Sequential Learning under Informational Ambiguity*

Jaden Yang Chen†

First version: Nov 2019 This version: Jan 2022

[Click here for the newest version]

Abstract

This paper studies a sequential learning problem where individuals are ambiguous about other people's data-generating processes. This paper finds that the occurrence of an information cascade can be interpreted as a result of ambiguity instead of details of the true data-generating process as suggested by the literature. When there is sufficient ambiguity, for *all* possible data-generating processes, an information cascade occurs almost surely. This paper further shows that many standard results may even represent a knife-edge case with respect to ambiguity. An arbitrarily small degree of ambiguity can produce a cascade when signals are bounded and destroy complete learning when signals are unbounded.

JEL Classification: D81, D83, C72
Keywords: Social learning, model uncertainty, ambiguity, information cascades, herding

*The paper was first circulated as "information cascades and ambiguity" in November 2019. I am indebted to my committee members, David Easley (chairperson), Larry Blume and Tommaso Denti, for their continuous guidance and encouragement. I thank Drew Fudenberg, Qingmin Liu, Omer Tamuz, and Peter Sorensen for comments and feedback. I received valuable feedback from seminar participants at Cornell, Oxford-Nuffield, Midwest Theory Conference 2019, Economics Graduate Student Conference 2020 and World Congress 2020. I thank the editor and four anonymous referees for excellent suggestions. All errors are mine.

†Department of Economics, Cornell University, Ithaca, NY 14850, USA; E-mail address: yc2325@cornell.edu

1 Introduction

1.1 Overview of the Results

In many economic problems, individuals need to make decisions according to some information. One standard assumption is that individuals entertain a specific data-generating process (or model) and interpret information according to it. However, in many interesting situations, individuals are unable to pin down a specific model and may face ***model uncertainty***. One good example is social learning. In social learning, individuals observe information from others, but the observations can be both imperfect (e.g., individuals only observe actions but not signals) and scarce (e.g., individuals may not repeatedly observe from the same individual), so individuals usually can not specify a unique model to interpret all information at hand. Moreover, due to the lack of prior knowledge, individuals may even be unable to assign a prior over models. In this paper, I introduce model uncertainty into a classical problem of social learning—the sequential learning model (SLM, henceforth), where every individual takes an action sequentially and can observe a private signal and previous actions. This paper finds that the social learning outcomes under model uncertainty and under model certainty have some interesting differences.

Under model certainty, the learning outcome depends on the details of the true data-generating process and its perception. In the pioneering work of Banerjee (1992) and Bikhchandani, Hirshleifer and Welch (1992) (BHW, henceforth) , individuals receive i.i.d. signals from a finite signal space according to a commonly known distribution. One important result is that an ***information cascade*** will arise with probability 1. That is, after some point, individuals will ignore their own signals and follow others even if the action is sub-optimal. Later findings suggest that the occurrence of a cascade relies on the finiteness of signals. Smith and Sørensen (2000) showed that: (i) when signals are unbounded, the society will settle on the correct action in the limit, so an incorrect herding can not occur; (ii) even if signals are bounded, the occurrence of an information cascade is not guaranteed for continuous signals; they showed a weaker result that herding occurs with probability 1, that is, individuals end up taking the same action but any individual could break the herd if he or she received a different signal. Herrera and Hörner (2012) showed that when the data-generating processes satisfy the increasing hazard ratio property (IHRP), an information cascade will not take place. Furthermore, if we allow individuals to perceive a misspecified model, the learning outcome depends more intricately on the details of the true data-generating process and its relation with the perceived data-generating process.[1]

One possible challenge of previous results is that it is empirically difficult to test the true data-generating process or model perceptions in most cases, so the it remains vague which result to expect in the social learning. [2]In this paper, I re-examine the SLM under the assumption that

[1]For example, we may have complete learning, information cascades or even action non-convergence as implied by Corollary 4. Similar results are also noted by recent works in misspecified learning, e.g., Bohren (2016), Bohren and Hauser (2021) and Frick, Iijima and Ishii (2020b), but with setups different from the standard SLM.

[2]For example, in terms of the boundedness condition, "whether bounded or unbounded beliefs provide a better approximation to reality is partly an interpretational and partly an empirical question" (Acemoglu et al. 2011), so it

individuals are uncertain about other people's data-generating processes. The paper finds that under sufficient model uncertainty, there exists a unique learning outcome—information cascade, with a positive probability of an incorrect cascade—for *all* possible true data-generating processes. Perhaps more surprisingly, this paper also shows that previous results featuring the absence of a cascade can only represent a knife-edge case from the perspective of model uncertainty.

The model setup is introduced in Section 2 and summarized as follows. This paper adopts the stylized SLM framework where there are binary states and actions. Individuals take actions sequentially to match the unknown state of the world. Every individual can observe all previous actions as well as a private signal. The only deviation from the standard SLM is that individuals are ***ambiguous*** about their predecessors' data-generating processes by perceiving a set of feasible data-generating processes, $\mathcal{F}_0$. The size of $\mathcal{F}_0$ intuitively measures the degree of ambiguity, where model certainty corresponds to the case where $\mathcal{F}_0$ is a singleton. The benchmark model assumes that individuals follow the max-min EU model to make choices. It turns out that for all large $\mathcal{F}_0$s, an information cascade occurs almost surely even when the true data-generating processes imply no cascades in the standard case.

At first glance, a large $\mathcal{F}_0$ can include models with different implications, so it is not straightforward which result to expect. The paper's result relies on the finding that the models encouraging an information cascade and the models discouraging it are *asymmetric*. For an individual who is called upon to take an action, the worst possibility if she were to break a herd happens when the herding individuals have highly precise information, whereas the worst possibility if she followed the herd happens when these predecessors have very imprecise information. The asymmetry in the worst cases implies that as $\mathcal{F}_0$ becomes larger, the cascading force can increase consistently, but the non-cascading force is always restricted. As a consequence, under sufficient ambiguity, the cascading force dominates, so we have an information cascade. Below is a simple example illustrating this idea.

Example 1. The state space $\Theta = \{0, 1\}$, and the prior is $\pi_0 = (1/3, 2/3)$. Every individual i takes action $a_i \in \{0, 1\}$ sequentially and receives a signal $s_i \in \{h, l\}$. The DGP $g_i(s|\theta)$ satisfies

$$\frac{g_i(h|1)}{g_i(l|1)} = \frac{g_i(l|0)}{g_i(h|0)} = \gamma_i \in (1, \infty) = \Gamma,$$

where $\gamma_i \overset{i.i.d.}{\sim} h \in \Delta(\Gamma)$ describes the signal precision. Individuals know their own precision but is ambiguous about h, so they are ambiguous about others' precision. Let's consider an extreme case, where $\mathcal{F}_0 = \Delta(\Gamma)$, so any distribution is possible for h. Suppose that the first individual (he) chose action 1. Denote by $V_2(a)$ the minimum EU of the second individual (she) if she takes action

remains vague when complete learning can be achieved and when it cannot. As for the more complicated properties, e.g., IHRP, it is harder to provide an intuitive explanation on why they can induce different results in terms of information cascades.

a. We have

$$V_2(1) = \begin{cases} \gamma_2/(\gamma_2+2) & s_2 = h \\ 2/(\gamma_2+2) & s_2 = l \end{cases} \text{ and } V_2(0) = 0.$$

If individual 2 followed the first individual and chose action 1, the minimum EU is obtained when she believes that individual 1 can only receive uninformative signals. In this case, individual 1's action contains no information which gives $V_2(1)$. In contrast, if individual 2 chose a different action, action 0, the the minimum EU is obtained when she believes that individual 1 can only receive the perfectly revealing signal, so acting against him yields a utility of 0, which gives $V_2(0)$.[3] As $V_2(1) > V_2(0)$, individual 2 will follow individual 1 for all possible private signals, so an information cascade occurs.

Notice that the occurrence of an action-1 cascade does not rely on the specific properties of h nor the true state, so an information cascade occurs for all $h \in \Delta(\Gamma)$, and an incorrect cascade occurs with a strictly positive probability.

The example shows that an information cascade arises under extreme ambiguity, but this paper further notes that to achieve a cascade, the condition can be much weaker. In many interesting situations, a cascade arises only when individuals face a *slight degree of ambiguity*. Section 5 characterizes conditions that ensure the almost sure occurrence of an information cascade when signals are **bounded**. Theorem 2 provides two sufficient conditions that guarantee a cascade. Intuitively, as long as individuals find it possible that other individuals may have a highly information data-generating process, an information cascade will occur almost surely, regardless of the details of the true data-generating process and other data-generating processes under consideration. The intuition is similar to that in Example 1. We first note that the perception of an informative data-generating process encourages individuals to follow a herd hence creates a cascading force. Due to the asymmetry, if the informativeness is sufficiently high, the presence of any other model is inadequately to offset the cascading force, so an information cascade always occur irrespective of what other models individuals may consider. Interestingly, the conditions proposed by Theorem 2 are very easy to hold in many situations. Suppose that the true model is $\overline{F}$, and there is no cascade when individuals correctly perceive $\overline{F}$. If individuals are ambiguous and consider all F such that $||F - \overline{F}|| \leq \varepsilon$, where $||.||$ is some metric (e.g., sup-norm metric), then an information cascade occurs almost surely for all $\varepsilon > 0$. It suggests that the standard non-cascade results are not robust, and the statistical properties relevant for a cascade (i.e., IHRP) only matter in knife-edge cases from the perspective of ambiguity.

Section 3 shows that similar results also exist for **unbounded** signals. For unbounded signals, an information cascade is more difficult to happen by definition, but this paper establishes that herding occurs almost surely, where an **incorrect herding** occurs with a strictly positive probability, even with arbitrarily small ambiguity. The result implies that the complete learning result in Smith and Sørensen (2000) can be non-robust. Theorem 3 provides conditions under which an in-

[3] More rigorously, the minimum EU is actually the infimum EU, since $\gamma \in (1, \infty)$.

correct herding occurs. The idea is similar to the bounded signal case—if individuals perceive some model that is adequately informative, the cascading force is so strong that it can not be outweighed by any other models, hence an incorrect herding can occur. In some interesting settings, Theorem 3 can be easily satisfied so that complete learning only represents a knife-edge case. In Section 7, I further provide a necessary and sufficient condition of complete learning under ambiguity for an important class of models—models with power tails (Theorem 4). To achieve complete learning, we must impose restrictions on $\mathcal{F}_0$ from both directions—$\mathcal{F}_0$ cannot be overly informative or overly uninformative, since the former will encourage an incorrect herding whereas the latter can lead to action non-convergence.

Section 8 extends the discussion to **general ambiguity preferences**. Under the max-min model, individuals are ambiguity-averse and are extremely ambiguity sensitive in the sense that decisions are only affected by the worst outcomes. Section 8 focuses on two common alternative models—the α-max-min model and the smooth ambiguity model—and has the following findings. First, a cascade does not require individuals to be ambiguity-averse, and it can also occur with ambiguity-seeking individuals. In the discussion of the α-max-min model, I show that an information cascade occurs for all $\alpha \in [0, 1]$, where $\alpha = 0$ means the max-max model and $\alpha = 1$ means the max-min EU model. Second, a cascade does not rely on the extreme ambiguity sensitivity as in the max-min EU model. As long as individuals are adequately ambiguity sensitive, an information cascade can occur. An example with smooth ambiguity preference shows that an information cascade arises when the curvature of the second-order utility function (i.e., measures the ambiguity sensitivity) is sufficiently large. The discussion implies that main results under the max-min EU model hold for general ambiguity preferences.

Techniques. The analysis under model uncertainty has the following challenges. First, individuals hold a set of posteriors, so we can not keep track of the posterior likelihood ratio as in the literature. To facilitate the analysis, this paper notices that there is a simple statistic—the average likelihood ratio—that captures all relevant information. Employing this fact, this paper is able to analyze the learning process by simply keeping track of this statistic.

Second, under model uncertainty, posteriors no longer exhibit the martingale property, which makes the analysis more difficult especially when signals are unbounded. This challenge is also present in the misspecified learning, where a common approach is to first analyze the local stability of each state and then to extend it globally. Following a similar idea, I re-define the notion of local stability using the average likelihood ratio. One challenge here is that standard techniques to analyze local stability are not applicable to the SLM. This paper adopts an alternative approach—**infinite series approach**—to facilitate the discussion. The key idea is that a state θ is locally stable (if) and only if the probability that all individuals take action θ is (uniformly) strictly positive when priors are near δ_θ. Besides, it can be further shown that the probability is (uniformly) strictly positive if and only if some infinite series is convergent. As a result, the discussion on local stability is transferred to a simpler problem of determining the convergence of the series. This approach is applied in Theorems 3 and 4. The relevant literature is discussed below.

1.2 Related Literature

This paper is among the few papers that study social learning under ambiguity, especially under model uncertainty. The most relevant paper is Ford, Kelsey and Pang (2013) which investigated a sequential trading model where traders face ambiguity and have neo-additive capacity EU preference.[4] They found that in the presence of ambiguity, informed traders can exhibit herding behavior, which is consistent with this paper in some sense. However, their setup and mechanism are different from this paper. First, in their paper, decisions are made in a specific market structure, where trading decisions are jointly determined by beliefs and ask-bid prices, but in this paper, there is no market and decisions are only determined by the beliefs. Second, in their paper, ambiguity also leads to both herding and contrarian, whereas in this paper, ambiguity only produces herding. [5]

Under model uncertainty, individuals will inevitably perceive some incorrect models, so the paper is closely related to the literature on misspecified social learning. The setup is most similar to that of Bohren and Hauser (2021) (also Bohren 2016), in which they also investigated a sequential learning problem with binary states. One of their main results is that complete learning is *robust* with respect to small misspecifications, which stands in contrast to this paper's finding. The difference arises from their assumption that the society has a positive fraction of autarkic agents who only act according to their private signals. This assumption plays an important role in establishing the local stability of the true state since it leads to a strict Berk-Nash equilibrium (Esponda and Pouzo 2016). However, their framework does not nest the standard SLM, where a strict Berk-Nash equilibrium can not be obtained. This paper employs a different approach to analyze the local stability and establishes that complete learning in the SLM is *non-robust* in many cases. Frick, Iijima and Ishii (2020a,b) also established that complete learning is not robust but in very different settings. Specifically, Frick, Iijima and Ishii (2020a) considered a social learning problem (not sequentially) where the state space is continuous and individuals with different preferences randomly meet with each other. Frick, Iijima and Ishii (2020b) proposed a local martingale-based approach to analyze misspecified learning and showed the fragility of sequential learning in an environment. Differently, their approach relies on the assumption is that signals are bounded and the fragility of the sequential learning relies on the assumption that risk preference is adequately hetergeneous.

One common feature of Bohren and Hauser (2021) and Frick, Iijima and Ishii (2020b) is that they demanded a strict dominance relation to establish local stability, e.g., strict Berk-Nash equilibrium, strict p-dominance. However, under SLM, the strict dominance is not satisfied. Technically, this paper complements the literature by employing an infinite series approach to establish local

[4]There are other papers related to learning under model uncertainty but not in social learning. Examples include Acemoglu, Chernozhukov, and Yildiz (2016), Battigalli et al. (2015), Battigalli, Francetich, et al. (2019), Chen (2021), Epstein and Schneider (2007), Fryer, Harms, and Jackson (2017), Marinacci (2002,2015), Marrinaci and Massari (2019).

[5]More technically, the occurrence of herding and contrarian in Ford, Kelsey and Pang (2013) comes from that posteriors are always bounded away from 0 and 1 under the neo-additive capacity preference. In contrast, this paper mainly works with the max-min EU preference, and the occurrence of herding or cascade does not rely on posteriors being bounded away from certainty.

stability. The most similar literature is Rosenberg and Vieille (2019), where they also employed an infinite series to characterize learning efficiency. The difference is that in their setup, individuals perceive a correct model, so complete learning occurs with unbounded signals; in this paper, individuals perceive multiple models, and the infinite series is useful in establishing incomplete learning. Furthermore, applying this paper's arguments, we can show that the efficient learning in their paper can be non-robust with respect to ambiguity (see Remark 2).

Under ambiguity, individuals no longer learn in a Bayesian manner, so this paper also belongs to the literature on non-Bayesian social learning. There are papers studying sequential learning with boundedly rational agents, for example, Eyster and Rabin (2010), Guarino and Jehiel (2013) and Dasaratha and He (2020), where individuals follow some naive rule to aggregate information from predecessors. Non-Bayesian social learning is also studied in general network structures, for example, DeMarzo et al. (2003), Golub and Jackson (2010), Li and Tan (2018), Molavi et al. (2018), where individuals apply a rule of thumb when aggregating information from others.

2 The Model

There are two possible states of world, $\Theta = \{0, 1\}$. A countably infinite set of individuals $N = \{1, 2, ...\}$ act sequentially. Each individual makes a binary choice $a \in A = \{0, 1\}$ and can observe the choices taken by all predecessors. Individuals have identical utility functions, which have a payoff of 1 when actions match the actual state and a payoff of 0 otherwise. Without loss of generality, the true state $\theta^* = 0$ and is not known to individuals.

Signal Structure

Individuals share a full-support common prior π_0. For simplicity, I assume that $\pi_0(0) = \pi_0(1) = \frac{1}{2}$. Each individual i, will receive a signal $s_i \in \mathcal{S} \subset \mathbb{R}$. Signals are independently (but not necessarily identically) distributed according to $\left\{\overline{G}_1^\theta, \overline{G}_2^\theta, ...\right\}$, where $\overline{G}_i^\theta : \mathcal{S} \to [0, 1]$ denotes the cumulative distribution function of s_i when the actual state is θ. I refer to $\overline{G}_i = \left(\overline{G}_i^0, \overline{G}_i^1\right)$ as individual i's *data-generating process*. No signal perfectly reveals the state; therefore, the probability measures induced by $\overline{G}_i^0$ and $\overline{G}_i^1$ are mutually absolutely continuous. Following the convention, I introduce the normalized signal, $\lambda_i(s)$, where $\lambda_i(s) = \frac{d\overline{G}_i^1(s)}{d\overline{G}_i^0(s)}$ denotes the likelihood ratio induced by each signal. The distribution of the likelihood ratio λ_i is denoted by $\overline{F}_i^\theta$, so $\overline{F}_i^\theta$ must satisfy $\lambda = \frac{d\overline{F}_i^1(\lambda)}{d\overline{F}_i^0(\lambda)}$ almost everywhere, which means that receiving a signal λ leads to a likelihood ratio equal to λ. For the rest of this paper, I focus on the normalized signal, λ, and the normalized data-generating processes, $\overline{F}_i^\theta$. For simplicity, I assume that: (i) all signals are continuous, that is, $\overline{F}_i^\theta$ is continuous for all i and θ, and (ii) signals are symmetric in the sense that $\overline{F}_i^1(\lambda) = 1 - \overline{F}_i^0(1/\lambda)$ for all i and λ. All results can be extended to cases where signals are discontinuous and asymmetric. All individuals' data-generating processes are assumed to have a common support, $co(\text{supp}(F_i)) = \left[\frac{1}{\gamma}, \gamma\right] \subset [0, \infty]$, where $\gamma > 1$, meaning that signals are informative. Signals are *bounded* if $\gamma < \infty$ and signals are

unbounded if $\gamma = \infty$. Let $\mathbb{P}^*$ denote the true probability measure, that is, the measure induced by data-generating processes, $\left\{\overline{F}_1^{\theta}, \overline{F}_2^{\theta}, ...\right\}$, conditional on the true state θ^*.

Belief Structure

Individuals are assumed to be *ambiguous* about their predecessors' data-generating processes, but they know that signals are symmetric and independently distributed. As a consequence of the symmetry, every data-generating process, $F = \left(F^0, F^1\right)$, is uniquely determined by one coordinate, so this paper keeps track of F^1 when characterizing every individual's belief set. Denote by $\mathcal{F}$ the set of all possible F^1, both discrete and continuous, with support in $\left[\frac{1}{\gamma}, \gamma\right]$. To be more precise, $\mathcal{F}$ consists of the set of F^1s that constitute some symmetric data-generating process with support in $\left[\frac{1}{\gamma}, \gamma\right]$. Further denote by $\mathcal{F}_{ij}$ the set of individual j's data-generating processes considered possible by individual i, where $\mathcal{F}_{ij} \subset \mathcal{F}$ and $j \neq i$, and refer to $B_i \coloneqq \{\mathcal{F}_{ij} : j \neq i\}$ as individual i's *belief structure*. With this, I make the following assumptions throughout this paper.

Assumption 1. *[Homogeneous Belief] There exists some $\mathcal{F}_0 \subset \mathcal{F}$ such that $\mathcal{F}_{ij} = \mathcal{F}_0$ for all $i, j \in N$ with $j \neq i$.*

Assumption 2. *[Common Knowledge] Individuals' belief structures are common knowledge.*

Assumption 1 is that individuals have homogeneous belief structures. The homogeneity is reflected in the following two aspects. First, for any given individual i, this individual is identically ambiguous about every other individual's data-generating process. Second, all individuals have identical belief structures. The homogeneous belief assumption is a simplifying assumption and can be extended to the heterogeneous belief setting. Assumption 2 is a standard assumption, with which individuals are able to make iterated inferences based on other individuals' actions.[6] These two assumptions hold in situations where the set of data-generating processes are public information, but individuals lack the information to determine the specific data-generating processes. For simplicity, I would also use the notation $F \in \mathcal{F}_0$, and here $F = \left(F^0, F^1\right)$ is a data-generating process. It actually means that the second coordinate $F^1 \in \mathcal{F}_0$.

Belief-updating Process

Let $h_i = (a_1, ..., a_{i-1})$ be the history observed by individual i. Denote $I_i = \{\lambda_i, h_i\}$, where I_i represents the information available to individual i, which consists of her private signal λ_i and history h_i. Let $\mathcal{I}_i$ be the set of all possible information available to individual i, where $\mathcal{I}_i = \Lambda \times \{0,1\}^{i-1}$. A (pure) strategy for individual i is a mapping $\sigma_i : \mathcal{I}_i \to \{0,1\}$, which maps individual i's information set, I_i, to an action $a_i \in \{0,1\}$. When the actual state is θ, for any

[6] Assumption 2 is commonly adopted in the literature of SLM. In the standard SLM, it is assumed that individuals correctly understand the true DGP and the true DGP is common knowledge (Banerjee 1992, BHW 1992, Smith and Sorensen 2000). In the SLM with model misspecification, individuals may perceive an incorrect DGP, but how individuals perceive the model is commonly known (e.g., Bohren 2016, Bohren and Hauser 2021). Similarly, this paper assumes that individuals are uncertain about the DGP and the feasible model set $\mathcal{F}_0$ is commonly known.

individual i, given the predecessors' strategy profile $\sigma_{-i} = (\sigma_1, ..., \sigma_{i-1})$, and their data-generating processes profile $F_{-i} = (F_1, ..., F_{i-1})$, the observed history $h_i = (a_1, ..., a_{i-1})$ is a stochastic process with a probability measure $\mathbb{P}_{F_{-i}}(.|\theta; \sigma_{-i})$. Let $\Pi(h_i, \sigma_{-i})$ denote the set of beliefs over the state space Θ, given history h_i, and strategy profile σ_{-i}, which I refer to as a *public belief set*. It is easy to see that

$$\Pi(h_i, \sigma_{-i}) = \left\{\pi \in \Delta(\Theta) : \pi(\theta) = \mathbb{P}_{F_{-i}}(\theta|h_i; \sigma_{-i}), F_{-i} \in \mathcal{F}_0^{i-1}\right\}$$

where $\mathbb{P}_{F_{-i}}(\theta|h_i; \sigma_{-i})$ is the conditional probability on θ derived from $\mathbb{P}_{F_{-i}}(.|\theta; \sigma_{-i})$, and $\mathcal{F}_0^{i-1}$ is $i-1$ copies of $\mathcal{F}_0$. The public belief set consists of conditional probabilities generated by all possible $F_{-i} \in \mathcal{F}_0^{i-1}$ for which the conditional probabilities are well-defined. Based on the public beliefs and private signal λ_i, individual i will form a belief set, $\Pi_i(I_i, \sigma_{-i})$, which I refer to as a *posterior set*. Assuming that individuals use the full Bayesian updating rule (axiomatized by Pires (2002)) to update beliefs, thus:

$$\Pi_i(I_i, \sigma_{-i}) = \left\{\pi \in \Delta(\Theta) : \pi = BU\left(\pi'; \lambda_i\right), \pi' \in \Pi(h_i, \sigma_{-i})\right\}$$

where $BU(\pi'; \lambda_i)$ denotes the Bayesian update of belief π' based on signal λ_i. In other words, individuals update the public belief set prior-by-prior using Bayes' rule.

This updating rule has the advantage of being straightforward and it has been adopted in many applications (e.g., Bose and Renou (2014)). Two major criticisms of it, however, are: first, the size of the belief set remains unchanged even after learning new information; second, it can lead to dynamic inconsistency (see Machina and Siniscalchi's (2013) survey). My responses are: the results in this paper are still robust to other updating rules, for example individuals can update the set of data-generating processes based on their observations; besides, in the setting of this paper, individuals only need to make a once-in-a-lifetime decision; therefore, dynamic inconsistency is not relevant here[7].

Equilibrium Concept

Assume that individuals are ambiguity averse and have max-min expected utility (MEU) preferences as in Gilboa and Schmeidler (1989). An equilibrium concept is defined as the following:

Definition 1. (Equilibrium) A strategy profile $\sigma^* = (\sigma_i^*)_{i \in N}$ constitutes an *equilibrium* if for all $i \in N$ and all information sets $I_i \in \mathcal{I}_i$, we have:

$$\sigma_i^*(I_i) \in \arg\max_{a \in \{0,1\}} \inf_{\pi \in \Pi_i\left(I_i, \sigma_{-i}^*\right)} \mathbb{E}_\pi U(a, \theta) \tag{1}$$

, where $U(a, \theta)$ is the utility function which equals 1 if $a = \theta$ and equals 0 if $a \neq \theta$.

[7]It remains a question whether dynamic consistency should be maintained in the presence of ambiguity. Notice that Bayes' updating rule comes from Savage's sure-thing principle, which implies both consequentialism and dynamic consistency. Because most ambiguity preferences assume a violation of the sure-thing principle, it becomes hard to maintain both properties. Many papers hence drop dynamic consistency to retain consequentialism (e.g., Pires (2002) and Eichberger et al. (2007)).

Where no confusion would exist, I omit the equilibrium strategy notation σ^* and denote $\Pi(h_i)$ and $\Pi_i(I_i)$ as the equilibrium public belief set and posterior set. To address the tie case, I assume the following "tie-breaking rule": when indifferent, individual i chooses action 1 if $\lambda_i > 1$ and action 0 if $\lambda_i \leq 1$. With the rule, Definition 1 provides a unique pure-strategy equilibrium.

It remains questionable whether the pure-strategy equilibrium is a reasonable concept to consider when mixed strategies are also allowed. The answer is straightforward in the standard model because individuals with expected utility preferences are indifferent to randomization over choices; thus, restricting attention to pure-strategy equilibriums is without loss of generality. However, with ambiguity, there needs to be slightly more justifications. It seems that ambiguity-averse individuals can make themselves better off by using randomizations, as suggested by the Uncertainty Aversion axiom in Gilboa and Schmeidler (1989). However, the Uncertainty Aversion axiom only assumes that individuals have incentives to engage in *ex-post randomization* instead of *ex-ante randomization*, as in the mixed-strategy case.[8] Although there still remains a question of whether individuals can benefit from ex-ante randomization, a growing body of literature suggests that indifference to ex-ante randomization seems a more reasonable assumption.[9]Hence, in this paper, I also assume that individuals are indifferent to ex-ante randomization. Under this assumption, individuals have no incentive to play mixed strategies, therefore Definition 1 accommodates the case where mixed strategies are allowed.

3 Equilibrium Strategies and Concepts

This section characterizes individuals' equilibrium strategies. Similar to the standard model, ambiguous individuals' equilibrium strategies can be decomposed into two parts representing information from private signals and public history. This section then characterizes cascade sets and formally defines an information cascade, which will be used in later discussions.

3.1 Characterizations of Equilibrium Strategies

In the standard model, after observing history h_i and private signal λ_i, individual i's posterior belief has a likelihood ratio equal to $\lambda_i \cdot l_i$, where l_i denotes the likelihood ratio of the public belief after observing history h_i. In the equilibrium, individual i will choose action 1 if the product, $\lambda_i \cdot l_i$,

[8]The Uncertainty Aversion axiom is: for all acts f, g and $\alpha \in (0,1)$, so we have:

$$f \simeq g \Rightarrow \alpha f + (1-\alpha) g \succeq f$$

, where $[\alpha f + (1-\alpha) g](s) \equiv \alpha f(s) + (1-\alpha) g(s)$ for all possible states, $s \in S$. That is, for all states, acts f and g are mixed with a fixed proportion, α and $1-\alpha$, which is the ex-post randomization. Whereas the ex-ante randomization means that individuals first take a lottery with probability α on f and $1-\alpha$ on g, individuals' payoffs are only generated by f or g depending on the lottery outcome.

[9]For example, Saito (2012) suggests that individuals have no incentive to engage in ex-ante randomization when the Certainty Strategic Rationality axiom is assumed. Eichberger et al. (2015) shows that dynamic consistency implies that individuals are indifferent to ex-ante randomizations. Besides, indifference to ex-ante randomizations is also implicitly assumed in the smooth ambiguity model axiomatized by Klibanoff et al. (2005) (in the assumption that individuals have expected utilities on second-order acts).

is greater than 1 and choose action 0 otherwise. As a result, individuals' decision rules can be decomposed into two parts: the private information part, λ_i, and public information part, l_i.

When individuals are ambiguous, it seems difficult to have such a neat decomposition because the public belief is represented by a set. However, it turns out that we can extend the concept of likelihood ratio and represent the public belief set using the average likelihood ratio for the beliefs featured in in it. Based on this unique fact, we have a parallel characterization of individuals' equilibrium strategies under ambiguity. The idea of the average likelihood ratio is introduced below:

Definition 2. (Average Public Likelihood Ratio) Denote $L_i = \left\{\frac{\pi(1)}{\pi(0)} : \pi \in \Pi(h_i)\right\}$, where $\underline{l_i} = \inf L_i$ and $\overline{l_i} = \sup L_i$. Denote $r_i = \sqrt{\overline{l_i} \cdot \underline{l_i}}$, called the *average public likelihood ratio*, based on history h_i.

The average public likelihood ratio r_i is the geometric average of the highest and lowest likelihood ratios in the public belief set, which reflects how likely the public thinks state 1 is (relative to state 0) on average. Proposition 1 characterizes individuals' equilibrium strategies employing the average public likelihood ratios.

Proposition 1. (Characterizations of Equilibrium Strategies) *In the equilibrium, for any individual, $i \in N$, and information set, $I_i \in \mathcal{I}_i$, we have:*

(1) When $\lambda_i > 1$: $\sigma_i^*(I_i) = 1$ if and only if $\lambda_i \cdot r_i \geq 1$; $\sigma_i^*(I_i) = 0$ if and only if $\lambda_i \cdot r_i < 1$.

(2) When $\lambda_i \leq 1$: $\sigma_i^*(I_i) = 1$ if and only if $\lambda_i \cdot r_i > 1$; $\sigma_i^*(I_i) = 0$ if and only if $\lambda_i \cdot r_i \leq 1$.

Proof. Denote $\underline{\pi}_i(\theta) = \inf\{\pi(\theta) : \pi \in \Pi_i(I_i)\}$. Suppose that $\lambda_i > 1$, then $a_i = 1$ if and only if $\underline{\pi}_i(1) \geq \underline{\pi}_i(0)$. Note that:

$$\underline{\pi}_i(1) = \frac{\lambda_i \underline{l_i}}{1 + \lambda_i \underline{l_i}} \qquad \underline{\pi}_i(0) = \frac{1}{1 + \lambda_i \overline{l_i}}$$

by solving $\underline{\pi}_i(1) \geq \underline{\pi}_i(0)$, we have: $\lambda_i \geq \frac{1}{\sqrt{\overline{l_i} \cdot \underline{l_i}}} = \frac{1}{r_i}$. Other cases follow symmetrically. □

The average likelihood ratio is an extension of the likelihood ratio in the standard model. It acts as a sufficient statistic for the public history in cases where there are multiple beliefs. Proposition 1 shows that individuals' equilibrium strategies can also be represented as the product of two parts. The private information component is still the private signal, λ_i, whereas the public information is captured by the average public likelihood ratio, r_i. When the product, $\lambda_i \cdot r_i$, is greater than 1, reflecting that state 1 is more likely, individuals will choose action 1 and vice versa. For simplicity, "average public likelihood ratio" is sometimes referred to as "public belief" when there is no confusion.

Note that the representation of the average public likelihood ratio relies on individuals' ambiguity preferences. The representation in Definition 2 relies on the assumption that individuals have MEU preferences. When individuals have other ambiguity preferences (e.g., smooth ambiguity preferences), we may have different representations.

3.2 Herding, Cascades and Learning

Definition 3. [Herding and Information Cascades] In the equilibrium, we say that

(i) a ***herding*** occurs if there exists some $I < \infty$ and $a \in A$ such that for all $i \geq I$, $a_i = a$;

(ii) an ***information cascade*** occurs if there exists some $I < \infty$ and $a \in A$ such that for all $i \geq I$, we have $\mathbb{P}^*(a_i = a|h_i) = 1$;

A herding occurs if the society ends up taking the same action, and an information cascade occurs if after some point, individuals will only choose one action regardless of their private signals. An information cascade is stronger than a herding. During a herding, individuals would have acted differently if they received different signals, whereas during an information cascade, any realizations of private signals are unable to overturn the herd.[10] When signals are bounded and individuals correctly specify the model, a herding occurs almost surely and the herding can be incorrect, but an information cascade may or may not happen depending on the details of the data-generating processes.

Definition 4. ***Complete learning*** occurs if there exists some $I < \infty$ such that for all $i \geq I$, $a_i = \theta^*$ $\mathbb{P}^*$-almost surely.

In other words, complete learning occurs if the society eventually settles on the optimal action with probability 1. In the standard framework, complete learning occurs if and only if signals are unbounded.

4 Benchmark Case: Cascades under Extreme Ambiguity

To build intuition, it is better to discuss the benchmark case, where individuals are extremely ambiguous.

Assumption 3. $\mathcal{F}_0 = \mathcal{F}$.

This assumption says that individuals consider all feasible models with support in $[\gamma, 1/\gamma]$. It describes a situation where individuals only know the range of signals without further knowledge about the true data-generating processes. Under this assumption, we have the following theorem.

Theorem 1. *Under Assumption 3, an information cascade occurs $\mathbb{P}^*$-almost surely for* ***all*** *possible $\overline{F}_i$s in $\mathcal{F}$.*

One significance of Theorem 1 is that an information cascade occurs almost surely for all possible true data-generating processes, regardless of whether signals are bounded or unbounded, discrete or continuous. Besides, it is also true that an incorrect cascade occurs with a strictly positive

[10] These two concepts was distinguished by Smith and Sørensen (2000) and experimentally distinguished by Çelen and Kariv (2004), and Anderson and Holt (1997) also provided laboratory evidence of information cascades.

probability, so the society can settle on the incorrect action for all possible signals (even if signals are unbounded).[11] The intuition comes from the following arguments.

Intuition. Suppose that the first i individuals take action 1 (i.e., $a_1 = ... = a_i = 1$), and individual $i+1$ receives a signal $\frac{1}{\gamma}$, the strongest signal for state 0. Suppose that an information cascade did not occur when the first i individuals made decisions. Consider the decision problem of individual $i+1$. As she has max-min EU preference, her decision is determined by the worst scenarios:

- If she follows the herd and takes action 1, the worst case happens when the predecessors' data-generating processes are uninformative. In this case, $\lambda_1 = ... = \lambda_i = 1$. By following the herd, she will be acting against one signal, $\frac{1}{\gamma}$ (her own signal);
- If she breaks the herd and takes action 0, the worst case arises when the predecessors' data-generating processes are most precise (only generating signal γ and $1/\gamma$). In this case, the predecessors' actions perfectly reveal that their signals are all γs. Hence, by taking action 0, individual $i+1$ follows her own signal, but she will be acting against i signal γs.

As can be seen, the forces encouraging a cascade and discouraging it are **asymmetric**. As i increases, the cost of breaking the herd increases consistently as individual $i+1$ will be acting against more and more signal γs in the worst case. However, the cost of herding remains the same due to the fact that, in the worst case, individual i is always acting against one signal $\frac{1}{\gamma}$. When i is sufficiently large, the cost of breaking the herd is higher; therefore, all the following individuals will choose action 1 and an information cascade of action 1 occurs.

Sketch of the Proof. The proof of Theorem 1 is not difficult. First, from the equilibrium strategy, the average likelihood ratio, r_i, serves as a sufficient statistic for the history, so we only keep track of r_i. Second, it can be verified that whenever a cascade does not occur, the increment or the decrement of r_i is bounded away from 1 by some constant. To see that, suppose that $a_i = 1$. Then, the highest likelihood ratio increases by a factor of γ, which corresponds to the increment when individual i has the most precise data-generating process, but the lowest likelihood ratio does not decrease, since intuitively, an action 1 appears as positive news for state 1. Combining these two cases, the average public likelihood ratio should at least increase by a factor of $\sqrt{\gamma}$, which is greater than 1. The case where $a_i = 0$ is symmetric. The rest of the analysis is standard. From the second step, we know that finite number of consecutive actions will trigger a cascade, which further implies that at each period, the probability of a cascade is bounded away from 0, so a cascade must occur with probability 1.

5 Information Cascades with Bounded Signals

Last section establishes an information cascade when there is extreme ambiguity. It is natural to ask whether the result still holds in less extreme cases. It turns out that when signals are

[11] In this paper, whenever an information cascade arises, an incorrect cascade must arise with a strictly positive probability, so I will not state it explicitly in theorems.

bounded, an information cascade is easy to occur in the presence of ambiguity. Interestingly, in many situations, the non-cascade results only represent knife-edge cases. To see that, let's first look at two conditions that ensure a cascade.

Theorem 2. *If there exists some $F \in \mathcal{F}_0$ such that one of the following conditions holds:*

(1) F is discrete at γ;

(2) F is continuously differentiable on $(\gamma - \varepsilon, \gamma)$ for some $\varepsilon > 0$ with $f^1(\gamma) > \frac{2}{\gamma - 1}$,

where $f^1(\gamma) = \lim_{x \nearrow \gamma} \frac{dF^1}{dx}(x)$. Then, when signals are bounded, an information cascade occurs $\mathbb{P}^*$-almost surely.

Conditions (1) and (2) can be intuitively interpreted as that some DGP under consideration is sufficiently informative. Specifically, the DGP assigns sufficiently large weights to high-precision signals, that is, signals close to γ, and symmetrically, signals close to $1/\gamma$. Theorem 2 says that if individuals find it possible that other individuals can have a highly informative DGP, an information cascade emerges almost surely.

Theorem 1 imposes the following restrictions. First, it only requires $\mathcal{F}_0$ to contain one one such F but does not impose other restrictions on $\mathcal{F}_0$. The intuition is based on the observation that the forces encouraging a cascade and discouraging it are asymmetric. Due to the asymmetry, if $\mathcal{F}_0$ contains a highly informative F, the cascading force becomes so strong such that it cannot be offset by other models, so a cascade will occur almost surely regardless of what other models $\mathcal{F}_0$ may contain. Second, it only requires the F to place sufficient weights on the tails but does not impose any restrictions in the middle. It comes from the fact that beliefs will approach the boundary after many identical actions, so by restricting the tail properties of F, we can ensure the occurrence of an information cascade.

Remark 1. One conjecture is that F is the "essential model" under the max-min EU, so individuals would act as if the true model was F. This conjecture is incorrect. In fact, every model in $\mathcal{F}_0$ may affect the learning process. It is just that the cascading force from F is too strong such that other models cannot alter the occurrence of a cascade. [12]

5.1 Fragility of the Non-cascade Result

As can be seen, the restrictions in Theorem 1 can be very moderate in some sense, which implies that the standard result about cascades can be non-robust. Below is an example.

Example 2. Suppose that individuals perceive the following set of models

$$\mathcal{F}_0 = (1-\varepsilon) G \oplus \varepsilon \mathcal{F} \equiv \{F_0 : F_0 = (1-\varepsilon) G + \varepsilon F, \text{for } F \in \mathcal{F}\}.$$

[12] Even if other models do not alter the occurrence of a cascade, but they do affect belief dynamics, convergence speed and so on. As a result, learning under a non-degenerate $\mathcal{F}_0$ containing F is not observationally equivalent to learning under $\{F\}$. Moreover, if F is less informative, it is possible to construct an example where a cascade may or may not occur depending on other models in $\mathcal{F}_0$.

Recall that $\mathcal{F}$ denotes the set of all possible data-generating processes. The model set, $\mathcal{F}_0$, is constructed by making an $\varepsilon-$perturbation to G using set $\mathcal{F}$. Notice that when $\varepsilon > 0$, there exists some F_0 that is discrete at γ, so an information cascade occurs almost surely for all $\varepsilon > 0$ by Theorem 2.

Example 2 shows that any positive perturbation can produce an information cascade, so we can have an information cascade even if individuals are just "slightly" ambiguous. To characterize the degree of ambiguity explicitly, I follow a common approach in the literature and assume that individuals' model sets are generated by some distance function.

Assumption 4. *Suppose that $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$, and that individuals have the following belief set:*

$$\mathcal{F}_0 = \{F \in \mathcal{F} : d(F, G) \leq K\}$$

for some $K \geq 0$ and some $G \in \mathcal{F}$ with a strictly positive density on $[1/\gamma, \gamma]$.

Assumption 4 states individuals only consider the set of data-generating processes within distance K to the benchmark model, G. The requirement that G has a positive density function is to simplify the discussion by ruling out some irregular Gs. Here, K is referred to as individuals' *ambiguity level*. When $K = 0$, meaning that individuals are not ambiguous, the belief set only contains the benchmark model, G; when $K = \infty$, individuals are extremely ambiguous, as in Assumption 3. I focus on the following class of distribution function.

Definition 5. d is *consistent with weak convergence* if for any data-generating process, F, and sequence, $(F_n) \in \mathcal{F}$, satisfying $F_n \Rightarrow F$, we have $d(F_n, F) \to 0$, where "$\Rightarrow$" represents weak convergence.

Consistency with weak convergence has an intuitive interpretation: when the distribution functions of two data-generating processes are close to each other, individuals tend to think that they are similar. Many commonly used distances belong to this class, including the sup-norm metric, total-variation, and Lévy–Prokhorov metric. With this class of distance metrics, we have a similar result as in Corollary 1.

Corollary 1. *Under Assumption 4 and that d is consistent with weak convergence, when signals are bounded, an information cascade occurs $\mathbb{P}^*$-almost surely for all $K > 0$.*

The idea is similar to Example 2. Under a metric consistent with weak convergence, any continuous data-generating process can be approximated by a discrete one. Therefore, for all $K > 0$, there exists a discrete model that is within benchmark K to the benchmark model, so a cascade occurs almost surely.

Other Distance Concepts. There are some also distance concepts not consistent with weak convergence. With these distances, the distance between a discrete and a continuous model can be infinity. An interesting example is relative entropy, which was adopted by Hansen and Sargent

(2001). The following corollary shows that with relative entropy distance, an information cascade emerges almost surely when individuals are sufficiently ambiguous (but still with a finite degree of ambiguity).

Corollary 2. *Under Assumption 4 and that d is the relative entropy, where*

$$d(F,G) = \int_{\frac{1}{\gamma}}^{\gamma} \log\left(\frac{dF(\lambda)}{dG(\lambda)}\right) dF(\lambda),$$

then there exists some finite number, $\overline{K}$, such that when signals are bounded, an information cascade occurs $\mathbb{P}^*$-almost surely as long as $K > \overline{K}$.

The proof makes use of condition (2) in Theorem 2. One can show that there exists some continuous data-generating process, F, satisfying condition (2) and $d(F,G) < \infty$. Then, we simply need to set $\overline{K} = d(F,G)$. The following example depicts one such data-generating process (detailed analysis is provided in the Appendix).

Example 3. [A continuous DGP that induces a cascade] For simplicity in exposition, I deal with the nominal signal space $S = [0,1]$. Consider the following h:

$$h^1(s) = \begin{cases} 1 + 2\varepsilon(1+\gamma)\cdot s & s \in \left[0, \frac{1}{1+\gamma}\right] \\ 0 & s \in \left(\frac{1}{1+\gamma}, \frac{\gamma}{1+\gamma}\right), \\ 2\varepsilon(1+\gamma)\cdot s + \gamma - 2\varepsilon(1+\gamma) & s \in \left[\frac{\gamma}{1+\gamma}, 1\right] \end{cases} \qquad h^0(s) = h^1(1-s)$$

where $0 < \varepsilon < \frac{\gamma-1}{2\gamma+2}$. After making the transformation, $\lambda = \frac{h^1(s)}{h^0(s)}$, we can express each signal s in terms of likelihood ratios, λ. It can be seen that likelihood ratio $\lambda \in \left[\frac{1}{\gamma}, \frac{1+2\varepsilon}{\gamma-2\varepsilon}\right] \cup \left[\frac{\gamma-2\varepsilon}{1+2\varepsilon}, \gamma\right]$. When ε is smaller, the signals (λ's) are more concentrated around the two tails, meaning that the data-generating process is more "precise". When ε is sufficiently small, condition (2) is satisfied. *As long as the belief set, $\mathcal{F}_0$, contains one such data-generating process, an information cascade occurs almost surely.* Under the assumptions of benchmark G, it is easy to see that a model set with a finite radius includes such a data-generating process.

6 Incorrect Herding with Unbounded Signals

The last section shows that when signals are bounded, the non-cascade results in the literature can be fragile with respect to ambiguity. Similar non-robustness also exists when signals are unbounded. Recall that Smith and Sorensen (2000) showed that complete learning occurs whenever signals are unbounded. However, this section shows that complete learning may collapse even if there is a slight degree of ambiguity.

6.1 Incomplete Learning under Ambiguity

To be more specific, under ambiguity, incomplete learning will arise in the form that individuals can herd on the incorrect action with a strictly positive probability. The following theorem provides conditions for an incorrect herding to arise.

Theorem 3. *Suppose that for all i, $\overline{F}_i^0(x) \leq ax^\alpha$ with $a, \alpha > 0$ as $x \to 0$. If there exists some $F \in \mathcal{F}_0$ such that $x^p = o\left(F^0(x)\right)$ as $x \to 0$ for some $p \in (0, \alpha)$, herding occurs $\mathbb{P}^*$-almost surely, and an incorrect herding occurs with a $\mathbb{P}^*$-strictly positive probability.*

The condition $x^p = o\left(F^0(x)\right)$ means that the tail of $F^0(x)$ is sufficiently fat, or intuitively, it can be interpreted as that $F^0(x)$ is sufficiently informative. Therefore, Theorem 3 conveys two messages: first, when individuals perceive some adequately informative DGP, herding occurs with probability 1, so it is not possible for actions to oscillate; second, an incorrect herding occurs with a strictly positive probability, so complete learning does not hold. Theorem 3 can be viewed as a parallel statement of Theorem 2 when signals are unbounded. Theorem 2 implies that non-cascade is not robust in many cases, and similarly, Theorem 3 also implies that complete learning is also not robust in some interesting cases. Corollary 3 presents one possibility.

Corollary 3. *Suppose that signals are i.i.d. with $\overline{F}^0(x) = O(x^\alpha)$ with $\alpha > 0$ as $x \to 0$. If there exists some $F^0 \in \mathcal{F}_0$ such that $F^0 = O(x^{\alpha-\varepsilon})$ with $\varepsilon \in (0, \alpha)$ as $x \to 0$, then herding occurs $\mathbb{P}^*$-almost surely, and an incorrect herding occurs with a strictly positive probability.*

Corollary 3 says that any small ambiguity in the order of F^0 will destroy complete learning learning. Notice that Corollary 3 does not impose other restrictions on the belief set, so individuals can also perceive data-generating processes with an order greater than α. But as long as any data-generating process with an order less than α is considered possible, the society will not achieve complete learning. Below is a concrete example.

Example 4. [Ambiguity in Model Parameter] Consider the signal space $\mathcal{S} = (0, 1)$, signals are i.i.d. and the data-generating process takes the form of $g_m = \left(g_m^0, g_m^1\right)$, where

$$g_m^0(s) = (m+1)(1-s)^m \text{ and } g_m^1(s) = (m+1)s^m \qquad \text{for } s \in (0,1).$$

It is easy to see that signals are unbounded under g_m (in terms of likelihood ratios). Suppose that the true data-generating process is g_{m_0}. Individuals are ambiguous about the true parameter m_0 and perceive a set $M_\varepsilon = [m_0 - \varepsilon, m_0 + \varepsilon] \subset \mathbb{R}_+$. When $\varepsilon = 0$, there is no ambiguity, so complete learning occurs. However, for *all* $\varepsilon > 0$, complete learning does not occur, and the society will settle on an incorrect action with a strictly positive probability. This implies that any small vagueness around the true parameter can lead to incomplete learning.

Remark 2. The collapse of complete learning can imply a discontinuity in learning efficiency. For instance, suppose that $m_0 > 1$. When there is no ambiguity, the expected number of incorrect

actions is finite as shown by Rosenberg and Vieille (2019), so learning is efficient according to their definition. In contrast, for all $\varepsilon > 0$, an incorrect herding occurs with a strictly positive probability, so the expected number of incorrect actions becomes infinite.

6.2 Sketch of the Proof.

For simplicity in illustration, I sketch the proof under the assumptions of Corollary 3. The complete learning in Smith and Sørensen (2000) is a result of the martingale convergence theorem. However, under model ambiguity, posteriors are no longer martingales, so we cannot apply the same martingale technique. This paper adopts a new approach to analyze learning with unbounded signals. The proof consists of the following three steps.

(i) Incorrect herding $p > 0$. The most important step is to show that an incorrect herding occurs with a strictly positive probability. We note that the probability of an incorrect herding is

$$\lim_{i\to\infty} \mathbb{P}^* \left(a_1 = ... = a_i = 1\right) = \prod_{i=1}^{\infty} \left(1 - \overline{F}^0 \left(1/r_i\right)\right),$$

where r_i denotes the average public likelihood after $h_i = (1, ..., 1)$. The probability is positive if and only if the infinite series $\sum \overline{F}^0 \left(1/r_i\right) < \infty$, which is equivalent to $\sum \left(\frac{1}{r_i}\right)^{\alpha} < \infty$, where α is the order of $\overline{F}^0$ near the neighborhood of 0. Under the assumption of Corollary 3, we can find an $\varepsilon \in (0, \alpha)$ such that the growing speed of r_i satisfies $r_{I+t} \geq (t+1)^{\frac{1}{\alpha-\varepsilon}}$ for some I sufficiently large and for $t \geq 1$. It implies that

$$\sum_t \left(\frac{1}{r_{I+t}}\right)^{\alpha} \leq \sum_t \frac{1}{(t+1)^{\frac{\alpha}{\alpha-\varepsilon}}} < \infty,$$

which establishes that an incorrect herding occurs with a strictly positive probability. To get a sense of how the non-robustness arises, notice that if there is no ambiguity, $\varepsilon = 0$, so the infinite series on the RHS becomes $\sum \frac{1}{t+1}$, which diverges and corresponds to the absence of an incorrect herding. However, when $\varepsilon > 0$ even if it is arbitrarily small, the infinite series becomes convergent, so an incorrect herding arises.

(ii) Herding occurs w.p.1. A symmetric argument implies that a correct herding occurs with $p > 0$. It can be further shown that the probabilities of both herding has a *uniform* lower bound when average likelihood ratio r_i is sufficiently large or small, i.e., both states are locally stable. In other words, for all possible history h_i and for all i, the probability that herding eventually occurs is uniformly bounded from below, which implies that herding must occur almost surely.

7 Conditions for Complete Learning under Ambiguity

The previous discussion shows that complete learning can easily collapse under ambiguity. One natural question is that: when does complete learning hold under ambiguity (if it is possible)? This

section presents a **necessary and sufficient condition** for complete learning within the class of data-generating processes that have power tails.

Definition 6. A data-generating process $F \in \mathcal{F}$ has a *power tail* if there exists some $\alpha > 0$ such that $F^0(x) = O(x^\alpha)$ as $x \to 0$. The power of F, denoted by $\mathcal{P}(F)$, is defined to be α.

A data-generating process has a power tail if it can be approximated by a power function around its tail. It is easy to see that a power-tail data-generating process is unbounded. The power provides an intuitive measure of the signal informativeness, where a larger power means that signals are less informative (around the tails). To see this, note that if F has a larger power, it means that its tails are thinner, so it generates high-precision signals with a lower probability. This section focuses on the power-tail models and imposes the following assumptions.

Assumption 5. *$\overline{F}$ has a power tail, and $\mathcal{F}_0$ only contains models with power tails.*

Assumption 6. *$\mathcal{F}_0$ is finite where each $F \in \mathcal{F}_0$ has a different power and is differentiable.*

Assumption 5 says that the true DGP has a power tail, and individuals only perceive DGPs with power tails. Assumption 6 is imposed for simplicity and the discussion can be extended to situations where $\mathcal{F}_0$ is infinite and some models may have identical powers. Theorem 4 provides a necessary and sufficient condition for complete learning under these two assumptions.

Theorem 4. *Under Assumptions 5 and 6, complete learning occurs* ***if and only if*** *$\mathcal{F}_0$ satisfies*

(i) for all $F \in \mathcal{F}_0$, we have $\mathcal{P}(F) \geq \mathcal{P}\left(\overline{F}\right)$, and

(ii) there exists some $F \in \mathcal{F}_0$ such that $\mathcal{P}(F) < \mathcal{P}\left(\overline{F}\right) + 1$.

Theorem 4 says that to establish complete learning, we need to impose restrictions from two directions. On one hand, all perceived models can not be too informative. Specifically, they must be less informative than the true model in the sense they must have higher powers. On the other hand, some perceived model has to be sufficiently informative such that its power does not exceed that of the true model by 1. Before explaining the intuition, it helps to see what will happen if the conditions in Theorem 4 are violated.

Corollary 4. *Under Assumptions 5 and 6, (i) if there exists some $F \in \mathcal{F}_0$ such that $\mathcal{P}(F) < \mathcal{P}\left(\overline{F}\right)$, an incorrect herding occurs with a strictly positive probability; (ii) if for all $F \in \mathcal{F}_0$, $\mathcal{P}(F) \geq \mathcal{P}\left(\overline{F}\right) + 1$, actions do not converge with probability 1.*

First, when individuals perceive some highly informative model, learning is incomplete and takes the form of incorrect herding. Second, when all models considered by individuals are insufficiently informative, incomplete learning takes the form of action non-convergence. The first case has been explained. The second case comes from an intuitive argument that if individuals underestimate the informativeness of their predecessors, any herd will be overturned frequently, and as a consequence, the actions will not converge.

Intuition of Theorem 4. In order to achieve complete learning, we must exclude two sources of incomplete learning—incorrect herding and action non-convergence. Correspondingly, we also need to restrict $\mathcal{F}_0$ from two directions. To prevent incorrect herding, $\mathcal{F}_0$ must not contain highly informative data-generating processes, which correspond to Theorem 4 (i). To prevent action non-convergence, $\mathcal{F}_0$ must not only contain data-generating processes that are too uninformative, which Theorem 4 (ii).

8 Other Ambiguity Preferences

Many results of this paper can be extended to a wider class of ambiguity preferences. This section focuses on two important examples, the α-max-min EU preference and the smooth ambiguity preference. The discussion shows that results under the max-min EU preference are not as extreme as it appears.

8.1 $\alpha-$Max-Min EU Model

Consider first the case where individuals hold the α-maxmin expected utility ($\alpha-$MEU) preferences (Hurwicz 1951, Ghirardato, Maccheroni, Marinacci 2004). With this class of preferences, individual i's utility is

$$V_i(a) = \alpha \cdot \inf_{\pi \in \Pi_i} \mathbb{E}_\pi U(a, \theta) + (1-\alpha) \cdot \sup_{\pi \in \Pi_i} \mathbb{E}_\pi U(a, \theta)$$

where $\alpha \in [0,1]$. Here α represents the degree of an individual's pessimism, where $\alpha = 1$ corresponds to the MEU model, and $\alpha = 0$ corresponds to the max-max expected utility model.[13] Within this class of models, we have a surprising result as follows.

Proposition 2. *When individuals have α-MEU preferences, **all** previous results hold for all $\alpha \in [0,1]$.*

Proof. It can be verified that the equilibrium strategy takes the identical form as in the MEU case. □

Proposition 2 shows that this paper's results hold for all ambiguity attitudes captured by the $\alpha-$MEU model. The intuition is less surprising than it appears. Notice that the key force is the asymmetry between the models encouraging and discouraging a cascade, but the asymmetry does not rely on ambiguity aversion, for example, we can still apply the similar arguments under Theorem 1 when individuals have max-max EU preference. It is worth noting that Proposition 2 relies on the binary choice structure, and if we allow for multiple actions, ambiguity attitudes affect which action the society settles on, e.g., individuals will herd on a safer action when they are ambiguity-averse but not when they are ambiguity-loving.

[13] Notice that $\alpha = \frac{1}{2}$ does not represent the expected-utility model—it is only that individuals attach the same weight to the best and worst scenario.

8.2 Smooth Ambiguity Model

Consider next that individuals hold the smooth ambiguity model as axiomatized by Klibanoff, Marinacci and Mukerji (2005).

$$V_i(a) = \phi^{-1}\left(\int_{\Pi_i} \phi\left[\mathbb{E}_\pi U(a,\theta)\right] d\mu(\pi)\right).$$

where μ stands for the second-order belief, and the curvature of the ϕ function describes the ambiguity attitude. The analysis under the smooth model becomes more difficult as the equilibrium strategy no longer has a simple characterization. However, many results can still hold qualitatively when individuals are sufficiently **ambiguity sensitive**. Below is an example.

Example 5. Let's consider the setup Example 1. Recall that in Example 1, where $S = \{h, l\}$, and the data-generating process has precision γ. Each individual's signal precision $\gamma \in (1, \overline{\gamma}) = \Gamma$ is further drawn from a full-support distribution, h. Signals are bounded, so $\overline{\gamma} < \infty$. Every individual has the following preference.

$$V_i(a) = \left[\int_{\Gamma^{i-1}} \left[\mathbb{E}_{\gamma_1,...,\gamma_{i-1}} u(a,\theta)\right]^{1-\sigma} dh(\gamma_1,...,\gamma_{i-1})\right]^{\frac{1}{1-\sigma}}.$$

This preference is a reformulation of the preference with constant relative ambiguity aversion (CRAA). Parameter σ represents the ambiguity attitude, where $\sigma = 1$ corresponds to ambiguity neutrality, and as σ grows, individuals are becoming more ambiguity-averse.

(i) *When $\sigma = 0$, whether an information cascade can occur depends on the properties of h.*

Note that $\sigma = 0$ corresponds to the situation where individuals have EU preference and hold a correct model perception, h, so the occurrence of a cascade depends h.

(ii) *When $|\sigma|$ is sufficiently large, an information cascade occurs with a strictly positive probability for all possible h.*

As $\sigma \to +\infty$, the preference approaches the max-min EU model, and as $\sigma \to -\infty$, it approaches the max-max EU model. For any fixed $I < \infty$, the continuity in preference implies that belief dynamics before individual I can be arbitrarily close to to those under the max-min or max-max EU when $|\sigma|$ is sufficiently large, which also implies the occurrence of a cascade. More details can be found in Appendix A.7.

The example suggests that information cascades are not unique to the max-min EU model. It also emerges under the smooth ambiguity preference when individuals are sufficiently ambiguity sensitive. Example 5 only looks at the bounded-signal case, but it can be extended to unbounded signals with some restrictions on h (see Appendix A.7).

9 Discussion: Bayesian vs Ambiguity

From the previous discussion, we know an information cascade occurs under sufficient model ambiguity. This section summarizes how results change if we switch to the Bayesian case.

Bayesian Model Certainty

When individuals are Bayesian and certain about the data-generating processes, the learning outcome depends on the features of the model specifications. When individuals hold a **correctly specified** model perception, the learning outcomes depend on the details of the true data-generating processes as discussed earlier. When individuals hold a **incorrectly specified** model perception, the learning outcomes depend on the interplay between the true and perceived data-generating processes. For example, if individuals overestimate other people's informativeness, the society may herd on the incorrect action; if they underestimate the informativeness, the society may fail to settle on an action as implied by Corollary 4.

Bayesian Model Uncertainty

Another case is that individuals are Bayesian and uncertain about the data-generating processes. In this case, the learning outcome depends on the **priors** over the model space. Below are two examples.

Example 6. The model space is $\mathfrak{F} \coloneqq \mathcal{F}^{\infty}$, where a typical element $\boldsymbol{F} = (F_1, F_2, ...)$ describes a list of all data-generating processes. Individuals hold an identical prior $Q \in \Delta(\mathfrak{F})$. All signals are i.i.d. and unbounded. The true data-generating process sequence is denoted by $\overline{\boldsymbol{F}}$.

(i) *If* $Q\left(\overline{\boldsymbol{F}}\right) > 0$*, complete learning occurs almost surely.*

In other words, if the prior puts a strictly positive probability on $\overline{\boldsymbol{F}}$, or it contains a "grain of the truth", complete learning occurs. This is because when the true model path is assigned a strictly positive probability, limit beliefs will "merge to" the beliefs induced by the true models (Kalai and Lehrer 1993). Note that if individuals knew the true model, complete learning would occur. As a consequence, it can be verified that complete learning also occur if the prior contains a "grain of the truth". The more rigorous proof can be found in Appendix A.8.

(ii) *If* $Q\left(\overline{\boldsymbol{F}}\right) = 0$*, complete learning may not occur.*

If the "grain-of-the-truth" condition fails, limit learning can be different from that under the true model. One example is that Q features an independent distribution across all individuals. In this case, Q is not updated after observations, so the problem becomes learning under model certainty, where the model perception is $F_Q = \mathbb{E}_Q F$. From the previous discussion, learning outcome depends on F_Q and its relation with $\overline{F}$, so it is possible that complete learning does not occur (Appendix A.8 provides an example).

Remark 3. Example 6 shows that when signals are unbounded, the occurrence of complete learning depends on the priors. Similarly, when signals are bounded, the occurrence of a cascade also depends on the priors. Appendix A.8 presents one such example. In the example, there are two models, where the first implies a cascade but the second does not. This example shows that a cascade occurs if the prior assigns a sufficiently large weight to the first model but does not occur if the second model is assigned a large weight.

Ambiguous Model Uncertainty

In Bayesian learning, the learning outcome differs with individuals' model perceptions or priors, so we do know which outcome will arise without the knowledge of priors or model perceptions. This study complements the literature by showing that when individuals consider several models simultaneously, an information cascade can almost surely occur under sufficient ambiguity. The results are driven by the following differences.

1. Ambiguity. Under ambiguity and with max-min EU preference, individuals are unable to assign probabilities, and their decisions are determined by comparing the worst payoffs. In this case, an information cascade arises because of the following asymmetric effects—the worst case for breaking a herd is when the predecessors have precise data-generating processes, but the worst case for following the herd is only when predecessors have imprecise information. The first possibility encourages a cascade whereas the second discourages it. Under sufficient model uncertainty, the encouraging effect outweighs the discouraging effect, so individuals will follow the herd to avoid the larger ambiguity from acting against it, hence an information cascade arises.

2. Bayesian. When individuals are Bayesian, their priors play an important role in determining the learning outcome. It is true that highly informative data-generating processes encourage an information cascade, and the encouraging effect cannot be offset by other data-generating processes in terms of their induced payoffs. However, individuals make decisions based on the payoffs weighted by the posteriors on models, which depend on how their priors are formed. For some priors, the posteriors on these informative data-generating processes will become very small or even zero in the limit. In these cases, the encouraging effect is so small such that an information cascade or an incorrect herding does not occur. Similarly, for some other priors, the encouraging effect can be very large such that a cascade occurs.

Remark 4. Under model uncertainty, the Bayesian approach assumes a **prior** over models, based on which individuals can pin down a unique posterior, then a decision can be made according to that posterior. One challenge is that the learning outcome depends on the priors, and it is difficult to test the priors individuals hold or to justify which priors should be imposed in the model. Differently, the ambiguous approach allows individuals to hold multiple posteriors but imposes an **ambiguity preference** in the decision making. It turns out that when people are sufficiently ambiguity sensitive, and when there is sufficient model uncertainty, an information cascade will occur.

10 Conclusion

This paper studies sequential learning under the assumption that individuals face ambiguity about other people's data-generating processes. This paper finds that under sufficient ambiguity, an information cascade arises with probability 1. Interestingly, the results that feature non-cascades may only represent a knife-edge case from the perspective of ambiguity. This paper provides a new perspective into the mechanism behind cascades and herding. The paper adopts the most standard setup, where there are binary states, binary actions. The Appendix B shows by examples that qualitative results hold with multi-state, multi-action and general updating rules (i.e., α-maximum likelihood updating). This paper also relies on the linear network, so an interesting direction is to extend the result to general networks.

References

[1] Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, 78(4), 1201-1236.

[2] Acemoglu, D., Chernozhukov, V., & Yildiz, M. (2016). Fragility of asymptotic agreement under Bayesian learning. *Theoretical Economics*, 11(1), 187-225.

[3] Anderson, L. R., & Holt, C. A. (1997). Information cascades in the laboratory. *The American Economic Review*, 847-862.

[4] Arieli, I., & Mueller-Frank, M. (2018). Multidimensional Social Learning. *The Review of Economic Studies*, 86(3), 913-940.

[5] Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3), 797-817.

[6] Battigalli, P., Cerreia-Vioglio, S., Maccheroni, F., & Marinacci, M. (2015). Self-confirming equilibrium and model uncertainty. *American Economic Review*, 105(2), 646-77.

[7] Battigalli, P., Francetich, A., Lanzani, G., & Marinacci, M. (2019). Learning and self-confirming long-run biases. *Journal of Economic Theory*, 183, 740-785.

[8] Bikhchandani, S., Hirshleifer, D., Tamuz, O., & Welch, I. (2021). Information Cascades and Social Learning (No. w28887). National Bureau of Economic Research.

[9] Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992-1026.

[10] Bohren, J. A. (2016). Informational herding with model misspecification. *Journal of Economic Theory*, 163, 222-247.

[11] Bohren, J. A., & Hauser, D. N. (2021). Learning with heterogeneous misspecified models: Characterization and robustness. *Econometrica*, 89(6), 3025-3077.

[12] Bohren, J. A., Imas, A., & Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10), 3395-3436.

[13] Bose, S., & Renou, L. (2014). Mechanism design with ambiguous communication devices. *Econometrica*, 82(5), 1853-1872.

[14] Çelen, B., & Kariv, S. (2004). Distinguishing informational cascades from herd behavior in the laboratory. *American Economic Review*, 94(3), 484-498.

[15] Chen, J. Y. (2020). Biased Learning under Ambiguous Information. Available at SSRN.

[16] Dasaratha, K., & He, K. (2020). Network structure and naive sequential learning. *Theoretical Economics*, 15(2), 415-444.

[17] DeMarzo, P. M., Vayanos, D., & Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3), 909-968.

[18] Easley, D., & Kleinberg, J. (2010). Networks, crowds, and markets (Vol. 8). Cambridge: Cambridge university press.

[19] Eichberger, J., Grant, S., & Kelsey, D. (2007). Updating choquet beliefs. *Journal of Mathematical Economics*, 43(7-8), 888-899.

[20] Eichberger, J., Grant, S., & Kelsey, D. (2016). Randomization and dynamic consistency. *Economic Theory*, 62(3), 547-566.

[21] Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 643-669.

[22] Epstein, L. G., & Schneider, M. (2007). Learning under ambiguity. *The Review of Economic Studies*, 74(4), 1275-1303.

[23] Esponda, I., & Pouzo, D. (2016). Berk–Nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3), 1093-1130.

[24] Eyster, E., & Rabin, M. (2010). Naive herding in rich-information settings. *American Economic Journal: Microeconomics*, 2(4), 221-43.

[25] Ford, J. L., Kelsey, D., & Pang, W. (2013). Information and ambiguity: herd and contrarian behaviour in financial markets. *Theory and Decision*, 75(1), 1-15.

[26] Frick, M., Iijima, R., & Ishii, Y. (2020a). Misinterpreting others and the fragility of social learning. *Econometrica*, 88(6), 2281-2328.

[27] Frick, M., Iijima, R., & Ishii, Y. (2020b). Stability and robustness in misspecified learning models.

[28] Fryer Jr, R. G., Harms, P., & Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, 17(5), 1470-1501.

[29] Fudenberg, D., Lanzani, G., & Strack, P. (2021). Limit Points of Endogenous Misspecified Learning. *Econometrica*, 89(3), 1065-1098.

[30] Ghirardato, P., Maccheroni, F., & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory*, 118(2), 133-173.

[31] Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2), 141-153.

[32] Gilboa, I., & Schmeidler, D. (1993). Updating ambiguous beliefs. *Journal of Economic Theory*, 59(1), 33-49.

[33] Gilboa, I., & Marinacci, M. (2016). Ambiguity and the Bayesian paradigm. In *Readings in Formal Epistemology* (pp. 385-439). Springer, Cham.

[34] Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1), 112-49.

[35] Guarino, A., & Jehiel, P. (2013). Social learning with coarse inference. *American Economic Journal: Microeconomics*, 5(1), 147-74.

[36] Hanany, E., & Klibanoff, P. (2009). Updating ambiguity averse preferences. *The BE Journal of Theoretical Economics*, 9(1).

[37] Hansen, L., & Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review*, 91(2), 60-66.

[38] Herrera, H., & Hörner, J. (2012). A Necesssary and Sufficient Condition for Information Cascades. Working paper.

[39] Herrera, H., & Hörner, J. (2013). Biased social learning. *Games and Economic Behavior*, 80, 131-146.

[40] Hurwicz, L. (1951). Some specification problems and applications to econometric models. *Econometrica*, 19(3), 343-344.

[41] Kalai, E., & Lehrer, E. (1993). Rational learning leads to Nash equilibrium. *Econometrica*, 1019-1045.

[42] Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, 73(6), 1849-1892.

[43] Klibanoff, P., Marinacci, M., & Mukerji, S. (2009). Recursive smooth ambiguity preferences. *Journal of Economic Theory*, 144(3), 930-976.

[44] Li, W., & Tan, X. (2020). Locally Bayesian learning in networks. *Theoretical Economics*, 15(1), 239-278.

[45] Marinacci, M. (2002). Learning from ambiguous urns. *Statistical Papers*, 43(1), 143.

[46] Marinacci, M. (2015). Model uncertainty. *Journal of the European Economic Association*, 13(6), 1022-1100.

[47] Marinacci, M., & Massari, F. (2019). Learning from ambiguous and misspecified models. *Journal of Mathematical Economics*, 84, 144-149.

[48] Molavi, P., Tahbaz-Salehi, A., & Jadbabaie, A. (2018). A theory of non-Bayesian social learning. *Econometrica*, 86(2), 445-490.

[49] Machina, M. J., & Siniscalchi, M. (2014). Ambiguity and ambiguity aversion. In *Handbook of the Economics of Risk and Uncertainty* (Vol. 1, pp. 729-807). North-Holland.

[50] Pires, C. P. (2002). A rule for updating ambiguous beliefs. Theory and Decision, 53(2), 137-152.

[51] Rosenberg, D., & Vieille, N. (2019). On the efficiency of social learning. *Econometrica*, 87(6), 2141-2168.

[52] Saito, K. (2012). Subjective timing of randomization and ambiguity. mimeo.

[53] Siniscalchi, M. (2008). Ambiguity and ambiguity aversion. In S. N. Durlauf & L. Blume (Eds.), *The New Palgrave Dictionary of Economics*, Vol. 1 (2nd ed., pp. 138–142). NY: Palgrave Macmillan.

[54] Smith, L., & Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*, 68(2), 371-398.

A Proofs

A.1 Proof of Theorem 1

Lemma 1. *For all data-generating process $F \equiv \left(F^0, F^1\right)$, we have:*

(1) $F^0(r) > F^1(r)$ except when both are equal to 0 or 1;

(2) $\frac{F^0(r)}{F^1(r)} \geq \frac{1}{r}$ and $\frac{1-F^1\left(\frac{1}{r}\right)}{1-F^0\left(\frac{1}{r}\right)} \geq \frac{1}{r}$ for $r \in (0,\infty)$ (strictly when $F^1(r) > 0$ and $F^0\left(\frac{1}{r}\right) < 1$);

(3) $\frac{F^0(r)}{F^1(r)}$ and $\frac{1-F^1\left(\frac{1}{r}\right)}{1-F^0\left(\frac{1}{r}\right)}$ are weakly decreasing (strictly on $supp(F)$).

Proof. See Smith and Sørensen (2000)'s Lemma A.1. □

Lemma 2. *Define $C_0 = \left[0, \frac{1}{\gamma}\right]$ and $C_1 = [\gamma, \infty]$. Whenever $r_i \in C_\theta$, an information cascade of action θ occurs.*

Proof. It follows directly from the definition of the information cascade. □

Step 1: The increment (or the decrement) of r_i is bounded by some constant.

Lemma 3. *Under Assumption 3, for all $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$, we have*

$$\frac{r_{i+1}}{r_i} \begin{cases} \geq \sqrt{\gamma} & \text{if } a_i = 1 \\ \leq \frac{1}{\sqrt{\gamma}} & \text{if } a_i = 0 \end{cases}.$$

Proof. Suppose that $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$, hence a cascade does not occur. If $a_i = 1$, we have

$$r_{i+1} = \sqrt{\sup_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)} \times \inf_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)}} \times r_i.$$

Let F_γ be the data-generating process such that $\text{supp}(F_\gamma) = \left\{\gamma, \frac{1}{\gamma}\right\}$. Intuitively, F_γ is the "most informative" data-generating process that only generates signals with the highest and the lowest likelihood ratios. For all $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$, we have

$$\sup_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)} \geq \frac{1 - F_\gamma^1\left(\frac{1}{r_i}\right)}{1 - F_\gamma^0\left(\frac{1}{r_i}\right)} = \frac{\mathbb{P}_{F_\gamma}^1(\gamma)}{\mathbb{P}_{F_\gamma}^0(\gamma)} = \gamma, \tag{2}$$

where $\mathbb{P}_{F_\gamma}^\theta(\gamma)$ denotes the probability of observing a signal γ in state θ, the first equality comes from $\text{supp}(F_\gamma) = \left\{\gamma, \frac{1}{\gamma}\right\}$, and the last equality comes from the definition of normalized signals.

From Lemma 1, we know that

$$\inf_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)} \geq 1. \tag{3}$$

Combining (2) and (3), we obtain $r_{i+1} \geq \sqrt{\gamma} \times r_i$ when $a_i = 1$. The discussion for $a_i = 0$ is symmetric. □

Step 2: An information cascade occurs almost surely $\mathbb{P}^*$.

Unbounded Signals. When signals are unbounded, the occurrence of a cascade is easy to see. This is because when $\gamma = \infty$, which Lemma 3 implies that

$$r_1 = \begin{cases} \infty & \text{if } a_1 = 1 \\ 0 & \text{if } a_1 = 0 \end{cases},$$

so a cascade occurs immediately after the first action.

Bounded Signals. When signals are bounded, Lemma 3 also implies that we can find some $K < \infty$ such that for all $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$, K consecutive action θs will lead r_i to enter cascade set θ, hence triggering an information cascade of action θ. Specifically, whenever $r_i \geq 1$, K consecutive signals $\lambda_i, \lambda_{i+1}, ..., \lambda_{i+K-1} > 1$ lead to $a_i = a_{i+1} = ... = a_{i+K-1} = 1$ and result in a cascade of action 1 afterwards. Further note that the probability of receiving a signal $\lambda_i > 1$ satisfies

$$\frac{\mathbb{P}^*(\lambda_i > 1)}{1 - \mathbb{P}^*(\lambda_i > 1)} = \frac{1 - \overline{F}^0(1)}{\overline{F}^0(1)} = \frac{\overline{F}^1(1)}{\overline{F}^0(1)} \geq \lim_{r \to \frac{1}{\gamma}} \frac{\overline{F}^1(r)}{\overline{F}^0(r)} = \frac{1}{\gamma},$$

where the second equality comes from the symmetry of signals, and the inequality comes from the Lemma 1 (iii). As a result, we have $\mathbb{P}^*(\lambda_i > 1) \geq \frac{1}{1+\gamma}$, and

$$\mathbb{P}^*(\text{Cascade}|r_i \geq 1) \geq \mathbb{P}^*(\lambda_i, \lambda_{i+1}, ..., \lambda_{i+K-1} > 1 | r_i \geq 1) \geq \left(\frac{1}{1+\gamma}\right)^K > 0.$$

Symmetrically, we also have $\mathbb{P}^*(\text{Cascade}|r_i < 1) \geq \left(\frac{\gamma}{1+\gamma}\right)^K > 0$. Therefore, for all possible history h_i, we have $\mathbb{P}^*(\text{Cascade}|h_i) \geq \varepsilon$ for some $\varepsilon > 0$. Levy's 0-1 Law shows that as $i \to \infty$, we $\mathbb{P}^*$-almost surely have

$$\mathbb{P}^*(\text{Cascade}|h_i) \to \mathbb{P}^*(\text{Cascade}|h_\infty) = 1_{\text{Cascade}} \in \{0, 1\}.$$

Recall that $\mathbb{P}^*(\text{Cascade}|h_i) > \varepsilon > 0$ for all i, so $1_{\text{Cascade}} = 1$ $\mathbb{P}^*$-almost surely, in other words, an information cascade almost surely happens.

A.2 Proof of Theorem 2

Condition (1): Suppose that there exists some $F \in \mathcal{F}_0$ that is discrete at γ. Due to the symmetry, F^0 is discrete at $\frac{1}{\gamma}$. Denote $p = \mathbb{P}_{F^0}\left(\frac{1}{\gamma}\right) > 0$, which is the probability that F^0 puts on $\frac{1}{\gamma}$. Suppose that $a_i = 1$, for $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$, we have:

$$\bar{l}_{i+1} = \bar{l}_i \times \sup_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)} \geq \bar{l}_i \times \frac{1 - F^1\left(\frac{1}{r_i}\right)}{1 - F^0\left(\frac{1}{r_i}\right)} \geq \bar{l}_i \cdot \left[\lim_{r \to \gamma} \frac{1 - F^1\left(\frac{1}{r}\right)}{1 - F^0\left(\frac{1}{r}\right)}\right] = \bar{l}_i \cdot \frac{1 - \frac{1}{\gamma} \cdot p}{1 - p}, \quad (4)$$

where the inequality line comes from Property (3) in Lemma 1, and the last equality comes from the discreteness of signals. Besides, we have $\underline{l}_{i+1} \geq \underline{l}_i$, so

$$r_{i+1} \geq \sqrt{\frac{1 - \frac{1}{\gamma} \cdot p}{1 - p}} r_i \equiv \beta \times r_i$$

Symmetrically, when $a_i = 0$, we have $r_{i+1} \leq \frac{1}{\beta} \times r_i$. From the proof of Theorem 1, an information cascade occurs $\mathbb{P}^*$-almost surely.

Condition (2): Suppose that there exists some $F^1 \in \mathcal{F}_0$ such that F^1 is continuously differentiable on $(\gamma - \varepsilon, \gamma)$ with $F^{1'}(\gamma^-) > \frac{2}{\gamma - 1}$. When F^1 is discrete at γ, an information cascade occurs almost surely as implied by condition (1). I thus only focus on the case where F^1 is continuous at γ. Suppose that $a_i = 1$, we have:

$$r_{i+1} = r_i \cdot \sqrt{\sup_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)} \cdot \inf_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)}} \geq r_i \cdot \sqrt{\frac{1 - F^1\left(\frac{1}{r_i}\right)}{1 - F^0\left(\frac{1}{r_i}\right)}} \equiv I(r_i)$$

Let $I'(\gamma) \equiv \lim_{\delta \to 0} I'(\gamma - \delta)$ and $f^\theta(\gamma) \equiv F^{\theta'}(\gamma^-)$. It is easy to verify

$$I'(\gamma) = \gamma \cdot \left[\frac{1}{\gamma} + \frac{1}{2}\left(f^0(\gamma) - f^1(\gamma)\right)\right] = 1 - \left(\frac{\gamma - 1}{2}\right) f^1(\gamma) < 0,$$

where the last equality comes from $f^0(\gamma) = \frac{1}{\gamma} f^1(\gamma)$. Because F^1 is continuously differentiable on $(\gamma - \varepsilon, \gamma)$, there exists some $\varepsilon_0 > 0$ such that for all $r \in [\gamma - \varepsilon_0, \gamma)$, $I'(r) < 0$. Since $I(\gamma) = \gamma$, we have $I(r) \geq \gamma$ for all $r \in [\gamma - \varepsilon_0, \gamma]$. For all $r_i \in \left(\frac{1}{\gamma - \varepsilon_0}, \gamma - \varepsilon_0\right)$, if $a_i = 1$, we have:

$$\frac{r_{i+1}}{r_i} \geq \sqrt{\frac{1 - F^1\left(\frac{1}{r_i}\right)}{1 - F^0\left(\frac{1}{r_i}\right)}} \geq \sqrt{\frac{1 - F^1\left(\frac{1}{\gamma - \varepsilon_0}\right)}{1 - F^0\left(\frac{1}{\gamma - \varepsilon_0}\right)}} > 1.$$

So, for all r_i, there exists a $K < \infty$ such that after K action 1s, we have $r_i \geq \gamma - \varepsilon_0$. Also note that if $r_i \in [\gamma - \varepsilon_0, \gamma]$ and $a_i = 1$, we have $r_{i+1} \geq I(r_i) \geq \gamma$, so $K + 1$ consecutive action 1s will trigger a cascade of action 1. Similarly, $K + 1$ consecutive action 0s will trigger a cascade of action

0. Applying the proof of Theorem 1 again, we can show that r_i will enter the cascade set almost surely.

A.3 Proof of Corollary 1

The idea of the proof is to make use of condition (1) in Theorem 2. As mentioned in the main text: weak convergence implies that we can construct a F that is discrete at γ and is sufficiently close to G (under d). Below is the explicit construction.

Construction of a Discrete Approximation of G. Let $x_0^n \equiv 1$, $\Delta_n \equiv \frac{\gamma-1}{n}$, $x_i^n \equiv x_0^n + \Delta_n \cdot i$, note that $x_n^n = \gamma$. Consider the following partition:

$$\tau^n = \left\{ [\frac{1}{x_n^n}, \frac{1}{x_{n-1}^n}), ..., [\frac{1}{x_2^n}, \frac{1}{x_1^n}), [\frac{1}{x_1^n}, 1], (1, x_1^n], ..., (x_{n-1}^n, x_n^n] \right\}$$

Since the benchmark distribution $G \in \mathcal{F}$ (recall that $G(x) = \mathbb{P}_G(\lambda \leq x|1)$, that is G is the data-generating process in state 1 by definition). We have:

$$\begin{aligned} G(x_i^n) - G\left(x_{i-1}^n\right) &= G^0\left(\frac{1}{x_{i-1}^n}\right) - G^0\left(\frac{1}{x_i^n}\right) = \int_{\frac{1}{x_i^n}}^{\frac{1}{x_{i-1}^n}} dG^0(\lambda) \\ &= \int_{\frac{1}{x_i^n}}^{\frac{1}{x_{i-1}^n}} \frac{1}{\lambda} dG(\lambda) = \frac{1}{v_i^n} \cdot \left[G\left(\frac{1}{x_{i-1}^n}\right) - G\left(\frac{1}{x_i^n}\right) \right] \quad \text{for some } v_i^n \in [x_{i-1}^n, x_i^n) \end{aligned}$$

where G^0 is the corresponding data-generating process in state 0. So we have:

$$v_i^n = \frac{G\left(\frac{1}{x_{i-1}^n}\right) - G\left(\frac{1}{x_i^n}\right)}{G(x_i^n) - G\left(x_{i-1}^n\right)}$$

Define the following cutoff points:

$$\varrho^n = \left\{ \frac{1}{\gamma}, \frac{1}{v_{n-1}^n}, \frac{1}{v_{n-2}^n}, ..., v_{n-2}^n, v_{n-1}^n, \gamma \right\}$$

Construct a discrete distribution $\mathbb{P}_n(.|1)$ that puts all the mass on the elements of ϱ^n:

(1) For all $i \leq n-1$, let

$$\begin{aligned} \mathbb{P}_n(v_i^n|1) &= G\left(\frac{1}{x_{i-1}^n}\right) - G\left(\frac{1}{x_i^n}\right) \\ \mathbb{P}_n\left(\frac{1}{v_i^n}|1\right) &= G(x_i^n) - G\left(x_{i-1}^n\right) \end{aligned}$$

Let $\mathbb{P}_n(v_i^n|0) = \frac{1}{v_i^n} \cdot \mathbb{P}_n(v_i^n|1)$.

(2) For $i = n$, let

$$\mathbb{P}_n\left(\frac{1}{\gamma}|1\right) = \frac{\gamma}{1+\gamma}\left(1 - G\left(x_{n-1}^n\right) + G\left(\frac{1}{x_{n-1}^n}\right)\right)$$
$$\mathbb{P}_n\left(\gamma|1\right) = \frac{1}{1+\gamma}\left(1 - G\left(x_{n-1}^n\right) + G\left(\frac{1}{x_{n-1}^n}\right)\right)$$

Let $\mathbb{P}_n\left(\gamma|0\right) = \frac{1}{\gamma}\mathbb{P}_n\left(\gamma|1\right)$ and $\mathbb{P}_n\left(\frac{1}{\gamma}|0\right) = \mathbb{P}_n\left(\frac{1}{\gamma}|1\right)\gamma$.

The idea of this construction is: we assign all the weights of G in the interval $[x_{i-1}^n, x_i^n)$ (and $[\frac{1}{x_i^n}, \frac{1}{x_{i-1}^n})$) on the cutoff point v_i^n (or $1/v_i^n$). Let F_n be the c.d.f. of $\mathbb{P}_n\left(.|1\right)$. It can be verified that:

(1) $F_n \in \mathcal{F}$, since by construction, $\mathbb{P}_n$ is a symmetric signal-generating process (with the data-generating process on state 0 given by $\mathbb{P}_n\left(.|0\right)$;

(2) $F_n \Rightarrow G$, since as $n \to \infty$, the division becomes finer and finer, the distance between $F_n\left(.\right)$ and $G\left(.\right)$ shrinks to 0 (i.e., $d\left(F_n, G\right) \to 0$).

A.4 Proof of Corollary 2

Proof. The idea of this proof makes use of condition (2) in Theorem 2. I prove this corollary by constructing a $F \in \mathcal{F}$ satisfying condition (2) and satisfy $d\left(F, G\right) < \infty$. Then we just need to set $\overline{K} \equiv d\left(F, G\right)$. As in the example 2, I deal with a signal space $S \equiv [0, 1]$ and consider the following h:

$$h^1\left(s\right) = \begin{cases} 1 + 2\varepsilon\left(1+\gamma\right)\cdot s & s \in \left[0, \frac{1}{1+\gamma}\right] \\ 2\varepsilon\left(1+\gamma\right)\cdot s + \left(1-2\varepsilon\right)\gamma - 2\varepsilon & s \in \left[\frac{\gamma}{1+\gamma}, 1\right] \end{cases}$$
$$h^0\left(s\right) = \begin{cases} 2\varepsilon\left(1+\gamma\right)\cdot\left(1-s\right) + \left(1-2\varepsilon\right)\gamma - 2\varepsilon & s \in \left[0, \frac{1}{1+\gamma}\right] \\ 1 + 2\varepsilon\left(1+\gamma\right)\cdot\left(1-s\right) & s \in \left[\frac{\gamma}{1+\gamma}, 1\right] \end{cases}$$

where $h^\theta\left(s\right)$ is the p.d.f. of the data-generating process in state θ and $\varepsilon > 0$. For a signal s, the likelihood ratio induced by it is $\lambda\left(s\right) = \frac{h^1(s)}{h^0(s)}$. By changing the variable (from s to λ), we can equivalently express $h^1\left(s\right)$ as a p.d.f. $f^1\left(\lambda\right)$, where f has support $\left[\frac{1}{\gamma}, \frac{1+2\varepsilon}{\gamma-2\varepsilon}\right] \cup \left[\frac{\gamma-2\varepsilon}{1+2\varepsilon}, \gamma\right]$. Besides $F \in \mathcal{F}$ because it represents a symmetric data-generating process. Due to the full supportness of G, we have: $d\left(F, G\right) < \infty$ for all $\varepsilon > 0$.

For $s \in \left[\frac{\gamma}{1+\gamma}, 1\right]$, the normalized signal is

$$\lambda = \frac{h^1\left(s\right)}{h^0\left(s\right)} = \frac{2\varepsilon\left(1+\gamma\right)\cdot s + \gamma - 2\varepsilon\left(1+\gamma\right)}{1 + 2\varepsilon\left(1+\gamma\right)\cdot\left(1-s\right)}$$

so

$$s = \frac{[1+2\varepsilon(1+\gamma)]\lambda - \gamma + 2\varepsilon(1+\gamma)}{2\varepsilon(1+\gamma)(\lambda+1)} = \frac{(1+\rho)\lambda - \gamma + \rho}{\rho(\lambda+1)} \text{ where } \rho = 2\varepsilon(1+\gamma)$$
$$\frac{ds}{d\lambda} = \frac{(1+\rho)\rho(\lambda+1) - \rho[(1+\rho)\lambda - \gamma + \rho]}{[\rho(\lambda+1)]^2} = \frac{1+\gamma}{\rho(1+\lambda)^2}$$

the transformed PDF $f^1(\lambda)$ becomes:

$$f^1(\lambda) = \left[\rho \times \frac{(1+\rho)\lambda - \gamma + \rho}{\rho(\lambda+1)} + \gamma - \rho\right] \times \frac{1+\gamma}{\rho(1+\lambda)^2}$$
$$F'(\gamma-) = \lim_{\lambda\to\gamma} f^1(\lambda) = \frac{\gamma}{1+\gamma}\frac{1}{\rho} = \frac{\gamma}{2(1+\gamma)^2\varepsilon}$$

It is easy to see that there exists some $\overline{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \overline{\varepsilon})$, $F'(\gamma-) > \frac{2}{\gamma-1}$, so condition (2) of Theorem 2 is satisfied. Let's just set $\varepsilon = \overline{\varepsilon}/2$ and F is the corresponding distribution function. Let $\overline{K} = d(F, G)$, it is easy to verify that $d(F, G) < \infty$ under the assumptions of G. When $K \geq \overline{K}$, the belief set $\mathcal{F}_0$ satisfies the condition (2) thus an information cascade occurs almost surely. □

A.5 Proof of Theorem 3

Auxiliary Results: Local Stability under Ambiguity

I first introduce the notion of local stability under ambiguity. Following concepts and results are parallel to those in Bayesian learning, especially learning with misspecified model.

Definition 7. [Local (un)stability under Ambiguity]

(i) State 0 (or state 1) is *locally stable* if there exists some $r \in \mathbb{R}_{++}$ (or $R \in \mathbb{R}_{++}$) and $\varepsilon > 0$ such that $\mathbb{P}_{r_0}(r_i \to 0) > \varepsilon$ (or $\mathbb{P}_{r_0}(r_i \to \infty) > \varepsilon$) for all prior sets Π_0 with average likelihood ratio $r_0 < r$ (or $r_0 > R$).

(ii) State 0 (or state 1) is *locally unstable* if there exists some $r \in \mathbb{R}_{++}$ (or $R \in \mathbb{R}_{++}$) such that $\mathbb{P}_{r_0}(r_i > r) = 1$ (or $\mathbb{P}_{r_0}(r_i < R) = 1$) for all prior sets Π_0 with average likelihood ratio $r_0 < r$ (or $r_0 > R$).

Intuitively speaking, state θ is locally stable if beliefs will converge to δ_θ with a strictly positive probability for all priors within a neighborhood of δ_θ. Symmetrically, priors are locally unstable if beliefs will escape from a small neighborhood almost surely. We have the following results.

Lemma 4. *Under the assumptions of Theorem 3, a herding of action* 0 *(or* 1*) occurs if and only if* $r_i \to 0$ *(or* $r_i \to \infty$*).*

Proof. Due to the symmetry, I only prove the result for a herding of action 1.

"If" part. Suppose that $r_i \to \infty$, we must have a herding of action 1, since if an action 0 is observed, we have

$$r_{i+1} = r_i \times \sqrt{\sup_{F_i \in \mathcal{F}_0} \frac{F_i^1(1/r_i)}{F_i^0(1/r_i)} \times \inf_{F_i \in \mathcal{F}_0} \frac{F_i^1(1/r_i)}{F_i^0(1/r_i)}} \leq r_i \times \sqrt{\frac{1}{r_i} \times \frac{1}{r_i}} = 1,$$

where the inequality comes from Lemma 1 (2), contradicting $r_i \to \infty$.

"Only if" part. Suppose that a herding of action 1 occurs. Lemma 1 (1) implies that

$$r_{i+1} = r_i \times \sqrt{\sup_{F \in \mathcal{F}_0} \frac{1 - F_i^1(1/r_i)}{1 - F_i^0(1/r_i)} \times \inf_{F \in \mathcal{F}_0} \frac{1 - F_i^1(1/r_i)}{1 - F_i^0(1/r_i)}} \geq r_i,$$

so $\{r_i\}$ is an increasing sequence, hence it converges in $\mathbb{R} \cup \{+\infty\}$. If r_i does not converge to infinity, it must converge to some $R < \infty$. Let F be the crucial DGP that $\mathcal{F}_0$ contains. We then have

$$r_{i+1} \geq r_i \times \sqrt{\frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)}}. \tag{5}$$

Take limit on both sides of (5), we obtain $R \geq \sqrt{\frac{1-F^1(1/R)}{1-F^0(1/R)}} \times R$, so $\sqrt{\frac{1-F^1(1/R)}{1-F^0(1/R)}} \leq 1$. However, since F has unbounded signals, Lemma 1 (1) implies that $\sqrt{\frac{1-F^1(1/R)}{1-F^0(1/R)}} > 1$ when $R < \infty$, which is a contradiction. As a consequence, $r_i \to \infty$. □

Lemma 5. *If both* 0 *and* 1 *are locally stable, then (i) herding occurs almost surely, and (ii) an incorrect herding occurs with a strictly positive probability.*

Proof. (ii) follows directly from Lemma 4 and the definition of local stability. It remains to prove that herding occurs almost surely. Since state 1 and 0 are locally stable, once beliefs enter $C = \{r_i < r\} \cup \{r_i > R\}$, it will remain in C with a strictly positive probability. Denote by $H = \{r_i \to 0\} \cup \{r_i \to \infty\}$, which represents the event of herding by Lemma 4. Notice that whenever r_i is not in C, we know that $r_i \in [r, R]$ is bounded, so K consecutive actions lead beliefs to enter C, which is positive-probability event.[14] After beliefs enter C, with a strictly positive probability, we either have $r_i \to 0$ or $r_i \to \infty$ depending on which neighborhood r_i enters. In other words, a herding will occur with a strictly positive probability. As a result, we can find a constant $\varepsilon' > 0$ such that for all possible history h_i, $\mathbb{P}^*(H|h_i) > \varepsilon'$. Applying the Levy's 0-1 Law, $\mathbb{P}^*(H|h_i) \to \mathbb{P}^*(H|h_\infty) = 1_H \in \{0, 1\}$, so H is a probability-1 event. □

From Lemma 5, we know that we only need to establish the local stability of both states to prove Theorem 3.

[14] On $\{r_i \leq R\}$, we have $r_{i+1} \leq r_i \times \sqrt{\frac{F^1(1/r_i)}{F^0(1/r_i)}} \leq r_i \times \sqrt{\frac{F^1(1/R)}{F^0(1/R)}}$ for any $F \in \mathcal{F}_0$ after an action 0. Hence, $r_{i+1}/r_i \leq \sqrt{\frac{F^1(1/R)}{F^0(1/R)}} < 1$, so the decrement is bounded by some constant less than 1. Since $r_i < R < \infty$, finite steps will make $r_i < r$. The case for $\{r_i \geq r\}$ is symmetric.

Step 1: Establish the Local Stability of State 1

To show that state 1 is locally stable, we need to show that there exists some $R < \infty$ such that for all $r_0 \geq R$, the probability of an action-1 herding is greater than some $\varepsilon > 0$. Recall that

$$\mathbb{P}^0_{r_0}(Herd_1) = \lim_{i\to\infty} \mathbb{P}^0_{r_0}(a_1 = a_2 = ...a_i = 1) = \prod_{i=1}^{\infty}\left[1 - F_i^0\left(\frac{1}{r_i}\right)\right] \geq \prod_{i=1}^{\infty}\left[1 - a \times \left(\frac{1}{r_i}\right)^{\alpha}\right], \quad (6)$$

where r_i represents the average public likelihood ratio after $h_i = (1, 1, ..., 1)$. Recall that

$$r_{i+1} = r_i \times \sqrt{\sup_{F\in\mathcal{F}_0} \frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)} \times \inf_{F\in\mathcal{F}_0} \frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)}} \geq r_i \times \sqrt{\frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)}},$$

where F denotes the model in $\mathcal{F}_0$ such that $x^p = o\left(F^0(x)\right)$. We first state the following lemma.

Lemma 6. $\sqrt{G_F(1/x)} = \sqrt{\frac{1-F^1(x)}{1-F^0(x)}} \sim 1 + \frac{1}{2}F^0(x)$ *as* $x \to 0$.

Proof. In Rosenberg and Vieille (2019), they showed that

$$\frac{1 - F^1(x)}{1 - F^0(x)} = 1 + F^0(x) + o\left(F^0(x)\right)$$

or equivalently, $\frac{1-F^1(x)}{1-F^0(x)} \sim 1 + F^0(x)$, so $\sqrt{\frac{1-F^1(x)}{1-F^0(x)}} \sim \sqrt{1 + F^0(x)} = 1 + \frac{1}{2}F^0(x) + o\left(F^0(x)\right)$, which proves the lemma. □

Let $q \in (p, \alpha)$ and consider the following limit.

$$\begin{aligned}
\lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1(1/r)}{1-F^0(1/r)}} - 1}{\left(1 + \frac{1}{r^q}\right)^{1/q} - 1} &= \lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1(1/r)}{1-F^0(1/r)}} - 1}{\frac{1}{r^q}} \times \lim_{r\to\infty} \frac{\frac{1}{r^q}}{\left(1 + \frac{1}{r^q}\right)^{1/q} - 1} \\
&= \lim_{r\to\infty} \frac{\frac{1}{2}F^0(1/r)}{\frac{1}{r^q}} \times \lim_{r\to\infty} \frac{\frac{1}{r^q}}{\left(1 + \frac{1}{r^q}\right)^{1/p} - 1} \qquad (7) \\
&> \lim_{r\to\infty} \frac{\frac{1}{2}(1/r)^p}{\frac{1}{r^q}} \times q = \infty,
\end{aligned}$$

where (7) follows from Lemma 6. From the proof of Lemma 4, we know that $\{r_i\}$ is increasing during an action-1 herd, so $r_i \geq R$ for all i. Therefore, we can choose R to be sufficiently large such that for all $i \geq 0$,

$$\sqrt{\frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)}} \geq \left(1 + \frac{1}{r_i^q}\right)^{1/q},$$

which further implies that

$$r_{i+1} \geq r_i \times \sqrt{\frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)}} \geq r_i \times \left(1 + \frac{1}{r_i^q}\right)^{1/q} = \left(r_i^q + 1\right)^{1/q}.$$

After iterations, we can obtain

$$r_i \geq \left(r_0^q + i\right)^{1/q}, \qquad \forall i \geq 1. \tag{8}$$

After substituting (8) into (6), we know that for all $r_0 \geq R$,

$$\mathbb{P}_{r_0}^0\left(Herd_1\right) \geq \prod_{i=1}^{\infty}\left[1 - a \times \left(\frac{1}{r_i}\right)^{\alpha}\right] \geq \prod_{i=1}^{\infty}\left[1 - a \times \frac{1}{\left(r_0^q + i\right)^{\alpha/q}}\right] \geq \prod_{i=1}^{\infty}\left[1 - a \times \frac{1}{\left(R^q + i\right)^{\alpha/q}}\right].$$

Here, we also choose the R to be sufficiently large such that $1 - a \times \frac{1}{R^{\alpha}} > 0$, so $1 - a \times \frac{1}{(R^q+i)^{\alpha/q}} \in (0,1)$ for all $i \geq 1$. Notice that the infinite product $\prod_{i=1}^{\infty}\left[1 - a \times \frac{1}{(R^q+i)^{\alpha/q}}\right] > 0$ if and only if the infinite series $\sum a \times \frac{1}{(R^q+i)^{\alpha/q}} < \infty$. Since $q < \alpha$, we know that $\sum a \times \frac{1}{(R^q+i)^{\alpha/q}} < \infty$, hence convergent, so

$$\mathbb{P}_{r_0}^0\left(Herd_1\right) \geq \prod_{i=1}^{\infty}\left[1 - a \times \frac{1}{\left(R^q + i\right)^{\alpha/q}}\right] \equiv \varepsilon > 0,$$

which establishes the local stability of state 1.

Step 2: Establish the Local Stability of State 0

The case for state 0 is symmetric to Step 1. Let r_i denotes the average likelihood ratio after $h_i = (0, ..., 0)$. From symmetry, we have

$$\mathbb{P}_{r_0}^0\left(Herd_0\right) = \prod_{i=1}^{\infty} F^0\left(\frac{1}{r_i}\right) = \prod_{i=1}^{\infty}\left[1 - F^1\left(r_i\right)\right] \geq \prod_{i=1}^{\infty}\left[1 - F^0\left(r_i\right)\right] = \mathbb{P}_{1/r_0}^0\left(Herd_1\right).$$

Roughly speaking, this relation says that the probability of a correct herding is higher than that of an incorrect herding. The intuition is straightforward, as the society is receiving some information, so the action is more likely to be correct than incorrect. From Step 1, there exists R such that $\mathbb{P}_{1/r_0}^0\left(Herd_1\right) \geq \varepsilon > 0$ for all $1/r_0 > R$. Let $r = 1/R$, so we also have $\mathbb{P}_{r_0}^0\left(Herd_0\right) \geq \varepsilon > 0$ for all $r_0 < r$, which establishes the local stability of state 0.

A.6 Proof of Theorem 4

A.6.1 Local stability and complete learning

Lemma 7. *Complete learning occurs if and only if $r_i \to 0$ with probability 1.*

Proof. It follows by the definition of complete learning. First, during complete learning, there must be a herding of action 0 after some point, so $r_i \to 0$ with probability 1. Second, if $r_i \to 0$ with probability 1, a herding of action 0 will eventually occur, since an action 1 will lead to $r_i \geq 1$. □

Lemma 8. *Complete learning occurs if* 0 *is locally stable and state* 1 *is locally unstable, only if state* 0 *is not locally unstable and state 1 is not locally stable.*

Proof. The proof is similar to the proof of Lemma 5. (i) "if" part. Since state 1 is locally unstable, beliefs will enter $\{r_i < R\}$ infinitely many often. Whenever $r_i < R$, we can find a finite K such that K consecutive action 0s lead to $r_i < r$, and this probability is greater than some positive constant. From the facts that state 0 is locally stable and that the process $\{r_i\}$ is a Markov process, whenever $\{r_I < r\}$ for some I, $r_i \to 0$ with a probability greater than ε. As a consequence, we can find a constant $\varepsilon' > 0$ such that for all possible history h_i, $\mathbb{P}(r_t \to 0|h_i) > \varepsilon'$. Applying the Levy's 0-1 Law as in the proof of Lemma 5, we know that complete learning occurs. (ii) "only if" part. If state 0 is locally unstable, beliefs will escape from the neighborhood around $r = 0$ with probability 1, which is inconsistent with complete learning. If state 1 is locally stable, we have $r_i \to \infty$ with a positive probability, which also contradicts complete learning. □

A.6.2 Proof of Theorem 4

Proposition 3. *Under Assumptions 5 and 6, we have:*

(a) if for all $F \in \mathcal{F}_0$, $\mathcal{P}(F) \geq \mathcal{P}\left(\overline{F}\right)$, *state* 1 *is locally unstable;*

(b) if there exists some $F \in \mathcal{F}_0$ *such that* $\mathcal{P}(F) < \mathcal{P}\left(\overline{F}\right)$, *state 1 is locally stable;*

(c) if for all $F \in \mathcal{F}_0$, $\mathcal{P}(F) \geq \mathcal{P}\left(\overline{F}\right) + 1$, *state 0 is locally unstable;*

(d) if there exists some $F \in \mathcal{F}_0$ *such that* $\mathcal{P}(F) < \mathcal{P}\left(\overline{F}\right) + 1$, *state 0 is locally stable.*

From Lemma 8, we know that Proposition 3 implies Theorem 4, so I now prove Proposition 3 as follows. For simplicity in notation, I define $\overline{\alpha} \coloneqq \mathcal{P}\left(\overline{F}\right)$, $\alpha_{max} \coloneqq \max_{F \in \mathcal{F}_0} \mathcal{P}(F)$ and $\alpha_{min} \coloneqq \min_{F \in \mathcal{F}_0} \mathcal{P}(F)$. The data-generating processes with the maximum and minimum power are denoted by F_{max} and F_{min}.

Proof of Proposition 3 (a)

To show that state 1 is locally unstable, it suffices to show that a herding of action 1 cannot occur for all priors r_0 sufficiently large.[15] Given r_0, the probability of a herding of action 1 is as follows

$$\lim_{i \to \infty} \mathbb{P}^0_{r_0}(a_1 = a_2 = ...a_i = 1) = \prod_{i=1}^{\infty} \mathbb{P}^0_{r_0}(a_i = 1|h_i) = \prod_{i=1}^{\infty} \left[1 - F^0\left(\frac{1}{r_i}\right)\right],$$

where r_i represents the average likelihood ratio after $h_i = (1, 1, ..., 1)$. The probability is equal to 0 if and only if $\sum F^0\left(\frac{1}{r_i}\right) = \infty$, or equivalently, $\sum \frac{1}{r_i^{\overline{\alpha}}} = \infty$. Note that $\{r_i\}$ is determined by the following dynamics

$$r_{i+1} = r_i \times \sqrt{\max_{F \in \mathcal{F}_0} \frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)} \times \min_{F \in \mathcal{F}_0} \frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)}}.$$

[15] In other words, action 0 occurs infinitely many often. Recall that after an action 0, we must have $r_i \leq 1$, so beliefs cannot remain in a small neighborhood around δ_1.

When r_0 is sufficiently large, we have $\frac{1-F^1(1/r_i)}{1-F^0(1/r_i)} \sim 1 + F^0(1/r_i)$ for all i, since $r_i \geq r_0$ is also sufficiently large. Therefore, we have

$$r_{i+1} = r_i \times \sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)} \times \frac{1-F^1_{max}(1/r_i)}{1-F^0_{max}(1/r_i)}} \leq r_i \times \frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)}$$

for all r_0 sufficiently large. By the definition of F_{min}, we have $\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)} \sim 1 + F^0_{min}(1/r_i) \sim 1 + C_{min} \times \frac{1}{r_i^{\alpha_{min}}}$, for some constant $C_{min} > 0$.

Suppose that all $F \in \mathcal{F}_0$, $\mathcal{P}(F) \geq \mathcal{P}\left(\overline{F}\right)$, in other words, $\alpha_{min} \geq \overline{\alpha}$. We have

$$\begin{aligned}
\lim_{r\to\infty} \frac{\frac{1-F^1_{min}(1/r)}{1-F^0_{min}(1/r)} - 1}{\left(1 + \frac{2\overline{\alpha}C_{min}}{r^{\overline{\alpha}}}\right)^{1/\overline{\alpha}} - 1} &= \lim_{r\to\infty} \frac{\frac{1-F^1_{min}(1/r)}{1-F^0_{min}(1/r)} - 1}{\frac{2\overline{\alpha}C_{min}}{r^{\overline{\alpha}}}} \times \frac{\frac{2\overline{\alpha}C_{min}}{r^{\overline{\alpha}}}}{\left(1 + \frac{2\overline{\alpha}C_{min}}{r^{\overline{\alpha}}}\right)^{1/\overline{\alpha}} - 1} \\
&= \lim_{r\to\infty} \frac{C_{min} \times \frac{1}{r^{\alpha_{min}}}}{\frac{2\overline{\alpha}C_{min}}{r^{\overline{\alpha}}}} \times \overline{\alpha} \\
&= \frac{1}{2} \times \lim_{r\to\infty} \frac{1}{r^{\alpha_{min}-\overline{\alpha}}} = \begin{cases} 0 & \alpha_{min} > \overline{\alpha} \\ \frac{1}{2} & \alpha_{min} = \overline{\alpha} \end{cases} < 1,
\end{aligned}$$

which implies that $\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)} < \left(1 + \frac{2\overline{\alpha}C_{min}}{r_i^{\overline{\alpha}}}\right)^{1/\overline{\alpha}}$ when r_0 is sufficiently large. Therefore, for all $i \geq 0$,

$$\begin{aligned}
r_{i+1} &< \left(1 + \frac{2\overline{\alpha}C_{min}}{r_i^{\overline{\alpha}}}\right)^{1/\overline{\alpha}} \times r_i = \left(r_i^{\overline{\alpha}} + 2\overline{\alpha}C_{min}\right)^{1/\overline{\alpha}} \\
r_{i+1} &< \left(r_{i+1}^{\overline{\alpha}} + 2\overline{\alpha}C\right)^{1/\overline{\alpha}} < \left(r_i^{\overline{\alpha}} + 2\overline{\alpha}C_{min} \times 2\right)^{1/\overline{\alpha}} \\
&\ldots \\
r_{i+t} &< \left(r_i^{\overline{\alpha}} + 2\overline{\alpha}C_{min} \times t\right)^{1/\overline{\alpha}}.
\end{aligned}$$

As a consequence, when r_0 is sufficiently large,

$$\sum_{i=1}^{\infty} \frac{1}{r_i^{\overline{\alpha}}} > \sum_{i=1}^{\infty} \frac{1}{r_0^{\overline{\alpha}} + 2\overline{\alpha}C \times i} = \infty,$$

so an herding of action 1 occurs with probability 0, which implies that state 1 is locally unstable.

Proof of Proposition 3 (b)

To show that state 1 is locally stable, we need to show that the probability of an action-1 herding is greater than some $\varepsilon > 0$ when r_0 is large. Recall that

$$\mathbb{P}^0_{r_0}(Herd_1) = \lim_{i\to\infty} \mathbb{P}^0_{r_0}(a_1 = a_2 = ...a_i = 1) = \prod_{i=1}^{\infty}\left[1 - F^0\left(\frac{1}{r_i}\right)\right],$$

so in order to establish local stability, we need to find a *uniform* lower bound of the probability on the RHS for all large r_0s.

Suppose that $F^0(x) \sim \overline{C} \times x^{\overline{\alpha}}$ for some constant $\overline{C} > 0$. When $r_0 \geq R$ with R sufficiently large, we have $\frac{F^0(x)}{\overline{C}\times x^{\overline{\alpha}}} \in [1-\varepsilon_1, 1+\varepsilon_1]$ for some $\varepsilon_1 > 0$, so

$$\mathbb{P}^0_{r_0}(Herd_1) = \prod_{i=1}^{\infty}\left[1 - F^0\left(\frac{1}{r_i}\right)\right] \geq \prod_{i=1}^{\infty}\left[1 - (1+\varepsilon_1)\times\overline{C}\times\frac{1}{r_i^{\overline{\alpha}}}\right]. \tag{9}$$

Here, R is sufficiently large such that the infinite product on the RHS is strictly positive. On the other hand, when R is large, we have

$$r_{i+1} = r_i \times \sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)} \times \frac{1-F^1_{max}(1/r_i)}{1-F^0_{max}(1/r_i)}}.$$

Define $\beta = (1-\varepsilon)\frac{C_{min}\times\alpha_{min}}{2}$ for some small $\varepsilon > 0$. We have

$$\begin{aligned}
\lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)} \times \frac{1-F^1_{max}(1/r_i)}{1-F^0_{max}(1/r_i)}} - 1}{\left(1+\frac{\beta}{r^{\alpha_{min}}}\right)^{1/\alpha_{min}} - 1} &= \lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)} \times \frac{1-F^1_{max}(1/r_i)}{1-F^0_{max}(1/r_i)}} - 1}{\sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)}} - 1} \times \lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)}} - 1}{\left(1+\frac{\beta}{r^{\alpha_{min}}}\right)^{1/\alpha_{min}} - 1} \\
&= 1 \times \lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)}} - 1}{\left(1+\frac{\beta}{r^{\alpha_{min}}}\right)^{1/\alpha_{min}} - 1} \\
&= \lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1_{min}(1/r_i)}{1-F^0_{min}(1/r_i)}} - 1}{\frac{\beta}{r^{\alpha_{min}}}} \times \lim_{r\to\infty} \frac{\frac{\beta}{r^{\alpha_{min}}}}{\left(1+\frac{\beta}{r^{\alpha_{min}}}\right)^{1/\alpha_{min}} - 1} \\
&= \frac{C_{min}\times\alpha_{min}}{2\beta} = \frac{1}{1-\varepsilon} > 1.
\end{aligned}$$

When R sufficiently large, we have

$$r_{i+1} \geq r_i \times \left(1 + \frac{\beta}{r_i^{\alpha_{min}}}\right)^{1/\alpha_{min}} = \left(r_i^{\alpha_{min}} + \beta\right)^{1/\alpha_{min}} \Rightarrow r_i \geq \left(r_0^{\alpha_{min}} + \beta\times i\right)^{1/\alpha_{min}}. \tag{10}$$

Similarly, $r_i \geq (r_0 + \beta \times i)^{1/\alpha_{min}}$. Combining (9) and (10), we obtain

$$\begin{aligned}
\mathbb{P}^0_{r_0}(Herd_1) &\geq \prod_{i=1}^{\infty}\left[1-(1+\varepsilon_1)\times\overline{C}\times\frac{1}{r_i^{\overline{\alpha}}}\right] \\
&\geq \prod_{i=1}^{\infty}\left[1-(1+\varepsilon_1)\times\overline{C}\times\frac{1}{(r_0^{\alpha_{min}}+\beta\times i)^{\overline{\alpha}/\alpha_{min}}}\right] \\
&\geq \prod_{i=1}^{\infty}\left[1-(1+\varepsilon_1)\times\overline{C}\times\frac{1}{(R^{\alpha_{min}}+\beta\times i)^{\overline{\alpha}/\alpha_{min}}}\right]
\end{aligned}$$

for all $r_0 \geq R$. Here, R is chosen to be sufficiently large such that each term is strictly positive. Suppose that there exists some $F \in \mathcal{F}_0$ such that $\mathcal{P}(F) < \mathcal{P}\left(\overline{F}\right)$, which implies that $\alpha_{min} < \overline{\alpha}$, so

$$\sum \frac{1}{(R^{\alpha_{min}}+\beta\times i)^{\overline{\alpha}/\alpha_{min}}} < \infty,$$

which further implies that

$$\mathbb{P}^0_{r_0}(Herd_1) \geq \prod_{i=1}^{\infty}\left[1-(1+\varepsilon_1)\times\overline{C}\times\frac{1}{(R^{\alpha_{min}}+\beta\times i)^{\overline{\alpha}/\alpha_{min}}}\right] =: \delta > 0,$$

for all $r_0 \geq R$. In other words, the probability of an action-1 herding is greater than $\delta > 0$, which proves that state 1 is locally stable.

Proof of Proposition 3 (c) & (d)

The proofs of Proposition 3 (c) and (d) are almost identical to the proofs of (a) and (b). The only difference is that the cutoff value becomes $\mathcal{P}\left(\overline{F}\right)+1$. To see where the difference arises, we note that the probability of an action-0 herd is as follows.

$$\mathbb{P}^0_{r_0}(Herd_0) = \lim_{i\to\infty}\mathbb{P}^0_{r_0}(a_1 = a_2 = ...a_i = 0) = \prod_{i=1}^{\infty}F^0\left(\frac{1}{r_i}\right) = \prod_{i=1}^{\infty}\left[1-F^1(r_i)\right],$$

where r_i denotes the average likelihood ratio after $h_i = (0,...,0)$. An action-0 herd occurs with a strictly positive probability if and only if $\sum F^1(r_i) < \infty$. During a herd of action 0, we have $r_i \to 0$. Besides, it can be verified that $\overline{F}^1(x) = O\left(x^{\overline{\alpha}+1}\right)$ as $x \to 0$. [16]Therefore, an action-0 herd occurs with a strictly positive probability if and only if $\sum r_i^{\overline{\alpha}+1} < \infty$. The rest of the proofs are exactly symmetric to those of (a) and (b).

[16]Recall that $\overline{F}^0(x) \sim \overline{C}\times x^{\overline{\alpha}}$ as $x \to 0$, so

$$\lim_{x\to0}\frac{\overline{F}^1(x)}{x^{\overline{\alpha}+1}} = \lim_{x\to0}\frac{\overline{f}^1(x)}{(\overline{\alpha}+1)x^{\overline{\alpha}}} = \frac{1}{\overline{\alpha}+1}\lim_{x\to0}\frac{\overline{f}^0(x)}{x^{\overline{\alpha}-1}} = \frac{\overline{\alpha}}{\overline{\alpha}+1}\lim_{x\to0}\frac{\overline{F}^0(x)}{x^{\overline{\alpha}}} = \frac{\overline{\alpha}}{\overline{\alpha}+1}\overline{C},$$

hence $\overline{F}^1(x) = O\left(x^{\overline{\alpha}+1}\right)$ as $x \to 0$.

A.7 Discussion of Example 5

To show that a cascade occurs with a strictly positive probability, it suffices to construct an example. For simplicity, I assume $\pi_0 = (2/3, 1/3)$ and $\overline{\gamma} > 2$, but examples can be constructed for general π_0 and $\overline{\gamma}$.

Suppose that $a_1 = a_2 = a_3 = 1$. It can be shown that an information cascade occurs for individual 4 when $|\sigma|$ is sufficiently large. Let $\lambda_i(\sigma)$ be the normalized signal such that individual i is indifferent between two actions. When $i = 1$, $\lambda_1(\sigma) = 2$. When $i = 2$, $\lambda_2(\sigma)$ is the solution to

$$V_2(1) - V_2(0) = \left[\int_1^{\overline{\gamma}} \left(\frac{\gamma_1 \lambda_2(\sigma)}{2 + \gamma_1 \lambda_2(\sigma)}\right)^{1-\sigma} h(d\gamma_1)\right]^{\frac{1}{1-\sigma}} - \left[\int_1^{\overline{\gamma}} \left(\frac{2}{2 + \gamma_1 \lambda_2(\sigma)}\right)^{1-\sigma} h(d\gamma_1)\right]^{\frac{1}{1-\sigma}} = 0.$$

When $i = 3$, $\lambda_3(\sigma)$ is the solution to

$$V_2(1) - V_2(0) = \left[\int_{\lambda_2(\sigma)}^{\overline{\gamma}} \int_1^{\overline{\gamma}} \left(\frac{\gamma_1 \gamma_2 \lambda_3(\sigma)}{2 + \gamma_1 \gamma_2 \lambda_3(\sigma)}\right)^{1-\sigma} h(d\gamma_1 d\gamma_2) + \int_1^{\lambda_2(\sigma)} \int_1^{\overline{\gamma}} \left(\frac{\gamma_1 \lambda_3(\sigma)}{2 + \gamma_1 \lambda_3(\sigma)}\right)^{1-\sigma} h(d\gamma_1 d\gamma_2)\right]^{\frac{1}{1-\sigma}} \tag{11}$$

$$- \left[\int_{\lambda_2(\sigma)}^{\overline{\gamma}} \int_1^{\overline{\gamma}} \left(\frac{2}{2 + \gamma_1 \gamma_2 \lambda_3(\sigma)}\right)^{1-\sigma} h(d\gamma_1 d\gamma_2) + \int_{\frac{1}{r_2(\sigma)}}^{\overline{\gamma}} \int_1^{\overline{\gamma}} \left(\frac{2}{2 + \gamma_1 \lambda_3(\sigma)}\right)^{1-\sigma} h(d\gamma_1 d\gamma_2)\right]^{\frac{1}{1-\sigma}},$$

where (11) comes from that when $\gamma_2 > \lambda_2(\sigma)$, individual 2 must receive a signal $s_2 = h$, but when $\gamma_2 < \lambda_2(\sigma)$, both signals can justify $a_2 = 1$. The expression for $\lambda_4(\sigma)$ can be written analogously. It is easy to see that both $\lambda_i(\sigma)$ is continuous in σ. Denote by $\lambda_i(\infty) = \lim_{\sigma \to \infty} \lambda_i(\sigma)$, which corresponds to the average likelihood ratio under the max-min EU model. It can be verified that

$$\lambda_2(\infty) = 2/\sqrt{\overline{\gamma}}, \quad \lambda_3(\infty) = 2/\overline{\gamma}, \quad \lambda_4(\infty) = 2/\overline{\gamma}^2,$$

so when σ is sufficiently large, $\lambda_4(\sigma)$ is sufficiently close to $2/\overline{\gamma}^2 < 1/\overline{\gamma}$, so an information cascade occurs.

When signals are unbounded, i.e., $\overline{\gamma} = \infty$.

The occurrence of a cascade also exists for unbounded signals, or $\overline{\gamma} = 0$. We have the following fact.

Fact 1. *Suppose that $h(\gamma) \sim C \times \frac{1}{\gamma^\alpha}$ for some $C, \alpha > 0$ as $\gamma \to \infty$.*

(i) When $\sigma = 0$, complete learning occurs.

(ii) When σ is sufficiently large, an information cascade occurs almost surely.

Proof. Suppose that $a_1 = 1$, and that individual 2 received an opposite signal, signal l, and that

her signal precision is γ_2. Her utility of each action is (the prior is assumed to be flat).

$$V_2(0) = \left[\int_1^\infty \left[\mathbb{P}_{\gamma_1}(\theta=0|I_2)\right]^{1-\sigma} h(\gamma_1)\, d\gamma_1\right]^{\frac{1}{1-\sigma}} = \left[\int_1^\infty \left[\frac{\gamma_2}{\gamma_1+\gamma_2}\right]^{1-\sigma} h(\gamma_1)\, d\gamma_1\right]^{\frac{1}{1-\sigma}}$$

$$V_2(1) = \left[\int_1^\infty \left[\mathbb{P}_{\gamma_1}(\theta=1|I_2)\right]^{1-\sigma} h(\gamma_1)\, d\gamma_1\right]^{\frac{1}{1-\sigma}} = \left[\int_1^\infty \left[\frac{\gamma_1}{\gamma_1+\gamma_2}\right]^{1-\sigma} h(\gamma_1)\, d\gamma_1\right]^{\frac{1}{1-\sigma}}.$$

Individual 2 will break the herd only if her signal precision γ_2 satisfies $V_2(0) \geq V_2(1)$. When σ is sufficiently large, or more specifically, when $\sigma > \alpha + 1$, we have

$$V_2(0) \leq \left[M + \int_R^\infty \left[\frac{\gamma_2}{\gamma_1+\gamma_2}\right]^{1-\sigma} \frac{2C}{\gamma_1^\alpha} d\gamma_1\right]^{\frac{1}{1-\sigma}} = \left[M + \int_R^\infty \frac{2C}{\gamma_2} \times \frac{(\gamma_1+\gamma_2)^{\sigma-1}}{\gamma_1^\alpha} d\gamma_1\right]^{-\frac{1}{\sigma-1}} = 0,$$

for some $M, R < \infty$. It comes from that when $\sigma > \alpha + 1$, we have $\frac{(\gamma_2\gamma_1+1)^{\sigma-1}}{\gamma_1^\alpha} \to \infty$ as $\gamma_1 \to \infty$, so the integral diverges. Further notice that $\frac{\gamma_1\gamma_2}{\gamma_1\gamma_2+1} \geq \frac{1}{2}$, so

$$V_2(1) = \left[\int_1^\infty \left[\frac{\gamma_1}{\gamma_1+\gamma_2}\right]^{1-\sigma} h(\gamma_1)\, d\gamma_1\right]^{\frac{1}{1-\sigma}} \geq \frac{1}{1+\gamma_2} > 0.$$

To sum up, $V_2(1) > V_2(0)$ for all $s \in S$ and for all $\gamma_2 \in (1, \infty)$, this implies that individual 2 will choose action 1 regardless of her private signal, so an information cascade occurs. □

A.8 Discussion of Example 6: Bayesian Model Uncertainty

Recall that in Example 6, the set of model paths is $\mathfrak{F} = \mathcal{F}^\infty$, the prior is $Q \in \Delta(\mathfrak{F})$, all signals are i.i.d. and unbounded, and the true model paths is $\overline{\boldsymbol{F}} = (\overline{F}, \overline{F}, ...)$.

Example 6 (i): *If $Q(\overline{\boldsymbol{F}}) > 0$, then complete learning occurs.*

Proof. Denote by $\mathbb{P}_{\overline{F}}$ the belief that individuals would form if they knew the true model, and by $\mathbb{P}_Q$ individuals' subjective beliefs under Q. Denote by $l_i^{\overline{F}} = \frac{\mathbb{P}_{\overline{F}}(\theta=1|h_i)}{\mathbb{P}_{\overline{F}}(\theta=0|h_i)}$ the likelihood ratio based on the true model $\overline{F}$, and by l_i^Q the likelihood ratio based on Q. First note that $\left\{l_i^{\overline{F}}\right\}$ is a martingale under the true measure $\mathbb{P}^*$, where $\mathbb{P}^* = \mathbb{P}^0_{\overline{F}}$. The Martingale Convergence Theorem implies that there exists a random variable $l_\infty^{\overline{F}}$ such that $l_i^F \to l_\infty^F$ and $l_\infty^F < \infty$ $\mathbb{P}^*$-almost surely. Since $Q(\overline{\boldsymbol{F}}) > 0$, Kalai and Lehrer (1993) implies that $\mathbb{P}_Q$ merges to $\mathbb{P}_{\overline{F}}$ almost surely ($\mathbb{P}_{\overline{F}}$), that is, for all $\varepsilon > 0$, there exists some $I < \infty$ such that

$$l_i^{\overline{F}}/l_i^Q \in (1-\varepsilon, 1+\varepsilon) \quad \text{for all } i \geq I \qquad \mathbb{P}_{\overline{F}} - a.s.. \tag{12}$$

Note that $\mathbb{P}_{\overline{F}} \gg \mathbb{P}^* = \mathbb{P}^0_{\overline{F}}$, the previous relation also holds $\mathbb{P}^*$-almost surely. It implies that l_i^Q

converges to some limit l_∞^Q, and that $l_\infty^Q = l_\infty^{\overline{F}}$ $\mathbb{P}^*$-almost surely. The dynamics of $l_i^{\overline{F}}$ is given by

$$l_{i+1}^{\overline{F}} = \begin{cases} l_i^{\overline{F}} \times \frac{1-\overline{F}^1\left(1/l_i^Q\right)}{1-\overline{F}^0\left(1/l_i^Q\right)} & \text{if } a_i = 1 \\ l_i^{\overline{F}} \times \frac{\overline{F}^1\left(1/l_i^Q\right)}{\overline{F}^0\left(1/l_i^Q\right)} & \text{if } a_i = 0 \end{cases}.$$

In the limit, we have $l_\infty^Q = l_\infty^{\overline{F}} =: l_\infty$, where l_∞ satisfies

$$l_\infty = \begin{cases} l_\infty \times \frac{1-\overline{F}^1(1/l_\infty)}{1-\overline{F}^0(1/l_\infty)} & \text{if } a_\infty = 1 \\ l_\infty \times \frac{\overline{F}^1(1/l_\infty)}{\overline{F}^0(1/l_\infty)} & \text{if } a_\infty = 0 \end{cases}.$$

All these claims hold $\mathbb{P}^*$-almost surely. From Lemma 1, we know that $l_\infty \in \{0, \infty\}$. Since l_∞ is almost surely finite, so we must have $l_\infty = 0$, which means that complete learning occurs $\mathbb{P}^*$-almost surely. □

Example 6 (ii): *If $Q\left(\overline{\boldsymbol{F}}\right) = 0$, complete learning may not occur.*

Proof. Suppose that Q features an independent distribution across individuals, so $Q\left(F_1, ..., F_n\right) = q\left(F_1\right) \times ... q\left(F_n\right)$ for all possible F_is and all n, where q is a distribution over models. As explained in Example6, the problem is identical to that individuals perceive $F_Q = \mathbb{E}_Q F = \sum_{F \in \text{supp}(q)} F \times q\left(F\right)$. For convenience, I assume that q has a finite support. Suppose that $\overline{F}^0\left(x\right) \sim x^{\overline{\alpha}}$ as $x \to 0$, and that there is some $F \in \text{supp}\left(q\right)$ such that $F^0\left(x\right) \sim x^\alpha$ as $x \to 0$, where $\alpha < \overline{\alpha}$. In this case, $F_Q^0\left(x\right) \geq x^\alpha > x^{\overline{\alpha}}$ as $x \to 0$. In other words, the perceived model F_Q is more informative that the true model $\overline{F}$, so an incorrect herding occurs with a strictly positive probability as implied by Corollary 4. □

A.8.1 Examples of Bounded Signals

Example6 assumes that signals are unbounded. It is also true that learning outcome depends on the prior when signals are bounded. Below is an example.

Example 7. Suppose that there are two possible data-generating processes, F and G. In Figure 1, L_G and L_F represent the public likelihood ratios when individuals perceive the true model as G and F respectively.

(i) A cascade may or may not occur depending on the prior.

For instance, let Q features an independent distribution with a marginal distribution q. If G is assigned a large weight by q, a cascade occurs; if F is assigned a large weight, a cascade does not occur. To see this, suppose that F is assigned a small weight, $q\left(F\right) = \varepsilon$, hence G is assigned a large weight. Then, L_Q is very close to L_G, so it enters the cascade set, which implies that an information cascade will occur. On the contrary, suppose that F is assigned a small weight,

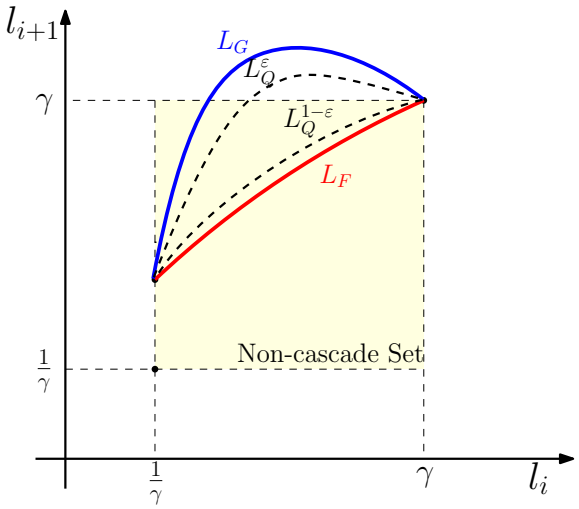

Figure 1: Information Cascades under Different q

$q(F) = 1 - \varepsilon$. In this case, L_Q is close to L_F, so it is trapped in the non-cascade set, implying that a cascade will not occur.

B Supplementary Materials

B.1 Multiple States and Actions

Multiple States

Multiple States. The analysis of multiple states becomes more complicated as the equilibrium strategy does not have a simple characterization. It is conceivable that qualitative results still hold. Below is a simple example.

Example 8. [Multi-state Case] Suppose that the state space $\Theta = \{0, 1, ..., K\}$, and the action space $A = \Theta$. Similarly, individuals get a payoff of 1 if the action matches the true state and a payoff of 0 if otherwise, and all priors are flat. Individual i has a data-generating process g_i with the following form

$C_i \times g_i(s\|\theta)$	s_0	s_1	...	s_K
0	γ_i	1	...	1
1	1	γ_i	...	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$
K	1	1	...	γ_i

, where $\gamma_i \in [1, \overline{\gamma}]$,

and C_i is a normalized term to ensure that all probabilities add up to 1. In this example, s_θ represents the good news for state θ, and all signals are symmetric. Individuals are ambiguous about the γ_is and consider any element in $[1, \overline{\gamma}]$ as possible.

(i) From $a_1 = \theta_1$, we know that $\mathfrak{s}_1 = s_{\theta_1}$, where $\mathfrak{s}_i$ denotes the signal of individual i. It is easy

to derive that for all $\mu \in \Pi_1$, we have

$$\mu(\theta_1) = \frac{\gamma_1}{\gamma_1 + K}, \text{ for some } \gamma_1 \in [1, \overline{\gamma}],$$

and $\mu(\theta) = \frac{1}{\gamma_1 + K}$ for all $\theta \neq \theta_1$.

(ii) From $a_2 = \theta_1$, we know that the data-generating process, g_2 (featured by γ_2), and the signal, $\mathfrak{s}_2$ must satisfy the following inequality:

$$\min_{\pi \in \Pi_1} \frac{\pi(\theta_1) g_2(\mathfrak{s}_2|\theta_1)}{\sum \pi(\theta) g_2(\mathfrak{s}_2|\theta)} \geq \min_{\pi \in \Pi_1} \frac{\pi(\theta') g_2(\mathfrak{s}_2|\theta')}{\sum \pi(\theta) g_2(\mathfrak{s}_2|\theta)} \quad \forall \theta' \in \Theta.$$

If $\mathfrak{s}_2 = s_{\theta_1}$, the signal increases the likelihood of state θ_1 and decreases the likelihood of all other states, so γ_2 can be any element in $[1, \overline{\gamma}]$. If $\mathfrak{s}_2 = s_\theta$ with $\theta \neq \theta_1$, we have

$$\min_{\pi \in \Pi_1} \frac{\pi(\theta_1)}{\pi(\theta) \times \gamma_2 + \sum_{\theta' \neq \theta} \pi(\theta')} \geq \min_{\pi \in \Pi_1} \frac{\pi(\theta) \times \gamma_2}{\pi(\theta) \times \gamma_2 + \sum_{\theta' \neq \theta} \pi(\theta')},$$

which implies that

$$\frac{1}{\gamma_2 + K} \geq \frac{\gamma_2}{\gamma_2 + \overline{\gamma} + K - 1},$$

so $\gamma_2 \leq \overline{\gamma}_2$ for some $\overline{\gamma}_2 \in (1, \overline{\gamma}]$. To sum up, when $\gamma_2 > \overline{\gamma}_2$, a_2 perfectly reveals that $\mathfrak{s}_2 = s_{\theta_1}$, but when $\gamma_2 \leq \overline{\gamma}_2$, a_2 is consistent with all s_θ thus is uninformative. As a consequence, for all $\mu \in \Pi_2$, we have

$$\mu(\theta_1) = \begin{cases} \frac{\gamma_1 \gamma_2}{\gamma_1 \gamma_2 + K} & \gamma_1 \in [1, \overline{\gamma}], \gamma_2 \in [\overline{\gamma}_2, \overline{\gamma}] \\ \frac{\gamma_1}{\gamma_1 + K} & \gamma_1 \in [1, \overline{\gamma}], \gamma_2 \in [1, \overline{\gamma}_2] \end{cases},$$

and $\mu(\theta) = \frac{1 - \mu(\theta_1)}{K}$ for all $\theta \neq \theta_1$.

(iii) From $a_3 = \theta_1$, we know that $\mathfrak{s}_3 = s_\theta$ with $\theta \neq \theta_1$ is consistent with γ_3 satisfying

$$\min_{\pi \in \Pi_2} \frac{\pi(\theta_1)}{\pi(\theta) \times \gamma_3 + \sum_{\theta' \neq \theta} \pi(\theta')} \geq \min_{\pi \in \Pi_2} \frac{\pi(\theta) \times \gamma_3}{\pi(\theta) \times \gamma_3 + \sum_{\theta' \neq \theta} \pi(\theta')},$$

which implies that

$$\frac{1}{\gamma_3 + K} \geq \frac{\gamma_3}{\gamma_3 + \overline{\gamma}^2 + K - 1},$$

so $\gamma_2 \leq \overline{\gamma}_3$ for some $\overline{\gamma}_3 \in (1, \overline{\gamma}]$, where $\overline{\gamma}_3 > \overline{\gamma}_2$. Therefore, for all $\mu \in \Pi_3$, we have

$$\mu(\theta_1) = \begin{cases} \frac{\gamma_1 \gamma_2 \gamma_3}{\gamma_1 \gamma_2 + K} & \gamma_1 \in [1, \overline{\gamma}], \gamma_2 \in [\overline{\gamma}_2, \overline{\gamma}], \gamma_3 \in [\overline{\gamma_3}, \overline{\gamma}] \\ \frac{\gamma_1 \gamma_2}{\gamma_1 \gamma_2 + K} & \gamma_1 \in [1, \overline{\gamma}], \gamma_2 \in [\overline{\gamma}_2, \overline{\gamma}], \gamma_3 \in [1, \overline{\gamma_3}] \\ \frac{\gamma_1 \gamma_3}{\gamma_1 \gamma_3 + K} & \gamma_1 \in [1, \overline{\gamma}], \gamma_2 \in [1, \overline{\gamma}_2], \gamma_3 \in [\overline{\gamma_3}, \overline{\gamma}] \\ \frac{\gamma_1}{\gamma_1 + K} & \gamma_1 \in [1, \overline{\gamma}], \gamma_2 \in [1, \overline{\gamma}_2], \gamma_3 \in [1, \overline{\gamma_3}] \end{cases}.$$

By induction, for all $i \geq 3$, $\mathfrak{s}_i = s_\theta$ with $\theta \neq \theta_1$ is consistent with γ_i satisfying

$$\frac{1}{\gamma_i + K} \geq \frac{\gamma_i}{\gamma_i + \overline{\gamma}^{i-1} + K - 1}.$$

When i is sufficiently large, the RHS is smaller than the LHS for all $\gamma_i \in [1, \overline{\gamma}]$. That is, for all possible data-generating processes and for all signals, individuals will find it optimal to follow the herd and choose action θ_1. In other words, an information cascade can arise after finite number of individuals for all possible combinations of data-generating processes.

Multiple Actions

When there are multiple actions, we will still have an information cascade in situations where a cascade is absent. However, with multiple actions, the learning outcomes depend more intricately on the ambiguity attitudes. As shown in an earlier version of this paper, if there exists a safe action and if individuals are ambiguity averse, there will be an information cascade on the safe action. In contrast, if individuals are ambiguity loving, they will only settle on the uncertain actions.

B.2 Other Updating Rules

Suppose that individuals hold follow the **α-maximum likelihood rule** as in Epstein and Schneider (2007) and update the model set $\mathcal{F}_0$ over time. That is,

$$\mathcal{F}_{-i} \mid h_i = \left\{ F_{-i} : \mathbb{P}_{F_{-i}}(h_i|\sigma_{-i}) \geq \alpha \cdot \sup_{F_{-i} \in \mathcal{F}_{-i}} \mathbb{P}_{F_{-i}}(h_i|\sigma_{-i}) \right\}$$

where $\alpha \in [0,1]$. Notice that $\alpha = 1$ corresponds to the maximum likelihood updating, and $\alpha = 0$ corresponds to the full Bayesian updating.

Proposition 4. *Suppose that individuals use $\alpha - MLU$ to update their beliefs. Under Assumption 3, for all $\alpha \in [0,1)$, an information cascade occurs with strictly positive probability.*

Proof. Notice that

$$\mathbb{P}_{F_{-i}}(h_i) = \mathbb{P}_{F_{-i}}(a_1)\,\mathbb{P}_{F_{-i}}(a_2|a_1)\,...\,\mathbb{P}_{F_{-i}}(a_{i-1}|a_1, a_2, ..., a_{i-2})$$

Consider the action profile where $a_1 = a_2 = ... = a_{i-1} = 1$, which is a positive probability event for any finite i. Suppose $F^*_{-i} = \left(F^*_1, ..., F^*_{i-1}\right) \in \arg\max \mathbb{P}_{F_{-i}}(h_i)$. Notice that the maximum can be obtained. Since $\mathbb{P}_{F^*_1}(a_1) = \frac{1}{2}$, which holds for all F_1 continuous at 1. We can just let $F^*_2 = ... = F^*_{i-1}$ be uninformative data-generating process. In this case $\mathbb{P}_{F_2}(a_2|a_1) = ... = \mathbb{P}_{F_{-i}}(a_{i-1}|a_1, a_2, ..., a_{i-2}) = 1$. The maximum is obtained. I then define $F_{-i} \equiv \left(F^*_1, ..., F^*_{i-2}, F_{i-1}\right)$,

then $F_{-i} \in \mathcal{F}_{-i} \mid h_i$ only if $\mathbb{P}_{F_{-i}}(h_i) \geq \alpha \cdot \mathbb{P}_{F^*_{-i}}(h_i)$ or

$$\mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1}) \geq \alpha \mathbb{P}_{F^*_{-i}}(a_{i-1}|h_{i-1}) = \alpha$$
$$\mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=0)\,\mathbb{P}_{F_{-i}}(\theta=0|h_{i-1}) + \mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=1)\,\mathbb{P}_{F_{-i}}(\theta=1|h_{i-1}) \geq \alpha \quad (13)$$

Since $a_1 = ... = a_{i-2} = 1$ or $h_{i-1} = \{1, ..., 1\}$, it is easy to verify that: $\mathbb{P}_{F_{-i}}(\theta=1|h_{i-1}) \geq \mathbb{P}_{F_{-i}}(\theta=0|h_{i-1})$ for all $F_{-i} \in \mathcal{F}_{-i}$, which means that a sequence of action 1 reveals that state 1 is more likely. Since $a_{i-1} = 1$,we also have $\mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=1) \geq \mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=0)$. So

$$\begin{aligned}&\mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=0)\,\mathbb{P}_{F_{-i}}(\theta=0|h_{i-1}) + \mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=1)\,\mathbb{P}_{F_{-i}}(\theta=1|h_{i-1})\\ \geq&\mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=0)\,\frac{1}{2} + \mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=1)\,\frac{1}{2}\end{aligned}$$

So inequality (13) is true when

$$\mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=0)\,\frac{1}{2} + \mathbb{P}_{F_{-i}}(a_{i-1}|h_{i-1};\theta=1)\,\frac{1}{2} \geq \alpha \quad (14)$$

Denote r_i as the average public likelihood ratio after observing h_i. Assume that there is no information cascade yet, suppose that $i \geq 2$. So $r_i \in (1, \gamma)$. From individuals' equilibrium strategies, we have: $\mathbb{P}_{F_i}(a_i|h_i;\theta) = 1 - F_i^\theta\left(\frac{1}{r_i}\right)$. Consider the following F_i where $\text{supp}(F_i) = \left\{\frac{1}{\gamma}, 1, \gamma\right\}$. Let f_i^θ be the p.m.f. of F_i^θ. Suppose that $f_i^0(\gamma) = f_i^1\left(\frac{1}{\gamma}\right) = p$ thus $f_i^0\left(\frac{1}{\gamma}\right) = f_i^1(\gamma) = p\gamma$, where $p \in \left[0, \frac{1}{\gamma+1}\right]$. We have:

$$\mathbb{P}_{F_i}(a_i|h_i;0) = 1 - p\gamma$$
$$\mathbb{P}_{F_i}(a_i|h_i;1) = 1 - p$$

Then (14) gives: $p \leq \frac{2-2\alpha}{1+\gamma}$. Then I just take $p = \frac{2-2\alpha}{1+\gamma}$, the F_i constructed with this p belongs to $\mathcal{F}_{-i} \mid h_i$. We have as long as $r_i \in (1, \gamma)$,

$$\frac{r_{i+1}}{r_i} = \frac{1-p\gamma}{1-p} > 1 \text{ when } \alpha < 1$$

So an information cascade occurs after finite steps thus with strictly positive probability. □

Remark 5. Notice that a cascade does not occur at $\alpha = 1$, the maximum likelihood updating. This is because under MLU, there exists an "over-fitting problem". A herding can be best justified when all followers have uninformative data-generating processes. As such, under MLU, individuals will only keep very uninformative models in $\mathcal{F}_0$, so beliefs stop updating after the first person in the herd.