A Geometric Model of Opinion Polarization

Jan Hązła\(^\ast\) Yan Jin\(^\dagger\) Elchanan Mossel\(^\ddagger\) Govind Ramnarayan\(^\S\)

Abstract

We introduce a simple, geometric model of opinion polarization. It is a model of political persuasion, as well as marketing and advertising, utilizing social values. It focuses on the interplay between different topics and persuasion efforts. We demonstrate that societal opinion polarization often arises as an unintended byproduct of influencers attempting to promote a product or idea. We discuss a number of mechanisms for the emergence of polarization involving one or more influencers, sending messages strategically, heuristically, or randomly. We also examine some computational aspects of choosing the most effective means of influencing agents, and the effects of those strategic considerations on polarization.

1 Introduction

Opinion polarization is a widely acknowledged social phenomenon, especially in the context of political opinions [FA08, SH15, IW15], leading to recent concerns over “echo chambers” created by mass media [Pri13] and social networks [CRF\(^+\)11, Par11, BMA15, BAB\(^+\)18, Gar18]. The objective of this paper is to propose a simple, multi-dimensional geometric model of the dynamics of polarization where the evolution of correlations between opinions on different topics plays a key role.

Many models have been proposed to explain how polarization arises, and this remains an active area of research [NSL90, Axe97, Noa98, HK02, MKFB03, BB07, DGL13, DVSC\(^+\)17, KP18, PY18, SCP\(^+\)19]. Our attempt aims at simplicity over complexity. As opposed to a large majority of previous works, our model does not require social network-based mechanism. Instead, we focus on influences of advertising or political campaigns that reach a wide segment of the population.

We develop a high-dimensional variant of biased assimilation [LRL79] and use it as our main behavioral assumption. The bias assimilation for one topic states that people tend to be receptive to opinions they agree with, and antagonistic to opinions they disagree with.

The multi-dimensional setting reflects the fact that campaigns often touch on many topics. For example, in the context of American politics, one might wonder why there exists a significant correlation between opinions of individuals on, say, abortion access, gun rights and urgency of climate change [Pew14]. Our model attempts to illustrate how such correlations between opinions can arise as a (possibly unintended) effect of advertising exploiting different topics and social values.

In mathematical terms, we consider a population of agents with preexisting opinions represented by vectors in \(\mathbb{R}^d\). Each coordinate represents a distinct topic, and the value of the coordinate

\(^\ast\)Email: jan.hazla@epfl.ch. Partially supported by DMS-1737944.
\(^\dagger\)Email: yjin1@mit.edu. Partially supported by ARO MURI W911NF1910217.
\(^\ddagger\)Email: elmos@mit.edu. Partially supported by by Simons Investigator in Mathematics award (622132), NSF DMS award 1737944 and CCF award 1918421, ARO MURI grant W911NF1910217 and by a Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826.
\(^\S\)Email: govind@mit.edu. Partially supported by DMS-1737944 and ARO MURI W911NF1910217.
reflects the agent’s opinion on the topic, which can be positive or negative. As discussed more fully in Section 1.4, we assume that all opinions lie on the Euclidean unit sphere. This reflects an assumption that each agent has the same “budget of importance” of different topics. We then consider a sequence of interventions affecting the opinions. An intervention is also a unit vector in \( \mathbb{R}^d \), representing the set of opinions expressed in, e.g., an advertising campaign or “news cycle”.

We model the effect of intervention \( v \) on an agent’s opinion \( u \) in the following way. Supposing an agent starts with opinion \( u \in \mathbb{R}^d \), after receiving an intervention \( v \) it will update the opinion to the unit vector proportional to

\[
    w = u + \eta \cdot \langle u, v \rangle \cdot v ,
\]

where \( \eta > 0 \) is a global parameter that controls the influence of an intervention. Most of our results do not depend on a choice of \( \eta \) and in our examples we often take \( \eta = 1 \) for the sake of simplicity. Smaller values of \( \eta \) could model campaigns with limited persuasive power. This and other design choices are discussed more extensively in Section 1.4.

Intuitively, the agent evaluates the received message in context of its existing opinion, and assimilates this message weighted by its “agreement” with it. Our model exhibits biased assimilation in that if the intervening opinion \( v \) is positively correlated with an agent’s opinion \( u \), then after the update the agent opinion moves towards \( v \), and conversely, if \( v \) is negatively correlated with \( u \), then the update moves \( u \) away from \( v \) and towards the opposite opinion \( -v \).

One way to think of the intervention is as an exposure to persuasion by a political actor, like a political campaign message. A different way, in the context of marketing, is a product advertisement that exploits values besides the quality of the product. In that context, we can think of one of the \( d \) coordinates of the opinion vector as representing opinion on a product being introduced into the market and the remaining coordinates as representing preexisting opinions on other (e.g., social or political) issues. Then, an intervention would be an advertising effort to connect the product with a certain set of opinions or values [VSL77]. Some examples are corporate advertising campaigns supporting LGBT rights [Sny15] or gun manufacturers associating their products with patriotism and conservative values [SVS04]. Another scenario of an intervention is a company (e.g., a bank or an airline [For18]) announcing its refusal to do business with the gun advocacy group NRA. Such advertising strategies can have a double effect of convincing potential customers who share relevant values and antagonizing those who do not.

Our main results show that such interventions, even if intending mainly to increase sales and without direct intention to polarize, can have a side effect of increasing the extent of polarization in the society. For example, it might be that, in a population with initial opinions distributed uniformly, a number of interventions introduces some weak correlations. In our model, these correlations can be profitably exploited by advertisers in subsequent interventions. As a side effect, the interventions strengthen the correlations and increase polarization.

For example, suppose that after various advertising campaigns, we observe that people who tend to like item A (say, electric cars) tend to be liberal, and people who like a seemingly unrelated item B (say, firearms) tend to be conservative. This may result from the advertisers exploiting some obvious connections, e.g., between electric cars and responding to climate change, and between firearms and respect for the military. Subsequently, future advertising efforts for electric cars may feature other values associated with liberals in America to appeal to potential consumers: an advertisement might show a gay couple driving to their wedding in an electric car. Similarly, future advertisements for firearms may appeal to conservative values for similar reasons. The end result can be that the whole society becomes more polarized by the incorporation of political topics into advertisements.
Throughout the paper, we analyze properties of our model in a couple of scenarios. With respect to the interventions, we consider two scenarios: either there is one entity (an influencer) trying to persuade agents to adopt their opinion or there are two competing influencers pushing different agendas. With respect to the time scale of interventions, we also consider two cases: the influencer(s) can apply arbitrarily many interventions, i.e., the asymptotic setting, or they need to maximize influence with a limited number of interventions, i.e., the short-term setting. The questions asked are: (i) What sequence of interventions should be applied to achieve the influencer’s objective? (ii) What are the computational resources needed to compute this optimal sequence? (iii) What are the effects of applying the interventions on the population’s opinion structure? We give partial answers to those questions. The gist of them is that in most cases, applying desired interventions increases the polarization of agents.

1.1 Model definition

The formal definition of our model is simple. We consider a group of \( n \) agents, whose opinions are represented by \( d \)-dimensional unit vectors, where each coordinate corresponds to a topic. We will look into how those opinions change after receiving a sequence of interventions. Each intervention is also a unit vector in \( \mathbb{R}^d \), representing the opinion contained in a message that the influencer (e.g., an advertiser) broadcast to the agents. Our model features one parameter: \( \eta > 0 \), signifying how strongly an intervention influences the opinions.

The interventions \( v^{(1)}, \ldots, v^{(t)}, \ldots \) divide the process into discrete time steps. Initially, the agents start with opinions \( u^{(1)}_i, \ldots, u^{(1)}_n \). Subsequently, applying intervention \( v^{(t)} \) updates the opinion of agent \( i \) from \( u^{(t)}_i \) to \( u^{(t+1)}_i \).

After each intervention, the agents update their opinions by moving towards or away from the intervention vector, depending on whether or not they agree with it (which is determined by the inner product between the intervention vector \( v^{(t)} \) and the opinion vector), and normalizing suitably. The update rule is given by

\[
 u^{(t+1)}_i = \frac{w^{(t+1)}_i}{\|w^{(t+1)}_i\|}, \quad \text{where} \quad w^{(t+1)}_i = u^{(t)}_i + \eta \langle u^{(t)}_i, v^{(t)} \rangle \cdot v^{(t)}. \tag{2}
\]

We note that, by expanding out the definition of \( w^{(t+1)}_i \),

\[
 \|w^{(t+1)}_i\|^2 = \langle w^{(t+1)}_i, w^{(t+1)}_i \rangle = 1 + (2\eta + \eta^2)(u^{(t)}_i, v^{(t)})^2 \tag{3}
\]

In particular, this implies that \( \|w^{(t+1)}_i\| \geq 1 \), and consequently that \( u^{(t+1)}_i \) is well-defined. The norm in (2) and everywhere else throughout is the standard Euclidean norm. Note that applying \( v^{(t)} \) or \(-v^{(t)}\) to an opinion \( u^{(t)}_i \) results in the same updated opinion \( u^{(t+1)}_i \).

1.2 Example

To illustrate our model, let us consider an empirical example with \( \eta = 1 \). Suppose an advertiser is marketing a new product. The opinion of the population has four dimensions. The population consists of 500 agents, each with initial opinions \( u^{(1)}_i = (u^{(1)}_{i,1}, u^{(1)}_{i,2}, u^{(1)}_{i,3}, 0) \in \mathbb{R}^4 \) subject to \( u^{2}_{i,1} + u^{2}_{i,2} + u^{2}_{i,3} = 1 \). The opinion on the new product is represented by the fourth coordinate, which is initially set to zero for all agents. These starting opinions are sampled independently at random from the uniform distribution on the sphere. A typical arrangement of initial opinions is shown under \( t = 1 \) in Figure 1.
Suppose the advertiser chooses to repeatedly apply an intervention that couples the product with the preexisting opinion on the first coordinate. More concretely, let the intervention vector be

\[ v = (\beta, 0, 0, \alpha) \]

where \( \alpha = \frac{3}{4}, \beta = \sqrt{1 - \alpha^2} \).

In that case, an application of the intervention \( v \) to an opinion \( u_1^{(1)} = (u_{i,1}, u_{i,2}, u_{i,3}, 0) \) results in

\[ \langle u_1, v \rangle = \beta u_{i,1} \]

and

\[ u_2^{(2)} = \frac{w_2^{(2)}}{\|w_2^{(2)}\|}, \quad w_2^{(2)} = ((1 + \beta^2)u_{i,1}, u_{i,2}, u_{i,3}, \beta \alpha u_{i,1}), \quad \|w_2^{(2)}\|^2 = 1 + 3 \beta^2 u_{i,1}^2. \]

Note that after applying the intervention the first and last coordinates have the same sign. In subsequent time step, the intervention \( v \) is applied again to the updated opinions \( u_2^{(2)} \) and so on.

The evolution of opinions over five consecutive applications of \( v \) in this process is illustrated in Figure 1. The interventions increase the affinity for the product for some agents while antagonizing others. Furthermore, they have a side effect of polarizing the agents’ opinions also on the first three coordinates. A similar example is included in Appendix B.

1.3 Outline of our results

We analyze the strategy of influencers in several settings.

In an “asymptotic scenario”, the influencer wants to apply an infinite sequence of interventions \( v^{(1)}, v^{(2)}, \ldots \), that maximizes how many out of the \( n \) agent opinions converge to the target vector \( v \). As is standard, we say that a sequence of vectors \( u^{(1)}, \ldots, u^{(t)}, \ldots \) converges to a vector \( v \) if \( \lim_{t \to \infty} \|u^{(t)} - v\| = 0 \). One way to interpret this scenario is that a campaigner wants to establish a solid base of support for their party platform.

In a “multiple-influencer scenario”, two influencers (such as two companies or two parties) who have different objectives apply their two respective interventions on the population in a certain order. We ask how the opinions change under such competing influences. This scenario can be interpreted as two parties campaigning their agendas to the population.

In a “short-term scenario”, the influencer is advancing a product/subject which is expressed in the last coordinate of opinion vectors \( u_{i,d} \). The influencer assumes some fixed threshold \( 0 < T < 1 \) and an upper bound \( K \) on the number of interventions, and asks, given \( n \) opinions \( u_1, \ldots, u_n \), how to choose \( v^{(1)}, \ldots, v^{(K)} \) in order to maximize the number of time-\((K+1)\) opinions \( u_{1}^{(K+1)}, \ldots, u_{n}^{(K+1)} \) with \( u_{i,d}^{(K+1)} > T \). One interpretation is that advertisers only have a limited number of opportunities to publicize their products to consumers, and consumers with \( u_{i}^{(K+1)} > T \) will decide to buy the product after the interventions \( v^{(1)}, \ldots, v^{(K)} \) are applied.

We briefly summarize our results for these scenarios. In Section 3 we start by showing that random interventions lead to a strong form of polarization. More precisely, assuming uniformly distributed initial opinions, we prove that applying an independent uniformly random intervention at each time step leads the opinions to form two equally-sized clusters converging to a pair of (moving) antipodal points.

In Section 4 we consider the asymptotic scenario, where there is one influencer with a desired campaign agenda \( v \) and unlimited numbers of interventions at its disposal. We ask which sequence of interventions maximizes the number of opinions that converge to the agenda \( v \). Somewhat surprisingly, we show that such optimal strategy does not necessarily promote the campaign agenda
Figure 1: Graphical illustration of the example discussed in Section 1.2. Since we are working in $d = 4$, we illustrate the first three dimensions as spatial positions and the fourth dimension with a color scale. Initially the opinions are uniformly distributed on the sphere, with the fourth dimension equal to 0 (no opinion) everywhere. Consecutive applications of the intervention $v = (\sqrt{7}/4, 0, 0, 3/4)$ in $\mathbb{R}^4$ result in polarization both in spatial dimensions and in the color scale.
directly at every step. Instead, it finds a hemisphere containing the largest number of initial opinions, concentrates the opinions in this hemisphere around an arbitrary point, and only in the last stage nudges them gradually towards the target agenda. We then show that it is computationally hard to approximate this densest hemisphere (and therefore the optimal strategy) to any constant factor. Again, strong polarization emerges from our dynamic: there exists a pair of antipodal points such that all opinions converge to one of them.

In Section 5 we study the short-term scenario where one influencer is allowed only one intervention. In Section 5.1, we describe a case study with one influencer and two agents in the population. We assume that the influencer wants to increase the correlations of agent opinions with the target opinion $v$ above a given threshold $T > 0$. We show consequences of optimal interventions depending on if the influencer can achieve this objective for one or both agents. In Section 5.2, we consider a similar scenario, but with a large number of agents. In that case, it surprisingly turns out that the problem of finding optimal intervention in this short-term setting is related to the problem analyzed in the asymptotic setting. Finding the optimal intervention is equivalent to finding a spherical cap containing the largest number of initial opinions.

In Section 6, we study two competing influencers. At each time step, one of the influencers is selected at random to apply its intervention. One might hope that having multiple advertisers can make the resulting opinions more spread-out, but we prove that this not the case. We show that, as time goes to infinity, all opinions converge to the convex cone between the two intervention vectors. Furthermore, we show that the if the correlation between the interventions is high enough, the strong form of polarization emerges: the opinions of the population concentrate around two antipodes moving around in the convex cones of the two interventions.

1.4 Design choices

Our goal in this work is to provide a simple, elegant and analyzable model demonstrating how correlations between different topics and natural interventions lead to polarization. That being the case, there are many societal mechanisms related to polarization that we do not discuss here.

First, in contrast to majority of existing literature, we present a mechanism independent from opinion changes induced by interactions between individuals. Second, we do not address aspects such as replacement of the population or unequal exposure and effects of the interventions. We do not consider any external influences on the population in addition to the interventions. Our model does not align with (limited) theoretical and empirical research suggesting that in certain settings exposure to conflicting views can decrease polarization [PT06, MS10, GMGM17, GGPT17] or works that question the overall extent of polarization in the society [FAP05, BG08].

In general we assume that the influencers have full knowledge of the agent opinions. This is not a realistic assumption and in fact our results in Section 4 show that in some settings the optimal influencer strategy is infeasible to compute even with the full knowledge of opinions. On the other hand, we observe polarization also in settings where the influencers apply interventions that are agnostic to the opinions, for example with purely random interventions in Section 3 or competing influencers in Section 6.

We sometimes discuss the uniform distribution of initial opinions on $\mathbb{R}^d$. We do this as the uniform distribution may be viewed as the most diverse and establishing polarization starting from the uniform distribution hints that we are modelic a generic phenomenon. Most of our results do not make assumptions about the initial distribution.

We assume that any group of topics can be combined into an intervention with the effect given by (1). A more plausible model might feature some “internal” (content) correlations between topics in addition to “external” (social) correlations arising out of the agents’ opinion structure. For
example, topics may have innate connections, causing inherent correlations between corresponding opinions (e.g., being positive on renewable energy and recycling). Furthermore, there are certain topics (e.g., undesirability of murder) on which nearly all members of the population share the same inclination. As a matter of fact, it is common for marketing strategies to exploit unobjectionable social values (see, e.g., [VSL77]). However, we presume that under suitable circumstances (e.g., due to inherent correlations we just mentioned) the “polarizing” topics might present a more appealing alternative for a campaign. Our model concerns such a case, where the “unifying” topics might be excluded from the analysis. We note that other works have also suggested that focusing on polarizing topics may be appealing for campaigns [PY18].

Below we discuss a couple of specific design choices in more detail:

**Euclidean unit ball** We make an assumption that all opinions and interventions lie on the Euclidean unit ball. Note that the interpretation of this representation is somewhat ambiguous. The magnitude of an opinion on a given subject $u_{i,k}$ might signify the strength of the opinion, the confidence of the agent or relative importance of the subject to the agent. While these are different measures, there are psychological reasons to expect that, e.g., “issue interest” and “extremity of opinion” are correlated [LBS00, Bal07, BB07]. Especially taking the magnitudes as signifying the relative importance, we believe that the assumption that this “budget of importance” for any given agent is fixed is quite natural. That being said, we are also motivated by simplicity and tractability. Multiple ways of relaxing or modifying this assumption are possible. While we do not study these variants in this paper, we now discuss them very briefly. At least empirically, our basic findings about ubiquity of polarization seem to remain valid for those modified models.

Perhaps the simplest modification is to use the same update rule as in (2) with a different norm (e.g., $\ell_1$ or $\ell_\infty$ norm). Such variant would also assume that opinions and interventions lie on the unit sphere of the respective norm. Our experiments suggest that, qualitatively, both $\ell_1$ and $\ell_\infty$ variants behave similarly to the Euclidean norm.

In another direction, rather than having all opinions on the unit sphere, fixed, but different norms $z_i$ can be specified for different agents. Then, the update rule (2) could be modified as

$$w_i^{(t+1)} = u_i^{(t)} + \eta \cdot \left( \frac{u_i^{(t)}}{z_i}, v(t) \right) \cdot v(t),$$

with normalization preserving $\|u_i^{(t+1)}\| = z_i$. As long as the values of $z_i$ are bounded from below and above, the resulting dynamic is essentially identical and our results carry over to this more general setup.

Yet another possibility is to consider opinion unit vectors $u \in \mathbb{R}^{d+1}$ with $u_{d+1} \geq 0$ and interpret the first $d$ coordinates as opinions and the last coordinate as “unused budget”. Therefore, large values of $u_{d+1}$ signify generally uncertain opinions and small values of $u_{d+1}$ correspond to strong opinions. There are multiple possible rules for interventions, where an intervention can have $d$ or $d+1$ coordinates, and with different treatments of the last coordinate. We leave the details for another time.

**Effects of applying $v$ and $-v$** In our model, an effect of an intervention $v$ is exactly the same as for the opposite intervention $-v$. This might look like a cynical assumption about human nature, but arguably it is not entirely inaccurate. For example, experiments on social media show that not only exposure to similar ideas (the “echo chamber” effect), but also exposure to opposing opinions causes beliefs to become more polarized [BAB+18]. This is even more apparent if a broader
notion of an intervention is considered. Using a recent example, social media platforms banning or disassociating from certain statements can have a polarizing effect [BBC20]. Furthermore, in our model this effect occurs only if all the components of an opinion are negated.

A related, more general objection is that direct persuasion is not possible in our model. If an agent has an opinion \( u \) with \( (u, v) < 0 \), directly applying \( v \) only makes the situation worse. Instead, an effective influencer needs to apply interventions utilizing different subjects to gradually move \( u \) through a sequence of intermediate positions towards \( v \). Our answer is that we posit that a lot of, if not all, persuasion actually works that way: to convince that “\( x \) is good”, one argues that “\( x \) is good, since it is quite like \( y \), which we both already agree is good”.

Notions of polarization While the notion of polarization is clear when discussing one topic, it is not straightforward to interpret in higher dimensions. Let \( S \subseteq \mathbb{R}^d \) be a set of \( n \) opinions. Writing \( u = (u_1, \ldots, u_d) \) for \( u \in S \), a natural measure of polarization of \( S \) on a single topic \( i \) is

\[
\rho_i(S) = \frac{1}{|S|^2} \max_{T \subseteq S} \sum_{u \in T, u' \in S \setminus T} (u_i - u'_i)^2,
\]

and we may generalize it to higher dimensions by measuring the polarization as:

\[
\rho(S) = \frac{1}{|S|^2} \max_{T \subseteq S} \sum_{u \in T, u' \in S \setminus T} \|u - u'\|^2.
\]

It is clear from the definition that

\[
\max_i \rho_i(S) \leq \rho(S) \leq \sum_i \rho_i(S).
\]

If we consider an example set \( S_1 \) with \( n/2 \) opinions at \( u \) and \( n/2 \) opinions at \(-u\), then clearly \( \rho(S_1) = \sum_i \rho_i(S_1) \), but in any other example, the upper bound will not be tight. For example, if \( S_2 \) is the set of the \( 2^d \) vertices of a hypercube, i.e., \( S_2 = 1/\sqrt{d} \cdot \{-1,1\}^d \), then \( \rho_i(S_2) = 1/d \) for all \( i \), but \( \rho(S_2) \) converges to \( 1/2 \) as \( n \to \infty \). This corresponds to the fact that while the society is completely polarized on each topic, two random individuals will agree on about half of the topics. In Section 2 we refer to such a situation as exhibiting issue radicalization, but no issue alignment.

Ultimately, in many of our results we do not worry about these issues, since we observe a strong form of polarization, where all opinions converge to two antipodal points.

1.5 Other variants

Other than discussed above, there are many possible variants that can lead to interesting future work. These include:

- “Targeting”, where the influencer can select subgroups of the population and apply interventions groupwise.
- Models where the strength of an intervention \( \eta \) varies across agents and/or time steps.
- Perturbing preferences with noise after each step.
- Replacement of the population, e.g., introducing new agents with “fresh” opinions or removing agents that stayed in the population for a long time or who already “bought” the product, i.e., exceeded the threshold \( u_{i,d} > T \). For example, this could correspond to ”one-time” purchase product like a house or a fridge, or situations where the customer’s opinion is more difficult to change as time passes.
• Models where the initial opinions are not observable or partially observable.

• Expanding the model by adding peer effects and social network structure and exploring the resulting dynamics of polarization and opinion formation. This can be done in different ways and we expect that polarization will feature in many of them. For example, [GKT21] show polarization for random interventions in what they term the “party model”.

• Strategic competing influencers: in the studied scenarios with competing influencers, we assume that they apply fixed interventions. One can ask: suppose the influencers have their own target opinions, what is each campaigner’s optimal sequence of messages in face of the other campaigner? Then, resulting equilibrium of opinion formation could be analyzed. This can be modeled as a dynamic game where the game state is the opinion configuration and optimal strategies may be derived using sequential planning and control.

2 Related works

As mentioned, there is a multitude of modeling and empirical works studying opinion polarization in different contexts [NSL90, Axe97, BG98, Noa98, HK02, MKFB03, MS05, BB07, DGL13, DVSC+17, KP18, SCP+19, PY18, BAB+18]. Broadly speaking, previous works have proposed various possible sources for polarization, including peer interactions, bias in individuals’ perceptions, and global information outlets.

There is an extensive line of models of opinion exchange on networks with peer interactions, where individuals encounter neighboring individuals’ opinions and update their own opinions based on, e.g., pre-defined friend/hostile relations [SPJ+16], or the similarity and relative strength of opinions [MS10], etc. This branch of work often attributes polarization to homophily of one’s social network [DGL13] that is induced by the self-selective nature of social relations and segregation of like-minded people [WMKL15] and exacerbated by the echo chamber effect of social media [Par11].

A parallel proposed mechanism points to psychological biases in individuals’ opinion formation processes. One example is biased assimilation [LRL79, DGL13, BB07, BAB+18]: the tendency to reinforce one’s original opinions regardless if other encountered opinions align with them or not. For example, [BAB+18] observed that even when social media users are assigned to follow accounts that share opposing opinions, they still tend to hold their old political opinions and often to a more extreme degree. On the modeling side, [DGL13] showed that DeGroot opinion dynamics with the biased assimilation property on a homophilous network may lead to polarization.

Existing works have also proposed models where polarization occurs even when information is shared globally [Zal92, MS05]. For example, [MS05] propose a model where competition for readership between global information outlets causes news to become polarized in a single-dimensional setting. Another example is [Zal92], a classical work on the formation of mass opinion. It theorizes that each individual has political dispositions formed in their own life experience, education and previous encounters that intermediate between the message they encounter and the political statement they make. Therefore, hearing the same political message can cause different thinking processes and changes in political preferences in different individuals.

It is noteworthy that the majority of previous work focuses on polarization on a single topic dimension. Two exceptions are [BB07], which studies biased assimilation with opinions on multiple topics and [BG08] that observed non-trivial correlations between people’s attitudes on different issues. We note that [BB07] uses a different updating rule to observe dynamics that differ from our work: in their simulations, polarization on one issue typically does not result in polarization on others. There is also a class of models [Axe97, Noa98, MKFB03] that concern multi-dimensional opin-
ions where an opinion on a given topic takes one of finitely many values (e.g., + or −). These models do not seem to have a geometric structure of opinion space similar to ours and usually focus on formation of discrete groups in the society rather than total polarization. Another model [PPTF17] uses a geometric (affine) rule of updating multi-dimensional opinions. Unlike us, they seem to be modeling pre-existing, “intrinsic” correlations between topics rather than the emergence of new ones and they are concerned mostly with convergence and stability of their dynamics.

A related paper [PY18] contains a geometric model of opinion (preference) structures. Both this and our model propose mechanisms through which information outlets acting for their own benefit can lead to increased disagreement in the society. The key difference to our model is that their population’s preferences are static and do not update, but the outlets are free to choose what information to offer to their customers. By contrast, in our model, the influencers have pre-determined ideologies and compete to align agents’ opinions with their own. In other words, [PY18] focuses on modeling of competitive information acquisition, and our paper on modeling the influence of marketing on the public opinion.

Our model suggests that under the conditions of biased assimilation, opinion manipulation by one or several global information outlets can unintentionally lead to a strong form of polarization in multi-dimensional opinion space. Not only do people polarize on individual issues, but also their opinions on previously unrelated issues become correlated. This form of polarization is known as issue alignment [BG08] in political science and sociology literature. Issue alignment refers to an opinion structure where the population’s opinions on multiple (relatively independent) issues correlate. It is related to issue radicalization, where the opinions polarize for each issue separately. Compared to issue radicalization, issue alignment is theorized to pose more constraints on the opinions an individual can take, resulting in polarized and clustered mass opinions even when the public opinions are not extreme in any single topic, and presenting more obstacles for social integration and political stability [BG08]. In light of this, one way to view our model is as a mathematical mechanism by which this strong form of polarization can arise and worsen due to companies’, politicians’, and the media’s natural attempts to gain support from the public.

On the more technical side, we note that our update equation bears similarity to Kuramoto model [JMB04] for synchronization of oscillators on a network in the control literature. In this model, each oscillator $i$ is associated with the point $\theta_i$ on the two-dimensional sphere, and $i$ updates its point continuously as a function of its neighbors’ points $\theta_j$:

$$\dot{\theta}_i = \omega_i + \frac{K}{N} \sin(\theta_j - \theta_i),$$

where $K$ is the coupling strength and $N$ is the number of nodes in the network. In two dimensions, our model can be compared to Kuramoto model with $\omega_i = 0$ on a star graph, with the influencers at the center of the star connected to the entire population, where the influencers’ opinions do not change and the update strength is qualitatively similar to $\sin((\theta_v - \theta_u)/2)$ (see (15)). However, we note a crucial difference: in the Kuramoto dynamic, $\theta_i$ always moves towards $\theta_j$, i.e. nodes always move towards synchronization, but in our dynamic, opinions $\theta_i$ are allowed to move further away from $\theta_j$ when the angle between their opinions are obtuse. In addition, the central node in our model can be strategic in choosing its positions, while the central node in Kuramoto model follows the synchronization dynamics of the system. We think this property provides a better model for opinion interactions.

**Subsequent work** A work by Gaitonde, Kleinberg and Tardos [GKT21], announced after we posted the preprint of this paper, proposes a framework that generalizes our random interventions
scenario from Section 3. They prove several interesting results, including a strong form of polarization under random interventions in some related models. They also shed more light on the scenario of dueling influencers from Section 6, showing that in case the dueling interventions are orthogonal, the resulting dynamics exhibits a weaker kind of polarization.

3 Asymptotic scenario: random interventions polarize opinions

In this section, we analyze the long-term behavior of our model in a simple random setting. We assume that, for given dimension $d$ and parameter $\eta$, at the initial time $t = 1$ we are given $n$ opinion vectors $u_1^{(1)}, \ldots, u_n^{(1)}$. Subsequently, we sample a sequence of interventions $v^{(1)}, v^{(2)}, \ldots$, each $v^{(t)}$ iid from the uniform distribution on the unit sphere $S^{d-1}$. At time $t$ we apply the random intervention $v^{(t)}$ to every opinion vector $u_i^{(t)}$, obtaining a new opinion $u_i^{(t+1)}$.

We want to show that the opinions $\{u_i^{(t)}\}$ almost surely polarize as time $t$ goes to infinity. We need to be careful about defining the notion of polarization: since the interventions change at every time step, the opinions cannot converge to a fixed vector. Instead, we show that for every pair of opinions the angle between them converges either to $0$ or to $\pi$. More formally:

**Theorem 3.1.** Consider the model of iid interventions described above for some $d \geq 2$, $\eta > 0$ and initial opinions $u_1^{(1)}, \ldots, u_n^{(1)}$. For any $1 \leq i < j \leq n$ and $t \to \infty$, 
\[ \Pr \left[ \|u_i^{(t)} - u_j^{(t)}\| \to 0 \lor \|u_i^{(t)} + u_j^{(t)}\| \to 0 \right] = 1. \]

This leads to a corollary which follows by applying the union bound (with probability 0 in each term) for each pair of opinions $u_i^{(t)}, u_j^{(t)}$:

**Corollary 3.1.** For any $d \geq 2$, $\eta > 0$, initial opinions $u_1^{(1)}, \ldots, u_n^{(1)}$ and a sequence of uniform iid interventions, almost surely, there exists $J \subseteq \{1, \ldots, n\}$ such that the diameter of the set \[ \left\{ (-1)^{l \in J} \cdot u_i^{(t)} : i \in \{1, \ldots, n\} \right\} \]
converges to zero.

**Remark 3.1.** Consider initial opinions of $n$ agents that are independently sampled from a distribution $\Gamma$ that is symmetric around the origin, in the sense that $\Gamma(-A) = \Gamma(A)$ for every set $A \subseteq S^{d-1}$. Then, with high probability, the opinions converge to two polarized clusters of size roughly $n/2$. Indeed, consider sampling $n$ independent vectors $u_1, \ldots, u_n$ from $\Gamma$ and $n$ independent signs $\sigma_1, \ldots, \sigma_n \in \{\pm 1\}$. Then $\sigma_1 u_1, \ldots, \sigma_n u_n$ are independent samples from $\Gamma$. Moreover, if the sizes of the two clusters for $u_1, \ldots, u_n$ are $r$ and $n-r$ then the size of each cluster for $\sigma_1 u_1, \ldots, \sigma_n u_n$ is distributed according to $\text{Bin}(r, 1/2) + \text{Bin}(n-r, 1/2) = \text{Bin}(n, 1/2)$ (this is due to the observation that $u_i$ and $-u_i$ converge to the opposite clusters).

**Remark 3.2.** For simplicity we do not elaborate on this later, but we note that, both empirically and theoretically, the convergence in our results is quite fast. This concerns Theorem 3.1, as well as the results presented in the subsequent sections.

We now proceed to the proof of Theorem 3.1:
3.1 Notation and main ingredients

Before we proceed with explaining the proof, let us make a general observation that we will use frequently. Let \( d \geq 2 \) and \( \eta > 0 \) and let \( f : S^{d-1} \times S^{d-1} \to S^{d-1} \) be the function mapping an opinion \( u \) and an intervention \( v \) to an updated opinion \( f(u, v) \), according to (2) and (3). It should be clear that this function is invariant under isometries: namely, for any real unitary transformation \( A : S^{d-1} \to S^{d-1} \) we have

\[
f(Au, Av) = Af(u, v) .
\] (4)

In our proofs we will be often using (4) to choose a convenient coordinate system.

Let us turn to Theorem 3.1. Again, let \( d \geq 2 \) and \( \eta > 0 \). Without loss of generality we will consider only two starting opinions called \( u_1^{(1)} \) and \( u_2^{(1)} \). To prove Theorem 3.1, we need to show that almost surely one of the vectors \( u_1^{(t)} - u_2^{(t)} \) and \( u_1^{(t)} + u_2^{(t)} \) vanishes.

We proceed by using martingale convergence. Specifically, let

\[
\alpha_t := \arccos(u_1^{(t)}, u_2^{(t)}) .
\]

That is, \( 0 \leq \alpha_t \leq \pi \) is the primary angle between \( u_1^{(t)} \) and \( u_2^{(t)} \).

The proof rests on two claims. First, \( \alpha_t \) is a martingale:

**Claim 3.1.** \( E[\alpha_{t+1} | \alpha_t] = \alpha_t \).

Second, we show a property which has been called “variance in the middle” [BGN+18]:

**Claim 3.2.** For every \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that,

\[
\varepsilon \leq \alpha_t \leq \pi/2 \implies \Pr[\alpha_{t+1} < \alpha_t - \delta | \alpha_t] > \delta ,
\] (5)

and, symmetrically,

\[
\pi/2 \leq \alpha_t \leq \pi - \varepsilon \implies \Pr[\alpha_{t+1} > \alpha_t + \delta | \alpha_t] > \delta .
\] (6)

These two claims imply Theorem 3.1 by standard tools from the theory of martingales (eg., [Wil91]):

**Claims 3.1 and 3.2 imply Theorem 3.1.** As a consequence of applying Claim 3.2 \([\pi/\delta]\) times, we obtain that for every \( \varepsilon > 0 \) there exist \( k_0 \in \mathbb{N} \) and \( \eta < 1 \) such that

\[
\varepsilon \leq \alpha_t \leq \pi - \varepsilon \implies \Pr[\forall 1 \leq k \leq k_0 : \varepsilon \leq \alpha_{t+k} \leq \pi - \varepsilon | \alpha_t] \leq \eta .
\]

Subsequently, it follows that for any fixed \( \varepsilon > 0 \) and \( T \in \mathbb{N} \),

\[
\Pr[\forall t \geq T : \varepsilon \leq \alpha_t \leq \pi - \varepsilon] = 0 .
\] (7)

By Claim 3.1, the sequence of random variables \( \alpha_t \) is a bounded martingale and therefore almost surely converges. Accordingly, let \( \alpha^* := \lim_{t \to \infty} \alpha_t \). We now argue that \( \Pr[0 < \alpha^* < \pi] = 0 \). To that end,

\[
\Pr[0 < \alpha^* < \pi] \leq \sum_{s=1}^{\infty} \Pr\left[ \frac{1}{s} < \alpha^* < \pi - \frac{1}{s} \right] \leq \sum_{s=1}^{\infty} \Pr\left[ \exists T : \forall t \geq T : \frac{1}{2s} < \alpha_t < \pi - \frac{1}{2s} \right] \\
\leq \sum_{s=1}^{\infty} \sum_{T=1}^{\infty} \Pr\left[ \forall t \geq T : \frac{1}{2s} < \alpha_t < \pi - \frac{1}{2s} \right] = 0 ,
\]

where we applied (7) in the last line. Hence, almost surely, either \( \alpha^* = 0 \), which is equivalent to \( \|u_1^{(t)} - u_2^{(t)}\| \to 0 \) or \( \alpha^* = \pi \), equivalent to \( \|u_1^{(t)} + u_2^{(t)}\| \to 0 \). \( \square \)
In the subsequent sections we proceed with proving Claims 3.1 and 3.2. In Section 3.2 we prove Claim 3.1 for \( d = 2 \). In Section 3.3 we show the same claim for \( d \geq 3 \) by a reduction to the case \( d = 2 \). Finally, in Section 3.4 we use a continuity argument to prove Claim 3.2.

In the following proofs, we fix \( d, \eta, \) time \( t \) and the opinions of two agents at that time. For simplicity, we will denote the relevant vectors as \( u := u_1^{(t)}, u' := u_2^{(t)} \) and \( v := v^{(t)} \).

### 3.2 Proof of Claim 3.1 for \( d = 2 \)

![Figure 2: On the left an illustration of the vectors and angles in the proof of Claim 3.1. On the right an illustration for the proof of Claim 3.2.](image)

It follows from (4) that we can assume wlog that \( u = (1,0) \) and \( u' = (\cos \alpha_t, \sin \alpha_t) \) (recall that by definition \( 0 \leq \alpha_t \leq \pi \) holds). Let us write the random intervention vector as \( v = (\cos \beta, \sin \beta) \), where the distribution of \( \beta \) is uniform in \([0,2\pi)\). We will also write (cf. Figure 2 for an overview)

\[
 f(u, v) = (\cos \gamma, \sin \gamma), \quad f(u', v) = (\cos(\alpha_t + \gamma'), \sin(\alpha_t + \gamma')) , \quad \gamma, \gamma' \in [-\pi, \pi).
\]

Note that \( \gamma \) is a function of the intervention angle \( \beta \), and it should be clear that \( \gamma(\beta) = -\gamma(-\beta) \). Accordingly, the distribution of \( \gamma \) is symmetric around zero and in particular \( \mathbb{E} \gamma = 0 \) (where the expectation is over \( \beta \)). Applying (4), it also follows \( \mathbb{E} \gamma' = 0 \).

Let \( \widehat{\alpha} := \alpha_t + \gamma' - \gamma \). Since \( \widehat{\alpha} \) is equal to the directed angle from \( f(u, v) \) to \( f(u', v) \), one might think that we have just established \( \mathbb{E}[\alpha_{t+1} | \alpha_t] = \alpha_t \). However, recall that we defined \( \alpha_{t+1} \) to be the value of the (primary) undirected angle between \( f(u, v) \) and \( f(u', v) \). In particular, it holds that \( 0 \leq \alpha_{t+1} \leq \pi \), but we cannot assume that about \( \widehat{\alpha} \). On the other hand, it is clear that if \( 0 \leq \widehat{\alpha} \leq \pi \), then indeed \( \alpha_{t+1} = \widehat{\alpha} \). Therefore, in the following we will show that \( 0 \leq \widehat{\alpha} \leq \pi \) always holds, which implies \( \mathbb{E}[\alpha_{t+1} | \alpha_t] = \mathbb{E}[\widehat{\alpha} | \alpha_t] = \alpha_t + \mathbb{E}\gamma' - \mathbb{E}\gamma = \alpha_t \).

To that end, we start with showing a weaker bound \(-\pi < \widehat{\alpha} < 2\pi \). To see this, we first establish that \(-\pi/2 < \gamma, \gamma' < \pi/2 \). The argument for \( \gamma \) is as follows: if \( \langle u, v \rangle \geq 0 \), then \( f(u, v) \) is a convex combination of \( u \) and \( v \). Therefore, an intervention cannot move \( f(u, v) \) away from \( u \) by an angle of more than \( \pi/2 \). If \( v \neq u \), then also \( f(u, v) \neq v \), so in fact the angle must be strictly less, that is \(-\pi/2 < \gamma < \pi/2 \). If \( \langle u, v \rangle < 0 \), then \(-\pi/2 < \gamma < \pi/2 \) follows from the same argument applied to \(-v \) (since the effect of both interventions is the same). Finally, \(-\pi/2 < \gamma' < \pi/2 \) holds by (4) and the same proof. Since we know \( 0 \leq \alpha_t \leq \pi \), we obtain \(-\pi < \widehat{\alpha} < 2\pi \).

Since we know \(-\pi < \widehat{\alpha} < 2\pi \), the inequality \( 0 \leq \widehat{\alpha} \leq \pi \) is equivalent to \( \sin \widehat{\alpha} \geq 0 \). Geometrically, this property means that the ordered pair of vectors \((u, v)\) has the same orientation as the pair \((f(u, v), f(u', v))\). To avoid case analysis, we prove this claim by a calculation:
Claim 3.3. \( \sin \hat{\alpha} \geq 0. \)

*Proof.* We defer the proof to Appendix A. \(\square\)

### 3.3 Proof of Claim 3.1 for \( d \geq 3 \)

In this case we will write the random intervention vector as \( v = v^\parallel + v^\perp \) where \( v^\parallel \) is projection of \( v \) onto the span of \( u \) and \( u' \). In particular, \( v^\parallel \) and \( v^\perp \) are orthogonal. We will now prove a stronger claim \( \mathbb{E} [\alpha_{t+1} | \alpha_t, \|v^\parallel\|] = \alpha_t \).

Accordingly, condition on the value \( \|v^\parallel\| = R \). Observe that, by symmetry, vector \( v^\parallel \) is distributed uniformly in the two-dimensional space \( \text{span}\{u, u'\} \) among vectors of norm \( R \). In other words, we can write \( v^\parallel = RV \), where \( V \) is a uniform two-dimensional unit length vector.

Denote the non-normalized vectors after intervention as

\[
\hat{u} := u + \eta(u, v^\parallel)(v^\parallel + v^\perp), \quad \hat{u}' := u' + \eta(u', v^\parallel)(v^\parallel + v^\perp).
\]

Let \( c := 2\eta + \eta^2 \). We proceed with calculations:

\[
\langle \hat{u}, \hat{u}' \rangle = \langle u, u' \rangle + c\langle u, v^\parallel \rangle \langle u', v^\parallel \rangle = \langle u, u' \rangle + cR^2\langle u, V \rangle \langle u', V \rangle,
\]

\[
\|\hat{u}\|^2 = 1 + c\langle u, v^\parallel \rangle^2 = 1 + cR^2\langle u, V \rangle^2,
\]

\[
\|\hat{u}'\|^2 = 1 + c\langle u', v^\parallel \rangle^2 = 1 + cR^2\langle u', V \rangle^2.
\]

Note that all these formulas are valid also for \( d = 2 \), with the only difference that \( R = 1 \) holds deterministically in that case.

Since \( c(\eta) = 2\eta + \eta^2 \) is a bijection on \( \mathbb{R}_{>0} \), there exists \( \hat{\eta} > 0 \) such that \( cR^2 = 2\hat{\eta} + \hat{\eta}^2 \). Accordingly, for any \( d \geq 3 \) and \( \eta > 0 \), the joint distribution of \( \langle \hat{u}, \hat{u}' \rangle \), \( \|\hat{u}\| \) and \( \|\hat{u}'\| \) conditioned on \( \alpha_t = \arccos(\langle u, u' \rangle) \) and \( \|v^\parallel\| = R \) is the same as their joint distribution for \( d = 2 \) and \( \hat{\eta} \), conditioned on the same value of \( \alpha_t \).

Since \( \alpha_{t+1} = \arccos \left( \frac{\langle \hat{u}, \hat{u}' \rangle}{\|u\|\|u'\|} \right) \), the same correspondence holds for the distribution of \( \alpha_{t+1} \) conditioned on \( \alpha_t \) and \( \|v^\parallel\| = R \). Therefore, \( \mathbb{E} [\alpha_{t+1} | \alpha_t, \|v^\parallel\| = R] = \alpha_t \) follows by Claim 3.1 for \( d = 2 \), which we already proved. \(\square\)

### 3.4 Proof of Claim 3.2

Again we use (4) to choose a coordinate system and assume wlog that \( u = (1, 0, \ldots, 0) \) and \( u' = (\cos \alpha_t, \sin \alpha_t, 0, \ldots, 0) \). Our objective is to show that, with probability at least \( \delta \), we will have \( \alpha_{t+1} - \alpha_t > \delta \) (in case \( \alpha_t \leq \pi/2 \)) or \( \alpha_{t+1} - \alpha_t < -\delta \) (in case \( \alpha_t \geq \pi/2 \)). To start with, we will show that by symmetry we need to consider only the first case \( \alpha_t \leq \pi/2 \).

Note that the intervention function \( f \) exhibits a symmetry \( f(-u, v) = -f(u, v) \). Furthermore, we also have \( \arccos \langle u, u' \rangle = \pi - \arccos \langle u, -u' \rangle \). Consequently,

\[
\alpha_{t+1} - \alpha_t = \arccos(f(u, v), f(u', v)) - \arccos(u, u') = -\left( \arccos(f(u, v), f(-u', v)) - (\pi - \arccos(u, -u')) \right).
\]

As a result, indeed it is enough that we prove (5) and then (6) follows by replacing \( u' \) with \( -u' \).

Consider vector \( v^* := (\cos \varepsilon, \sin \varepsilon, 0, \ldots, 0) \) (see Figure 2). We will now show that if \( \varepsilon \leq \alpha_t \leq \pi/2 \) and the intervention \( v \) is sufficiently close to \( v^* \), then \( v \) decreases the angle between \( u \) and \( u' \). To that end, let us use a metric on \( S^{d-1} \) given by

\[
D(u, v) := \arccos(u, v).
\]
Note that this metric is strongly equivalent to the standard Euclidean metric restricted to \( S^{d-1} \).

We can now use the triangle inequality to write
\[
\alpha_{t+1} = D(f(u, v), f(u', v)) \\
\leq D(f(u, v), f(u, v^*)) + D(f(u, v^*), v^*) + D(v^*, f(u', v^*)) + D(f(u', v^*), f(u', v)).
\]
(8)

Let us bound the terms in (8) one by one.

First, since, by (2), \( f(u', v^*) \) is a strict convex combination of \( u \) and \( v^* \) (note that in our coordinate system neither \( u \) nor \( v^* \) depends on \( \alpha_t \)), we have
\[D(f(u, v^*), v^*) = d(\varepsilon) < D(u, v^*) = \varepsilon.\]

Similarly,
\[D(v^*, f(u', v^*)) \leq D(v^*, u') = \alpha_t - \varepsilon.\]

Second, since \( f \) is continuous, \( D(v,v^*) < \delta' \) for small enough \( \delta' > 0 \) implies that both \( D(f(u,v), f(u,v^*)) \) and \( D(f(u', v^*), f(u', v)) \) are as small as needed (for example, less than \( (\varepsilon - d(\varepsilon))/4 \)).

All in all, we have that for some \( \delta' = \delta'(\varepsilon) > 0 \),
\[D(v,v^*) < \delta' \implies \alpha_{t+1} < \frac{\varepsilon - d(\varepsilon)}{4} + d(\varepsilon) + (\alpha_t - \varepsilon) + \frac{\varepsilon - d(\varepsilon)}{4}
= \alpha_t - \frac{\varepsilon - d(\varepsilon)}{2}.\]

However, clearly, the event \( D(v,v^*) < \delta' \) has some positive probability \( \delta'' \). Therefore, taking \( \delta := \min(\delta''/2, (\varepsilon - d(\varepsilon))/2) \), we have
\[
\Pr[\alpha_{t+1} < \alpha_t - \delta | \alpha_t] > \delta,
\]
as claimed in (5).

4 Asymptotic scenario: finding densest hemisphere

In this section we study the asymptotic scenario with one influencer who wishes to propagate a campaign agenda \( v^* \in \mathbb{R}^d \). We assume that the influencer can use an unlimited number of interventions and its objective is to make the opinions of as many agents as possible to converge to \( v^* \). More specifically, in this section we denote the initial opinions of agents at time \( t = 1 \) by \( u_1, \ldots, u_n \). Given these preexisting opinions of \( n \) agents, we want to find a sequence of interventions, \( v(1), v(2), v(3) \ldots \) that maximizes the number of agents whose opinions converge to \( v^* \).

The thrust of our results is that finding a good strategy for the influencer is computationally hard. However, both the optimal strategy and some natural heuristics result in the polarization of agents.

4.1 Equivalence of optimal strategy to finding densest hemisphere

We first argue that the problem of finding an optimal strategy is equivalent to identifying an open hemisphere that contains the maximum number of agents. An (open) hemisphere is an intersection of the unit sphere with a homogeneous open halfspace of the form \( \{ x \in \mathbb{R}^d : \langle x, a \rangle > 0 \} \) for some \( a \in \mathbb{R}^d \).
Theorem 4.1. For any \( v^* \), there exists a strategy to make at least \( k \) agents converge to \( v^* \) if and only if there exists an open hemisphere containing at least \( k \) of the opinions \( u_1, \ldots, u_n \).

A surprising aspect of Theorem 4.1 is that the maximum number of agents that can be persuaded does not depend on the target vector \( v^* \). As we argue in Remark 4.1, this is somewhat plausible in the long-term setting with unlimited number of interventions. We also note that the number of interventions required to bring the opinions up to a given level of closeness to \( v^* \) does depend on \( v^* \).

Proof of Theorem 4.1. First, we prove that the hemisphere condition is sufficient for the existence of a strategy to make the agents’ opinions converge (Claim 4.1). Then we prove the trickier direction: that the hemisphere condition is also necessary for the existence of such a strategy (Claim 4.5).

Claim 4.1. If opinions \( u_1, \ldots, u_k \) are contained in an open hemisphere, then there is a sequence of interventions making all of \( u_1, \ldots, u_k \) converge to \( v^* \).

Proof. By definition of open hemisphere, there is a vector \( a \in \mathbb{R}^d \) such that \( \langle u_i, a \rangle > 0 \) for every agent \( i = 1, \ldots, k \). By (2), it is clear that repeated application of \( a \) makes all the points converge to \( a \) as time \( t \to \infty \).

After all the points are clustered close enough to \( a \), by a similar argument they can be “moved around” together towards another arbitrary point \( v^* \). For example, if \( \langle v^*, a \rangle > 0 \), the intervention \( v^* \) can be applied repeatedly. If \( \langle v^*, a \rangle \leq 0 \), one can proceed in two stages, first applying an intervention proportional to \( (v^* + a)/2 \), and then applying \( v^* \).

Remark 4.1. As a possible interpretation of the mechanism in Claim 4.1, it is not unheard of in campaigns on political issues to use an analogous strategy. First, build a consensus around a (presumably compromise) opinion. Then, “nudge” it little by little towards another direction.

In an extreme case one can imagine this mechanism even flipping the opinions of two polarized clusters. One example of this could be the reversal of the opinions on certain issues of 20th century Republican and Democratic parties in the US (this particular phenomenon can be found in many texts, e.g. [KW18]).

To prove the other direction of Theorem 4.1, we will rely on the notions of conical combination and convex cone. A conical combination of points \( u_1, \ldots, u_n \in \mathbb{R}^d \) is any point of the form \( \sum_{i=1}^{n} \alpha_i u_i \) where \( \alpha_i \geq 0 \) for every \( i \). A convex cone is a subset of \( \mathbb{R}^d \) that is closed under finite conical combinations of its elements. Given a finite set of points \( S \subseteq \mathbb{R}^d \), the convex cone generated by \( S \) is the smallest convex cone that contains \( S \).

Claim 4.2. Suppose that for a given sequence of interventions, the opinions \( u_1, \ldots, u_n \) converge to the same point \( v^* \). Then, for any unit vector \( u_{n+1} \) that lies in the convex cone of \( u_1, \ldots, u_n \), we have that \( u_{n+1} \) also converges to \( v^* \).

Proof. It suffices to prove that if at time \( t \) an opinion \( u_{n+1}^{(t)} \) lies in the convex cone of other opinions \( u_1^{(t)}, \ldots, u_n^{(t)} \), then after applying one intervention \( v^{(t)} \) the new opinion \( u_{n+1}^{(t+1)} \) lies in the convex cone of \( u_1^{(t+1)}, \ldots, u_n^{(t+1)} \). Then the claim follows by induction.
To prove this, we can simply write out $u^{(t+1)}_{n+1}$, using the relation $u^{(t)}_{n+1} = \sum_{i=1}^{n} \lambda_{i} u^{(t)}_{i}$ (where we use the notation $u \propto v$ to mean that $u = c \cdot v$ for some constant $c > 0$):

$$u^{(t+1)}_{n+1} \propto u^{(t)}_{n+1} + \eta \left( u^{(t)}_{n+1}, v^{(t)} \right) \cdot v^{(t)}$$

$$= \sum_{i=1}^{n} \lambda_{i} u^{(t)}_{i} + \eta \sum_{i=1}^{n} \lambda_{i} \left( u^{(t)}_{i}, v^{(t)} \right) \cdot v^{(t)}$$

$$= \sum_{i=1}^{n} \lambda_{i} \left( u^{(t)}_{i} + \eta \cdot \left( u^{(t)}_{i}, v^{(t)} \right) \cdot v^{(t)} \right)$$

$$= \sum_{i=1}^{n} \lambda_{i} \cdot c_{i} u^{(t+1)}_{i}$$

(9)

where the constants in (9) are $c_{i} := \left\| u^{(t)}_{i} + \eta \cdot \left( u^{(t)}_{i}, v^{(t)} \right) \cdot v^{(t)} \right\|$. Specifically, they are all non-negative. □

**Claim 4.3.** Suppose there are two opinions $u_{1}, u_{2}$ that are antipodal, i.e., $u_{1} = -u_{2}$. Then these two opinions will remain antipodal in future time steps. In particular, they will never converge to a single point.

**Proof.** This follows directly from (2), noting that, for any intervention $v$, we have $u_{1} + \eta \cdot \left( u_{1}, v \right) \cdot v = - \left( u_{2} + \eta \cdot \left( u_{2}, v \right) \cdot v \right)$. □

We will also use the following consequence of the separating hyperplane theorem:

**Claim 4.4.** A collection of unit vectors $a_{1}, \ldots, a_{n}$ cannot be placed in an open hemisphere if and only if the zero vector lies in the convex hull of $a_{1}, \ldots, a_{n}$.

Now we are ready to establish the reverse implication in Theorem 4.1.

**Claim 4.5.** Suppose that we start with agent opinions $u_{1}, \ldots, u_{n}$ and that there is no hemisphere that contains $M$ of those opinions. Then, there is no strategy that makes $M$ of the opinions converge to the same point.

**Proof.** Assume towards contradiction that there exists a strategy that makes $M$ opinions converge to the same point, and assume wlog that they are $u_{1}, \ldots, u_{M}$. By assumption, we know that there is no hemisphere that contains all of $u_{1}, \ldots, u_{M}$, hence, by Claim 4.4, there is a convex combination of $u_{1}, \ldots, u_{M}$ that equals 0. Therefore, there is also a conical combination of $u_{1}, \ldots, u_{M-1}$ that equals $-u_{M}$, where wlog we assume that the coefficient on $u_{M}$ is initially nonzero. By Claim 4.2, we conclude that if $u_{1}, \ldots, u_{M-1}$ converge to the same point, then so does $-u_{M}$. But that means that $-u_{M}$ and $u_{M}$ converge to the same point, which is a contradiction by Claim 4.3. □

That concludes the proof of Theorem 4.1. □

**Remark 4.2.** One consequence of Theorem 4.1 is that if the agent opinions are initially distributed uniformly on the unit sphere, and if the number of agents $n$ is large compared to the dimension $d$, an optimal strategy converging as many opinions as possible to $v^{*}$ results, with high probability, in dividing the population into two groups of roughly equal size, where the opinions inside each group converge to one of two antipodal limit opinions (i.e., $v^{*}$ and $-v^{*}$). Furthermore, this optimal strategy, which, as discussed below, might not be easy to implement, will not perform significantly better than a very simple strategy of fixing a random intervention and applying it repeatedly. Of course the simple strategy will also polarize the agents into two approximately equally large groups.


4.2 Computational equivalence to learning halfspaces

Theorem 4.1 implies that an optimal strategy for the influencer is to compute the open hemisphere that is the densest, i.e., it contains the most opinions, and then apply the procedure from Claim 4.1 to converge the opinions from this hemisphere to $v^*$. In this section we study the computational complexity of this problem. While different approaches are possible, we focus on hardness of approximation and worst-case complexity.

**Definition 4.1** (Densest hemisphere). The input to the densest hemisphere problem consists of parameters $n$ and $d$ and a set of $n$ unit vectors $D = \{u_1, \ldots, u_n\}$ with $u_i \in S^{d-1}$. The objective is to find vector $a \in S^{d-1}$ maximizing the number of points from $D$ that belong to the open halfspace $\{x \in \mathbb{R}^d : \langle x, a \rangle > 0\}$.

We analyze the computational complexity of the densest hemisphere problem in terms of the number of vectors $n$, regardless of dimension $d$. In particular, the computationally hard instances that exist as we will show in Theorem 4.2 have high dimension, without any guarantees beyond $d \leq n$ (which can always be assumed wlog). On the other hand, the algorithms from Theorem 4.3 run in time polynomial in $n$ uniformly for all values $d \leq n$.

In contrast, the case of finding densest hemisphere in fixed dimension $d$ can be solved efficiently. For example, an optimal solution can be found by considering $O(n^d)$ halfspaces defined by $d$-tuples of input vectors. We omit further details.

Our main result in this section relies on equivalence of the densest hemisphere problem and the problem of learning noisy halfspaces. Applying a work by Guruswami and Raghavendra [GR09] we will show that it is computationally difficult to even approximate the densest hemisphere up to any non-trivial constant factor:

**Theorem 4.2.** Unless $P=NP$, for any $\varepsilon > 0$, there is no polynomial time algorithm $A_\varepsilon$ that distinguishes between instances of densest hemisphere problem such that, letting $D := \{u_1, \ldots, u_n\}$:

- Either there exists a hemisphere $H$ such that $|D \cap H|/n > 1 - \varepsilon$.
- Or for every hemisphere $H$ we have $|D \cap H|/n < 1/2 + \varepsilon$.

Consequently, unless $P=NP$, for any $\varepsilon > 0$ there is no polynomial time algorithm $A_\varepsilon$ that, given an instance $D$ that has a hemisphere with density more than $1 - \varepsilon$, always outputs a hemisphere with density more than $1/2 + \varepsilon$.

In other words, even if guaranteed the existence of an extremely dense hemisphere, no polynomial time algorithm can do significantly better than choosing an arbitrary hyperplane and outputting the one of its two hemispheres that contains the larger number of points. At the same time, [BDS00] (relying on earlier work [BDES02]) shows that there exists an algorithm that finds a dense hemisphere provided that this hemisphere is stable in the sense that it remains dense even after a small perturbation of its separating hyperplane:

**Theorem 4.3** ([BDS00]). For every $\mu > 0$, there exists a polynomial time algorithm $A_\mu$, that, given an instance $D = \{u_1, \ldots, u_n\}$ of the densest hemisphere problem, provides the following guarantee:

Let $a \in S^{d-1}$ be the vector that maximizes the size of intersection $|D \cap H_a,\mu|$ for halfspace $H_{a,\mu} = \{x : \langle x, a \rangle > \mu\}$. Then, the algorithm $A_\mu$ outputs a hemisphere corresponding to a homogeneous halfspace $H_{a'} = \{x : \langle x, a' \rangle > 0\}$ such that $|D \cap H_{a'}| \geq |D \cap H_{a,\mu}|$. 

18
We emphasize that the only inputs to the algorithms are \( n, d \) and the set of vectors \( D \), and that the complexity is measured as a function of \( n \). For example, the algorithm \( A_\mu \) runs in polynomial time for every \( \mu > 0 \), but the running time is not uniformly polynomial in \( 1/\mu \).

In the remainder of this section we elaborate on how to obtain Theorem 4.2 from known results. To that end, we start with defining the related problem of finding maximum agreement halfspace.

**Definition 4.2 (Maximum Agreement Halfspace).** In the problem of maximum agreement halfspace, the inputs are parameters \( n \) and \( d \), and a labeled set of points \( D = \{(x_1,y_1), \ldots, (x_n,y_n)\} \in \mathbb{R}^d \times \{\pm 1\} \). The objective is to find a halfspace \( H = \{ x: \langle x,a \rangle > c \} \) for some \( a \in \mathbb{R}^d \) and \( c \in \mathbb{R} \) which maximizes the agreement

\[
A(D,H) = \frac{\sum_{i=1}^{n} 1 [y_i \cdot x_i \in H]}{n}.
\]

There is a strong hardness of approximation result for maximum halfspace agreement [GR09] (see also [FGKP06, BB06, BDEL03, AK98] for related work):

**Theorem 4.4 ([GR09]).** Unless \( P=NP \), for any \( \varepsilon > 0 \), there is no polynomial time algorithm \( A_\varepsilon \) that distinguishes the following cases of instances of maximum agreement halfspace problem:

- There exists a halfspace \( H \) such that \( A(D,H) > 1 - \varepsilon \).
- For every halfspace \( H \) we have \( A(D,H) < 1/2 + \varepsilon \).

As in Theorem 4.2, the hard instances are not guaranteed to have any dimension bounds beyond trivial \( d \leq n \).

As pointed out in [BDS00], there exists a reduction from the maximum agreement halfspace problem to the densest hemisphere problem that preserves the quality of solutions. Since this reduction is only briefly sketched in [BDS00], we describe it below.

The reduction proceeds as follows: Given a labeled set \( D = \{(x_1,y_1), \ldots, (x_n,y_n)\} \in \mathbb{R}^d \times \{\pm 1\} \), we map it to \( D' = \{x'_1, \ldots, x'_n\} \in \mathbb{R}^{d+1} \) using the formula

\[
x'_i = \frac{1}{\sqrt{1+\|x_i\|^2}} \cdot (y_i; x_i, 1).
\]

In other words, we proceed in three steps: first, we negate each point that came with negative label \( y_i = -1 \). Then, we add a new coordinate and set its value to 1 for every point \( x_i \). Finally, we normalize each resulting point so that it lies on the unit sphere in \( S^d \).

This is a so-called “strict reduction”, which is expressed in the following claim:

**Claim 4.6.** The solutions (halfspaces) for an instance \( D \) of Maximum Agreement Halfspace are in one-to-one correspondence with solutions (hemispheres) for the reduced instance \( D' \) of Densest Hemisphere. Furthermore, for a corresponding pair of solutions \((H,H')\) the agreement \( A(D,H) \) is equal to the density \( |D' \cap H'|/n \).

**Proof.** It is more convenient to think of solutions for \( D' \) as homogeneous, open halfspaces \( H' = \{x \in \mathbb{R}^{d+1}: \langle x,a \rangle > 0 \} \).

With that in mind, we map a solution to the maximum agreement halfspace problem \( H = \{x \in \mathbb{R}^d: \langle x,a \rangle > c \} \) to a solution to the densest hemisphere problem \( H' = \{(x,x_{d+1}) \in \mathbb{R}^{d+1}: \langle \left( x,x_{d+1}, (a,-c) \right) \rangle > 0 \} \). Clearly, this is a one-to-one mapping between open halfspaces in \( \mathbb{R}^d \) and homogeneous open halfspaces in \( \mathbb{R}^{d+1} \).

Furthermore, it is easy to verify that \( y_i \cdot x_i \in H \) if and only if \( x'_i \in H' \) and therefore \( A(D,H) = |D' \cap H'|/n \). \( \Box \)
Theorem 4.2 follows from Theorem 4.4 and Claim 4.6 by standard (and straightforward) arguments from complexity theory.

5 Short-term scenario: polarization as externality

The analysis of the asymptotic setting with unlimited interventions tells us what is feasible and what is not. A fundamentally different question is how to persuade as many as possible with a limited number of interventions. This is motivated by bounded resources or time that usually allow only limited placements of campaigns and advertisements. Furthermore, arguably only the initial interventions can be considered effective: in the long run the opinions might shift due to external factors and become more unpredictable and harder to control. Therefore, in this section we discuss strategies where the influencer has only one intervention at its disposal, and its goal is to get as many agents as possible to exceed certain “threshold of agreement” with its preferred opinion. Throughout this section, we fix \( \eta = 1 \) in Equation 1, so an opinion \( u \) is updated to be proportional to \( w = u + \langle u, v \rangle \cdot v \).

Both scenarios we discuss in this section describe a situation where a “new” product or idea is introduced. Therefore, we assume that the agents have some preexisting opinions in \( \mathbb{R}^{d-1} \) and that they are neutral as to the new idea, with the \( d \)-th coordinate set to zero for every agent. Our results indicate significant potential for polarization in such a situation. This is in spite of the fact that the influencer might only care about persuading a number of agents towards the new subject, without intention to polarize.

Since we are dealing with scenarios with only one intervention, we use the following notational convention: an initial opinion of agent \( i \) is denoted \( u_i \) and the opinion after intervention is denoted \( \tilde{u}_i \).

5.1 One intervention, two agents: polarization costs

We consider a simple example that features only two agents and one influencer who is allowed one intervention. We imagine a new product, such that the agents are initially agnostic about it, i.e., \( u_{i,d} = 0 \) for \( i = 1, 2 \). Given an intervention \( v \), we are interested in two issues: First, what will be new opinions of agents about the product \( \tilde{u}_{i,d} \)? Second, assuming that the initial correlation between opinions is \( c = \langle u_1, u_2 \rangle \), what will be the new correlation \( \tilde{c} = \langle \tilde{u}_1, \tilde{u}_2 \rangle \)? We think of the correlation as a measure of agreement between the agents and therefore interpret differences in correlation as changes in the extent of polarization.

In order to answer these questions, we introduce notions of two- and one-agent interventions corresponding to two natural strategies:

Definition 5.1. The two-agent intervention is an intervention that maximizes \( \min(\tilde{u}_{1,d}, \tilde{u}_{2,d}) \). The one-agent intervention maximizes \( \max(\tilde{u}_{1,d}, \tilde{u}_{2,d}) \).

The motivation for this definition is as follows. Assume that there exists a threshold \( T > 0 \) such that agent \( i \) is going to make a positive decision (e.g., buy the product or vote a certain way) if its coordinate \( \tilde{u}_{i,d} \) exceeds \( T \). Then, if the influencer cares only about inducing agents to make the decision, it has two natural choices for the intervention. One option is the case where it is possible to induce two decisions, i.e., achieve \( \tilde{u}_{1,d}, \tilde{u}_{2,d} > T \). By continuity considerations, it is not difficult to see that an intervention that achieves this can be assumed to maximize \( \min(\tilde{u}_{1,d}, \tilde{u}_{2,d}) \) with \( \tilde{u}_{1,d} = \tilde{u}_{2,d} \) (such intervention is also optimal if the influencer bets on convincing both agents without knowing \( T \)). The other case is to appeal only to one of the agents, disregarding the second agent and concentrating only on achieving, say, \( \tilde{u}_{1,d} > T \).
Let $c = \langle u_1, u_2 \rangle$ be the initial correlation between opinions and let $c_{\text{two}}$ and $c_{\text{one}}$ be the correlations after applying, respectively, the two- and one-agent interventions. Our main result in this section is:

**Proposition 5.1.** Let $\rho := c_{\text{two}} - c_{\text{one}}$ be a value that we call the polarization cost. Then, we always have $\rho \geq 0$ with exact values given as

\[ c_{\text{two}} = 1 - \frac{\sqrt{2}(1 - c)}{\sqrt{3c + 5}}, \quad c_{\text{one}} = \frac{c\sqrt{2}}{\sqrt{c^2 + 1}}. \tag{10} \]

The values of $\rho, c_{\text{two}}$ and $c_{\text{one}}$ as functions of $c$ are illustrated in Figure 3. Proposition 5.1 states that the one-agent intervention always results in smaller correlation than the two-agent intervention. Note that we made a modeling assumption that the influencer will always choose an intervention as opposed to doing nothing. This is consistent with a scenario where the influencer’s objective is to increase the opinions above the threshold $T$. In that case doing nothing is certain to give no gain to the influencer.

The main conclusion of this theorem is consistent with our other results. In the setting we consider, in the absence of any external mitigation, the self-interested influencer without direct intention to polarize might be incentivized to choose the intervention that increases polarization. If polarization is regarded as undesirable, the polarization cost can be thought of as the externality imposed on the society.

Looking at Figures 3 and 4, this effect seems most pronounced for initial correlation around $c \approx -0.5$, where the one-agent intervention increases polarization, the polarization cost is large and the range of thresholds $T$ for which the influencer profits from the one-agent strategy is relatively large. This suggests that a situation where the society is already somewhat polarized is particularly vulnerable to spiraling out of control. It also suggests that situations where the level of commitment required for the decision (i.e., the threshold $T$) is large increase the risk of polarization.

We also note that this overall picture is complicated by the case of positive initial correlation $c > 0$. In that case both two- and one-agent interventions actually increase the correlation between the agents, even though the two-agent intervention does so to a larger extent. The analysis leading to the proof of Proposition 5.1 is contained in Appendix C.

### 5.2 One intervention, many agents: finding the densest spherical cap

A more general version of the problem of persuading with limited number of interventions features $n$ agents with opinions $u_1, \ldots, u_n \in \mathbb{R}^d$. The influencer is given a threshold $0 \leq T < 1$ and can apply one intervention $v$ with the objective of maximizing the number of agents such that $\tilde{u}_{i,d} > T$. As before, we assume that initially $u_{i,d} = 0$ and that $T$ can be interpreted as a threshold above which a consumer decides to buy the newly advertised product, or more generally take a desired action, such as voting, donating, etc.

Interestingly, we show that this problem is equivalent to a generalization of the densest hemisphere problem from the long-term scenario discussed in Section 4. More precisely, it is equivalent to finding a densest spherical cap of a given radius (that depends on the threshold $T$) in $d - 1$ dimensions.

We give the technical statement in the proposition below. We make an assumption $0 \leq T < 1/3$, since $1/3$ is the maximum value that can be achieved in the $d$-th coordinate by a single intervention, cf. Figure 4. In order to state Proposition 5.2, we slightly abuse notation and write vectors $u \in \mathbb{R}^d$ as $u = (u^*, u_d)$ for $u^* \in \mathbb{R}^{d-1}$, $u_d \in \mathbb{R}$.
Figure 3: Illustration of the polarization cost as a function of the initial correlation $c$. The dashed line is the initial correlation included as a reference point. The red and blue lines are correlations after applying two- and one-agent interventions respectively. The green line shows the polarization cost $c_{\text{two}} - c_{\text{one}}$.

Figure 4: The after-intervention opinions of both agents $\tilde{u}_{i,d}$ as functions of initial correlation $c$. The red line represents the opinion of either agent after applying the two-agent intervention. The blue line is the opinion of the second agent after the one-agent intervention. For reference, the dashed line ($1/3$) shows the opinion of the first agent in the one-agent intervention (which does not depend on $c$). The grey area represents the range of thresholds $T$ where it is preferable for the influencer to apply the one-agent intervention.

**Proposition 5.2.** In the setting above, let

$$c := \frac{2T}{1 - 3T^2}, \quad z := \sqrt{\frac{\sqrt{1 + 3c^2} - 1}{3c}}, \quad \beta := \arccos(z).$$

Then, the number of agents with $\tilde{u}_{i,d} > T$ is maximized by applying an intervention

$$v := (\cos \beta \cdot \nu^*, \sin \beta)$$

for a unit vector $\nu^* \in \mathbb{R}^{d-1}$ that maximizes the number of agents satisfying

$$\langle u_i^*, v^* \rangle > c.$$

The proof of Proposition 5.2 is contained in Appendix D. Note that the solution to this short-term problem for $T$ going to zero approaches the densest hemisphere solution to the long-term problem discussed in Section 4.
6 Asymptotic effects of two dueling influencers: two randomized interventions polarize

Finally, we analyze a scenario where there are two influencers with differing agendas, represented by different intervention vectors $v$ and $v'$. We consider the randomized setup, where at each time step, one of the influencers is randomly chosen to apply their intervention. We demonstrate that this setting also results, in most cases and in a certain sense, in the polarization of agents.

Recall that a convex cone of two vectors $v$ and $v'$ is the set $\{\alpha v + \beta v' : \alpha, \beta \geq 0\}$. A precise statement that we prove is:

**Theorem 6.1.** Let $\langle v, v' \rangle > 0$ and let a starting opinion $u^{(1)}$ be such that $\langle u^{(1)}, v \rangle \neq 0$ or $\langle u^{(1)}, v' \rangle \neq 0$. Then, as $t$ goes to infinity and almost surely, either the Euclidean distance between $u^{(t)}$ and the convex cone generated by $v$ and $v'$ or between $u^{(t)}$ and the convex cone generated by $-v$ and $-v'$ goes to 0.

In order to justify the assumptions of Theorem 6.1, note that if an agent starts with an opinion $u^{(1)}$ such that

$$\langle u^{(1)}, v \rangle = \langle u^{(1)}, v' \rangle = 0,$$

applying $v$ or $v'$ never changes their opinion. In Theorem 6.1 we show that if (12) does not hold and, additionally, $\langle v, v' \rangle > 0$, (if $\langle v, v' \rangle < 0$ we can exchange $v'$ with $-v'$ without changing the effects of any interventions), the opinion vector with probability 1 ends up either converging to the convex cone generated by $v$ and $v'$ or to the convex cone generated by $-v$ and $-v'$. In particular, since vectors $u$ for which (12) holds form a set of measure 0, if $n$ initial opinions are sampled iid from an absolutely continuous distribution, almost surely all opinions converge to the convex cones (which are themselves sets of measure 0 for $d > 2$).

Furthermore, this notion of polarization is strengthened if the correlation between the two interventions is large. As in Theorem 3.1, the best we can hope for is that for each pair of opinions either the distance between $u_1^{(t)}$ and $u_2^{(t)}$ or between $u_1^{(t)}$ and $-u_2^{(t)}$ converges to 0. Letting $V := \text{span}\{v, v'\}$ and $W := V^\perp$ and writing any vector $u$ as a sum of its respective projections $u = u_V + u_W$, we show:

**Theorem 6.2.** Suppose that $\langle v, v' \rangle > 1/\sqrt{2 + \eta}$ and let $u_1^{(1)}, u_2^{(1)}$ be such that $(u_1^{(1)})_V \neq 0$, $(u_2^{(1)})_V \neq 0$. Then, almost surely, either $\|u_1^{(t)} - u_2^{(t)}\|$ converges to 0, or $\|u_1^{(t)} + u_2^{(t)}\|$ converges to 0.

In other words, the stronger notion of convergence, same as in Section 3 with uniformly drawn random interventions, reappears in case the correlation between two interventions $v$ and $v'$ is larger than $1/\sqrt{2 + \eta}$. In particular, we have strong convergence for any $\eta > 0$ and $\langle v, v' \rangle \geq \sqrt{2}/2 \approx 0.71$, and for $\eta = 1$ for $\langle v, v' \rangle > \sqrt{3}/3 \approx 0.58$. Our experiments suggest that this convergence occurs also for other non-zero values of the correlation $\langle v, v' \rangle$, but we do not prove it here.

Also note that same in spirit as Remark 3.1, the usual argument from symmetry shows that if the initial opinions are independent samples from a symmetric distribution, then with high probability the opinions divide into two clusters of roughly equal size.

The case when $v$ and $v'$ are orthogonal is different. As we mentioned, if $\langle v, v' \rangle > 0$, i.e., the angle between $v$ and $v'$ is less than $\pi/2$, then all opinions converge to the two “narrow” convex cones, respectively between $v$ and $v'$ and between $-v$ and $-v'$ — namely, the pairs of vectors among

---

1We also assume that $v \neq -v'$, as otherwise the intervention effects are the same in our model.
v, v', −v, and −v' between which there are acute angles. Similarly, if ⟨v, v'⟩ < 0, then the opinions converge to two cones between v and −v' and between −v and v'. In case ⟨v, v'⟩ = 0 the four convex cones form right angles, so such a result is not possible.

However, we can still show that an initial opinion u(1) converges to the same quadrant in which it starts with respect to v and v'. Namely, for all t, we have that sgn(⟨u(t), v⟩) = sgn(⟨u(1), v⟩) and sgn(⟨u(t), v'⟩) = sgn(⟨u(1), v'⟩), and furthermore the distance between u(t) and the subspace V goes to 0 with t:

**Proposition 6.1.** Suppose that ⟨v, v'⟩ = 0 and let an initial opinion u(1) be such that ⟨u(1), v⟩ 0 and ⟨u(1), v'⟩ 0. Then, almost surely, the following facts hold:

1. \( ||u_W(t)|| \to 0 \) as \( t \to \infty \).
2. For all \( t \), sgn(⟨u(t), v⟩) = sgn(⟨u(1), v⟩) and sgn(⟨u(t), v'⟩) = sgn(⟨u(1), v'⟩).

Fascinatingly, Gaitonde, Kleinberg and Tardos [GKT21] showed subsequently to our initial preprint that strong polarization does not occur for orthogonal interventions. Specifically, they proved that two opinions in \( S^{d-1} \) with random interventions chosen iid from the standard basis \( \{e_1, \ldots, e_d\} \) do not polarize in the sense of \( u_1^{(t)} - u_2^{(t)} \) or \( u_1^{(t)} - u_2^{(t)} \) vanishing, but they do exhibit a weaker form of polarization. We refer to their paper for more details.

In order to prove Theorem 6.1, we first show that the distance between \( u(t) \) and V almost surely goes to 0 as \( t \to \infty \), by showing that the norm of the projection of \( u(t) \) onto W converges to 0. Then, we demonstrate that the convex cone spanned by v and v' is absorbing: when the projection of \( u(T) \) onto V falls in the cone, then the projections of \( u(t) \) for \( t \geq T \) always stay in the cone as well.

Finally, we show that almost surely the projection of \( u(t) \) onto V eventually enters either the cone spanned by v and v', or the cone spanned by −v and −v'. More concretely, we show that at any time \( t \), there is a sequence of \( T \) interventions that lands the projection of \( u(t+T) \) in one of the cones, for some \( T \) that is independent of \( t \). Since this sequence occurs with probability \( 2^{-T} \), which is independent of \( t \), the opinion almost surely eventually enters one of the cones.

### 6.1 Proofs of Theorem 6.1 and Proposition 6.1

We start with the fact the opinions converge to the subspace V spanned by the two intervention vectors. Recall that \( V = \text{span}\{v, v'\} \) and that \( W = V^⊥ \). In the following we will write \( \langle v, v' \rangle = \cos \theta \) for \( 0 < \theta \leq \pi/2 \).

**Proposition 6.2.** Let \( \langle v, v' \rangle \geq 0 \) and take an opinion vector \( u \) such that \( ||u_W|| = c \geq 0 \). Furthermore, let \( \tilde{u} \) be the vector resulting from randomly intervening on \( u \) with either \( v \) or \( v' \). Then:

1. \( ||\tilde{u}_W||^2 \leq ||u_W||^2 \).
2. With probability at least 1/2, \( ||\tilde{u}_W||^2 \leq ||u_W||^2 \cdot (1 - \xi) \), where

\[
\xi = \min\left(\frac{1}{2}, (\eta + \eta^2/2) \cdot \frac{e^2 \theta^2}{16}\right).
\]

**Proof.** Recall from (2)–(3) that if \( \bar{v} \in \{v, v'\} \) is the intervention vector, then

\[
\tilde{u} = k(u + \eta \langle u, \bar{v} \rangle \cdot \bar{v})
\]
where \( k = \sqrt{\frac{1}{1+(2\eta+\eta^2)\langle u,\bar{v} \rangle^2}} \) is the normalizing constant. Observe that when we project onto \( W \), the component in the direction of \( \bar{v} \) vanishes, so we have that
\[
\tilde{u}_W = k \cdot u_W,
\]
and the first claim easily follows since \( k \leq 1 \).

To establish the second point, we need to show that with probability \( 1/2 \) we have \( k^2 < 1 \) or, equivalently, \( \langle u,\bar{v} \rangle^2 = \langle u_V,\bar{v} \rangle^2 > 0 \). Since \( \theta \neq 0 \), the projected vector \( u_V \) cannot be orthogonal both to \( v \) and \( v' \) (cf. Figure 5). More precisely, for at least one of \( v \in \{v,v'\} \) the primary angle between \( u_V \) and \( v \) (or \( -v \)) must be at most \( \pi/2 - \theta/2 \) and consequently
\[
|\langle u_V, v \rangle| \geq \|u_V\| \cdot |\cos(\pi/2 - \theta/2)| \geq c \cdot \theta/4,
\]
resulting in
\[
k^2 = \frac{1}{1+(2\eta+\eta^2)\langle u_V,\bar{v} \rangle^2} \leq \max \left( \frac{1}{2} - \frac{1}{2} \frac{(\eta + \eta^2/2) \cdot c^2 \theta^2}{16} \right).
\]

![Figure 5: Projection onto the subspace \( V = \text{span}\{v,v'\} \).](image)

Next, we show that the convex cone of vectors \( v \) and \( v' \) is absorbing:

**Proposition 6.3.** Let \( \langle v,v' \rangle \geq 0 \) and take \( u \) to be an opinion vector and \( \tilde{u} \) to be a vector resulting from intervening on \( u \) with either \( v \) or \( v' \). If \( u_V \) is a conical combination of \( v \) and \( v' \), then also \( \tilde{u}_V \) is such a conical combination.

**Proof.** Assume w.l.o.g. that the vector applied is \( v \) and let \( k \) be the same constant as in the proof of Proposition 6.2. Then,
\[
\tilde{u}/k = u + \eta \cdot \langle u,v \rangle \cdot v = u_V + \eta \cdot \langle u_V,v \rangle \cdot v + u_W.
\]
Therefore, \( \tilde{u}_V \) can be written as a nonnegative linear combination of \( u_V \) and \( v \), where we use the fact that \( \langle u_V,v \rangle \) is nonnegative, which follows since \( u_V \) is a conical combination of \( v \) and \( v' \), and \( \langle v,v' \rangle \geq 0 \).

Next, we prove that when \( \langle v,v' \rangle > 0 \), the opinion \( u^{(t)} \) not only approaches subspace \( V \), but also a specific area of \( V \), namely, either \( \text{cone}(v,v') \) or \( \text{cone}(-v,-v') \).

**Proposition 6.4.** Let \( \langle v,v' \rangle > 0 \) and consider a vector \( u^{(t)} \) such that \( \|u_V^{(t)}\| \geq c > 0 \). Then, there exists \( T := T(c,\theta,\eta) \) such that with probability at least \( 2^{-T} \), vector \( u_V^{(t+T)} \) will either be a conical combination of \( v \) and \( v' \) or a conical combination of \( -v \) and \( -v' \).

25
Proof. First, for any vector $u^{(t)}$ such that $\|u^{(t)}_V\| \geq c > 0$, at least one of $v, v', -v, -v'$ has positive inner product with $u^{(t)}_V$ (and $u^{(t)}_W$) which can be lower bounded by a function of $c$ and $\theta$ (see Figure 5). Take such a vector and call it $\bar{v}$. By the argument from Proposition 6.2, applying $\bar{v}$ repeatedly will bring $u^{(t+T)}_V$ arbitrarily close to it. More precisely, for every $\varepsilon > 0$, there exists $T_1 = T_1(c, \theta, \eta, \varepsilon)$ such that $\|u^{(t+T_1)}_V - \bar{v}\| < \varepsilon$ and $\|u^{(t+T_1)}_V - \bar{v}\| < \varepsilon$ both hold.

Furthermore, since $(v, v') > 0$, there exists $\varepsilon > 0$ such that if $\|u^{(t)}_V - \bar{v}\| < \varepsilon$, then applying the other intervention vector $(v$ or $v')$ once guarantees that $u^{(t+1)}_V$ enters the convex cone between $v$ and $v'$ or, respectively, between $-v$ and $-v'$. In particular, if $u^{(1)}_V$ already is in the convex cone, then applying either intervention will keep it inside by Proposition 6.3. On the other hand, if $u^{(t)}_V$ is not yet in the cone, but at the distance at most $\varepsilon$ to $\bar{v}$, then applying the other intervention will bring it inside the cone (see Figure 5).

Therefore, there exists a sequence of $T(c, \theta, \eta) = T_1 + 1$ interventions that make $u^{(t+T)}_V$ enter cone$(v, v')$ or cone$(-v, -v')$. Clearly, this sequence occurs with probability $2^{-T}$. $\square$

We are now ready to prove Theorem 6.1.

Proof of Theorem 6.1. Let $\|u^{(1)}_V\| = c > 0$. Proposition 6.2 tells us that the squared norm of the projection $u^{(t)}_W$ onto subspace $W = V^\perp$ never increases, and with probability $1/2$ decreases by the multiplicative factor $1 - \xi(c, \eta, \theta) < 1$. By induction (note that $\xi$ increases with $c$), $u^{(t)}_W$ converges to 0, and consequently $\|u^{(t)}_V - u^{(t)}_W\|$ converges to 0, almost surely.

In order to show that convergence to one of the two convex cones occurs, we apply Proposition 6.4. Since at any time step $t$, there exists a sequence of $T$ choices that puts $u^{(t)}_V$ in one of the convex cones, and since $T$ depends only on the starting parameters $c$, $\theta$, and $\eta$, we get that $u^{(t)}_V$ almost surely eventually enters one of the cones. By Proposition 6.3 and induction, once $u^{(t)}_V$ enters a convex cone, it never leaves.

Proposition 6.1 follows as a corollary of Propositions 6.2 and 6.3:

Proof of Proposition 6.1. The first statement is an inductive application of Proposition 6.2, exactly the same as in the proof of Theorem 6.1.

The second statement follows from noting that out of four orthogonal pairs of vectors $\{v, v'\}$, $\{v, -v'\}$, $\{-v, v'\}$, or $\{-v, -v'\}$, there is exactly one such that $u^{(1)}_V$ is a (strict) conical combination of this pair (by assuming $\langle u^{(1)}, v \rangle \neq 0$ and $\langle u^{(1)}, v' \rangle \neq 0$ we avoid ambiguity in case $u^{(1)}_V$ is parallel to $v$ or $v'$). By the same argument as in Proposition 6.3 and by induction, if the initial projection $u^{(1)}_V$ is strictly inside one of the convex cones, the projection $u^{(t)}_V$ remains strictly inside forever. $\square$

6.2 Proof of Theorem 6.2

Consider the subspace $V = \text{span}\{v, v'\}$ with some coordinate system (cf. Figure 5) imposed on it. As is standard, a unit vector $u \in V$ can be represented in this system by its angle $\alpha(u) \in [0, 2\pi)$ as measured counterclockwise from the positive $x$-axis.

Given a unit vector $\bar{v} \in V$, let $f_{\bar{v}} : [0, 2\pi) \rightarrow [0, 2\pi)$ be the function with the following meaning: given a unit vector $u \in V$ with angle $\alpha = \alpha(u)$, the value $f_{\bar{v}}(\alpha) = \alpha(\bar{u})$ represents the angle of vector $\bar{u}$ resulting from applying intervention $\bar{v}$ to vector $u$. Note that $\alpha(\bar{v})$ is a fixed point of $f_{\bar{v}}$. Also, by Proposition 6.3, both functions $f_v$ and $f_{v'}$ map the interval corresponding to cone$(v, v')$ to itself.

The main part of our argument is the following lemma, which we prove last:
Lemma 6.1. If $\langle v, v' \rangle = \cos \theta > 1/\sqrt{2 + \eta}$, then functions $f_v$ and $f_{v'}$ restricted to the convex cone of $v$ and $v'$ are contractions, i.e., there exists $k = k(\theta, \eta) < 1$ such that for all vectors $u, u' \in \text{cone}(v, v')$, letting $\alpha := \alpha(u), \beta := \alpha(u'), \overline{v} \in \{v, v'\}$, we have

$$|f_\overline{v}(\beta) - f_\overline{v}(\alpha)| \leq k \cdot |\beta - \alpha|,$$  \hspace{1cm} (13)

where the distances $|f_\overline{v}(\beta) - f_\overline{v}(\alpha)|$ and $|\beta - \alpha|$ are in the metric induced by $S^1$, i.e., “modulo $2\pi$”.

Proof of Theorem 6.2. Lemma 6.1 implies that the angle distance between two opinions $u_1(t), u_2(t) \in V$ starting in the convex cone deterministically converges to 0 as $t$ goes to infinity. Of course, this is equivalent to their Euclidean distance $\|u_1(t) - u_2(t)\|$ converging to 0. We now make a continuity argument to show that such convergence almost surely occurs also for general $u_1(t), u_2(t) \in S^{d-1}$. To this end, we let $g_v, g_{v'} : S^{d-1} \to [0, 2\pi)$ as natural extensions of $f_v, f_{v'}$: the value $g_\overline{v}(u)$ denotes the value $g_\overline{v}(u)$ denotes the angle of the projection $\overline{u}_V$ of the new opinion onto $V$, after applying $\overline{v}$ on opinion $u$ (cf. Figure 5). Note that the value $g_\overline{v}(u)$ depends only on the angle $\alpha(u)$ and the orthogonal projection length $\|u_W\|:

$$g_\overline{v}(u) = g_\overline{v}(\alpha(u)), \|u_W\|).$$

In this parametrization, for $u \in V$ we have $f_\alpha(u) = g_\alpha(u) = g_\alpha(\alpha(u), 0)$.

By Theorem 6.1, for any starting opinions $u_1(t)$ and $u_2(t)$ having non-zero projections onto $V$, almost surely there exists a $t$ such that $(u_1(t))_V$ and $(u_2(t))_V$ end up inside (possibly different) convex cones. We consider the case of $u_1(t)$ and $u_2(t)$ both in cone$(v, v')$, other three cases being analogous. Furthermore, almost surely, $\|u_1(t)\|_W$ and $\|u_2(t)\|_W$ converge to 0. Hence, it is enough that we show that almost surely $|\alpha((u_1(t))_V) - \alpha((u_2(t))_V)|$ (in $S^1$ distance) converges to zero.

To this end, let $\delta > 0$. By uniform continuity of $g_v$, we know that for small enough value of $r$, we have

$$|g_v(\alpha, r) - g_v(\alpha, 0)| < \frac{1 - k}{2} \cdot \delta$$

for every $\alpha \in [0, 2\pi)$, where $k$ is the Lipschitz constant from (13). Therefore, almost surely, for $t$ large enough, for $u_1(t)$ and $u_2(t)$ parameterized as $u_1(t) = (\alpha_1, r_1)$ and $u_2(t) = (\alpha_2, r_2)$ we have

$$|g_v(\alpha_1, r_1) - g_v(\alpha_2, r_2)| \leq |g_v(\alpha_1, r_1) - g_v(\alpha_1, 0)| + |g_v(\alpha_1, 0) - g_v(\alpha_2, 0)| + |g_v(\alpha_2, 0) - g_v(\alpha_2, r_2)|$$

$$\leq \frac{1 - k}{4} \cdot \delta + k \cdot |\alpha_1 - \alpha_2| + \frac{1 - k}{4} \cdot \delta \leq \left(k + \frac{1 - k}{2}\right) \cdot \max(|\alpha_1 - \alpha_2|, \delta).$$

Since $k + (1 - k)/2 < 1$, and applying the same argument to $f_{v'}$, we conclude by induction that the distance $|\alpha_1(t) - \alpha_2(t)|$ must decrease and stay below $\delta$ in a finite number of steps. Since $\delta > 0$ was arbitrary, it must be that $|\alpha_1(t) - \alpha_2(t)|$ converges to 0, concluding the proof of Theorem 6.2. \qed

It remains to prove Lemma 6.1:

Proof. Proof of Lemma 6.1: Recall that we assumed a two-dimensional coordinate system on $V$. Let $f := f_{(1, 0)}$, i.e., $f$ corresponds to the intervention along the $x$-axis in this coordinate system. Clearly, functions $f_v$ and $f_{v'}$ are cyclic shifts of $f$ modulo $2\pi$. More precisely, we have

$$f_\overline{v}(\alpha) = \alpha(\overline{v}) + f(\alpha - \alpha(\overline{v})), \hspace{1cm} (14)$$

where arithmetic in (14) is modulo $2\pi$. Furthermore, $f$ is symmetric around the intervention vector, i.e., $f(\alpha) = 2\pi - f(2\pi - \alpha)$ for $0 < \alpha \leq \pi$. Hence, to prove that $f_v$ and $f_{v'}$ restricted to cone$(v, v')$
are contractions, it is enough that we show that $f$ restricted to the interval $[0, \theta]$ is a contraction (recall that we assumed $\cos^2(\theta) > 1/(2 + \eta))$.

To that end, we use (2) to calculate the formula for $f$ for $0 \leq \alpha \leq \pi/2$ as

$$f(\alpha) = \arccos \left( \frac{(1 + \eta) \cos \alpha}{\sqrt{1 + (2\eta + \eta^2) \cos^2 \alpha}} \right).$$ (15)

More computation using elementary calculus (we omit the details) establishes that, additionally, for every $0 \leq \alpha < \beta \leq \pi/2$:

1. $f(\alpha) \leq \alpha$. In other words, applying the intervention brings vector $u$ closer to the intervention vector.

2. $f(\alpha) < f(\beta)$, i.e., applying the intervention does not change relative ordering of vectors wrt the intervention vector.

3. If $\beta \leq \theta^* := \arccos \left( \sqrt{\frac{1}{2 + \eta}} \right)$, then $0 \leq \alpha - f(\alpha) < \beta - f(\beta)$, i.e., in absolute terms, the “pull” on a vector is stronger the further away it is from the intervention vector (until the correlation reaches the threshold $1/\sqrt{2 + \eta}$, cf. Figure 6).

The preceding items taken together imply that for every $0 \leq \alpha < \beta \leq \theta^*$ we have $0 < f(\beta) - f(\alpha) < \beta - \alpha$. To conclude that $f$ is a contraction, we observe that $f$ and its derivative $f'$ are continuous on the interval $[0, \theta^*]$. If there exist sequences $(\alpha_k)$ and $(\beta_k)$ in $[0, \theta]$ for $\theta < \theta^*$ such that $|f(\alpha_k) - f(\beta_k)|/|\beta_k - \alpha_k|$ converges to 1, then, by compactness, there exist convergent sequences $\alpha_k \to \alpha^*$ and $\beta_k \to \beta^*$ such that $|f(\alpha_k) - f(\beta_k)|/|\beta_k - \alpha_k| \to 1$. Then,

1. Either $\alpha^* \neq \beta^*$ and by continuity we get $f(\beta^*) - f(\alpha^*) = \beta^* - \alpha^*$, contradicting the third property above.

Figure 6: The graph of the “pull function” $\alpha - f(\alpha)$ in case $\eta = 1$. 

\[ \begin{array}{cccccccc}
\pi/16 & \pi/8 & 3\pi/16 & \pi/4 & 5\pi/16 & 3\pi/8 & 7\pi/16 & \pi/2 \\
0 & 0.1 & 0.2 & 0.3 & & & & \\
\end{array} \]
2. Or \( \alpha^* = \beta^* \), which by continuity of \( f' \) implies \( f'(\alpha^*) = 1 \) for some \( 0 \leq \alpha^* < \theta^* \). But that would imply that the derivative of \( \alpha - f(\alpha) \), i.e., \( 1 - f'(\alpha) \), vanishes at \( \alpha^* < \theta^* \), again contradicting the third property above (see also Figure 6).

A Proof of Claim 3.3

Let us embed our underlying space \( \mathbb{R}^2 \) in \( \mathbb{R}^3 \) by setting the last coordinate to zero. Letting \( \times \) denote the cross product, we have

\[
u \times \nu' = (0, 0, \sin \alpha_t), \quad f(u, v) \times f(u', v) = (0, 0, \sin \hat{\alpha}) .
\]

Since the case \( \alpha_t \in \{0, \pi\} \) is easily handled by noticing that \( \hat{\alpha} = \alpha_t \), we can assume that \( 0 < \alpha_t < \pi \). In that case, it is enough that we prove

\[
\langle u \times u', f(u, v) \times f(u', v) \rangle = \sin \alpha_t \sin \hat{\alpha} \geq 0 .
\]  

(16)

Setting \( C(w) := \sqrt{1 + (2\eta + \eta^2)(w, v)^2} \), we apply (2) and bilinearity of cross product to compute

\[
f(u, v) \times f(u', v) = \frac{1}{C(u)C(u')} \left( u \times u' + \eta (u, v)(v \times u') + \langle u', v \rangle (u \times v) \right)
\]

\[
= \frac{1}{C(u)C(u')} \left( u \times u' + \eta (u \times u' + (u, v)v - u) \times (u - \langle u', v \rangle v) \right)
\]

(17)

\[
= \frac{1 + \eta}{C(u)C(u')} \cdot u \times u' ,
\]  

(18)

where in (17) we used the identity \( a \times b + c \times d = a \times d + c \times b + (a - c) \times (b - d) \), and in (18) we used that both \( (u, v)v - u \) and \( u' - \langle u', v \rangle v \) are projections of vectors onto the line orthogonal to \( v \), and therefore they are parallel and their cross product vanishes.

Consequently, we conclude that \( f(u, v) \times f(u', v) \) is parallel to \( u \times u' \) with a positive proportionality constant, which implies (16) and concludes the proof.

B Example with two advertisers

For another slightly more involved example, suppose there are two advertisers marketing their products. Agents’ opinions now have five dimensions (\( d = 5 \)) with the fourth and fifth coordinates corresponding to the opinions on these two products. Initially, 500 opinions on the first three coordinates are distributed randomly and uniformly on a three-dimensional sphere, and the last two coordinates are equal to zero:

\[
u_i = (u_{i,1}, u_{i,2}, u_{i,3}, 0, 0) \quad \text{subject to} \quad u_{i,1}^2 + u_{i,2}^2 + u_{i,3}^2 = 1 .
\]

Suppose the two advertisers apply interventions \( v_1 \) and \( v_2 \) in an alternating fashion. We take \( v_1 \) and \( v_2 \) to be orthogonal, letting

\[
v_1 = (\beta, 0, 0, \alpha, 0) , \quad v_2 = (0, \beta, 0, 0, \alpha) , \quad \alpha = \frac{3}{4} , \beta = \sqrt{1 - \alpha^2} .
\]

We proceed to apply \( v_1 \) and \( v_2 \) in an alternating fashion. In Figure 7 we illustrate the agents’ opinions after each advertiser applied their intervention two, four and six times (so the total of, respectively, four, eight and twelve interventions have been applied). A pattern of polarization on
the fourth and fifth coordinates can be observed. At the same time, the pattern on the first three coordinates is more complicated. The opinions on these dimensions are scattered around a circle on the plane spanned by the first two coordinates. This is a somewhat special behavior that arises because vectors $v_1$ and $v_2$ are orthogonal. It is connected to the difference between Theorem 6.1 and Proposition 6.1 discussed in Section 6.

C Proof of Proposition 5.1

Recall that the two-agent intervention maximizes $\min(\tilde{u}_{1,d}, \tilde{u}_{2,d})$. Due to symmetry, we will consider wlog the one-agent intervention that maximizes $\tilde{u}_{1,d}$. Substituting into (2), we get that applying an intervention $v$ results in

$$\tilde{u}_{i,d} = \frac{\langle u_i, v \rangle \cdot v_d}{\sqrt{1 + 3\langle u_i, v \rangle^2}}. \quad (19)$$

Recalling (4), we can apply any unitary transformation on the opinions without changing the correlations, and hence assume that

$$u_1 := (\sin \alpha, \cos \alpha, 0, \ldots, 0), \quad u_2 := (-\sin \alpha, \cos \alpha, 0, \ldots, 0) \quad (20)$$

for some $0 \leq \alpha \leq \pi/2$ and accordingly, $c = \cos^2 \alpha - \sin^2 \alpha = \cos(2\alpha)$. In particular, $\alpha = 0$ means that the agents are in full agreement, $\alpha = \pi/4$ corresponds to the case of orthogonal opinions and $\alpha = \pi/2$ is the case where the opinions are antipodal.

Assuming (20), once we fix the first two coordinates of the intervention $v_1$ and $v_2$, also the values of $\langle u_1, v \rangle$ and $\langle u_2, v \rangle$ become fixed. Therefore, due to (19), the values of $\tilde{u}_{i,d}$ depend only on $v_d$ in a linear fashion. Accordingly, the influencer should place as much weight as possible on the last coordinate and we can conclude that both two- and one-agent interventions have $v_j = 0$ for $2 < j < d$. Hence, in the following we will assume wlog that $d = 3$, $u_1 = (\sin \alpha, \cos \alpha, 0)$ and $u_2 = (-\sin \alpha, \cos \alpha, 0)$ (see Figure 8).

First, consider the one-agent intervention maximizing $\tilde{u}_{1,3}$. Clearly, the intervention should be of the form

$$v_{one} = \cos \beta \cdot u_1 + \sin \beta \cdot (0, 0, 1)$$

for some $0 \leq \beta \leq \pi/2$. Substituting in (19), we compute

$$(\tilde{u}_{1,3})^2 = \frac{\cos^2 \beta \sin^2 \beta}{1 + 3\cos^2 \beta}. \quad (21)$$

Maximizing (21), we get the maximum at $\cos \beta = \sqrt{3}/3$ and

$$v_{one} = \frac{\sqrt{3}}{3} \cdot u_1 + \frac{\sqrt{6}}{3} \cdot (0, 0, 1),$$

resulting in $\tilde{u}_{1,3} = 1/3$. The value $1/3$ is the benchmark for what can be achieved by one intervention. It is a maximum value for $\tilde{u}_{1,3}$ attainable provided that initially $u_{1,3} = 0$.

What is the effect of this intervention on the other opinion $u_2$? Since $\langle u_2, v_{one} \rangle = \sqrt{3}c/3$, substituting into (19) we get

$$\tilde{u}_{2,3} = \frac{\sqrt{3}c/3 \cdot \sqrt{6}/3}{\sqrt{1 + c^2}} = \frac{c\sqrt{2}}{3\sqrt{1 + c^2}}.$$
Figure 7: Illustration of the process described in Appendix B. This time we need to visualize five dimensions. This is done with spatial positions for the first three dimensions $j = 1, 2, 3$ and two different color scales for $j = 4, 5$. Accordingly, two figures are displayed for each time step $t = 5, 9, 13$. In each pair of figures the points in the left figure have the same spatial positions as in the right figure and the colors illustrate dimensions $j = 4$ (on the left) and $j = 5$ (on the right).
The value of $\tilde{u}_{2,3}$ as a function of the correlation $c \in [-1, 1]$ is shown in blue in Figure 4. In particular, it increases from $-1/3$ to $1/3$, passing through 0 for $c = 0$.

Moving to the two-agent intervention, in this case it is not difficult to see (cf. Figure 8) that the intervention vector should be of the form

$$v_{\text{two}} = (0, \cos \beta, \sin \beta)$$

for some $0 \leq \beta \leq \pi/2$. A computation in a computer algebra system (CAS) establishes that $\tilde{u}_{1,3} = \tilde{u}_{2,3}$ is maximized for

$$\cos^2 \beta = \frac{\sqrt{2}(\sqrt{3}c + 5 - \sqrt{2})}{3(c + 1)},$$

yielding an expression

$$\tilde{u}_{1,3} = \tilde{u}_{2,3} = \sqrt{\frac{3c + 7 - 2\sqrt{6c + 10}}{9(c + 1)}}.$$

This function is depicted in Figure 4 in red. In particular, for $c \in [-1, 1]$, it increases from 0 to 1/3 and its value at $c = 0$ is approximately 0.27. Furthermore, its growth close to $c = -1$ is of the square-root type.

Turning to the new correlation values $c_{\text{one}}$ and $c_{\text{two}}$, another CAS computation using the formulas for $v_{\text{one}}$ and $v_{\text{two}}$ gives

$$c_{\text{one}} = \frac{c\sqrt{2}}{\sqrt{c^2 + 1}}, \quad c_{\text{two}} = 1 - \frac{\sqrt{2}(1 - c)}{\sqrt{3c + 5}},$$

establishing (10). To conclude the proof we need another elementary calculation showing that $c_{\text{two}} \geq c_{\text{one}}$ always holds. We omit the details, referring to Figure 3 and noting that in the critical region for $c = 1 - \varepsilon$ we have

$$c_{\text{two}} = 1 - \frac{1}{2}\varepsilon - \frac{3}{32}\varepsilon^2 + O(\varepsilon^3) \geq c_{\text{one}} = 1 - \frac{1}{2}\varepsilon - \frac{3}{8}\varepsilon^2 + O(\varepsilon^3).$$

\section{Proof of Proposition 5.2}

Let us write a generic intervention vector as

$$v = (\cos \beta \cdot v^*, \sin \beta),$$

Figure 8: The projection of one-agent (left) and two-agent (right) interventions onto the first two dimensions.
where $0 \leq \beta \leq \pi/2$, $v^* \in \mathbb{R}^{d-1}$ and $\|v^*\| = 1$. If $v$ is applied to an opinion vector $u_i = (u_i^*, 0)$ and we let $c_i := \langle u_i^*, v^* \rangle$, substituting into (2) we can compute
\[
u_i + \langle v_i, v \rangle \cdot v = (u_i^*, 0) + c_i \cos \beta (\cos \beta \cdot v^*, \sin \beta) = (u_i^* + c_i \cos^2 \beta \cdot v^*, c_i \cos \beta \sin \beta),
\]
and therefore, using (3),
\[
\tilde{u}_{i,d} = \frac{c_i \cos \beta \sin \beta}{\sqrt{1 + 3c_i^2 \cos^2 \beta}} = \frac{c_i z \sqrt{1 - z^2}}{\sqrt{1 + 3c_i^2 z^2}},
\]
where we let $z := \cos \beta$.

Consider a fixed unit vector $v^* \in \mathbb{R}^{d-1}$. In order to maximize $\tilde{u}_{i,d}$ for an opinion $u_i$ with $\langle u_i^*, v^* \rangle = c_i$, we need to optimize over $z$ in (22), resulting in $z = \sqrt{1 + 3c_i^2 - 1/(\sqrt{3}c_i)}$ and, substituting,
\[
\tilde{u}_{i,d} = \frac{1 + 3c_i^2 - 1}{3c_i}.
\]
The right-hand side of (22) is easily seen to be increasing in $c_i > 0$ for a fixed $z$. Therefore, in order to maximize the number of points with $\tilde{u}_{i,d} > T$ for a fixed $v^*$, we solve the equation $T = \frac{\sqrt{1 + 3c_i^2 - 1}}{3c_i}$ for $c_i$, resulting in $c_i = \frac{2T}{1 - \beta T}$ and apply the intervention
\[
v = (\cos \beta \cdot v^*, \sin \beta),
\]
just as claimed in (11). This intervention results in $\tilde{u}_{i,d} > T$ for all opinions satisfying $\langle u_i^*, v^* \rangle > c$. In other words, the objective $\tilde{u}_{i,d} > T$ is achieved for exactly those opinions contained in the spherical cap $\{x \in \mathbb{R}^{d-1} : \langle x, v^* \rangle > c\}$. Maximizing over all directions $v^* \in \mathbb{R}^{d-1}$ completes the proof.

\section*{References}

[AK98] Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. \textit{Theoretical Computer Science}, 209(1–2):237–260, 1998.

[Axe97] Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. \textit{Journal of Conflict Resolution}, 41(2):203–226, 1997.

[BAB+18] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. \textit{Proceedings of the National Academy of Sciences}, 115(37):9216–9221, 2018.

[Bal07] Delia Baldassarri. \textit{Crosscutting Social Spheres? Political Polarization and the Social Roots of Pluralism}. PhD thesis, Columbia University, 2007.

[BB06] Nader H. Bshouty and Lynn Burroughs. Maximizing agreements and coagnostic learning. \textit{Theoretical Computer Science}, 350(1):24–39, 2006.
[BB07] Delia Baldassarri and Peter Bearman. Dynamics of political polarization. *American Sociological Review, 72*(5):784–811, 2007.

[BBC20] BBC. Trump signs executive order targeting Twitter after fact-checking row, 2020. 29 May 2020. https://www.bbc.com/news/technology-52843986.

[BDEL03] Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences, 66*(3):496–514, 2003.

[BDES02] Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. The computational complexity of densest region detection. *Journal of Computer and System Sciences, 64*(1):22–47, 2002.

[BDS00] Shai Ben-David and Hans-Ulrich Simon. Efficient learning of linear perceptrons. In *Advances in Neural Information Processing Systems (NIPS)*, pages 189–195, 2000.

[BG98] Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *The Review of Economic Studies, 65*(3):595–621, 1998.

[BG08] Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology, 114*(2):408–446, 2008.

[BGN+18] Jarosław Błasiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan. General strong polarization. In *Symposium on Theory of Computing (STOC)*, pages 485–492, 2018.

[BMA15] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239):1130–1132, 2015.

[CRF+11] Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on Twitter. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 89–96, 2011.

[DGL13] Pranav Dandekar, Ashish Goel, and David T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences, 110*(2):5791–5796, 2013.

[DVSC+17] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. Modeling confirmation bias and polarization. *Scientific Reports, 7*:40391, 2017.

[FA08] Morris P. Fiorina and Samuel J. Abrams. Political polarization in the American public. *Annual Review of Political Science, 11*:563–588, 2008.

[FAP05] Morris P. Fiorina, Samuel J. Abrams, and Jeremy C. Pope. *Culture War? The Myth of a Polarized America*. Pearson-Longman, 2005.

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Symposium on Foundations of Computer Science (FOCS)*, pages 563–574, 2006.
[For18] Jacey Fortin. A list of the companies cutting ties with the N.R.A., 2018. The New York Times website, 24 February 2018. https://www.nytimes.com/2018/02/24/business/nra-companies-boycott.html.

[Gar18] Kiran Garimella. Polarization on Social Media. PhD thesis, Aalto University, 2018. 20/2018.

[GGPT17] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. Balancing information exposure in social networks. In Advances in Neural Information Processing Systems (NIPS), pages 4663–4671, 2017.

[GKT21] Jason Gaitonde, Jon Kleinberg, and Éva Tardos. Polarization in geometric opinion dynamics. To appear in ACM Conference on Economics and Computation (EC), 2021.

[GMGM17] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In International Conference on Web Search and Data Mining (WSDM), pages 81–90. ACM, 2017.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. SIAM Journal on Computing, 39(2):742–765, 2009.

[HK02] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. Journal of Artificial Societies and Social Simulation, 5(3), 2002.

[IW15] Shanto Iyengar and Sean J. Westwood. Fear and loathing across party lines: New evidence on group polarization. American Journal of Political Science, 59(3):690–707, 2015.

[JMB04] Ali Jadbabaie, Nader Motee, and Mauricio Barahona. On the stability of the Kuramoto model of coupled nonlinear oscillators. In American Control Conference (ACC), volume 5, pages 4296–4301, 2004.

[KP18] Stefan Krasa and Mattias K. Polborn. Political competition in legislative elections. American Political Science Review, 112(4):809–825, 2018.

[KW18] Ilyana Kuziemko and Ebonya Washington. Why did the democrats lose the south? Bringing new data to an old debate. American Economic Review, 108(10):2830–67, 2018.

[LBS00] Howard Lavine, Eugene Borgida, and John L. Sullivan. On the relationship between attitude involvement and attitude accessibility: Toward a cognitive-motivational model of political information processing. Political Psychology, 21(1):81–106, 2000.

[LRL79] Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37(11):2098–2109, 1979.

[MKFB03] Michael W. Macy, James A. Kitts, Andreas Flache, and Steve Benard. Polarization in dynamic networks: A Hopfield model of emergent structure. Dynamic Social Network Modeling and Analysis, pages 162–173, 2003.
Sendhil Mullainathan and Andrei Shleifer. The market for news. *American Economic Review*, 95(4):1031–1053, 2005.

Elchanan Mossel and Grant Schoenebeck. Reaching consensus on social networks. In *Innovations in Computer Science (ITCS)*, 2010.

Mark Noah. Beyond individual differences: Social differentiation from first principles. *American Sociological Review*, 63(3):309, 1998.

Andrzej Nowak, Jacek Szamrej, and Bibb Latané. From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, 97(3):362, 1990.

Eli Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, New York, 2011.

Pew Research Center. Political polarization in the American public: How increasing ideological uniformity and partisan antipathy affect politics, compromise and everyday life, 2014. [https://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/](https://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/).

Sergey E. Parsegov, Anton V. Proskurnikov, Roberto Tempo, and Noah E. Friedkin. Novel multidimensional models of opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285, May 2017.

Markus Prior. Media and political polarization. *Annual Review of Political Science*, 16:101–127, 2013.

Thomas F. Pettigrew and Linda R. Tropp. A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90(5):751–783, 2006.

Jacopo Perego and Sevgi Yuksel. Media competition and social disagreement, 2018. Working Paper.

Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. On the inevitability of online echo chambers. arXiv:1905.03919, 2019.

John Sides and Daniel J. Hopkins. *Political polarization in American politics*. Bloomsbury Publishing USA, 2015.

Brendan Snyder. LGBT advertising: How brands are taking a stance on issues. Think with Google, 2015.

Guodong Shi, Alexandre Proutiere, Mikael Johansson, John S. Baras, and Karl H. Johansson. The evolution of beliefs over signed social networks. *Operations Research*, 64(3):585–604, 2016.

Elizabeth A. Saylor, Katherine A. Vittes, and Susan B. Sorenson. Firearm advertising: Product depiction in consumer gun magazines. *Evaluation Review*, 28(5):420–433, 2004.

Donald E. Vinson, Jerome E. Scott, and Lawrence M. Lamont. The role of personal values in marketing and consumer behavior. *Journal of Marketing*, 41(2):44–50, 1977.
[Wil91] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[WMKL15] Hywel T.P. Williams, James R. McMurray, Tim Kurz, and F. Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, 2015.

[Zal92] John R. Zaller. *The Nature and Origins of Mass Opinion*. Cambridge University Press, 1992.