4

C.1 Results on under/overinference

Lemma 9. *Fix $j > 0$, $k > j$, $\tau \in (0,1)$, Then,*

$$\mathbb{E}_{\nu|1}(k,j,\tau,c) = \frac{\int_\tau^1 x^{jc}(1-x)^{(k-j)c-1}\,dx}{\int_\tau^1 x^{jc-1}(1-x)^{(k-j)c-1}\,dx}$$

is strictly decreasing in c on $(0,\infty)$.

Proof.

Step 1: Reduction and exponential form.

For $n \in \{0,1\}$, write

$$Z_n := \int_\tau^1 x^{cj+n-1}(1-x)^{c(k-j)-1}\,dx, \quad \text{so} \quad \mathbb{E}_{\nu|1} = \frac{Z_1}{Z_0}.$$

Integrating Z_1 by parts gives

$$\mathbb{E}_{\nu|1} = \frac{j}{k} + \phi_\tau(c), \quad \text{with} \quad \phi_\tau(c) := \frac{\tau^{cj}(1-\tau)^{c(k-j)}}{c\,k\,Z_0}.$$

Thus to prove the lemma it suffices to prove that $\phi_\tau(c)$ is strictly decreasing in c. It will help to write

$$\phi_\tau(c) = \frac{e^{c\,T(\tau)}}{k\,c\,Z_0} \quad \text{and} \quad Z_0 = \int_\tau^1 \frac{e^{c\,T(x)}}{x(1-x)}\,dx$$
$$\text{where} \qquad T(x) := j\,\log x + (k-j)\,\log(1-x).$$

Step 2: Differentiate $\phi_\tau(c)$ w.r.t. c.

Differentiating using the quotient rule

$$\phi'_\tau(c) = \frac{e^{c\,T(\tau)}\,k\big[c\,Z_0(c)\,T(\tau) - Z_0(c) - c\,Z_0'(c)\big]}{\big(k\,c\,Z_0(c)\big)^2} = \phi_\tau(c)\left(T(\tau) - \frac{1}{c} - \frac{Z_0'(c)}{Z_0(c)}\right),$$

which has the same sign as

$$c\,T(\tau)\,Z_0(c) - Z_0(c) - c\,Z_0'(c) = \int_\tau^1 \left(c\,T(\tau) - 1 - c\,T(x)\right)\,\frac{e^{\,cT(x)}}{x(1-x)}\,dx =: W(c,\tau). \tag{1}$$

We will show that $W(c,\tau) < 0$, thus proving that $\phi_\tau(c)$ is strictly decreasing in c.

Step 3: Derivative of $W(c,\tau)$ w.r.t. τ is strictly positive.

Using the Leibniz integral rule for differentiation, we obtain

$$\frac{d}{d\tau}W(c,\tau) = \frac{Y(\tau)}{\tau(1-\tau)}, \quad \text{where} \quad Y(\tau) := e^{cT(\tau)} + c\,(j-k\tau)\int_\tau^1 \frac{e^{cT(x)}}{x(1-x)}\,dx.$$

To show that $Y(\tau) > 0$, observe that

$$\frac{d}{dx}e^{cT(x)} = c\,T'(x)\,e^{cT(x)} = c\,\frac{(j-kx)}{x(1-x)}e^{cT(x)} \underbrace{\leq}_{\forall x>\tau} c\,\frac{(j-k\tau)}{x(1-x)}\,e^{cT(x)},$$

then integrate both sides w.r.t. x from τ to 1,

$$-e^{cT(\tau)} < c(j-k\tau)\int_\tau^1 \frac{e^{cT(x)}}{x(1-x)}\,dx,$$

therefore $Y(\tau) > 0$.

Step 4: $W(c,\tau) \to 0$ as $\tau \to 1$.

For $\tau > \frac{1}{2}$, the term x^{cj-1} is bounded on $[\tau,1]$ by some constant $C < \infty$ that is independent of τ. Hence

$$\frac{e^{cT(x)}}{x(1-x)} = x^{cj-1}(1-x)^{c(k-j)-1} \leq C(1-x)^{c(k-j)-1}.$$

Moreover, for $x \in [\tau,1]$,

$$|T(\tau) - T(x)| \leq j\log\frac{1}{\tau} + (k-j)\log\frac{1-\tau}{1-x}.$$

Therefore

$$|W(c,\tau)| \leq \int_\tau^1 \left(1 + c|T(\tau) - T(x)|\right)\frac{e^{cT(x)}}{x(1-x)}\,dx$$

$$\leq C \int_\tau^1 \left(1 + cj \log \frac{1}{\tau} + c(k-j) \log \frac{1-\tau}{1-x}\right) (1-x)^{c(k-j)-1}\, dx.$$

The two elementary integrals are

$$\int_\tau^1 (1-x)^{c(k-j)-1}\, dx = \frac{(1-\tau)^{c(k-j)}}{c(k-j)}$$

and

$$\int_\tau^1 \log \frac{1-\tau}{1-x} (1-x)^{c(k-j)-1}\, dx = \frac{(1-\tau)^{c(k-j)}}{c^2(k-j)^2}.$$

Thus $|W(c,\tau)| \to 0$, and hence $W(c,\tau) \to 0$, as $\tau \to 1$.

Step 5: Conclusion.

Step 3 showed $W(c,\tau)$ is strictly increasing in τ and Step 4 showed $W(c,\tau) \to 0$ as $\tau \to 1$. Therefore $W(c,\tau) < 0$ for all $\tau < 1$. Step 2 showed that this suffices for $\phi_\tau(c)$ to be strictly decreasing in c. This completes the proof. □

Proof of Theorem 2.

Recall that $\tau^*(k,j,\xi,c)$ is the lowest value of τ that gives $\mathbb{E}_{\nu|1} \geq \xi$.

Observe that $E_{\nu|1}$ is continuous in τ and c.

If $j = k$, beliefs are unaffected by τ or c. $\mathbb{E}_\nu = \mathbb{E}_{\nu|1} = 1$, so $\tau^*(k,j,\xi,c) = 0$.

If $0 < j < k$, $E_{\nu|1}$ is strictly increasing in τ and,

(i) for $\tau = 0$, independent of c and equal to j/k.

(ii) for $\tau \in (0,1)$, strictly decreasing in c (by Lemma 9).

Therefore,

(i) If $\tau^*(k,j,\xi,c) = 0$ for some c, then $\tau^*(k,j,\xi,c) = 0$ for all c.

(ii) Otherwise, $\tau^*(k,j,\xi,c) > 0$ for all c. Continuity of $E_{\nu|1}$ in τ implies that $E_{\nu|1} = \xi$ at $\tau = \tau^*(k,j,\xi,c)$. Therefore, maintaining $E_{\nu|1} = \xi$ for a higher value of c requires a higher value of $\tau^*(k,j,\xi,c)$.

□

Proof of Corollary 2.

Part (i): $\tau^+(k,\xi,c)$ is weakly increasing in c.

By Theorem 2, $\tau^*(k,j,\xi,c)$ is weakly increasing in c.

$x^*(k,j)$ does not depend on c. Combining,

$$\tau^+(k,\xi,c) \;=\; \min_{j:\,0<j\le k} \max\{\tau^*(k,j,\xi,c),\; x^*(k,j)\}$$

is also weakly increasing in c.

Part (ii): $\xi_{\frac{1}{2}}(k,c)$ is weakly decreasing in c.

Consider

$$\mathcal{S}_{k,c} := \left\{\xi \in \left(\tfrac{1}{2},1\right) :\; \tau^+(k,\xi,c) \le \tfrac{1}{2}\right\}.$$

By part (i), if $c_2 > c_1$ then $\tau^+(k,\xi,c_2) \ge \tau^+(k,\xi,c_1)$ for all ξ, hence $\mathcal{S}_{k,c_2} \subseteq \mathcal{S}_{k,c_1}$. Taking maxima yields

$$\xi_{\frac{1}{2}}(k,c_2) = \max \mathcal{S}_{k,c_2} \;\le\; \max \mathcal{S}_{k,c_1} = \xi_{\frac{1}{2}}(k,c_1),$$

so $\xi_{\frac{1}{2}}(k,c)$ is weakly decreasing in c. □

D Binary signals and threshold disclosures

The signal structure used in the main analysis is binary and takes threshold form. Formally, a deterministic binary signal corresponds to an event $S \subseteq [0,1]$, with

$$\sigma(x) = \mathbf{1}_{\{x \in S\}}.$$

A threshold disclosure is the special case in which $S = [\tau, 1]$. The first restriction that the signal has only two messages is without loss in the single-belief persuasion problem with two actions (see footnote 9). The second restriction that $S = [\tau, 1]$ is a substantive restriction on the planner's disclosure technology, as examples below demonstrate.

Non-threshold signal can persuade an agent at more states

Consider an agent whose prior before observing planner's signal is uniform on $[0,1]$. This is the belief $\nu = \text{Beta}[1,1]$, which arises in the baseline model when $k = 2$ and the agent samples $j = 1$ adopters. Let the adoption threshold be

$$\xi = \frac{3}{4},$$

so that the agent adopts B if and only if the posterior expectation of x is at least $3/4$.

First consider threshold signal with $S = [\tau, 1]$. In this case, after message 1 the posterior expectation is

$$\mathbb{E}[x \mid x \geq \tau] = \frac{\int_\tau^1 x\, dx}{\int_\tau^1 dx} = \frac{1+\tau}{2}.$$

Thus, message 1 persuades the agent to adopt B only if

$$\frac{1+\tau}{2} \geq \frac{3}{4},$$

or equivalently $\tau \geq 1/2$. Therefore, no threshold signal can persuade this agent to adopt B at state $x_0 = 2/5$.

Now consider the deterministic binary but non-threshold signal $\sigma(x) = \mathbf{1}_{\{x \in S\}}$ with

$$S = \left[\frac{2}{5}, \frac{41}{100}\right] \cup \left[\frac{4}{5}, 1\right].$$

After observing message 1, the agent conditions on $x \in S$, which under the uniform prior, implies

$$\nu(S) = \left(\frac{41}{100} - \frac{2}{5}\right) + \left(1 - \frac{4}{5}\right) = \frac{1}{100} + \frac{1}{5} = \frac{21}{100}.$$

Moreover,

$$\int_S x\, dx = \int_{2/5}^{41/100} x\, dx + \int_{4/5}^1 x\, dx = \frac{81}{20000} + \frac{9}{50} = \frac{3681}{20000}.$$

Therefore,

$$\mathbb{E}[x \mid x \in S] = \frac{3681/20000}{21/100} = \frac{1227}{1400} \approx 0.876 > \frac{3}{4}.$$

Thus, with this signal, the agent takes action B after message 1. Since $x_0 = 2/5 \in S$, message 1 is sent at this x_0 inducing the agent to adopt B at a state where no threshold disclosure can persuade her to do so.

Non-threshold signal can enlarge the basin of full adoption

Even though non-threshold signals can persuade agents to adopt B at lower states, one might conjecture that threshold disclosure is optimal in our dynamic environment. In particular, persuading marginal agents when current adoption is low may seem insufficient to ignite a virtuous growth cycle, because subse-

quent samples will continue to contain few adopters. This intuition, however, presumes that at every state, the growth in the adoption rate is sustained by the same message. Proposition 3 demonstrates that this does not have to be the case. A non-threshold signal can use its two messages as a relay: message 1 initiates growth by persuading relatively pessimistic agents at low states, while message 0 persuades only more optimistic agents at higher states, thus sustaining growth once adoption becomes sufficiently widespread.

Proposition 3 (A non-threshold disclosure can outperform threshold disclosure)**.** *Suppose that $k = 3$ and $\xi = 3/5$. Among threshold disclosures, the smallest initial adoption level from which the planner can induce convergence to full adoption is $2/5$. There exists a deterministic binary non-threshold disclosure that induces convergence to full adoption from every initial state*

$$x(0) > \frac{19}{48} < \frac{2}{5}.$$

Proof. First consider threshold disclosures with $S = [\tau, 1]$. An agent who observes one adopter in a sample of three has belief Beta$[1, 2]$, with density

$$f_1(x) = 2(1 - x).$$

After receiving message 1, this agent's posterior expectation is

$$\mathbb{E}[x \mid x \geq \tau, j = 1] = \frac{\int_\tau^1 2x(1-x)\,dx}{\int_\tau^1 2(1-x)\,dx} = \frac{1+2\tau}{3}.$$

Thus, message 1 persuades this agent if and only if

$$\frac{1+2\tau}{3} \geq \frac{3}{5},$$

or equivalently, $\tau \geq 2/5$.

If choosing B is a best response only for agents with at least two sampled adopters, then

$$BR(x) = G_2(x) = 1 - BinCDF(1; 3, x) = 3x^2 - 2x^3,$$

and hence, by the best response dynamic, whether x grows is determined by the sign of

$$G_2(x) - x = x(1-x)(2x-1). \tag{2}$$

This expression is negative for x below $1/2$. Thus, to generate growth at states below $1/2$, the threshold disclosure must persuade the agents with $j = 1$.

By choosing $\tau = 2/5$, the planner just persuades agents with $j = 1$, thus achieving full adoption from any $x_0 \geq 2/5$, and no threshold disclosure can do better.

Now consider the deterministic binary non-threshold disclosure

$$\sigma(x) = \begin{cases} 1, & x \in S, \\ 0, & x \notin S, \end{cases}$$

where

$$S = \left[\frac{19}{48}, \frac{1}{2} + \varepsilon\right] \cup \left[\frac{3}{5}, \frac{9}{10}\right],$$

where $\varepsilon > 0$ is arbitrarily small.[13]

For an agent with $j = 1$, the posterior expectation after message 1 is

$$\mathbb{E}[x \mid x \in S, j = 1] = \frac{\int_S 2x(1-x)\,dx}{\int_S 2(1-x)\,dx} = \frac{3303863}{5495400}.$$

Moreover,

$$\frac{3303863}{5495400} - \frac{3}{5} = \frac{6623}{5495400} > 0.$$

Thus, message 1 persuades agents with $j = 1$. By monotone likelihood ratio property, it also persuades every type with $j \geq 1$.

Next consider agents with $j = 2$. Their belief is Beta$[2, 1]$, with density

$$f_2(x) = x.$$

After message 0, these agents' posterior expectation is

$$\mathbb{E}[x \mid x \notin S, j = 2] = \frac{\int_{S^c} 2x^2\,dx}{\int_{S^c} 2x\,dx} = \frac{5861663}{9469800}.$$

Since

$$\frac{5861663}{9469800} - \frac{3}{5} = \frac{179783}{9469800} > 0,$$

message 0 persuades agents with $j = 2$ to adopt B. It therefore also persuades agents with $j \geq 2$.

[13] Small $\varepsilon > 0$ ensures that after the switch to message 0, adoption is already strictly above $1/2$.

Since when message 1 is sent, all agents with at least one sampled adopter choose B, the aggregate best response is

$$G_1(x) = 1 - BinCDF(0; 3, x) = 1 - (1 - x)^3.$$

Hence, for every $x \in (0, 1)$, best response dynamic reduces to

$$G_1(x) - x = x(1 - x)(2 - x) > 0$$

and x grows.

Since when message 0 is sent, all agents with at least two sampled adopters choose B, by (2) x grows whenever $x > 1/2$.

Message 1 is sent throughout interval

$$\left(\frac{19}{48}, \frac{1}{2} + \varepsilon\right],$$

and so any trajectory beginning at a state above $19/48$ passes $x = 1/2$. Once the adoption rate is above $1/2$, either message produces strictly positive growth in x. Therefore,

$$x(0) > \frac{19}{48} \quad \Longrightarrow \quad \lim_{t \to \infty} x(t) = 1.$$

□

Recovering threshold optimality through monotone disclosure. Threshold disclosure is optimal if admissible disclosures are restricted to be monotone in the state — that is, satisfy

$$x' < x'' \quad \Longrightarrow \quad \sigma(x') \leq \sigma(x'').$$

Every deterministic binary signal satisfying this monotonicity condition is a threshold disclosure:

$$\sigma(x) = \mathbf{1}_{\{x \geq \tau\}}$$

for some $\tau \in [0, 1]$. Thus, the results in this paper pertain to optimal monotone disclosure.

A parameter region in which threshold disclosure is optimal. Threshold disclosure is also optimal when strategic resistance to B is sufficiently strong.

Proposition 4 (Threshold optimality under strong strategic resistance)**.** *Suppose that*

$$\xi > \frac{k-1}{k}.$$

If a deterministic binary disclosure induces convergence to full adoption from every initial state above some level a, then a threshold disclosure with $S = [a, 1]$ also induces convergence from every initial state a or above.

Proof. Because for $j = k$, inequality (Gr) is never satisfied, to ignite growth in x some agents with heterogeneous samples must be persuaded to adopt B.

Before observing the planner's signal, the prior expectation of an agent with $j = k - 1$ is

$$\frac{k-1}{k} < \xi.$$

Since the posteriors sandwich the prior, both planner's messages cannot induce a posterior expectation weakly above ξ — that is, only one message can persuade an agent with $j = k - 1$ to adopt B.

If a message persuades agents with some heterogeneous sample, monotone likelihood ratio property implies that it also persuades agents with $j = k - 1$. Hence, at most one of the two messages can persuade any heterogeneous-sample agent.

Consequently, if a disclosure induces convergence from every initial state $x(0) > a$, the persuasive message must be sent at every state in $(a, 1)$ and its event S contains $(a, 1)$. Removing any states below a from S weakly raises the posterior expectation after the persuasive message for every sample composition. The persuasive event can therefore be replaced by $[a, 1]$ to result in a threshold disclosure with the same or a weakly larger basin of attraction. □

Since $(k-1)/k = 1/2$ when $k = 2$, Proposition 4 implies that threshold disclosure is optimal among deterministic binary disclosures for every $\xi > 1/2$ when agents sample two peers.

ONLINE APPENDICES

E Best response threshold under assortativity

Let $\phi \in [0,1)$ denote a level of assortativity in interaction. This can be interpreted in a number of ways. For example, as self-generated payoffs, assortative matching, homo moralis preferences, coalitional updating, or rational anticipation based on symmetry.[14] In any case, players act to maximize a convex combination of the payoff from playing against their own strategy and the payoff from playing against the strategy of the population as a whole. Playing A, this convex combination is

$$\pi_\phi(A,x) := \phi(1+\beta) + (1-\phi)\Big((1-x)(1+\beta) + x\cdot 0\Big), \tag{3}$$

and playing B, it is

$$\pi_\phi(B,x) := \phi\cdot 1 + (1-\phi)\Big((1-x)(\beta-\alpha) + x\cdot 1\Big), \tag{4}$$

so that an updating player would like to choose B if and only if

$$x \geq \frac{(1-\phi)(1+\alpha)+\phi\beta}{(1-\phi)(2+\alpha)} =: \xi_\phi. \tag{5}$$

If $\phi = 0$, there is no assortativity and ξ_ϕ reduces to $\xi = {}^{(1+\alpha)}/_{(2+\alpha)} \in ({}^1/_2, 1)$ from the main body of the paper. For general ϕ, further conditions ensure the threshold is $> {}^1/_2$ (i.e. A has an advantage) and < 1 (i.e. persuasion to B is possible).

$\xi_\phi > {}^1/_2$ requires

$$(1-\phi)\alpha + 2\phi\beta > 0, \tag{6}$$

which holds unless we have a combination of negative β and large enough

[14] Perhaps, with probability ϕ, a player plays against himself (*self-generated payoffs*), or is matched to someone who plays the same strategy (*assortative matching*). Perhaps players update in a coordinated manner as part of coalitions comprising a share ϕ of their interaction partners. Perhaps players' preferences differ from payoffs, with a weight of ϕ on a Kantian component and a weight of $1-\phi$ on a homo economicus component (*homo moralis preferences*). Perhaps players conjecture that other players, being in a similar situation, will adjust their strategies in a similar manner to themselves, with a share ϕ expected to adjust before payoffs are generated (*rational anticipation based on symmetry*).

ϕ. In the conventions literature, this is the condition for asymmetry towards A when players follow payoff-difference based behavioral rules with homo moralis preferences (Newton, 2021, Appendix G).

$\xi_\phi < 1$ requires

$$\phi(1+\beta) < 1. \tag{7}$$

which holds unless ϕ and β are large enough. If $\phi > 0$ and $\beta \geq 0$, then we can interpret the condition as stating that a set of players with a share of interactions ϕ between one another do not benefit from switching to A from a status quo at which the entire population is playing B. This is identical to the condition for a set of players with *cohesion* (Morris, 2000) of ϕ not being *agency autonomous* (Newton and Sercombe, 2020). From a status quo at which everyone plays B, not enough additional payoff β would be generated within this set by a joint switch to A to offset the downside of miscoordinating with the remainder of the population.

F General priors, i.e. beyond $Beta[0,0]$

In our baseline model, after j draws of B from a total of k draws, beliefs followed a $Beta[j, k-j]$ distribution. Implicitly, this worked from an improper $Beta[0,0]$ prior. What about alternative priors? Let us consider an arbitrary prior density $\rho(x)$ that is continuous on $[0,1]$ and strictly positive on $(0,1)$. Following sampling and a likelihood $x^j(1-x)^{k-j}$, we then have an intermediate posterior density of

$$\nu(x) = \frac{\rho(x)x^j(1-x)^{k-j}}{\int_0^1 \rho(x)x^j(1-x)^{k-j}dx},$$

and following the signal $\bar{\sigma} = 1$,

$$\nu(x \mid 1) = \frac{\rho(x)x^j(1-x)^{k-j}\mathbf{1}_{\{x\geq\tau\}}}{\int_\tau^1 \rho(x)x^j(1-x)^{k-j}dx}, \qquad \mathbb{E}_{\nu|1} := \int_0^1 x\,\nu(x \mid 1)\,dx.$$

Given $\mathbb{E}_{\nu|1}$, thresholds can be calculated as in the main text, the exact expressions depending on the prior $\rho(x)$. Furthermore, the comparative statics of under/overinference when persuading a majority extend to general priors in

the following way. Consider $\tau = 1/2$ and $j/k = 1/2$. We have

$$E_{\nu|1} = \frac{\int_{\frac{1}{2}}^{1} x\rho(x)x^{c\frac{k}{2}}(1-x)^{c\frac{k}{2}}dx}{\int_{\frac{1}{2}}^{1} \rho(x)x^{c\frac{k}{2}}(1-x)^{c\frac{k}{2}}dx}.$$

We show that this decreases in c and therefore underinference makes persuasion easier, while overinference makes it harder.

Proposition 5. *If $\tau = j/k = 1/2$, then $E_{\nu|1}$ is strictly decreasing in c.*

Proof. Let

$$D(c) = \int_{\frac{1}{2}}^{1} \rho(x)\, x^{\frac{k}{2}c}(1-x)^{\frac{k}{2}c}\, dx, \quad N(c) = \int_{\frac{1}{2}}^{1} x\,\rho(x)\, x^{\frac{k}{2}c}(1-x)^{\frac{k}{2}c}\, dx.$$

Therefore,

$$\log E_{\nu|1} \;=\; \log N(c) \;-\; \log D(c).$$

By the dominated convergence theorem, we may differentiate under the integral,

$$N'(c) = \tfrac{k}{2}\int_{\frac{1}{2}}^{1} x\,\rho(x)\, x^{\frac{k}{2}c}(1-x)^{\frac{k}{2}c}\log\big(x(1-x)\big)\, dx,$$

$$D'(c) = \frac{k}{2}\int_{\frac{1}{2}}^{1} \rho(x)\, x^{\frac{k}{2}c}(1-x)^{\frac{k}{2}c}\log\big(x(1-x)\big)\, dx.$$

Therefore,

$$\frac{d}{dc}\log E_{\nu|1} = \frac{N'(c)}{N(c)} - \frac{D'(c)}{D(c)} = \frac{k}{2}\Big(\int_{\frac{1}{2}}^{1} \log(x(1-x))\,\widetilde{P}_c(dx) - \int_{\frac{1}{2}}^{1} \log(x(1-x))\,P_c(dx)\Big),$$

where

$$P_c(dx) = \frac{\rho(x)\, x^{\frac{k}{2}c}(1-x)^{\frac{k}{2}c}}{D(c)}\, dx, \quad \widetilde{P}_c(dx) = \frac{x\,\rho(x)\, x^{\frac{k}{2}c}(1-x)^{\frac{k}{2}c}}{N(c)}\, dx.$$

Note that $\widetilde{P}_c$ is the x–tilt of P_c, hence stochastically larger, while $\log(x(1-x))$ is strictly decreasing on $[\frac{1}{2}, 1]$. It follows that

$$\int_{\frac{1}{2}}^{1} \log(x(1-x))\,\widetilde{P}_c(dx) < \int_{\frac{1}{2}}^{1} \log(x(1-x))\,P_c(dx),$$

so $\frac{d}{dc}\log E_{\nu|1} < 0$ and therefore $\frac{d}{dc}E_{\nu|1} < 0$. □

Remark 12 (Odd k)**.** A proviso is that the proof of Proposition 5 relies on $j/k = 1/2$. If we consider odd k, then the pivotal sample is $j = (k+1)/2 \neq k/2$ and the proof no longer applies. The likelihood $x^j(1-x)^{k-j}$ increases on $[\frac{1}{2}, \frac{j}{k}]$ and decreases on $[\frac{j}{k}, 1]$, so for a prior $\rho(x)$ concentrated around $1/2$, it is possible that sampling $j = (k+1)/2$ actually gives a conditional mean $\mathbb{E}_{\nu|1}$ that is higher than the conditional mean $\mathbb{E}_{\rho|1}$ that would have been arrived at with no sampling at all. Thus, increasing c and strengthening the sampling effect may actually increase $\mathbb{E}_{\nu|1}$ and make persuasion easier. ◁

Example 1. Consider $k = 3$, so the pivotal sample is $j = (k+1)/2 = 2$. Let $\rho(\cdot)$ follow a Beta$[a, a]$ distribution. For low a and $c = 1$, sampling $j = 2$ decreases the conditional posterior expectation compared to the hypothetical case with no sampling. That is, $\mathbb{E}_{\nu^{c=1}|1} < \mathbb{E}_{\rho|1}$. In contrast, for high a, sampling $j = 2$ gives $\mathbb{E}_{\nu^{c=1}|1} > \mathbb{E}_{\rho|1}$. The critical value of a is calculated numerically to be ≈ 4.745. ◁

However, despite the proviso of Remark 12, for any given prior, we can choose large enough odd k such that, for the pivotal sample $j = (k+1)/2$ and $c = 1$, we have $\mathbb{E}_{\nu^{c=1}|1} < \mathbb{E}_{\rho|1}$. This is possible because $\mathbb{E}_{\rho|1} > 1/2$, and $\mathbb{E}_{\nu^{c=1}|1} \to 1/2$ as $k \to \infty$. Weak enough inference from the pivotal sample then gives a conditional expectation close to $\mathbb{E}_{\rho|1}$, higher than the conditional expectation under ordinary Bayesian inference ($c = 1$). Thus, for large odd sample sizes, sufficient underinference still facilitates persuasion relative to the Bayesian benchmark. The qualitative implications of our baseline model continue to hold when k is even, and also when k is odd but not too small.

Remark 13 (Mixing beta distributions)**.** An interesting observation is that, compared to $\mathbb{E}_{\rho|1}$, the conditional mean with sampling $\mathbb{E}_{\nu|1}$ is not necessarily closer to $\frac{j}{k}$. Now, for a beta distribution following binomial updating, the posterior mean is a convex combination of the prior mean and the mean of the observations. Of course, a truncated beta distribution is not a beta distribution, but it, indeed any continuous prior, can be approximated by a convex combination of beta distributions (Diaconis and Ylvisaker, 1983, Theorem 1) .

$$\sum_i w_i \, x^{a_i - 1}(1-x)^{b_i - 1} \tag{8}$$

for $a_i, b_i \geq 0$ and weights $w_i > 0$. Therefore, after observing j of k successes from binomial draws, the posterior mean for every beta distribution in the mixture will move towards j/k. However, the overall mean need not. This is because the weights given to each distribution will change as a result of

renormalization following updating. An example of this is the beta prior of the previous subsection with $a = 4$. ◁

G Non-Bayesian inference from sampling

Inference here is non-Bayesian in the sense that it cannot be arrived at by a Bayesian update of a common prior. We show briefly how our analysis extends to such a setup. Note that belief updating after observing a signal is still a conditional expectation. That is, it is still Bayesian.

G.1 Logit inference

Assume that sampling is the same as in Section 2.5. However, agents form beliefs using **logit inference**,

$$\nu(x) = Z\, e^{-\frac{1}{\eta}|x-\frac{j}{k}|}$$

where $Z > 0$ is a normalizing constant so that $\int_0^1 \nu(x)dx = 1$ Parameter η measures dispersion of beliefs. As $\eta \to \infty$, beliefs become uniform. As $\eta \to 0$ beliefs put point mass on the maximum likelihood estimate j/k.

We consider conditions under which a majority of agents playing B suffices for persuasion to $x = 1$ to be possible. Assume that k is even (the odd case can be solved similarly). At $x = 1/2$, the planner has to persuade more than half the population to play B when he sets $\tau = 1/2$. This occurs if and only if agents with $j = k/2$ play B after observing the signal $\bar{\sigma} = 1$. Observing that

$$\frac{1}{Z}\int_{\frac{1}{2}}^1 x\nu(x)dx = \int_{\frac{1}{2}}^1 x\, e^{-\frac{1}{\eta}(x-\frac{1}{2})}dx = \left(\frac{\eta}{2}+\eta^2\right) - \left(\eta\, e^{-\frac{1}{2\eta}} + \eta^2\, e^{-\frac{1}{2\eta}}\right),$$

$$\text{and} \qquad \frac{1}{Z}\int_{\frac{1}{2}}^1 \nu(x)dx = \int_{\frac{1}{2}}^1 e^{-\frac{1}{\eta}(x-\frac{1}{2})}dx = \eta\,(1-e^{-\frac{1}{2\eta}}),$$

$$\text{we obtain} \qquad \mathbb{E}_{\nu|1} = 1+\eta - \frac{1}{2\left(1-e^{-\frac{1}{2\eta}}\right)} \geq \xi$$

as the condition for any majority to be persuadable. The conditional expectation $\mathbb{E}_{\nu|1}$ increases in η. That is, more dispersed inference by agents makes it easier

for the planner to persuade them towards B.

G.2 Uniform and triangular distributions on an interval

G.2.1 Uniform

Consider symmetric beliefs around the natural point estimate j/k. For $0 < \eta < 1/2$, set $\bar{\eta} = \min\{\eta, \frac{j}{k}, 1 - \frac{j}{k}\}$. After sampling j of k players playing B, let beliefs ν be uniform on the support $[\frac{j}{k} - \bar{\eta}, \frac{j}{k} + \bar{\eta}]$. Observe that beliefs are less concentrated for higher values of η.

Assume that k is even. At $x = 1/2$, the planner has to persuade more than half the population to play B when he sets $\tau = 1/2$. This occurs if and only if agents with $j = k/2$ play B after observing the signal $\bar{\sigma} = 1$. As ν is uniform on $[\frac{j}{k} - \bar{\eta}, \frac{j}{k} + \bar{\eta}] = [\frac{1}{2} - \bar{\eta}, \frac{1}{2} + \bar{\eta}] = [\frac{1}{2} - \eta, \frac{1}{2} + \eta]$, we have the condition

$$\mathbb{E}_{\nu|1} = \frac{1}{2}(1 + \eta) \geq \xi.$$

That is, more dispersed inference by agents makes it easier for the planner to persuade them towards B.

G.2.2 Triangular

Consider beliefs $\nu(x) = \frac{1}{\bar{\eta}^2}(\bar{\eta} - |x - \frac{j}{k}|)$ on $[\frac{j}{k} - \bar{\eta}, \frac{j}{k} + \bar{\eta}]$ for $j = 1, 2, \cdots, k - 1$. Let ν be point distributions on $x = 0$, $x = 1$ for $j = 0, k$ respectively.

Again, assume that k is even and consider $\tau = 1/2$, $j = k/2$. We obtain

$$\int_{\frac{1}{2}}^{1} x\nu(x)dx = \frac{\eta}{6} + \frac{1}{4}, \qquad \int_{\frac{1}{2}}^{1} \nu(x)dx = \frac{1}{2},$$

$$\text{thus} \qquad \mathbb{E}_{\nu|1} = \frac{\eta}{3} + \frac{1}{2} \geq \xi$$

as the condition for majority persuasion. Again, more dispersed inference by agents makes it easier for the planner to persuade them towards B.

References

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of*

Behavioral Economics: Applications and Foundations 2, volume 2, pages 69–186. North-Holland.

Bergin, J. and Lipman, B. L. (1996). Evolution with state-dependent mutations. *Econometrica*, 64(4):943–56.

Binmore, K. (2005). *Natural justice*. Oxford University Press, Oxford, UK.

Blume, L. E. (2003). How noise matters. *Games and Economic Behavior*, 44(2):251–271.

Carvalho, J.-P. (2013). Veiling. *Quarterly Journal of Economics*, 128(1):337–370.

Dahl, G. B., Løken, K. V., and Mogstad, M. (2014). Peer effects in program participation. *American economic review*, 104(7):2049–2074.

Diaconis, P. and Ylvisaker, D. (1983). Quantifying prior opinion. Technical Report EFS NSF 207, Stanford University, Department of Statistics.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics*, 95(3):537–557.

Gulesci, S., Jindani, S., La Ferrara, E., Smerdon, D., Sulaiman, M., and Young, P. (2025). A stepping stone approach to norm transitions. *American Economic Review*, 115(7):2237–2266.

Heiman, S. L., Claessens, S., Ayers, J. D., Guevara Beltrán, D., Van Horn, A., Hirt, E. R., Aktipis, A., and Todd, P. M. (2023). Descriptive norms caused increases in mask wearing during the covid-19 pandemic. *Scientific reports*, 13(1):11856.

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

Kolotilin, A. (2015). Experimental design to persuade. *Games and Economic Behavior*, 90:215–226.

Mensch, J. (2021). Monotone persuasion. *Games and Economic Behavior*, 130:521–542.

Monderer, D. and Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, 14(1):124 – 143.

Morris, S. (2000). Contagion. *Review of Economic Studies*, 67(1):57–78.

Murakami, M. (2021). The role of an expert in the evolution of conventions. Mimeo.

Naidu, S., Hwang, S.-H., and Bowles, S. (2017). The evolution of egalitarian sociolinguistic conventions. *American Economic Review: Papers & Proceedings*, 107(5):572–77.

Nax, H. H. and Newton, J. (2022). Deep and shallow thinking in the long run. *Theoretical Economics*, 17(4):1501–1527.

Newton, J. (2018). Evolutionary game theory: A renaissance. *Games*, 9(2):31.

Newton, J. (2021). Conventions under heterogeneous behavioural rules. *Review of Economic Studies*, (https://doi.org/10.1093/restud/rdaa063).

Newton, J. and Sercombe, D. (2020). Agency, potential and contagion. *Games and Economic Behavior*, 119:79–97.

Oyama, D., Sandholm, W. H., and Tercieux, O. (2015). Sampling best response dynamics and deterministic equilibrium selection. *Theoretical Economics*, 10(1):243–281.

Peski, M. (2010). Generalized risk-dominance and asymmetric dynamics. *Journal of Economic Theory*, 145(1):216 – 248.

Salant, Y. and Cherry, J. (2020). Statistical inference in games. *Econometrica*, 88(4):1725–1752.

Sandholm, W. H. (2005). Negative externalities and evolutionary implementation. *Review of Economic Studies*, 72(3):885–915.

Sandholm, W. H. (2007). Pigouvian pricing and stochastic evolutionary implementation. *Journal of Economic Theory*, 132(1):367–382.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. Economic learning and social evolution. Cambridge, Mass. MIT Press.

Sawa, R. and Wu, J. (2023). Statistical inference in evolutionary dynamics. *Games and Economic Behavior*, 137:294–316.

Weibull, J. (1995). *Evolutionary game theory*. MIT Press.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, 61(1):57–84.

Young, H. P. (1998). *Individual strategy and social structure*. Princeton University Press, Princeton, NJ, USA.

Young, H. P. and Burke, M. A. (2001). Competition and custom in economic contracts: A case study of illinois agriculture. *American Economic Review*, 91(3):559–573.