Polarization in Geometric Opinion Dynamics

JASON GAITONDE[*], Cornell University
JON KLEINBERG[†], Cornell University
ÉVA TARDOS[‡], Cornell University

In light of increasing recent attention to political polarization, understanding how polarization can arise poses an important theoretical question. While more classical models of opinion dynamics seem poorly equipped to study this phenomenon, a recent novel approach by Hązła, Jin, Mossel, and Ramnarayan (HJMR) [19] proposes a simple *geometric* model of opinion evolution that provably exhibits strong polarization in specialized cases. Moreover, polarization arises quite organically in their model: in each time step, each agent updates opinions according to their correlation/response with an issue drawn at random. However, their techniques do not seem to extend beyond a set of special cases they identify, which benefit from fragile symmetry or contractiveness assumptions, leaving open how general this phenomenon really is.

In this paper, we further the study of polarization in related geometric models. We show that the exact form of polarization in such models is quite nuanced: even when strong polarization does not hold, it is possible for *weaker* notions of polarization to nonetheless attain. We provide a concrete example where weak polarization holds, but strong polarization provably fails. However, we show that strong polarization provably holds in many variants of the HJMR model, which are also robust to a wider array of distributions of random issues—this suggests that the form of polarization introduced by HJMR is more universal than suggested by their special cases. We also show that the weaker notions connect more readily to the theory of Markov chains on general state spaces.

CCS Concepts: • **Theory of computation** → **Social networks**.

Additional Key Words and Phrases: opinion dynamics, polarization, network dynamics

1 INTRODUCTION

People's opinions naturally evolve in response to a variety of external factors, like interactions with other individuals, campaign messaging, media reports, political identities, and random events. Understanding how opinions evolve and the emergent qualitative features of these dynamics remains a subject of immense interest in the computer science, economics, and social science communities.

In recent years, a crucial phenomenon of interest is that of *polarization*, where agents roughly partition into groups holding diametric views. That is, rather than agents holding a rich spectrum of beliefs, individuals instead typically belong to opposite clusters even in cases where beliefs on separate topics ostensibly ought not be correlated. This phenomenon is somewhat widely observed to have accelerated in the last couple of decades; for instance, Gentzkow, Shapiro, and Taddy show that political partisanship is more easily inferrable in recent years than previously from textual analysis using machine learning methods [13]. While polarization and issue realignment is perhaps most familiarly observed along political dimensions (for instance, the correlation between beliefs on climate change and gun rights), these effects are not limited to just these facets [8].

A prototypical way to model the evolution of the beliefs of agents, and thereby address phenomena like polarization, is to assume that the opinion of each agent is scalar- or vector-valued and evolves according to some prescribed update rule. Many classical models, like the DeGroot and Friedkin-Johnsen models [7, 11], study opinion evolution as a form of social learning where agents update beliefs based on the graph-weighted opinions of their neighbors within an ambient network. However, the original forms of these models do not appear to provide a satisfactory mechanism to explain polarization, as the dynamics provably lead to *less* polarization than in the initial configuration (see Section 1.2 for a larger discussion).

To understand how polarization can arise in a more intrinsic way, Hązła, Jin, Mossel, and Ramnarayan (HJMR) recently introduced a novel and simple geometric model of opinion dynamics [19]. In their model, each agent's opinion is a unit length vector where each dimension represents a particular relevant political axis. Their dynamics work as follows: at each time, a new random direction representing a new issue, political figure, or influencer is drawn from some fixed distribution. In response, each agent evaluates the correlation between their current opinions and this new issue and moves either toward or away from this new direction depending on the strength and sign of this correlation. The resulting vector is then renormalized to unit length (see Section 2 for the exact model). These random directions model the natural intuition that opinion evolution is often driven by concrete, possibly random events. For instance, opinions potentially change dramatically in response to campaign messages or candidates during election seasons, when new political coalitions form and ideologies can themselves prove malleable to accommodate the new compositions of these groups. Because each vector is normalized, polarization has a very clean geometric interpretation: a set of opinion vectors is polarized if they are all equal up to sign.

One of the key findings of HJMR is a proof that these simple dynamics *intrinsically and strongly polarize* when restricted to a set of specialized settings. Concretely, when opinions are two-dimensional (so lie on the unit circle) and the new issues are drawn uniformly at random from the circle, or when vectors have arbitrary dimension but new issues correspond to one of two "duelling influencers" that are sufficiently close and drawn with equal probability, then almost surely the distance between any two starting opinions will converge to 0 or 2 via these random dynamics; in other words, almost surely any two agents will have opinions either converging to each other or to the negative of each other so that they are equal up to sign.

However, it is not at all clear whether or not this almost sure polarization is specific to their exact model formulation: for instance, their technique to show strong polarization in the two-dimensional case with uniform random directions does not obviously generalize to higher dimensions, nor to non-uniform distributions—without this high degree of symmetry, it is not clear whether the dynamics still model polarization (see Section 1.1 for more discussion). Moreover, their dynamics are completely *oblivious* in the sense that each agents' opinions evolve randomly in response to (the same) new issues but otherwise have no social influence over each other. A natural and important question in this framework is thus whether, or to what extent, polarization can occur more generally in related variants that do not satisfy these stringent constraints.

To address this, we can abstractly define geometric opinion dynamics more generally as follows: the key components of the HJMR model are that (i) the dynamics are random to model the impact of new political issues that arise, and in fact form a Markov chain on the hypersphere, and (ii) the set of polarized vectors is invariant under the dynamics. This latter property is of course necessary to prove that polarization occurs, while the former provides a convenient mathematical framework to analyze these dynamics. This motivates the following generalization of their dynamics:

Definition 1.1. Let $n \in \mathbb{N}$ be an arbitrary, but otherwise fixed parameter denoting the number of agents, and let d denote the dimension of opinions. In a general model of **geometric opinion dynamics**, the evolution of each agent's opinion follows a discrete-time Markov chain in $\mathbb{S}^{d-1}$ given by a recurrence of the following form:

$$\mathbf{X}_{t+1}^{(i)} \propto \mathbf{X}_t^{(i)} + f_i(\mathbf{X}_t^{(1)}, \dots, \mathbf{X}_t^{(n)}, \xi_t), \tag{1}$$

where $\mathbf{X}_0 = (\mathbf{X}_0^{(1)}, \dots, \mathbf{X}_0^{(n)}) \in (\mathbb{S}^{d-1})^n$ is a given starting configuration of opinions, $\xi_t \in \mathbb{S}^{d-1}$ is drawn i.i.d. over time (and independent of all other random variables) from some distribution $\mathcal{D}$, and each $f_i : (\mathbb{S}^{d-1})^{n+1} \to \mathbb{R}^d$ is an explicit, fixed updating function.

The implicit normalization in this formulation ensures that each opinion returns to the unit sphere; in particular, we assume that the unnormalized quantity above is nonzero so that the projection onto the sphere is well-defined. The joint dynamics of this system thus take place as a Markov chain in $\prod_{i=1}^n \mathbb{S}^{d-1}$. In such models, the random vector ξ_t again denotes a common, random stimulus that each agent in the system updates opinions with respect to, but now possibly also as a function of the opinions of the other agents in the system. For example, ξ_t may model a particular issue or figure during election season that splits society into two camps, "for" and "against," and opinions update according to these coalitions. Crucially, we also assume that the distribution $\mathcal{D}$ that these vectors are drawn from remains fixed throughout time so that these dynamics form a time-homogeneous Markov chain.

With this definition, we extend the results of HJMR in this paper by pursuing a more systematic study of polarization in these general geometric opinion dynamics models that take the form of Equation (1). The motivating questions we consider are: *does strong polarization hold more generally, or is it specific to the HJMR model formulation? Can strong polarization be shown with nontrivial network interactions, thereby incorporating a key feature of models like the DeGroot and Friedkin-Johnsen dynamics? Are there other interesting notions of polarization that hold in these settings, even if the strong form does not hold?*

We show that whether polarization holds in such models is rather surprisingly nuanced and requires extending the notion of polarization beyond what is proven in HJMR. In particular, we show that there exist models that *weakly* polarize in a formal sense, but provably do not satisfy the stronger form in HJMR. We show that these weaker forms of polarization are connected to the existence of nontrivial invariant distributions on the induced Markov chain. Nonetheless, we also prove that strong polarization holds for nontrivial variants of the HJMR model which are also robust to more general distributions of update vectors—in fact, we show that this holds in a model that has network interactions in addition to the random updates. We hope that some of our techniques will prove useful in studying polarization in such models more generally.

1.1 Overview of Results and Techniques

We contribute to the theoretical understanding of polarization in geometric opinion models initiated in [19]. In Section 2, we begin by identifying, though not necessarily requiring, several natural natural properties that such dynamics might satisfy and their relation to the problem of polarization.

These will be convenient in our later arguments. We also define formally several distinct geometric opinion models, including the HJMR model, that will be the subject of our analysis later on.

Our main results in Section 3 show that *strong polarization is more universal in such models than perhaps it would appear from the specialized cases studied by HJMR*. In particular, we show that the strong form of polarization exhibited by HJMR in restricted versions of their model holds in two other variants we consider. The first model we study, which we term the *signed HJMR model*, considers variant of the HJMR dynamics that replaces an inner product with a sign. With these altered dynamics, we show that strong polarization holds in any dimension and is robust to the choice of distribution on update vectors so long as it is sufficiently close to the uniform measure on the sphere. We also show that the same holds in the *party model*, which notably incorporates *network effects* where agents exert influence over each other—such effects are not present in HJMR and it is not clear how to apply their methods to this setting.

To prove our results in Section 3, we resort to conceptually different techniques from those of HJMR. Their work shows strong polarization primarily by appealing to either martingale convergence (in the case of $d = 2$ and with a uniform distribution), which relies on rather delicate symmetries, or deterministic contraction (in two-point models representing "duelling influences", where the challenge of proving contraction is geometric rather than probabilistic). To establish strong polarization in these variants, we appeal to more general zero-one laws to first simplify the problem to showing just that there is *some* nonzero probability of strongly polarizing. We obtain our main results from this quite general reduction by employing clean probabilistic and geometric estimates to give the required uniform bounds in our analysis to prove strong polarization.

Our next main contribution is to connect the polarization phenomenon to the general theory of Markov chains in uncountable state spaces in Section 4. In general, Markov chains in such spaces have considerably more delicate behavior than with countable or finite state spaces. To do so, we define weaker notions of polarization than the strong form considered previously; these notions are motivated by standard notions of convergence from probability theory. We first show that in most cases, whether these weaker forms of polarization hold in geometric opinion dynamics is in fact equivalent to the existence of nontrivial invariant distributions of the Markov chain defined by the dynamics. We then turn to showing the utility of these more nuanced notions of polarization with a concrete example: we conclude by proving that the original HJMR model with orthonormal updates satisfies our notion of weak polarization, but provably *does not* satisfy strong polarization—in fact, we show that almost surely, almost every starting configuration will not strongly polarize. We do so by connecting these dynamics to pathwise properties of infinite balls-in-bins processes, which we analyze by applying yet another zero-one law of Hewitt and Savage.

1.2 Related Work

Our work broadly relates to the vast body of theoretical work in opinion dynamics that studies concrete models wherein agents update numerical beliefs and the resulting properties of these dynamics; for a comprehensive survey of this style of work, see [4]. The particular geometric model of opinion dynamics considered by Hązła, Jin, Mossel, and Ramnarayan [19] motivates many of the considerations in this work. To our knowledge, their work is first to introduce the dynamics in Equation (1) in the specific form given below in Equation (2). We are not aware of related work exploring polarization in variants of their model.

Polarization more generally has been studied in other models of opinion dynamics, typically in the more well-understood DeGroot and Friedkin-Johnsen dynamics mentioned above. Both models assume that agents update opinions according to the graph-weighted average of their neighbors, and both of these models have been extensively studied for their analytical elegance, tractability, and deep connections to well-studied topics like finite-state Markov chains and random walks [14, 15].

However, an idiosyncratic feature of these models is that they are often inherently contractive: opinions get closer due to the dynamics—for instance, in the DeGroot model, agents will necessarily converge to a common limiting opinion under very mild connectedness assumptions. Therefore, the price of the analytic simplicity of this model is that no nontrivial polarization can occur. The Friedkin-Johnsen model typically will not exhibit perfect consensus, but the network effects of the process nonetheless cause initial opinions to contract closer to a common consensus. This suggests that these models simply are not well-equipped to offer a mechanism for this phenomenon.

Recent theoretical work has attempted to combine the analytically desirable features of these models with questions of polarization by incorporating new elements to the model. For instance, behavioral biases like *biased assimilation*, which underlies the intuition behind the HJMR model, have also been fruitfully studied in the context of DeGroot dynamics [6]. Motivated by recent events, a related line of work has also attempted to understand polarization by incorporating the influence of external actors or platforms into the dynamics. Polarization then arises either explicitly or implicitly because of the (possibly orthogonal) objectives of these external parties [1, 5, 12, 21]. A related model by Hegselmann and Krause [18] circumvents near-consensus by allowing agents to filter out overly dissimilar opinions before averaging at each step, but already this modest extension incurs a high cost: many basic questions about convergence remain open in this model, greatly complicating a more extensive understanding of how polarization arises [2]. Moreover, the polarization in this model is somewhat built-in by requiring that far opinions stop responding to each other. While these lines of work have shed significant light on how polarization can arise in such models, and certainly such forces can help explain an empirical increase in polarization in the last few decades, the polarization that arises does not arise organically via the original dynamics.

As we discuss in Section 4, geometric models of opinion dynamics can be viewed as a special case of the more general theory of Markov chains in non-discrete state spaces. The results in this setting are significantly more involved than corresponding results on discrete state spaces; an accessible reference to some important results in this area can be found in the surveys of Hairer [16, 17]. As we show, polarization is related to the set of invariant distributions on the Markov chain induced by the dynamics. Numerous techniques have been developed to determine the uniqueness of an invariant distribution in these settings, as well as the rate and mode of convergence to it if possible (see for instance, [3]). It would be interesting to find new ways to apply results from this area to provide quantitative bounds on the convergence to polarization when one can prove strong polarization holds.

2 PRELIMINARIES

As stated above, we are primarily interested in the *polarization* properties of these random processes given by Equation (1). Throughout this paper, we reserve n to denote the number of agents and d to denote the dimensionality of opinions. We write $\|\cdot\|_p$ for the standard ℓ_p-norm and $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$. We will write $P_{\mathbb{S}^{d-1}} : \mathbb{R}^d \setminus \{\mathbf{0}\} \to \mathbb{S}^{d-1}$ for the projection onto the unit sphere, i.e. $P_{\mathbb{S}^{d-1}}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$. We also write $\mathbf{e}_i$ for the ith standard basis vector in $\mathbb{R}^d$, where d will be clear from context. For a set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, we define $\text{cone}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{z} = \sum_{i=1}^n \alpha_i \mathbf{x}_i, \alpha_i \geq 0\}$. We write $\angle(\mathbf{x}, \mathbf{y})$ for the angle between $\mathbf{x}$ and $\mathbf{y}$.

For given n, d, we then define $D \subseteq \prod_{i=1}^n \mathbb{S}^{d-1}$ to be the set of diagonal opinion vectors, i.e. the set of elements of the form $(\mathbf{x}, \ldots, \mathbf{x})$ for a single vector $\mathbf{x} \in \mathbb{S}^{d-1}$. For a sign vector $\boldsymbol{\sigma} \in \{-1, 1\}^n$, define $\boldsymbol{\sigma}(D) := \{\mathbf{X} \in \prod_{i=1}^n \mathbb{S}^{d-1} : \mathbf{X} = (\sigma_1 \mathbf{x}, \ldots, \sigma_n \mathbf{x}), \mathbf{x} \in \mathbb{S}^{d-1}\}$. Finally, we define the set P of polarized vectors by $P = \bigcup_{\boldsymbol{\sigma} \in \{-1,1\}^n} \boldsymbol{\sigma}(D)$. In other words, P is the set of tuples of vectors such that each vector is equal to the rest up to sign. This definition allows for consensus as a special case, but in many settings, consensus or near-consensus is exponentially unlikely (see Remark 3.1).

For a subset A in some Euclidean space and a point $\mathbf{x}$, we define $\rho(\mathbf{x}, A) := \inf_{\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|_2$ to denote the distance from a point to a set with respect to the Euclidean metric. With this in mind, we define strong polarization of a geometric opinion dynamics model as follows:

Definition 2.1. Let $\mathbf{X}_t := (\mathbf{X}_t^{(1)}, \ldots, \mathbf{X}_t^{(n)})$ be a discrete-time Markov chain as given by Equation (1). Then $\mathbf{X}_t$ **strongly polarizes** (from $\mathbf{X}_0$) if almost surely, $\rho(\mathbf{X}_t, P) \to 0$; that is, the distance between $\mathbf{X}_t$ and the set of polarized vectors converges to zero almost surely.

There are several natural properties that one might desire in these dynamics. Below, we consider, though do not require, the following properties which abstracts those of the original HJMR model:

Definition 2.2. If each function $f_i(\cdot, \xi)$ is continuous from $\prod_{i=1}^n \mathbb{S}^{d-1}$ to $\mathbb{R}^d$ for each fixed ξ, then we say the dynamics are **continuous**.

We will show in Section 4 that continuity implies various desirable properties for the dynamics. Note that if this is the case, combined with the fact that we assume the unnormalized update rule is always nonzero, it follows that the dynamics are jointly continuous as transitions from $\prod_{i=1}^n \mathbb{S}^{d-1}$ to itself. This follows because the the map taking the joint opinion vector to the unnormalized opinion vectors is continuous and nonzero in every coordinate, which is then composed coordinatewise with a continuous map on $\mathbb{R}^d \setminus \{\mathbf{0}\}$, and so is continuous.

Definition 2.3. The dynamics are **sign-invariant** if, for all $i \in [n]$, the function f_i is *odd* with respect to $\mathbf{x}_i$, but *even* with respect to the other arguments (i.e. with respect to $\mathbf{x}_{-i}$ and ξ).

Sign-invariance implies each agent reacts to the random update vector and the others the same regardless if any are negated. This feature will be present in all of the models we consider below, including the original HJMR model. The intuition behind sign-invariance is that from the perspective of each agent, if he or she were to react "positively" to the new issue or a different opinion, she would react "negatively" to the negative of that issue or different opinion—sign-invariance thus asserts that these reactions are balanced.

Definition 2.4. If $f_i(\mathbf{X}_t^{(i)}, \mathbf{X}_t^{(-i)}, \xi) = f_j(\mathbf{X}_t^{(j)}, \mathbf{X}_t^{(-j)}, \xi)$ when $\mathbf{X}_t^{(i)} = \mathbf{X}_t^{(j)}$ and $\mathbf{X}_t^{(-i)} = \mathbf{X}_t^{(-j)}$, we say that the dynamics are **symmetric**, as the updates do not depend on the identities of the agents.

Definition 2.5. If each function f_i does not depend on $\mathbf{X}_t^{(-i)}$ for each i (so that f_i depends only on $\mathbf{X}_t^{(i)}$ and ξ_t), then we say the dynamics are **oblivious**.

In this case, each component of the above process follows a Markov chain, and the joint dynamics form a particular coupling where each component responds to the same update vector. However, any polarization that arises happens indirectly because agents do not influence each other.

In this work, we treat such models in relatively full generality and also specialize to particular models where more specific techniques can establish various desirable properties. The concrete examples we will consider are listed below:

Definition 2.6 (HJMR Model). In the **HJMR Model** [19], the update for each agent i takes the following form for some fixed scalar $\eta > 0$:

$$\mathbf{X}_{t+1}^{(i)} \propto \mathbf{X}_t^{(i)} + \eta \cdot \langle \mathbf{X}_t^{(i)}, \xi_t \rangle \xi_t, \tag{2}$$

where $\xi_t \sim \mathcal{D}$ is drawn i.i.d. over time from some distribution $\mathcal{D}$ on $\mathbb{S}^{d-1}$. In words, each agent moves in the (signed) direction of the random update vector proportionally to the correlation with their current opinion, and then renormalizes.

Note that this model satisfies continuity, sign-invariance, obliviousness, and symmetry (assuming η is a constant over all agents).

Definition 2.7 (Signed HJMR Model). In the **signed HJMR model**, the update rule in Equation (2) is amended to

$$\mathbf{X}_{t+1}^{(i)} \propto \mathbf{X}_t^{(i)} + \eta \cdot \mathrm{sgn}(\langle \mathbf{X}_t^{(i)}, \xi_t \rangle) \xi_t. \tag{3}$$

Here, we define $\mathrm{sgn}(0) = 0$, but in our applications below, we will assume ξ_t is drawn from a continuous distribution so that almost surely $\langle \mathbf{X}_t^{(i)}, \xi_t \rangle \neq 0$. In this case, the amount the vector updates does not depend on the correlation. This choice is intended to model elections, where one is in favor either towards or against a particular candidate and is drawn "all-or-nothing" towards or against the views of this candidate. This model is *not* continuous due to the sign function, but is still sign-invariant, oblivious, and symmetric (assuming η is a constant over all agents).

Definition 2.8 (Party Model). Suppose each agent i has multipliers $(\eta_1^{(i)}, \ldots, \eta_n^{(i)})$ where $\eta_j^{(i)} \geq 0$ measures the influence of agent j on agent i. The **party model** is defined by:

$$\mathbf{X}_{t+1}^{(i)} \propto \mathbf{X}_t^{(i)} + \left(\sum_{j \in [n]: \mathrm{sgn}(\langle \mathbf{X}_t^{(j)}, \xi_t \rangle) = \mathrm{sgn}(\langle \mathbf{X}_t^{(i)}, \xi_t \rangle)} \eta_j^{(i)} \mathbf{X}_t^{(j)} - \sum_{j \in [n]: \mathrm{sgn}(\langle \mathbf{X}_t^{(j)}, \xi_t \rangle) \neq \mathrm{sgn}(\langle \mathbf{X}_t^{(i)}, \xi_t \rangle)} \eta_j^{(i)} \mathbf{X}_t^{(j)} \right). \tag{4}$$

That is, each agent moves towards the vectors that came on the same "side" of the random issue ξ_t, and away from those on the opposite "side" of the issue. While this latter assumption may appear non-obvious, there is considerable empirical evidence for the sociological principle that "out-group conflict builds in-group solidarity" [20]: when a binary issue creates disagreement within a collection of people, the people on each side of the disagreement move toward those they agree with and away from those they disagree with [10, 22]. Once again, this model is not continuous due to the sign function, and also is *not* oblivious as clearly the update rule depends on the values of the other agents. However, it remains sign-invariant, as negating one's own opinions interchanges the sums, and flipping either ξ_t or any other agents opinions only permutes summands.

For any sort of polarization to arise, it is natural to ensure that the dynamics are such that *if* the vector are completely polarized, then they will surely remain so. We provide one simple condition that is easily verified[1]:

LEMMA 2.9. *Suppose that some geometric opinion dynamics satisfy symmetry and sign-invariance. Then $\sigma(D)$ is invariant under the transitions for every $\sigma \in \{-1, 1\}^n$.*

Finally, note that if the dynamics are oblivious and strongly polarizes for $n = 2$ agents, then it must do so for any finite n by a simple union bound. Note that oblivious and symmetric dynamics ensure that the process is well-defined for any number of agents.

LEMMA 2.10. *Suppose that the opinion dynamics are symmetric and oblivious. Then if any form of convergence holds for $n = 2$ agents, the same holds for any finite number of agents.*

3 MODELS WITH STRONG POLARIZATION

In this section, we prove the strong polarization of the signed HJMR model and the party model. To do this, we first establish a simple, but powerful general principle that will significantly simplify the analysis that has the following simple intuition: suppose momentarily that geometric opinion dynamics were a *finite-state* Markov chain and that "polarization" is an absorbing state of the Markov chain. Then from standard and simple Markov estimates, so long as it is possible to reach the state "polarized" from any starting point in some fixed finite number of steps, an easy calculation shows that almost surely the Markov chain will become "polarized." In that case, it would suffice to

---

[1]Note that the converse of Lemma 2.9 trivially fails: one can simply ensure the dynamics are invariant on each such set and otherwise define them arbitrarily.

show that from any starting state, there is *some* nonzero probability of reaching "polarized" from any starting state in some finite number of steps.

In general, this idea is not so simple to formalize because the geometric opinion dynamics lie in a non-discrete state space, and moreover, it is often only possible to reach P asymptotically, not in any finite number of steps. However, by appealing to more general zero-one laws, we show that this intuition nonetheless holds:

THEOREM 3.1. *For any geometric model of opinion dynamics that satisfies the Markov property, the following are equivalent:*

(1) *For every choice of starting vector* $\mathbf{X}_0$, $\mathrm{Pr}_{\mathbf{X}_0}(\rho(\mathbf{X}_t, P) \to 0) = 1$.
(2) *For every choice of starting vector* $\mathbf{X}_0$, $\mathrm{Pr}_{\mathbf{X}_0}(\rho(\mathbf{X}_t, P) \to 0) \geq c$ *for some constant* $c > 0$.

PROOF. One direction is trivial, so we assume the second condition. Consider the dynamics started at any choice of starting vector $\mathbf{X}_0$, and consider the event $A = \{\rho(\mathbf{X}_t, P) \to 0\}$. Let $\mathcal{F}_t = \sigma(\xi_0, \ldots, \xi_{t-1})$ be the filtration generated by the random updates up to time t and let $\mathcal{F}_\infty = \sigma(\xi_0, \ldots)$ be the filtration generated by all of them. By standard arguments, $A \in \mathcal{F}_\infty$.

For each $T \geq 0$, define $A_T := \mathbb{E}[\mathbf{1}(A)|\mathcal{F}_T]$. As $\mathbf{1}(A)$ is an indicator random variable, Lévy's upward theorem (Theorem 4.2.11 of [9]) implies that $A_T = \mathbb{E}[\mathbf{1}(A)|\mathcal{F}_T] \to \mathbb{E}[\mathbf{1}(A)|\mathcal{F}_\infty] = \mathbf{1}(A)$ almost surely. But observe that by the Markov property, $A_T = \mathrm{Pr}_{\mathbf{X}_T}(\rho(\mathbf{Z}_t, P) \to 0)$, where $\mathbf{Z}_t$ gives the dynamics started at $\mathbf{Z}_0 = \mathbf{X}_T$. In particular, $A_T \geq c > 0$ pointwise. Because $A_T \to \mathbf{1}(A)$ almost surely and is bounded from below almost surely by a strictly positive quantity, the only way this can happen is if $\mathbf{1}(A) \equiv 1$ almost surely (over the realizations of the ξ_t), so that A holds almost surely. As $\mathbf{X}_0$ was arbitrary, this completes the harder direction. □

3.1 Strong Polarization in Signed HJMR Model

Our first main result is that strong polarization holds in the signed HJMR model with a common value of $\eta > 0$, and that this holds for a general class of distributions:

THEOREM 3.2. *Suppose there are* $n \geq 1$ *agents in the signed HJMR model given by Equation* (3) *where each* ξ_t *is drawn i.i.d. from a distribution* $\mathcal{D}$ *that is equivalent to the uniform (Haar) measure* μ, *i.e. there exists* $C, C' > 0$ *such that for every measurable set* $A \subseteq \mathbb{S}^{d-1}, C\mu(A) \leq \mathrm{Pr}(\xi_t \in A) \leq C'\mu(A)$. *Then this system strongly polarizes from any choice of starting vector* $\mathbf{X}_0$.

REMARK 3.1. *It can be shown that for any sign-invariant dynamics that strongly polarizes, if* $\mathbf{X}_0$ *is drawn uniformly from* $(\mathbb{S}^{d-1})^n$, *then each possible clustering is equally likely even conditioned on the sequence* $\{\xi_t\}_{t=0}^\infty$ *almost surely. In particular, the probability over starting configurations and the random updates of consensus is exponentially small in* n.

To set up this result, we establish a sequence of lemmas that will prove useful. We begin with a simple geometric fact:

LEMMA 3.3. *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ *and suppose* $\mathbf{z} \in \mathbb{R}^d$ *is such that* $\|\mathbf{x} + \mathbf{z}\|_2, \|\mathbf{y} + \mathbf{z}\|_2 \geq 1 + \epsilon$ *for some* $\epsilon \geq 0$. *Then* $\|P_{\mathbb{S}^{d-1}}(\mathbf{x} + \mathbf{z}) - P_{\mathbb{S}^{d-1}}(\mathbf{y} + \mathbf{z})\|_2 \leq \frac{\|\mathbf{x}-\mathbf{y}\|_2}{1+\epsilon}$.

PROOF. We may assume $\mathbf{x} \neq \pm\mathbf{y}$. Consider the arrangement of vectors $\mathbf{x} + \mathbf{z}, \mathbf{y} + \mathbf{z}, \mathbf{x} - \mathbf{y}$ in the plane spanned by $\{\mathbf{x} + \mathbf{z}, \mathbf{y} + \mathbf{z}\}$ forming a triangle with a vertex at the origin and adjacent sides $\mathbf{x} + \mathbf{z}, \mathbf{y} + \mathbf{z}$. By the assumption that these vectors have length at least $1 + \epsilon$, scaling this triangle by a factor of $r = (1 + \epsilon)^{-1}$ ensures that $r(\mathbf{x} + \mathbf{z})$ and $r(\mathbf{y} + \mathbf{z})$ continue to have at least unit norm, and the distance between them is at most $r(\|\mathbf{x} - \mathbf{y}\|_2)$. As projection onto $\mathbb{S}^{d-1}$ is a contraction in Euclidean distance for vectors of length at least 1, it follows that

$$\|P_{\mathbb{S}^{d-1}}(\mathbf{x} + \mathbf{z}) - P_{\mathbb{S}^{d-1}}(\mathbf{y} + \mathbf{z})\|_2 = \|P_{\mathbb{S}^{d-1}}(r(\mathbf{x} + \mathbf{z})) - P_{\mathbb{S}^{d-1}}(r(\mathbf{y} + \mathbf{z}))\|_2 \leq r\|\mathbf{x} - \mathbf{y}\|_2.$$

□

Next, we show that with some constant probability, a random vector drawn from $\mathcal{D}$ will not split two vectors that form an acute angle, and that the probability of splitting at all tends to zero as the distance tends to zero.

LEMMA 3.4. *There exists constants* $\beta, \gamma > 0$ *depending on d and the measure of equivalence such that the following holds: suppose that* $\mathbf{X}_0 = (\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)})$ *satisfies* $\langle \mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)} \rangle \geq 0$. *Then the probability that ξ_0 satisfies:*

(1) $sgn(\langle \mathbf{X}_0^{(1)}, \xi_0 \rangle) = sgn(\langle \mathbf{X}_0^{(2)}, \xi_0 \rangle)$, *and*

(2) $|\langle \mathbf{X}_0^{(i)}, \xi_0 \rangle| \geq \gamma$ *for $i = 1, 2$,*

is at least β.

PROOF. We first show this when ξ_0 is drawn uniformly from the sphere and then simply change constants when moving to any measure that is equivalent to Haar measure. But this is clear: the probability of ξ_0 satisfies the first property is at least $1/2$ under these assumptions, as the direction of ξ_0 in the plane spanned by $\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)}$ is itself uniform and using the acuteness of the two vectors. Moreover, the distribution of $|\langle \xi_0, \mathbf{z} \rangle|$ does not depend on $\mathbf{z}$ by uniformity, so we may choose $\gamma > 0$ small enough so that the probability of the second property is at least $7/8$ for any fixed $\mathbf{z}$. By a union bound, it follows that the probability ξ_0 has the desired properties is at least $1/4$ under Haar measure. Under the true, equivalent distribution, the probability is thus at least $\beta := C/4 > 0$. □

LEMMA 3.5. *Under the assumptions and notation of Lemma 3.4, the probability that $sgn(\langle \mathbf{X}_0^{(1)}, \xi_0 \rangle) \neq sgn(\langle \mathbf{X}_0^{(2)}, \xi_0 \rangle)$ is at most* $O(\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2)$ *and* $O(\angle(\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)}))$, *where the implicit constant depends only on d and the measure of equivalence.*

PROOF. We first show this for Haar measure. By an analogous argument, the set of vectors with the desired property have directions lie in a band of width $O(\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2)$ in the plane spanned by $\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)}$, and therefore has probability at most $O(\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2)$ and $O(\angle(\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)}))$ by uniformity. For the true distribution, this can increase by a factor of at most C', which we may absorb into the implicit constant. □

With this result in hand, we turn to the proof of the main theorem:

PROOF OF THEOREM 3.2. We begin with a series of reductions that simplifies the problem. First, because these dynamics are oblivious, we observe that by Lemma 2.10 it suffices to consider the case $n = 2$ with an arbitrary starting vector $\mathbf{X}_0 = (\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)})$. Next, by sign-invariance of these dynamics, we may assume that $\langle \mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)} \rangle \geq 0$ by possibly flipping the sign of one of the vectors and noting that both the dynamics and the set P are invariant under these sign changes. Note that this implies that the two starting vectors form an acute angle. Finally, by Theorem 3.1, it suffices to show that the probability that $\|\mathbf{X}_t^{(1)} - \mathbf{X}_t^{(2)}\|_2 \to 0$ is bounded below by some nonzero constant $c > 0$, uniformly over the choice of starting vector (though assuming nonnegative inner product).

We use the notation of Lemma 3.4. Note that the good event of Lemma 3.4 and the bad event of Lemma 3.5 are disjoint, though not mutually exhaustive. We claim that by the craps principle, the probability of encountering a random update ξ_t satisfying the good event before the bad event is at least $\beta/(\beta + O(\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2))$. Indeed, while an update need not satisfy either event, on the complement of the bad event, $\|\mathbf{X}_t^{(1)} - \mathbf{X}_t^{(2)}\|_2$ is nonincreasing so long as the bad event does not occur as the unnormalized lengths increase by sign-invariance with respect to ξ_t (so that we may

assume both signs are positive) and contractions decrease distances. The claim then follows from the crap's principle, Lemma 3.4, and Lemma 3.5.

Moreover, on this event, the distance between $\mathbf{X}_{t+1}^{(1)}$ and $\mathbf{X}_{t+1}^{(2)}$ decreases by a factor of $(1+\epsilon)^{-1}$ where $\epsilon = \Omega(\eta\gamma)$ by Lemma 3.3 (using $\mathbf{z} = \pm\eta \cdot \xi_t$) and the fact that the inner products are bounded below on the good event. It follows that on this event, the new distance between vectors is at most $\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2/(1+\epsilon)$. By the strong Markov property, we may iterate this argument to show that the probability of the good event occurring before the bad event is now at least $\beta/(\beta+O(\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2/(1+\epsilon))) \geq 1 - O\left(\frac{\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2}{1+\epsilon}\right)$ where we absorb the constant β. It follows that the probability that the good event occurs infinitely often without the bad event is at least

$$\prod_{k=0}^{\infty} \left(1 - O\left(\frac{\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2}{(1+\epsilon)^k}\right)\right) \geq \prod_{k=0}^{\infty} \left(1 - O\left(\frac{1}{(1+\epsilon)^k}\right)\right),$$

where we simply upper bound $\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2 \leq 2$ in the inequality. Note that if this occurs, then Lemma 3.3 show that this implies that $\|\mathbf{X}_t^{(1)} - \mathbf{X}_t^{(1)}\|_2 \to 0$ as the distance is nonincreasing and geometrically decreases infinitely often. From standard analysis,

$$\prod_{k=0}^{\infty} \left(1 - O\left(\frac{1}{(1+\epsilon)^k}\right)\right) > 0 \iff \sum_{k=0}^{\infty} O\left(\frac{1}{(1+\epsilon)^k}\right) < \infty,$$

and the latter is clearly true as a geometric series. Moreover, these lower bounds are uniform over the value of $\|\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)}\|_2$. By the reductions above, this completes the proof. □

3.2 Strong Polarization in the Party Model

We now turn to proving strong polarization in the party model. One complicating factor is that these dynamics are *not* oblivious, unlike the other models where strong polarization is known. Therefore, we have to reason about multiple vectors acting on each other at the same time.

To set up the formal statement of the result, we need the following definition: for any given set of nonnegative coefficients $\boldsymbol{\eta}$, let $A = A(\boldsymbol{\eta})$ be the (directed) $n \times n$ adjacency matrix defined by

$$A_{ij} = \begin{cases} 1 & \text{if } \eta_j^{(i)} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, we consider the directed graph with a directed edge from i to j if j influences i. We say that A is irreducible if $A^n > 0$ componentwise. This is equivalent to the existence of a directed path between any two agents i and j. Our main result of this section is as follows:

THEOREM 3.6. *Suppose there are $n \geq 1$ agents in the party model where each ξ_t is drawn i.i.d. from a distribution $\mathcal{D}$ that is equivalent to the uniform (Haar) measure μ, i.e. there exists $C, C' > 0$ such that for every measurable set $B \subseteq \mathbb{S}^{d-1}$, $C\mu(B) \leq \Pr(\xi_t \in B) \leq C'\mu(B)$. Moreover, suppose that $A = A(\boldsymbol{\eta})$ is irreducible. Then this system strongly polarizes from any choice of starting vector $\mathbf{X}_0$.*

To prove this, we follow a similar high-level plan as that of Theorem 3.2. By sign-invariance, we will assume via Lemma 3.7 that each component of $\mathbf{X}_0$ lies on one side of a hyperplane with margin strictly bounded below by zero regardless of the individual configuration. We then construct a potential function that is equivalent to the maximum angle between any two agents with the property that, assuming the dynamics do not split up the vectors on the next n iterations, it is guaranteed to decay by some factor strictly bounded above by 1. Since, as we will again see, the probability that a random vector splits up any two components is essentially bounded by the

maximum angle between any two components and this quantity is decreasing geometrically, it will follow that the probability of strong polarization is bounded below uniformly. We may then again conclude via Theorem 3.1 that the system strongly polarizes from any starting configuration.

We now carry out this plan. First, we show that there always exists a signing of the starting configuration such that all vectors lie on one side of some hyperplane with nontrivial margin:

LEMMA 3.7. *For all $d, n \geq 1$, there exists a constant $\lambda = \lambda(n, d) > 0$ such that for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{S}^{d-1}$, there exists $\mathbf{z} \in \mathbb{S}^{d-1}$ such that $|\langle \mathbf{z}, \mathbf{x}_i \rangle| \geq \lambda$ for all $i \in [n]$.*

PROOF. Simply choose $\lambda > 0$ such that the probability of a random unit vector drawn from Haar measure does not satisfy the condition for a given i is at most $1/(n + 1)$ and then apply a union bound to conclude there exists such a vector. Note that this choice of λ indeed depends on d, n, but not on the choice of vectors as the distribution of the inner product of a uniformly random vector on the sphere with a fixed vector does not depend on the identity of this fixed vector. □

Next, we proceed with several purely geometric results that will enable us to show contraction of a suitable potential function.

LEMMA 3.8. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{S}^{d-1}$ all lie strictly on one side of a hyperplane in $\mathbb{R}^d$. Define $\mathbf{y}$ by*

$$\mathbf{y} \in \arg \min_{\mathbf{v} \in \mathbb{S}^{d-1}} \max_{j \in [n]} \angle(\mathbf{v}, \mathbf{x}_j).$$

Note $\mathbf{y}$ exists as the objective function is continuous and the constraint set is compact. Let $H = \mathbf{y}^\perp$ be the orthogonal subspace to $\mathbf{y}$ and let P_H be the orthogonal projection onto H. Then $\mathbf{0} \in \mathrm{conv}(P_H(\mathbf{x}_1), \ldots, P_H(\mathbf{x}_n))$.

PROOF. Observe that the desired claim is implied by $\mathbf{y} \in \mathrm{cone}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. To see this, suppose that this holds: suppose there exists $\alpha_i \geq 0$ such that $\mathbf{y} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$. Note that not all α_i are zero as $\mathbf{y}$ has unit norm by construction. Then applying P_H to both sides, we deduce that

$$\mathbf{0} = P_H(\mathbf{y}) = \sum_{i=1}^n \alpha_i P_H(\mathbf{x}_i).$$

As $\sum_{i=1}^n \alpha_i > 0$, we may renormalize these coefficients so that their sum is one to deduce that $\mathbf{0} \in \mathrm{conv}(P_H(\mathbf{x}_1), \ldots, P_H(\mathbf{x}_n))$. Moreover, note that $\mathbf{y}$ can equivalently be defined via $\mathbf{y} \in \arg \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \min_{j \in [n]} \langle \mathbf{y}, \mathbf{x}_j \rangle$. This holds as the inner product is a monotonically decreasing function of angle on $[0, \pi]$ and so the optimization problems are equivalent.

Therefore, suppose that $\mathbf{y} \notin \mathrm{cone}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. By the separating hyperplane theorem, there exists a unit vector $\mathbf{u}$ such that $\langle \mathbf{u}, \mathbf{x}_i \rangle > 0$ for all $i \in [n]$ but $\langle \mathbf{u}, \mathbf{y} \rangle < 0$. Write $\mathbb{R}^d$ in an orthogonal basis that includes $\mathbf{u}$. Then the coordinate with respect to $\mathbf{u}$ for each of the $\mathbf{x}_i$ is strictly positive, while the coordinate for $\mathbf{y}$ is strictly negative. Therefore, by reflecting $\mathbf{y}$ about $\mathbf{u}$, we obtain unit $\tilde{\mathbf{y}}$ such that $\langle \tilde{\mathbf{y}}, \mathbf{x}_i \rangle > \langle \mathbf{y}, \mathbf{x}_i \rangle$ for all $i \in [n]$, contradicting the optimality of $\mathbf{y}$. □

LEMMA 3.9. *Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ satisfy $\|\mathbf{z}_i\| \leq 1$ for all $i \in [n]$ and $\mathbf{0} \in \mathrm{conv}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$. Then*

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \right\|_2 \leq 1 - \frac{1}{n}.$$

PROOF. We consider two cases:

(1) First, suppose that $\mathbf{0} \in \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$; without loss of generality, suppose $\mathbf{z}_1 = \mathbf{0}$. Then

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \right\|_2 = \left\| \frac{1}{n} \sum_{i=2}^n \mathbf{z}_i \right\|_2$$

As $\mathbf{0}$ is trivially in the convex hull no matter the choice of $\mathbf{z}_2, \ldots, \mathbf{z}_n$, the convexity of the Euclidean norm implies that this latter quantity is optimized for $\mathbf{z}_2 = \ldots = \mathbf{z}_n = \mathbf{u}$ for some unit vector $\mathbf{u}$. This has norm $1 - \frac{1}{n}$ as needed.

(2) Now suppose that $\mathbf{0} \notin \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. We claim that there exists some $j \in [n]$ such that setting $\mathbf{z}_j = \mathbf{0}$ does not decrease the desired quantity, thereby reducing to the previous case.

To see this, suppose that this is false. In particular, for every $j \in [n]$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \right\|_2 > \left\| \frac{1}{n} \sum_{i \neq j} \mathbf{z}_i \right\|_2 \iff \left\langle \mathbf{z}_j, \sum_{i \neq j} \mathbf{z}_i \right\rangle > 0.$$

This in turn clearly implies that

$$\left\langle \mathbf{z}_j, \sum_{i=1}^{n} \mathbf{z}_i \right\rangle > 0. \tag{5}$$

By assumption, we may write $\mathbf{0} = \sum_{i=1}^{n} \alpha_i \mathbf{z}_i$ for some nonnegative scalars summing to 1. Multiplying Equation (5) by α_j for each $j \in [n]$ and summing, we obtain

$$0 = \left\langle \mathbf{0}, \sum_{i=1}^{n} \mathbf{z}_i \right\rangle = \left\langle \sum_{j=1}^{n} \alpha_j \mathbf{z}_j, \sum_{i=1}^{n} \mathbf{z}_i \right\rangle > 0,$$

a contradiction.

$\square$

Next, we show that the irreducibility of A implies that there is geometric decay in the minimal angle in each n steps that the dynamics do not split the vectors. We start with the following crude, but intuitive lemma:

LEMMA 3.10. *Let $\mathbf{x}_0 = (\mathbf{x}_0^{(1)}, \ldots, \mathbf{x}_0^{(n)})$ be any set of vectors all lying strictly on one side of a hyperplane with $\lambda = \lambda(n, d) > 0$ margin (i.e. there exists unit $\mathbf{v}$ satisfying $\langle \mathbf{v}, \mathbf{x}_0^{(i)} \rangle > \lambda$ for all $i \in [n]$), and consider the update rule given by Equation (4) where the second summand is empty.*

Suppose that we apply this update rule n times to obtain $\mathbf{x}_n$. Then, there exists $\epsilon = \epsilon(n, d, \boldsymbol{\eta}) > 0$ such that $\mathbf{x}_n \propto M\mathbf{x}_0$ for some row-stochastic matrix M depending on $\mathbf{x}_0$ satisfying

$$M = \epsilon \Pi + (1 - \epsilon)Q, \tag{6}$$

where every entry of Π is equal to $1/n$ and Q is some arbitrary stochastic matrix.

REMARK 3.2. *In the above matrix equation, we interpret $\mathbf{x}_t$ as a $n \times d$ matrix where the ith row is $\mathbf{x}_t^{(i)}$. Moreover, the point of the lemma is that ϵ depends on $n, d, \boldsymbol{\eta}$, but not on the starting configuration.*

PROOF. We show the following claim by induction: for each $1 \leq k \leq n$, there exists $\epsilon_k = \epsilon(n, d, \boldsymbol{\eta}, k)$ such that $\mathbf{x}_k \propto M_k \mathbf{x}_0$ where M_k is a row-stochastic matrix such that $(M_k)_{ij} > \epsilon_k$ if $j \in \Gamma^k(i)$, where $\Gamma^k(i)$ is the set of nodes reachable from i in k steps in the directed graph induced by $A = A(\boldsymbol{\eta})$ above. This clearly implies the claim by setting $\epsilon = \epsilon_n$ noting that $\Gamma^k(i) \subseteq \Gamma^{k+1}(i)$ by the fact that this matrix has ones on the diagonal.

For the base case $k = 1$, by definition (and absorbing the term $\mathbf{x}_0^{(i)}$ into the $\eta_i^{(i)}$ multiplier): $\mathbf{x}_1^{(i)} \propto \sum_{j=1}^{n} \eta_j^{(i)} \mathbf{x}_0^{(j)}$. By dividing by $\eta^{(i)} \triangleq \sum_{j=1}^{n} \eta_j^{(i)}$, we clearly obtain the claim with $\epsilon_1 \triangleq \min_{i \in [n]} \min_{j \in [n]:\eta_j^{(i)} > 0} \eta_j^{(i)}/\eta^{(i)}$. Note that this is independent of $\mathbf{x}_0$.

Now suppose it holds for some $k \geq 1$ so that $\mathbf{x}_k \propto M_k \mathbf{x}_0$. By applying the base case to $\mathbf{x}_{k+1}$ (noting that $\mathbf{x}_k$ still lies on the same side of the hyperplane by convexity with same margin),

$$\mathbf{x}_{k+1} \propto M_1' \mathbf{x}_k \propto (M_1' M_k) \mathbf{x}_0.$$

By the induction hypothesis, $(M_1')_{ij} \geq \epsilon(n, d, \boldsymbol{\eta}, 1)$ for all $j \in \Gamma(i)$, while $(M_k)_{ij} \geq \epsilon(n, d, \boldsymbol{\eta}, k)$ for all $j \in \Gamma^k(i)$. For any $j \in \Gamma^{k+1}(i)$, there exists some $j' \in [n]$ such that $j' \in \Gamma^k(i)$ and $j \in \Gamma(j')$. From the definition of matrix multiplication, if $j \in \Gamma^{k+1}(i)$, it follows that $(M_1' M_k)_{i,j} \geq \epsilon_1 \cdot \epsilon_k$. Letting $\epsilon_{k+1} \triangleq \epsilon_1 \cdot \epsilon_k$ and noting the product of stochastic matrices is stochastic, the claim follows. □

Finally, we show that one can get the geometric rate of convergence with respect to a natural potential function.

LEMMA 3.11. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{S}^{d-1}$ *all lie strictly on one side of a hyperplane in* $\mathbb{R}^d$ *with margin at least* $\lambda = \lambda(n, d) > 0$. *Define* $\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *by*

$$\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \min_{\mathbf{v} \in \mathbb{S}^{d-1}} \max_{i \in [n]} \angle(\mathbf{v}, \mathbf{x}_i). \tag{7}$$

Let $\mathbf{x}_1', \ldots, \mathbf{x}_n'$ *be the updated vectors after* n *iterations of the update rule in Equation* (4) *when the second summand is empty. Then there exists* $c = c(n, d, \boldsymbol{\eta}) > 0$ *such that*

$$\Phi(\mathbf{x}_1', \ldots, \mathbf{x}_n') \leq (1 - c)\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n).$$

PROOF. Let $\mathbf{y}$ be an optimizer in Equation (7) and let H be the orthogonal complement. Let P_H be the corresponding orthogonal projection onto H and $P_{\mathbf{y}}$ be the orthogonal projection onto $\mathbf{y}$. Note that for any vector $\mathbf{z}$, $\mathbf{z} = P_H \mathbf{z} + P_{\mathbf{y}} \mathbf{z}$. By Lemma 3.10, we have $\mathbf{x}' \propto M\mathbf{x}$ for some stochastic matrix M such that $M = \epsilon \Pi + (1 - \epsilon)Q$.

Now, observe that by the definition of Φ and from elementary geometry, $\|P_H(\mathbf{x}_i)\|_2 \leq \sin(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_1))$ while $\langle \mathbf{y}, \mathbf{x}_i \rangle = \langle \mathbf{y}, P_{\mathbf{y}} \mathbf{x}_i \rangle \geq \cos(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_1))$ for all $i \in [n]$. Now by definition,

$$\Phi(\mathbf{x}_1', \ldots, \mathbf{x}_n') \leq \max_{i \in [n]} \angle(\mathbf{y}, \mathbf{x}_i') = \max_{i \in [n]} \angle(\mathbf{y}, (M\mathbf{x})_i),$$

as angles do not change under positive scaling.

Therefore, we consider the (unnormalized) set of vectors $(M\mathbf{x})_i$. By linearity and the fact M is row-stochastic, we still have $\|P_{\mathbf{y}}(M\mathbf{x})_i\|_2 = \langle \mathbf{y}, (M\mathbf{x})_i \rangle \geq \cos(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n))$. On the other hand,

$$\|P_H(M\mathbf{x})_i\| = \|\epsilon P_H(\overline{\mathbf{x}}) + (1 - \epsilon) P_H \tilde{\mathbf{x}}_i\|, \tag{8}$$

where $\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ is some arbitrary convex combination. We thus have by linearity

$$\|P_H \tilde{\mathbf{x}}_i\| \leq \sin(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n)). \tag{9}$$

Moreover, again by linearity, $P_H \overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n P_H \mathbf{x}_i$. Recall that by Lemma 3.8, $\mathbf{0} \in \text{conv}(P_H \mathbf{x}_1, \ldots, P_H \mathbf{x}_n)$ and therefore by scaling and applying Lemma 3.9 with $\mathbf{z}_i = P_H \mathbf{x}_i$,

$$\|P_H \overline{\mathbf{x}}\|_2 \leq \left(1 - \frac{1}{n}\right) \sin(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n)). \tag{10}$$

Putting it all together, we now have by monotonicity of the function $\tan(\alpha)$ on $[0, \pi/2]$ and the triangle inequality with Equations (8) to (10) that

$$\begin{aligned}
\tan\left(\Phi(\mathbf{x}_1', \ldots, \mathbf{x}_n')\right) &\leq \max_{i \in [n]} \frac{\|P_H(M\mathbf{x})_i\|_2}{\|P_{\mathbf{y}}(M\mathbf{x})_i\|} \\
&\leq \frac{\epsilon \left(1 - \frac{1}{n}\right) \sin(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n)) + (1 - \epsilon) \sin(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n))}{\cos(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n))} \\
&= \left(1 - \frac{\epsilon}{n}\right) \tan(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n)).
\end{aligned}$$

We may assume that $\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n) < \pi/2 - \delta$ for some small enough $\delta = \delta(n, d) > 0$ by the assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ lie on one side of a hyperplane with margin $\lambda = \lambda(n, d) > 0$. The function $\tan(\cdot)$ has derivative bounded between 1 and some constant depending on n, d on the interval $[0, \pi/2 - \delta]$ and thus by Lemma 3.12, the above geometric decay of $\tan \circ \Phi$ implies that $\Phi(\mathbf{x}_1', \ldots, \mathbf{x}_n') \leq (1 - c) \cdot \Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for some constant $c = c(n, d, \boldsymbol{\eta}) > 0$. □

LEMMA 3.12. *Let $f : [0, a] \to [0, b]$ be a differentiable function such that $f(0) = 0$ and $1 \leq f' \leq K$ for some constant K. If $f(x) \leq (1 - c)f(y)$, then $x \leq (1 - c')y$ for $c' = c/2K$.*

PROOF. Observe from the Mean Value Theorem, the assumption, and the fact $f(y) \geq f(x) \geq x$ from the derivative condition that

$$y - x \geq \frac{f(y) - f(x)}{K} \geq \frac{cf(y)}{K} \geq \frac{cx}{K}.$$

It immediately follows that $y \geq \left(1 + \frac{c}{K}\right) x$. As $(1 + z)^{-1} \leq 1 - z/2$ for $z \leq 1$, we obtain the claim. □

Finally, we can return to the proof of Theorem 3.6:

PROOF OF THEOREM 3.6. By sign-invariance and Lemma 3.7, we may assume that each component of $\mathbf{X}_0 = (\mathbf{X}_0^{(1)}, \ldots, \mathbf{X}_0^{(n)})$ lies strictly on one side of a hyperplane in $\mathbb{R}^d$ with margin at least $\lambda = \lambda(n, d) > 0$; this follows because we may sign the starting configuration arbitrarily, run the dynamics, and then undo the signing by sign-invariance without affecting the polarization properties of the dynamics. We now show that with some constant (independent of $\mathbf{X}_0$) probability, the dynamics monotonically decrease $\Phi(\mathbf{X}_t)$ to zero. Moreover, it is easy to see by the triangle inequality that for any vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$,

$$\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n) \leq \max_{i,j} \angle(\mathbf{x}_i, \mathbf{x}_j) \leq 2\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n). \tag{11}$$

Therefore, $\Phi(\mathbf{X}_t) \to 0$ implies the same of the maximum angle between components and thus implies polarization.

Observe that if the dynamics do not split the vectors on any of the next n iterations, Lemma 3.11 implies that $\Phi(\mathbf{X}_n) \leq (1 - c)\Phi(\mathbf{X}_0)$ for some constant $c > 0$ independent of $\mathbf{X}_0$. This occurs with at least some nonzero constant probability δ (again, depending on $n, d, \boldsymbol{\eta}$, but not on $\mathbf{X}_0$) by the margin condition, Lemma 3.5, and the equivalence with the standard Haar measure.

Now, recall from Lemma 3.5 and a union bound that the probability that the dynamics split up any given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that all lie strictly on one side of a hyperplane with is $O(\max_{i,j} \angle(\mathbf{x}_i, \mathbf{x}_j)) = O(\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n))$ by Equation (11), where the implicit constant depends on n, d and the measure of equivalence. The probability that the sequence $\Phi(\mathbf{X}_{t \cdot n}) \to 0$ is at least the probability that $\Phi(\mathbf{X}_{(t+1) \cdot n}) \leq (1 - c)\Phi(\mathbf{X}_{t \cdot n})$ for each $t \geq 0$. As we have shown that the probability that this fails decays geometrically, it follows that this latter probability is at least

$$\prod_{t=0}^{\infty} \left(1 - \Theta_{n,d,\boldsymbol{\eta}}\left((1-c)^t\right)\right) > c'$$

for some constant $c' > 0$ that does not depend on $\mathbf{X}_0$, where the inequality follows from the same standard analysis argument as in Theorem 3.2. This immediately implies that strong polarization holds with constant probability on the restriction of the sequence to each n steps. To extend this to the whole sequence, by inspecting the proof of Lemma 3.11, it is easy to see that if the vectors are not split at some time $t + 1$, then $\Phi(\mathbf{X}_{t+1}) \leq \Phi(\mathbf{X}_t)$ for *every* $t \geq 0$, not just on the subsequence (though we may not have strict contraction). In particular, the event that the dynamics never split up the vectors implies $\Phi(\mathbf{X}_t) \to 0$, and this holds with constant nonzero probability depending on just $n, d, \boldsymbol{\eta}$. By Equation (11) and Theorem 3.1, we conclude the result. □

4 WEAK POLARIZATION AND MARKOV CHAINS

In the previous section, we showed that the strong polarization observed by HJMR extends to nontrivial variants of geometric opinion dynamics. However, we caution that it is *not* generally true that just any geometric opinion model, even one that is continuous and P-invariant, will strongly polarize from an arbitrary configuration of starting opinions. As a trivial example, suppose that the updates are such that they simply apply a common orthogonal transformation to each vector at each time. This is clearly continuous and P-invariant, but obviously does not lead polarization except in very pathological examples.

While these trivial examples show that strong polarization does not necessarily hold in such models, it is natural to wonder if it is possible for other dynamics to satisfy other forms of polarization even if they do not satisfy strong polarization. In this section, we consider weaker forms of polarization than the strong form from Definition 2.1 (restated as part (1) in the definition below):

Definition 4.1. Let $\mathbf{X}_t := (\mathbf{X}_t^{(1)}, \ldots, \mathbf{X}_t^{(n)})$ be a discrete-time Markov chain as given by Equation (1).

(1) $\mathbf{X}_t$ **strongly polarizes** (from $\mathbf{X}_0$) if almost surely, $\rho(\mathbf{X}_t, P) \to 0$.

(2) $\mathbf{X}_t$ **weakly polarizes** (from $\mathbf{X}_0$) if, for any fixed $\epsilon > 0$, $\Pr(\rho(\mathbf{X}_t, P) \geq \epsilon) \overset{t \to \infty}{\to} 0$.

(3) $\mathbf{X}_t$ **weakly polarizes on average** (from $\mathbf{X}_0$) if, for any fixed $\epsilon > 0$,

$$\lim_{T \to \infty} \frac{\mathbb{E}\left[\sum_{t=1}^{T} I_{P_\epsilon}(\mathbf{X}_t)\right]}{T} = 1,$$

where I_{P_ϵ} is the indicator of the set $P_\epsilon \triangleq \{\mathbf{z} \in (\mathbb{S}^{d-1})^n : \rho(\mathbf{z}, P) \leq \epsilon\}$.

It is not difficult to see that these forms of convergence are listed in decreasing order of strength:

PROPOSITION 4.2. *Strong polarization implies weak polarization, which in turn implies weak polarization on average.*

PROOF. Suppose that $\mathbf{X}_t$ strongly polarizes. Then by definition, for any fixed $\epsilon > 0$, there exists an almost surely finite (but random) T such that for $t \geq T$, $\rho(\mathbf{X}_t, P) < \epsilon$. For any fixed $\delta > 0$, there exists $K = K(\delta)$ large enough such that $\Pr(T \geq K) < \delta$. It follows that for $t \geq K$,

$$\Pr(\rho(\mathbf{X}_t, P) \geq \epsilon) \leq \Pr(T \geq K) < \delta.$$

As $\delta > 0$ is arbitrary, it follows that the probability of being at least ϵ-far from P tends to zero. As $\epsilon > 0$ is arbitrary, this gives weak polarization.

For the second implication, fix $\delta > 0$ and let K be large enough so that for all $t \geq K$, $\Pr(\rho(\mathbf{X}_t, P) \geq \epsilon/n) \leq \delta/2$. Then for all $T \geq 2K/\delta$,

$$\frac{\mathbb{E}\left[\sum_{t=1}^{T} I_{P_\epsilon}(\mathbf{X}_t)\right]}{T} \geq \frac{\sum_{t=1}^{T} \Pr(\rho(\mathbf{X}_t, P) \leq \epsilon/n)}{T} \geq (1 - \delta/2) - K/T \geq 1 - \delta.$$

As this holds for all large enough T and δ was arbitrary, the claim follows. □

4.1 Polarization and Invariant Distributions

With these weaker forms of polarization, we now proceed to show their connection to the theory of invariant distributions in general Markov chains. To begin, we will actually consider the simplest case of $n = 1$ agent, where clearly all forms of polarization are trivial.

PROPOSITION 4.3. *Consider $n = 1$ agent updating opinions via dynamics that are continuous in the current state, i.e. the function $P_{\mathbb{S}^{d-1}}(\mathbf{x} + f_1(\mathbf{x}, \xi))$ is continuous in $\mathbf{x} \in \mathbb{S}^d$ for every fixed realization of ξ. Then there exists an invariant distribution on $\mathbb{S}^{d-1}$ for these dynamics.*

Moreover, assume

(1) *There exists some point* $\mathbf{y} \in \mathbb{S}^{d-1}$ *such that for any open neighborhood* U *of* $\mathbf{y}$, *and any choice of starting point* $x \in \mathbb{S}^{d-1}$ *for the dynamics, there exists some* $t = t(\mathbf{x}, U)$ *such that* $\Pr_{\mathbf{x}}(\mathbf{X}_t^{(1)} \in U) > 0$.

(2) *The dynamics admit a continuous density with respect to the standard Haar measure on* $\mathbb{S}^{d-1}$, *i.e. distribution of* $\mathbf{X}_1^{(1)}$ *given* $\mathbf{X}_0^{(1)} = \mathbf{x}$ *has a density with respect to standard Haar measure.*

Then the above invariant distribution is unique.

PROOF. From our continuity assumption, it follows that the dynamics are *Feller*, i.e. for any continuous function $h : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$, the map from $\mathbf{x}$ to $\mathbb{E}[h(\mathbf{x}')]$ where $\mathbf{x}'$ takes one step of the dynamics, remains continuous (Definition 2.36 of [16]). Because $\mathbb{S}^{d-1}$ is compact and the dynamics are Feller, it follows that there exists at least one invariant measure (Corollary 4.18 of [16]).

For the second part, the first additional assumption implies there exists an *accessible* point $\mathbf{y} \in \mathbb{S}^{d-1}$, while the second implies that the dynamics satisfy the *strong Feller* property (Definition 2.5 of [17]), i.e. the transition operator maps bounded measurable functions to continuous functions. From Corollary 2.7 of [17], this implies that the dynamics have a unique invariant distribution. □

We now consider invariant distributions on *the joint opinion vector* and then compare to the above results.

PROPOSITION 4.4. *Consider* $n \geq 1$ *agents updating opinions, and suppose the dynamics are jointly continuous for every fixed realization of* ξ. *Then there exists an invariant distribution for the vector* $(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)})$ *under these dynamics. Moreover, if the dynamics preserve* $\sigma(D)$ *for each* $\sigma \in \{-1, 1\}^n$, *there exists such a distribution supported in* $\boldsymbol{\sigma}(D)$ *for every* $\boldsymbol{\sigma} \in \{-1, 1\}^n$.

If the dynamics are oblivious and satisfy the stronger assumptions of the previous proposition componentwise, then the ith marginal distribution of every such invariant distribution is unique.

PROOF. The first part is identical to the proof of Proposition 4.3 using the Feller property. The second claim holds from the assumed invariance of $\boldsymbol{\sigma}(D)$ and noting that the same argument holds when restricting to this invariant, compact set. The last claim follows from the obliviousness, as then the restriction of any invariant distribution to any component must yield an invariant distribution of the $n = 1$ dynamics. This distribution is unique from Proposition 4.3. □

REMARK 4.1. *Note that by Proposition 4.3, it is* not *true that there exists a unique invariant distribution for* $n > 1$ *agents for most natural dynamics. The failure of the uniqueness of the invariant distribution arises because the joint dynamics typically will not have either accessible points nor satisfy the strong Feller property, as the set of possible updated values forms a lower-dimensional subset of* $\prod_{i=1}^n \mathbb{S}^{d-1}$ *and so cannot admit a density.*

With these results in hand, we can finally turn to the connection with polarization, which *a priori* is a statement about whether dynamics starting at arbitrary points converge to the polarized set in some suitable sense. The connection is the following:

THEOREM 4.5. *Suppose that every invariant distribution of the joint dynamics with continuous updates has support contained in* P. *Then* $\mathbf{X}_t$ *weakly polarizes on average from any choice of starting vectors.*

PROOF. The proof of this result is implicit in the proof of the Krylov-Bogolubov theorem (Theorem 4.17 of [16]). Fix any choice $\mathbf{X}_0$ of starting vectors for the joint dynamics. Define the family of probability measures for $T = 1, 2, \dots$ by $Q_T(A) := \frac{1}{T} \sum_{t=0}^{T-1} \Pr(\mathbf{X}_t \in A)$. Then because the state space is compact, Prokhorov's Theorem (Theorem 4.15 of [16]) asserts that this sequence of measures contains a weakly convergent subsequence Q_{T_k} for $k = 1, \dots$, with weak limit we denote μ^*;

moreover, from the proof of the Krylov-Bogolubov theorem and the fact that we assumed the dynamics are Feller, μ^* is invariant.

Now, by our assumption, every invariant measure for these dyamics is concentrated on P. For any fixed $\epsilon > 0$, let P_ϵ be as defined above. Because μ^* is supported on P by assumption, $\mu^*(\partial P_\epsilon) = 0$, where ∂A denotes the boundary of a set A. Recall that weak convergence of measures is equivalent to convergence on all continuity sets A that satisfy $\mu^*(\partial A) = 0$. It follows immediately that

$$Q_{T_k}(P_\epsilon) = \frac{\sum_{t=0}^{T_k-1} \Pr(\mathbf{X}_t \in P_\epsilon)}{T_k} \to \mu^*(P_\epsilon) = 1.$$

Moreover, this same argument holds for *any* subsequence of $\{Q_T\}_{T=0}^\infty$, possibly with a different invariant measure which is nonetheless supported on P by assumption. In particular, for any subsequence of the scalar sequence $\{Q_T(P_\epsilon)\}_{T=1}^\infty$, there exists a further subsequence which converges to 1. From standard analysis, any sequence such that every subsequence contains a further subsequence that converges to a fixed number c must itself converge to c. Therefore, it follows that $Q_T(P_\epsilon) \to 1$, which by linearity of expectation is the statement of weak polarization on average. □

The converse always holds even when the dynamics are not continuous (though we do need to assume that there exists *some* invariant distribution).

THEOREM 4.6. *Suppose that under the geometric opinion dynamics, there exists an invariant distribution μ^* that puts nontrivial mass on the set of nonpolarized vectors, i.e. $\mu^*(P^c) > 0$. Then the opinion dynamics do not weakly converge on average. In particular, if the dynamics are continuous, then the dynamics weakly polarize on average if and only if every invariant distribution is supported on P.*

PROOF. The assumption and taking limits for $\epsilon \to 0$ with continuity of measure implies that there exists some $\epsilon > 0$ such that $\mu^*(P_\epsilon^c) > 0$. By invariance, this means that for every $t \geq 0$, $\Pr_{\mathbf{X}_0 \sim \mu^*}(\mathbf{X}_t \in P_\epsilon) = \mu^*(P_\epsilon^c) > 0$, which trivially implies that for every $T \geq 0$,

$$\mathbb{E}_{\mathbf{X}_0 \sim \mu^*}\left[\mathbb{E}\left[\frac{\sum_{t=0}^{T-1} I_{P_\epsilon}}{T}\right]\right] = 1 - \mu^*(P_\epsilon^c),$$

where the inner expectation is over the sequence of random vectors $\{\xi_t\}_{t=1}^T$. By Fatou's lemma,

$$\mathbb{E}_{\mathbf{X}_0 \sim \mu^*}\left[\liminf_{T \to \infty} \mathbb{E}\left[\frac{\sum_{t=0}^{T-1} I_{P_\epsilon}}{T}\right]\right] \leq \liminf_{T \to \infty} \mathbb{E}_{\mathbf{X}_0 \sim \mu^*}\left[\mathbb{E}\left[\frac{\sum_{t=0}^{T-1} I_{P_\epsilon}}{T}\right]\right] = 1 - \mu^*(P_\epsilon^c).$$

This implies that there exists some fixed $\mathbf{X}_0 \in \prod_{i=1}^n \mathbb{S}^{d-1}$ such that

$$\liminf_{T \to \infty} \mathbb{E}\left[\frac{\sum_{t=0}^{T-1} I_{P_\epsilon}}{T}\right] \leq 1 - \mu^*(P_\epsilon^c) < 1,$$

when the process is started at $\mathbf{X}_0$. By definition, it follows that the process does not weakly polarize on average with this choice of starting vector. The last claim then follows from combining the above with Theorem 4.5. □

As an application, HJMR proved that for $d = 2$ and ξ_t is uniform on the unit circle, every finite set of vectors strongly polarizes. By Proposition 4.2, this implies weak polarization on average. Because their dynamics are oblivious, admit a density for each single agent system, and every point is accessible, the unique invariant measure for the dynamics of a single agent is clearly uniform. Theorem 4.6 shows that every invariant measure on the *joint* system must be supported on P.

To summarize, under very natural conditions, these geometric dynamics have invariant distributions in P, essentially by P-invariance of the dynamics. In the oblivious case, the projections of *any*

invariant distribution on each component will be unique under natural regularity assumptions and the only ambiguity is in how these distributions are coupled together. If one can show that these P-invariant measures are the *only* such measures, or contrapositively that no invariant measure can put nontrivial mass on the complement of P, then the previous result immediately supplies weak polarization on average.

4.2 HJMR Model and Weak Polarization

Given the results in the previous sections, it is perhaps tempting to believe that under some necessary form of irreducibility, strong polarization will arise. In this section, we exhibit a model where *weak polarization* on restricted subsets of initial configurations holds, but strong polarization provably does not. We do so in a very simple version of the original HJMR model by showing how the dynamics in this case relate to basic properties of simple random walks in $\mathbb{Z}$ and more generally, infinite balls-in-bins processes.

In this section, we will consider the HJMR dynamics where the update vector is drawn uniformly from a complete orthonormal set of basis vectors in $\mathbb{R}^d$; because the HJMR dynamics commute with orthogonal transformations, we will assume that these are the standard basis vectors. Towards showing that weak, but not strong, polarization holds, we need the following simple calculation showing that the random updates in this process are *commutative*, greatly simplifying updating.

LEMMA 4.7. *Let* $\mathbf{v} \in \mathbb{S}^{d-1}$ *be an arbitrary starting vector. Then for any sequence* $\xi_1, \ldots, \xi_T$ *of update vectors drawn from* $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$, *the vector* $\mathbf{z}$ *obtained after* T *steps of the dynamics given by Equation* (2) *and starting at* $\mathbf{v}$ *satisfies*

$$\mathbf{z} \propto \left((1 + \eta)^{N_T^{(1)}} v_1, \ldots, (1 + \eta)^{N_T^{(d)}} v_d) \right),$$

where $N_T^{(i)}$ *is the number of times* $\xi_t = \mathbf{e}_i$ *up to time* T.

In particular, with orthonormal update vectors, the vector obtained after T *steps of the dynamics depends only on the multiset* $\{\xi_1, \ldots, \xi_T\}$ *and not on the order they arrive.*

PROOF. Observe that the functional form of the update rule in Equation (2) is homogeneous in the starting vector in that scaling the starting vector leads to the same update. As a consequence, one can perform the dynamics as stated by renormalizing after every step, or by applying the update rule without renormalizing until the end.

As a result of this observation, it suffices to show that for any vector $\mathbf{v} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, the effect of applying the update vector $\xi = \mathbf{e}_i$ is proportional to $(v_1, \ldots, (1+\eta)v_i, \ldots, v_d)$. This is sufficient as then we may iterate these updates without renormalizing until the only the end of the T unnormalized updates to deduce commutativity. This is now immediate to see from Equation (2). □

Notice that the vector $(N_t^{(1)}, \ldots, N_t^{(d)})$ is precisely a balls-in-bins process with d bins and t balls. The identity of Lemma 4.7 shows that to understand the behavior of the process with orthonormal update vectors, we need to understand the sample path properties of an infinite balls-in-bins process. This is done in the following lemma:

LEMMA 4.8. *Consider an infinite balls-in-bins process with* d *labeled bins* $\{1, \ldots, d\}$. *Let* $N_t^{(i)}$ *denote the number of balls in bin* i *at time* t. *Then almost surely, the following holds:*

(1) *For each fixed* $i \in [d]$, $N_t^{(i)} = \max_{j \in [d]} N_t^{(j)}$ *for infinitely many* t.
(2) *For each fixed* $i, j \in [d]$, $N_t^{(i)} = N_t^{(j)} = \max_{k \in [d]} N_t^{(k)}$ *for infinitely many* t.

PROOF. Recall from the Hewitt-Savage zero-one law that any event depending on a sequence of i.i.d. random variables that is invariant under finite permutations has probability 0 or 1 (Theorem

2.5.4 of [9]). The first listed event evidently has this property. Clearly by the pigeonhole principle, there surely exists a (random) index $i^* \in [d]$ that is infinitely often the largest, and so by symmetry, the probability that any fixed $i \in [d]$ has this property is at least $1/d$ by a union bound. The Hewitt-Savage zero-one law then implies that the probability is one.

For the second claim, note that the previous claim and the fact that each $N_t^{(i)}$ is nondecreasing in time and changes by at most one in each round implies that infinitely often, there must be at least two indices of bins that are at least as large as the rest (if not, then with positive probability a single index is the unique largest element for all but finitely many times). Again by the pigeonhole principle and symmetry, each fixed pair (i, j) has probability at least $\binom{d}{2}^{-1}$ of having this property, hence by the Hewitt-Savage zero-one law has probability one. □

LEMMA 4.9. *Consider an infinite balls-in-bins process with d labeled bins $\{1, \ldots, d\}$. Let $N_t^{(i)}$ denote the number of balls in bin i at time t. Then for any fixed constant $K \in \mathbb{N}$, the probability that there exists two bins $i \neq j$ such that $|N_t^{(i)} - N_t^{(j)}| \leq K$ is $O(1/\sqrt{t})$, where the implicit constant depends on K, d, but not t.*

PROOF. Observe that the distribution of $N_t^{(i)} - N_t^{(j)}$ for any fixed $i \neq j$ is equal to that of a lazy, simple random walk on $\mathbb{Z}$ that updates with probability $2/d$. It is well-known (for instance, by the central limit theorem or Littlewood-Offord lemma) that for non-lazy simple random walks on $\mathbb{Z}$, the probability that the walk has magnitude at most some fixed constant K at time t is $O\left(\frac{1}{\sqrt{t}}\right)$, where the implicit constant depends on K. To account for laziness, note that with all but exponentially small probability, the number of updates to the i or jth bin at time t will exceed, say, t/d. We can then apply the result for non-lazy simple random walks and absorb the error term, just changing the implicit constant to depend on d as well to account for the smaller number of steps. We may then apply a union bound over such pairs. □

We can now turn to a full analysis of the polarization behavior of this system.

THEOREM 4.10. *Consider the HJMR dynamics in $\mathbb{S}^{d-1}$ given by Equation (2), where ξ_t is drawn i.i.d. and uniformly from $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$. Let $\mathbf{X}_0 = (\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)}) \in (\mathbb{S}^{d-1})^2$ be such that the two vectors have equal supports but are not equal up to sign. Then:*

(1) *The vector $\mathbf{X}_t$ weakly polarizes.*
(2) *With probability 1, $\|\mathbf{X}_t^{(1)} - \mathbf{X}_t^{(2)}\|_2 \not\to 0, 2$. In particular, no nontrivial pair of starting vectors polarizes with nonzero probability.*

REMARK 4.2. *If the distribution of random updates is not uniform, but instead has strictly unequal probabilities associated to each basis vector, then it is easy to amend the argument to show that now* strong *polarization holds between any two starting vectors with the same support, simply because both vectors will get concentrated about the basis vector in their support with the largest probability.*

Moreover, note that there do *exist invariant distributions in this system, considered as $\prod_{i=1}^n \mathbb{S}^{d-1}$, that are not supported on P. However, this does not contradict Theorem 4.6 because Theorem 4.10 only applies to vectors with equal supports, a full measure set.*

PROOF OF THEOREM 4.10. Let $\mathbf{X}_0 = (\mathbf{X}_0^{(1)}, \mathbf{X}_0^{(2)}) \in (\mathbb{S}^{d-1})^2$ be as stated. Recall that to show weak polarization, we must show that for any fixed $\epsilon > 0$,

$$\Pr\left(\|\mathbf{X}_t^{(1)} - \mathbf{X}_t^{(2)}\|_2 \wedge \|\mathbf{X}_t^{(1)} + \mathbf{X}_t^{(2)}\|_2 > \epsilon\right) = o(1).$$

Recall that from Lemma 4.7, the (unnormalized) ith component of each vector is amplified multiplicatively by a factor of $(1 + \eta)^{N_t^{(i)}}$ before renormalization. From this, it is easy to see that for any

fixed $\mathbf{X}_0$ and $\epsilon > 0$, there exists some finite $K = K(\mathbf{X}_0, \epsilon) \in \mathbb{N}$ such that if $N_t^{(i)} \geq N_t^{(j)} + K$ for some $i \in [d]$ and all $j \neq i$, then both components of $\mathbf{X}_t$ are individually $\epsilon/2$-close to $\pm \mathbf{e}_i$. By the triangle inequality, it would then follow that $\mathbf{X}_t$ satisfies $\|\mathbf{X}_t^{(1)} - \mathbf{X}_t^{(2)}\|_2 \wedge \|\mathbf{X}_t^{(1)} + \mathbf{X}_t^{(2)}\|_2 \leq \epsilon$. Applying Lemma 4.9 with this value of K, we see that the probability of this event is $1 - O\left(\frac{1}{\sqrt{t}}\right)$, where the constant depends only on K, d and hence tends to 1 with t as needed. As $\epsilon > 0$ was arbitrary, this proves the first claim.

For the second, we rely on the other properties we have already shown of the infinite balls-in-bins process. By assumption, as $\mathbf{X}_0^{(1)}$ and $\mathbf{X}_0^{(2)}$ are not equal up to sign, there exists a pair of indices such that the vectors restricted to these indices are not equal up to a constant. Without loss of generality, we assume that these indices are $1, 2$. By Lemma 4.8, there almost surely exist infinitely many times t such that $N_t^{(1)} = N_t^{(2)} = \max_{k \in [d]} N_t^{(k)}$. By the unnormalized form of the directions at each such time t given by Lemma 4.7, it follows that at each such time, the first and second components of each vector remain proportional to their value at the beginning of this process but are now scaled by a factor that is at least 1 after normalization. In particular, at each such time t where this holds, the distance between $\mathbf{X}_t^{(1)}$ and $\mathbf{X}_t^{(2)}$ exceeds some strictly positive constant. Because this occurs infinitely often almost surely, the claim follows. □

5 CONCLUSIONS

In this paper, we have attempted to more closely study the polarization phenomenon that arises in geometric models as considered by HJMR. We introduce different forms of polarization and consider how these relate to more general Markov principles. Finally, we consider these various forms of polarization in concrete models to extend the class of natural dynamics that exhibit this property. We hope that some of the techniques we introduce here to complement the original results of HJMR prove useful in understanding the probabilistic properties of these models.

However, the most general version of this question remains open: are there very general, *easily verified*, conditions on the update rule that ensures that some form of polarization arises (perhaps under some sort of irreducibility assumption)? Are there other interesting models of geometric opinion formation under which strong polarization holds? Understanding which sorts of dynamics lead to polarization, and also a better sense of why, may shed significant insight into the theoretical underpinnings of this phenomenon.

REFERENCES

[1] Jerry Anunrojwang, Ozan Candogan, and Nicole Immorlica. 2020. Social Learning Under Platform Influence: Extreme Consensus and Persistent Disagreement. *SSRN* (2020).

[2] Arnab Bhattacharyya, Mark Braverman, Bernard Chazelle, and Huy L. Nguyen. 2013. On the convergence of the Hegselmann-Krause system. In *Innovations in Theoretical Computer Science, ITCS '13, Berkeley, CA, USA, January 9-12, 2013*, Robert D. Kleinberg (Ed.). ACM, 61–66. https://doi.org/10.1145/2422436.2422446

[3] Oleg Butkovsky. 2014. Subgeometric rates of convergence of Markov processes in the Wasserstein metric. *Annals of Applied Probability* 24, 2 (2014), 526–552.

[4] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of modern physics* 81, 2 (2009), 591.

[5] Uthsav Chitra and Christopher Musco. 2020. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 115–123. https://doi.org/10.1145/3336191.3371825

[6] Pranav Dandekar, Ashish Goel, and David T. Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proc. Natl. Acad. Sci. USA* 110, 15 (2013), 5791–5796. https://doi.org/10.1073/pnas.1217220110

[7] Morris H DeGroot. 1974. Reaching a consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121.

[8] Daniel DellaPosta, Yongren Shi, and Michael Macy. 2015. Why do liberals drink lattes? *Amer. J. Sociology* 120, 5 (2015), 1473–1511.

[9] Rick Durrett. 2019. *Probability: theory and examples.* Vol. 49. Cambridge university press.

[10] Ronald J Fisher. 2016. Towards a social-psychological model of intergroup conflict. In *Ronald J. Fisher: A North American Pioneer in Interactive Conflict Resolution.* Springer, 73–86.

[11] Noah E Friedkin and Eugene C Johnsen. 1990. Social influence and opinions. *J. Math. Soc.* 15, 3-4 (1990), 193–206.

[12] Jason Gaitonde, Jon M. Kleinberg, and Éva Tardos. 2020. Adversarial Perturbations of Opinion Dynamics in Networks. In *EC '20: The 21st ACM Conference on Economics and Computation, Virtual Event, Hungary, July 13-17, 2020*, Péter Biró, Jason D. Hartline, Michael Ostrovsky, and Ariel D. Procaccia (Eds.). ACM, 471–472. https://doi.org/10.1145/3391403.3399490

[13] Matthew Gentzkow, Jesse M Shapiro, Matt Taddy, et al. 2016. *Measuring polarization in high-dimensional data: Method and application to congressional speech.* National Bureau of Economic Research.

[14] Aristides Gionis, Evimaria Terzi, and Panayiotis Tsaparas. 2013. Opinion Maximization in Social Networks. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA.* SIAM, 387–395. https://doi.org/10.1137/1.9781611972832.43

[15] Benjamin Golub and Matthew O Jackson. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics* 2, 1 (2010), 112–49.

[16] Martin Hairer. 2006. Ergodic properties of Markov processes. (2006).

[17] Martin Hairer. 2010. Convergence of Markov processes. (2010).

[18] Rainer Hegselmann and Ulrich Krause. 2002. Opinion Dyanamics and Bounded Confidence Models, Analysis, and Simulation. *Journal of Artifical Societies and Social Simulation (JASSS) vol* 5, 3 (2002).

[19] Jan Hązła, Yan Jin, Elchanan Mossel, and Govind Ramnarayan. 2019. A Geometric Model of Opinion Polarization. *CoRR* abs/1910.05274 (2019). arXiv:1910.05274 http://arxiv.org/abs/1910.05274

[20] Michael J McCallion. 2007. In-Groups and Out-Groups. *The Blackwell encyclopedia of sociology* (2007).

[21] Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. 2018. Minimizing Polarization and Disagreement in Social Networks. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 369–378. https://doi.org/10.1145/3178876.3186103

[22] Muzafer Sherif, O.J. Harvey, Jack B. White, William R. Hood, and Carolyn W. Sherif. 1961. *Intergroup Conflict and Cooperation. The Robbers Cave Experiment [1954].* University of Oklahoma Book Exchange, Norman.