Sequential Learning under Informational Ambiguity*

Jaden Yang Chen†

First version: Nov 2019 This version: Aug 2022

Abstract

This paper studies a sequential social learning problem in which 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 or its perception, as suggested by the literature. When there is sufficient ambiguity, an information cascade occurs almost surely for all possible data-generating processes. This paper further shows that standard results that feature no cascade may be fragile to arbitrarily small ambiguity.

JEL Classification: D81, D83, C72

Keywords: Social learning, 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 encouragement and guidance. I thank Gary Biglaiser, Simon Board, Larry Epstein, Drew Fudenberg, Fei Li, Qingmin Liu, Erik Madsen, Suraj Malladi, Peter Norman, Seth Sanders, Peter Sorensen, Omer Tamuz, Jianrong Tian, Alex Wolitzky and many conference participants for discussion and comments. I thank the editor and four anonymous referees for excellent suggestions. Any errors are mine.

†Department of Economics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27517, USA; E-mail address: yangch@unc.edu

1 Introduction

Herding behavior is an important feature of real life. One of the most influential models to explain herding behavior is the sequential social learning model (SSLM) introduced by Banerjee (1992) and Bikhchandani et al. (1992). Under this framework, a sequence of individuals take actions to match the actual state. Individuals receive i.i.d. private signals with a commonly known data-generating process (DGP), which generates a finite number of signals. The authors showed that an ***information cascade*** will arise with probability 1 and an incorrect cascade occurs with a positive probability. In the end, all individuals will choose to ignore their own signals and follow one action, even if that action is suboptimal.

The theory of information cascades was later challenged by the following two findings by Smith and Sørensen (2000). First, when signals are unbounded, the society will eventually settle on the correct action, which implies that an incorrect herding cannot occur. Second, even if signals are bounded, an information cascade may or may not occur depending on the statistical properties of the actual DGPs. More precisely, when the DGPs satisfy the increasing hazard ratio property (IHRP) or the log-concavity property, an information cascade will not take place (Herrera and Hörner, 2012; Smith et al., 2021). Because many common distributions satisfy these properties, the SSLM cannot generate information cascades or incorrect herding in many interesting situations. Recent works on misspecified learning further suggest that if individuals consider an incorrect DGP, the learning outcome would also depend on the model perception, and it is possible that actions will oscillate forever (Bohren, 2016; Bohren and Hauser, 2021).[1] In summary, the literature concludes that the social learning outcome relies on fine details of the true model and model perceptions, which leaves the question of which learning outcome is robust to alternative specifications of those details.[2]

This paper revisits the SLLM under the assumption that individuals are ***ambiguous*** about predecessors' DGPs by considering a set of DGPs as possible. This assumption describes a situation in which individuals face model uncertainty and cannot determine a specific model that explains the information of interest.[3] To motivate the assumption, let's consider the classical restaurant example, in which customers observe all previous choices and then decide which restaurant to go to (Banerjee, 1992). The standard setup implies that customers understand everyone's signal precision, so they are able to interpret all actions with certainty. In reality, customers' information quality often varies a lot—some

[1]Their setups do not nest the standard SSLM, but the insight still holds in the standard model.

[2]Bikhchandani et al. (2021) and Golub and Sadler (2017) provide excellent surveys on SSLM.

[3]See Hansen et al. (2014); Marinacci (2015); and Hansen and Marinacci (2016) for surveys on model uncertainty.

customers may be “experts,” e.g., food critics for the *New York Times*, who know well which restaurant serves the best food; however, some customers may be “laymen,” who have little information about each restaurant, so they mostly follow others. Also, customers only have very limited observations from every other customer, so they often lack the information to determine who is an “expert” and who is a “layman,” or the distribution of the two types. As a consequence, they are likely to entertain both possibilities simultaneously, which leads to ambiguous information. I also assume that individuals have the ***max-min*** expected utility (MEU) preferences as in Wald (1950) and Gilboa and Schmeidler (1989). The MEU preference features ambiguity aversion, which is compatible with experimental evidence that violates the independence axiom of the expected utility preference; e.g., Ellsberg (1961). An individual with the MEU preference chooses an action that maximizes the expected utility in the worst-case scenario.

I find that when there is sufficient ambiguity, i.e., when individuals consider adequately many models, an information cascade occurs almost surely. This implies that in an adequately ambiguous environment, information cascade is robust with respect to many details of the environment—e.g., (i) the fine properties of the true DGPs, and (ii) whether a particular model is perceived by individuals. The intuition comes from a simple fact, whereby the DGPs that encourage an information cascade and discourage it are *asymmetric*. To see that, suppose that all customers went to restaurant A, and that the next customer received a signal supporting restaurant B. If the next customer went to restaurant B, the worst-case scenario is that all predecessors are experts, their signals are all very precise, and their actions reveal strong information supporting restaurant A. On the other hand, if the customer went to restaurant A, the worst-case scenario is that all predecessors are laymen, their signals are uninformative, and their actions do not reveal any information about each restaurant. As ambiguity increases, the customer would be more concerned about breaking the herd, because she could not rule out the possibility that she might act against some highly precise signals; in contrast, the concern about following the herd is more controllable, because the customer would only act against her private signal, whose precision is certain to her. The asymmetry in the worst-case scenarios pushes the customer to follow and creates a cascade force. Below is a simple example.

Example 1. The state space $\Theta = \{0, 1\}$. The true state is unknown. Individuals share a common prior $\pi_0 = (1/2, 1/2)$. Every individual i takes action $a_i \in \{0, 1\}$. The utility is 1 if the action matches the state and 0 otherwise. Each individual i receives a signal $s_i \in \{H, L\}$ and has DGP $g_i(s|\theta)$ with

$$\frac{g_i(H|1)}{g_i(L1)} = \frac{g_i(H|0)}{g_i(L|0)} = \gamma_i \in (1,\infty) \equiv \Gamma,$$

where $g_i(s|\theta)$ denotes the conditional probability of signal s in state θ, and γ_i describes individual i's signal precision. Suppose that $\gamma_i \overset{I.I.D.}{\sim} f$, where $f \in \Delta(\Gamma)$. Individuals only know their own signal precision but are ambiguous about others' and believe that all $\gamma_i \in \Gamma$ are possible. Suppose that the first individual's (his) action is $a_1 = 1$. Denote by $V_2(a)$ the minimum EU of the second individual (she) if she takes action a. We have

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

To see that, the worst-case scenario for $a_2 = 1$ is that individual 1 only received uninformative signals, so individual 2's utility is $\frac{\gamma_2}{\gamma_2+1}$ if her signal is H and $\frac{1}{\gamma_2+1}$ if her signal is L. On the other hand, the worst-case scenario for $a_2 = 0$ is that individual 1 received the perfectly revealing signal, so individual 2's minimum utility is 0.[4] Since $V_2(1) > V_2(0)$, individual 2 will always follow individual 1's action, so an information cascade occurs. Note that the argument does not rely on f, so a cascade occurs under *all* possible $f \in \Delta(\Gamma)$. In contrast, if individuals have EU preferences and correctly understand f, the outcome would depend on f—e.g., if f has an unbounded support, correct learning occurs f-almost surely. Furthermore, if individuals misspecify f as $\hat{f}$, the outcome would then depend on both f and $\hat{f}$.

The paper extends this example to a more general setting, in which (i) signals can come from a general class of distribution, and (ii) individuals can perceive a general model set. Theorem 2 provides two sufficient conditions for an information cascade to arise when signals are bounded. The theorem says that whenever individuals find it possible that others may have an adequately informative DGP, a cascade will occur almost surely, where "adequately informative" means that the DGP's tail is sufficiently thick—i.e., high-precision signals will be generated with a sufficiently large probability. The intuition is that the perception of a highly informative DGP would encourage individuals to follow a herd, and hence create a cascade force. Due to the aforementioned asymmetric effect, if individuals consider a sufficiently informative model, the perception of any other model cannot outweigh the cascade force, so an information cascade would always occur without any regard to other details of the model set. Perhaps surprisingly, these conditions imply that non-cascade results may represent knife-edge cases in some interesting situations. For example, an information

[4] To be more precise, it is actually the infimum utility because γ_i cannot be 1 or infinity.

cascade can almost surely occur when individuals are just slightly ambiguous and consider an ε-neighborhood centered at $\overline{F}$, even if $\overline{F}$ does not produce a cascade in the standard case. Similar results also hold for unbounded signals (Theorem 3). I also find that incorrect herding can emerge when individuals entertain a slight degree of ambiguity, so the complete learning result in Smith and Sørensen (2000) may be non-robust in the view of model ambiguity.

The paper is organized as follows. Sections 2 and 3 lay out the model setup and characterize the equilibrium. Sections 4 to 6 present the paper's main results. Section 7.2 discusses how the paper's main results can be extended to other ambiguity preferences. Section 8 discusses some important topics and extensions, and Section 9 reviews related literature. Other topics and extensions are relegated to the Supplementary Material.

2 The Model

States and Actions. There are two possible states of world, $\Theta = \{0, 1\}$. Without loss of generality, the true state $\theta^* = 0$. A countably infinite set of individuals $N = \{1, 2, ...\}$ act sequentially. Each individual makes a choice $a \in A$ and can observe the choices taken by all predecessors. The paper focuses on the binary action case, $A = \{0, 1\}$, in which individuals get a payoff of 1 when their actions match the actual state and a payoff of 0 otherwise. Multiple actions and states are discussed in the Supplementary Material.

Information structures. Individuals do not know the true state and share a common prior π_0. For simplicity, $\pi_0(0) = \pi_0(1) = \frac{1}{2}$, but any full-support prior or a prior set does not change the result. 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. No signal perfectly reveals the state, so 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, λ_i, where $\lambda_i(s) = \frac{d\overline{G}_i^1(s)}{d\overline{G}_i^0(s)}$ denotes the likelihood ratio induced by signal s. The distribution of the likelihood ratio λ_i is denoted by $\overline{F}_i^\theta$, which must satisfy $\lambda = \frac{d\overline{F}_i^1(\lambda)}{d\overline{F}_i^0(\lambda)}$ almost everywhere. The rest of the paper focuses on the normalized signal, λ, and the normalized DGP, $\overline{F}_i^\theta$. All normalized DGPs have a common support, $\Lambda = \left[\frac{1}{\gamma}, \gamma\right]$, where $\gamma > 1$. Signals are ***bounded*** if $\gamma < \infty$ and signals are ***unbounded*** if $\gamma = \infty$. For notional simplicity, I assume: (i) all signals are continuous, that is, $\overline{F}_i^\theta$ is continuous for all i and θ, and (ii) signals are symmetric $\overline{F}_i^1(\lambda) = 1 - \overline{F}_i^0(1/\lambda)$ for all i and λ. The paper's results can be extended to discontinuous and asymmetric signals. Let $\mathbb{P}^*$ denote the objective probability measure, which is the measure induced by $\left\{\overline{F}_1^\theta, \overline{F}_2^\theta, ...\right\}$, conditional on the true state θ^*.

Ambiguous Information and Beliefs. Individuals know their own DGPs and that all signals are independently distributed, but they are ***ambiguous*** about others' DGPs by considering a set of models. Specifically, individuals share a common set of models, $\mathcal{F}_0$, and believe that every other individual's DGP belongs to $\mathcal{F}_0$ but do not know which is the true DGP. The ambiguity assumption describes a situation in which individuals lack the knowledge to pin down the society's signal structure. It is believed to emerge naturally in sequential social learning, in which individuals only observe one action from every other individual, so the information is often insufficient to determine a specific DGP. As a benchmark, I also assume that the model set is homogeneous. The homogeneity assumption describes a situation in which all individuals share the same information about the society's signal structures. For example, suppose that all DGPs are drawn from the same model set, and individuals only know the realizations of their own DGPs but do not know others', then a possible assumption is that all individuals understand the model set and consider every feasible model to be possible. The paper's discussion can be extended to heterogeneous model sets, as shown in Section 6 of the Supplementary Material.

Belief-updating Rule. Due to the informational ambiguity, individuals will form ambiguous beliefs in social learning. Denote by $h_i = (a_1, ..., a_{i-1})$ the history observed by individual i and by $I_i = \{\lambda_i, h_i\}$ the information available to individual i—that is, her private signal λ_i and history h_i. Let $\mathcal{I}_i$ be the set of all possible information available to i, and denote by $\sigma_i : \mathcal{I}_i \to A$ the (pure) strategy of individual i. Given strategy profile $\sigma_{-i} = (\sigma_1, ..., \sigma_{i-1})$, DGP profile $F_{-i} = (F_1, ..., F_{i-1})$ and conditional on state θ, the observed history $h_i = (a_1, ..., a_{i-1})$ is a stochastic process with a probability measure $\mathbb{P}_{F_{-i}}(.|\theta; \sigma_{-i})$. Given history h_i and strategy profile σ_{-i}, denote by $\Pi(h_i, \sigma_{-i})$ the set of beliefs generated by DGPs in $\mathcal{F}_0$, which I refer to as a *public belief set*. That is,

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

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 *private belief set*. Assuming that individuals use the full Bayesian rule (axiomatized by Pires (2002)) to update beliefs,

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

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. The full Bayesian rule is commonly adopted in applications, but two major criticisms of it are (i) the set of models remains unchanged after learning new information, and (ii) it can lead to dynamic inconsistency. These criticisms are of less concern in this paper, because (i) individuals observe one action from every other individual, so there is often very limited information to be learned about others' DGPs, and (ii) individuals make a once-in-a-lifetime decision, so dynamic inconsistency is not relevant here. For these reasons, the paper's main result also holds for many alternative updating rules.[5]

Equilibrium Concept. Assume that individuals have ***max-min expected utility*** (MEU) preferences, as in Gilboa and Schmeidler (1989). The equilibrium is defined as follows.

Definition 1. 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 that equals 1 if $a = \theta$ and 0 if $a \neq \theta$.

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. Individuals follow some tie-breaking rule when indifferent, so Definition 1 provides a unique pure-strategy equilibrium. The choice of tie-breaking rule is not essential to the result, so I do not specify it in the paper. By focusing on pure strategies, the paper implicitly assumes that agents cannot be better off by playing mixed strategies. Note that this assumption is not contradictory to ambiguity-aversion, which says that individuals have incentives to engage in ex post randomization instead of ex ante randomization, as in the mixed-strategy case.[6]

[5] For example, the α-maximum likelihood rules as discussed in Section 7 of the Supplementary Material. Also, when individuals have smooth ambiguity preference, the main result can hold under *any* updating rule (of the second-order belief) that produces a full-support posterior after finite history (Section 7).

[6] Although it still remains a question of whether individuals can benefit from ex ante randomization, the literature seems to suggest that indifference to ex ante randomization seems a reasonable assumption; e.g., in Seo (2009) and Klibanoff et al. (2005), individuals have expected utility preferences over second-order acts; Saito (2012) suggests that individuals have no incentive to engage in ex ante randomization under some axiom; Eichberger et al. (2016) show that dynamic consistency implies that individuals are indifferent to ex ante randomization.

3 Equilibrium Strategies and Learning Concepts

This section first characterizes individuals' equilibrium strategies under ambiguity and then defines some learning concepts that will be used later.

3.1 Characterizations of Equilibrium Strategies

When individuals are ambiguous, it seems difficult to analyze the learning dynamics because individuals now form a set of posteriors instead of a single posterior, as in the Bayesian case. Fortunately, the max-min model enables us to extend the concept of likelihood ratio and represent the posterior set using the ***average likelihood ratio*** of beliefs featured in it. This property leads to a simple equilibrium characterization, which enhances the tractability.

Definition 2. (Average Public Likelihood Ratio) Denote $L_i = \left\{\frac{\pi(1)}{\pi(0)} : \pi \in \Pi\left(h_i\right)\right\}$, 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. It reflects how likely the public thinks state 1 is (relative to state 0) on average. Proposition 1 characterizes individuals' equilibrium strategies by employing 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*

$$\sigma_i^*\left(I_i\right) = \begin{cases} 1 & \text{if } \lambda_i \cdot r_i > 1 \\ 0 & \text{if } \lambda_i \cdot r_i < 1 \end{cases},$$

and the strategy at $\lambda_i \cdot r_i = 1$ is determined by the tie-breaking rule.

Proof. Denote by $\underline{\pi}_i\left(\theta\right) = \inf\left\{\pi\left(\theta\right) : \pi \in \Pi_i\left(I_i\right)\right\}$, then $a_i = 1$ if $\underline{\pi}_i\left(1\right) > \underline{\pi}_i\left(0\right)$. Note that

$$\underline{\pi}_i\left(1\right) = \frac{\lambda_i \underline{l}_i}{1 + \lambda_i \underline{l}_i} \qquad \underline{\pi}_i\left(0\right) = \frac{1}{1 + \lambda_i \overline{l}_i}.$$

By solving $\underline{\pi}_i\left(1\right) > \underline{\pi}_i\left(0\right)$, we have $\lambda_i > \frac{1}{\sqrt{\overline{l}_i \cdot \underline{l}_i}} = \frac{1}{r_i}$. The other case follows 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 in which there are multiple beliefs. Proposition 1 shows that individuals' equilibrium strategies can be decomposed into two parts. The private information component is the private signal, λ_i, whereas 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.

3.2 Information Cascades and Some Learning Concepts

Definition 3. In the equilibrium, an ***information cascade*** occurs if there exists some $I < \infty$ and $a \in A$ such that for all $i \geq I$, we have $\mathbb{P}^* (a_i = a | h_i) = 1$.

An information cascade occurs if, after some point, individuals will only choose one action regardless of their private signals. During a cascade, information stops aggregating and the society may settle on an incorrect action, albeit with infinitely many signals. Using Proposition 1, information cascades can be described using the average likelihood ratio.

Lemma 1. *Denoting by* $C_0 = \left[0, \frac{1}{\gamma}\right]$ *and* $C_1 = [\gamma, \infty]$*, an information cascade of action* a *occurs when there exists some* $I < \infty$ *such that* $r_i \in C_a$ *for all* $i \geq I$*.*

In the literature, C_a is referred to as the *cascade set* of action a. Whenever $r_i \in C_a$, the public belief that favors action a becomes so strong such that individuals will choose action a regardless of their private signals; that is, an information cascade takes place. In classical models of finite signals, an information cascade will almost surely occur, e.g., Banerjee (1992) and Bikhchandani et al. (1992). However, with more general signal structures, the occurrence of a cascade is not always guaranteed. The following outcomes are also possible.

Definition 4. In the equilibrium, we say that

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

(iii) ***complete learning*** occurs if the correct-action herding occurs $\mathbb{P}^*$-almost surely;

(ii) ***action non-convergence*** occurs if a_i fails to converge.

The concept of herding is often confused with information cascade, since both imply the conformity of actions. The difference is that individuals during a herd would have acted differently if they received different signals, but individuals during an information cascade will choose the same action and ignore their private signals, so cascades are more stable than herding.[7] When signals are continuous and satisfy IHRP, the society achieves herding but not

[7] The concept of information cascade was proposed by Bikhchandani et al. (1992) and tested in a lab experiment by Anderson and Holt (1997). The difference between information cascade and herding was distinguished by Smith and Sørensen (2000) and tested by Çelen and Kariv (2004) in an experimental environment.

information cascades. In addition to the absence of a cascade, Smith and Sørensen (2000) show that complete learning occurs when signals are unbounded, so the SSLM does not explain the persistence of incorrect herds in this case. Furthermore, if individuals misspecify the actual DGPs, action non-convergence may emerge, in which case the SLLM does not necessarily lead to a consensus of actions.

4 Benchmark Case: Cascades under Extreme Ambiguity

The variety of learning outcomes leads to a natural question: what is the robust outcome in sequential social learning? This section considers a benchmark case with extreme ambiguity. Denote by $\mathcal{F}$ the set of all DGPs on $[\gamma, 1/\gamma]$, where γ denotes the highest signal. We have the following theorem.

Theorem 1. *When $\mathcal{F}_0 = \mathcal{F}$, an information cascade occurs $\mathbb{P}^*$-almost surely.*

The condition $\mathcal{F}_0 = \mathcal{F}$ describes a situation in which individuals only understand the true support of signals and consider all DGPs on this support to be possible. Theorem 1 shows that in this benchmark case, an information cascade occurs almost surely. This finding is different from the standard findings in the following respects. First, in the standard literature, the occurrence of a cascade relies on specific properties of true DGPs. However, Theorem 1 does not impose any restrictions, so a cascade can occur under all DGPs on $[\gamma, 1/\gamma]$. Second, in the misspecified learning literature, the learning outcome depends on the model specification; however, Theorem 1 shows that when individuals consider multiple model specifications simultaneously, information cascades will emerge as the only outcome. To summarize, Theorem 1 shows that under high ambiguity, a cascade almost surely arises without regard to many details that would matter in the standard case.

4.1 Intuition

When signals are unbounded, an information cascade occurs immediately after the first individual, as in Example 1. Below, I focus on the bounded-signal case $\gamma < \infty$. Suppose that the first i individuals took action 1 and that individual $i+1$ received 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. Let's 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 is that the predecessors' DGPs are uninformative. In this case, $\lambda_1 = ... = \lambda_i = 1$. By following the herd, she would act against her private signal, $\frac{1}{\gamma}$.

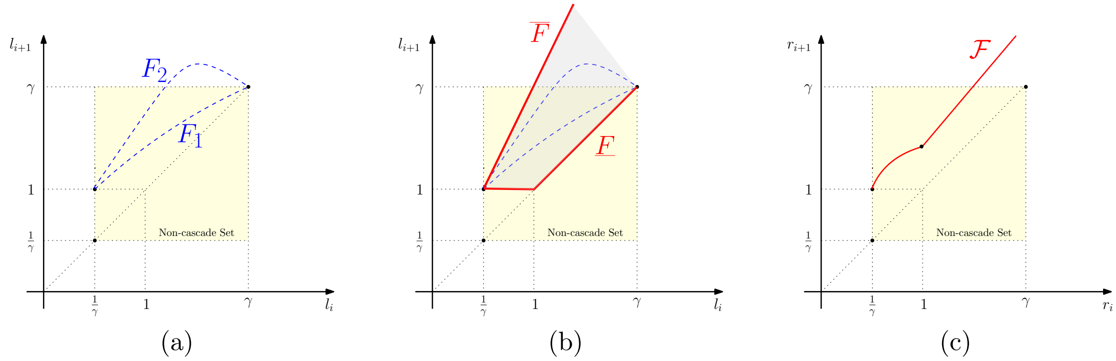

Figure 1: Information Cascades under Ambiguity

Note: The horizontal axis represents the prior likelihood ratio (between states 1 and 0), and the vertical axis represents the posterior likelihood ratio after observing an action 1 (the dynamics after an action 0 are symmetric). The yellow area represents the non-cascade region. Figure 1a depicts the likelihood curves under F_1 and F_2. Figure 1b depicts the set of likelihood curves under all DGPs in $\mathcal{F}$ (marked by the gray shaded area). Figure 1c depicts the average likelihood curve under $\mathcal{F}$.

- If she breaks the herd and takes action 0, the worst case is that every predecessor's DGP has the most precise DGP—i.e., the DGP that only generates signals γ and $1/\gamma$. In this case, the predecessors' actions reveal that their signals must be γs.[8] Hence, by taking action 0, individual $i+1$ would act against i signal γs.

As can be seen, the forces that encourage a cascade and discourage it are **asymmetric**. As i increases, individual $i+1$ would act against increasingly more signals γs in the worst case if she broke the herd; however, she would only act against one signal—her private signal—in the worst case if she followed the herd. When i is sufficiently large, individual $i+1$ would find it optimal to follow, which creates an information cascade.

Graphic illustration. Figure 1a illustrates that the occurrence of a cascade depends on the DGP if there is no ambiguity. In the figure, F_1 satisfies the IHRP, but F_2 does not. As can be seen, posteriors under F_1 are trapped in the non-cascade set, so a cascade never occurs; however, posteriors under F_2 can extend into the cascade set, so a cascade can emerge.

Figure 1b illustrates that under ambiguity, the forces that encourage a cascade and discourage it are asymmetric. The upper envelope of all likelihood curves under $\mathcal{F}$ (marked by $\overline{F}$) has a slope of γ, which means that observing an action 1 can at most increase the likelihood of state 1 (relative to state 0) by a factor of γ. However, the lower envelope (marked by $\underline{F}$) is always bounded from below by the 45-degree line, which means that observing an action 1 cannot decrease the likelihood of state 1.[9] The asymmetry of these two curves

[8] If any of them received a signal $1/\gamma$, that individual would have taken action 0 given the fact that a cascade did not occur.

[9] The upper envelope is obtained at the DGP that only generates the most precise signals, γ and $1/\gamma$.

corresponds to the asymmetry between cascade and non-cascade. The worst-case scenario for breaking a herd happens when beliefs are updated according to the upper envelope, in which case predecessors have the most precise DGP and individuals would act against a sequence of signal γs. In contrast, the worst-case scenario for following a herd happens when beliefs are updated according to the lower envelope, in which case the minimum expected utility is bounded from below by the expected utility when predecessors have uninformative DGPs.

Figure 1c shows how a cascade is triggered by the asymmetry. Recall that under max-min preferences, it suffices to keep track of the average likelihood ratio to determine a cascade (Lemma 1). Figure 1c depicts the average likelihood curve (marked by $\mathcal{F}$), which is obtained by averaging the two envelope curves. As can be seen, the average likelihood curve extends to the cascade set, so an information cascade emerges. To fully prove Theorem 1, it remains to show that the probability of a cascade is 1, which can be established using the standard 0-1 arguments (see the Appendix).

5 Information Cascades with Bounded Signals

The previous section focuses on the extreme case in which individuals consider all DGPs on the actual support to be possible. This leaves the question of whether an information cascade only occurs under extreme ambiguity. This section provides more general conditions for information cascades when signals are bounded.

5.1 Sufficient Conditions for Cascades

Throughout this section, I focus on the bounded-signal case, i.e., $\gamma < \infty$. The following theorem provides sufficient conditions for a cascade to arise.

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

(1) F is discrete at γ;

(2) F^1 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.

The lower envelope has a kink at 1. When $l_i > 1$, the lower envelope is obtained at the uninformative DGP. When $l_i < 1$, individual i's prior favors state 0, so she must have received some minimum information to take action 1. In this case, the lower envelope is obtained at the DGP that only generates signals $1/l_i$ and l_i.

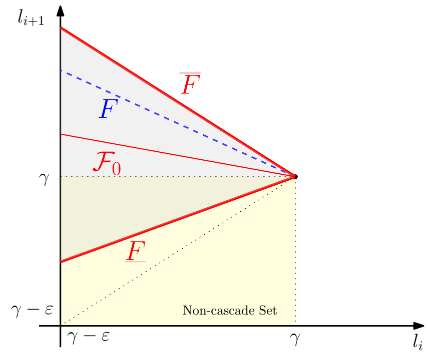

Figure 2: Sufficient Conditions for Information Cascades

Note: Figure 2 depicts the posterior likelihood curves in a small neighborhood near γ. The gray shaded area represents all likelihood curves generated by models in $\mathcal{F}_0$, and the yellow area represents the non-cascade set.

The two conditions say that individuals consider a DGP that assigns adequately large weights to high-precision signals, i.e., the tail is adequately heavy. With some abuse of language, I refer to DGPs that satisfy similar heavy-tail conditions as "highly informative." Theorem 2 therefore says that if individuals find it possible that other individuals may have a highly informative DGP, an information cascade will occur almost surely.

The conditions in Theorem 2 are not very restrictive: First, it only requires that $\mathcal{F}_0$ contain one one such F but does not impose other restrictions on $\mathcal{F}_0$; second, it only imposes restrictions on the F's tail but does not impose any restrictions in the middle. The basic intuition is similar to that in the benchmark case: If a highly informative DGP is considered possible, it will create a strong cascade force—which, due to the asymmetry, cannot be mitigated by any other model perception, so an information cascade will take place regardless of what other models are included in $\mathcal{F}_0$.

The idea can be further illustrated using Figure 2, which depicts the likelihood curves in the corner of the non-cascade set, i.e., $r_i \in (\gamma - \varepsilon, \gamma)$. By definition, the upper envelope of likelihood curves under $\mathcal{F}_0$ (marked by $\overline{F}$) must lie above the likelihood curve under F (marked by F) for all $F \in \mathcal{F}_0$. From the previous discussion, the lower envelope curve (marked by $\underline{F}$) is always bounded from below by the 45-degree line. As a consequence, when the likelihood curve under F is adequately high, which can be guaranteed by the "highly informative" conditions in Theorem 2, the averaged likelihood curve (marked by $\mathcal{F}_0$) will enter the cascade set. This implies that a cascade will be triggered by an action 1 whenever beliefs enters the corner $(\gamma - \varepsilon, \gamma)$. It can be further verified that the corner cascade also implies a global cascade. Therefore, we only need to focus on the corner and restrict the tail property. It is worth noting that Theorem 2 does not mean that F is the only DGP that affects the learning dynamics. Recall that the role of F is to provide a lower bound for the upper envelope, but the exact learning dynamics depend on the structure of $\mathcal{F}_0$. As a consequence, learning under $\mathcal{F}_0$ is usually not observationally equivalent to learning under F or under an arbitrary model set that also contains F.

5.2 Examples: Fragility of Non-cascade Results

Theorem 2 implies that non-cascade results may represent knife-edge cases in some interesting examples.

Example 2. (ε-perturbation set) Suppose that individuals perceive the following set of models

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

where $\mathcal{F}$ denotes all DGPs on $[\gamma, 1/\gamma]$, and $G \in \mathcal{F}$ denotes the benchmark DGP. When $\varepsilon = 0$, it becomes the Bayesian case, so the occurrence of a cascade depends on the properties of G and the true DGP. However, when $\varepsilon > 0$, an information cascade occurs almost surely for all possible Gs and true DGPs.

Example 3. (ε-ambiguity ball) Suppose that individuals consider the following model set:

$$\mathcal{F}_0 = \left\{F \in \mathcal{F} : \sup_{\theta,x} ||F^{\theta}(x) - G^{\theta}(x)|| \leq \varepsilon\right\},$$

where G has full support on $[\gamma, 1/\gamma]$. Similarly, when $\varepsilon = 0$, the learning outcome depends on G, but whenever $\varepsilon > 0$, an information cascade occurs almost surely. This comes from the fact that when $\varepsilon > 0$, there exists a discrete distribution F that approximates G very well under the sup-norm. The result also holds for other norms that allow similar approximations.

In previous examples, whenever individuals face the slightest degree of ambiguity, an information cascade occurs with probability 1. Therefore, an information cascade does not always require substantial ambiguity, as in the benchmark case; moreover, non-cascade results can be non-robust in some sense.

5.3 Discussion: Necessary Conditions for Cascades

The conditions in Theorem 2 are sufficient but not necessary. A simple necessary condition for cascades is that $\mathcal{F}_0$ must contain a DGP whose likelihood curve enters the cascade set, and hence violates the IHRP; e.g., F_2 in Figure 1a. This is because if all DGPs are trapped in the non-cascade set, the average likelihood curve is also trapped in the set, so a cascade will not occur. This condition, however, is not sufficient. In Figure 3a, the likelihood curve under $\hat{F}$ enters the cascade set, so an information cascade occurs when individuals only perceive $\hat{F}$;

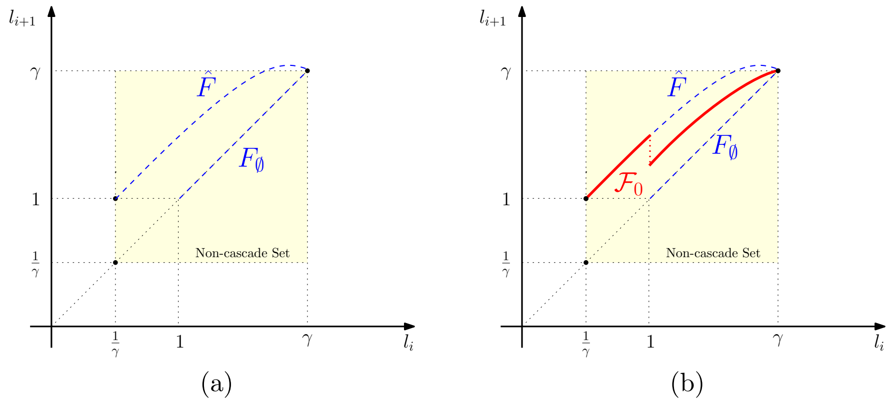

Figure 3: Necessary Conditions for Cascades

however, when individuals consider $\mathcal{F}_0 = \left\{\hat{F}, F_\emptyset\right\}$, where $F_\emptyset$ denotes the uninformative DGP, an information cascade will not occur, as shown in Figure 3b. Intuitively, $\hat{F}$ is inadequately informative in Theorem 2's sense, so when it is perceived with a highly uninformative DGP, the non-cascade force dominates and a cascade will not occur. This example suggests that the detailed structure of $\mathcal{F}_0$ can matter for a cascade, which also illustrates how learning under ambiguity differs from learning under a specific model. To obtain tighter conditions for information cascades, we need to impose restrictions on the structure of $\mathcal{F}_0$, which are discussed in more detail in the Supplementary Material.[10]

6 Incorrect Herding with Unbounded Signals

This section extends Theorem 1 to unbounded signals. Note that information cascade is a restrictive concept for unbounded signals, so it is difficult to occur under small ambiguity as in the last section.[11] This section finds that we can still establish results parallel to the bounded-signal case by focusing on a weaker but qualitatively similar concept: herding. This section shows that under moderate conditions, herding occurs almost surely, and an incorrect herd occurs with a strictly positive probability. In some interesting situations, the complete learning result featured by Smith and Sørensen (2000) no longer holds as long as there is a grain of ambiguity.

[10] In Section 2 of the Supplementary Material, I provide a condition parallel to the IHRP in the literature. The condition says that if the *average* hazard ratio of models in $\mathcal{F}_0$ is strictly increasing, an information cascade will not occur. In other words, this condition must be violated for a cascade to arise.

[11] When signals are unbounded, information cascade requires individuals to ignore arbitrarily strong signals. To have a cascade, individuals' model set must also be "unbounded" in the sense of Example 1 and Theorem 1, in which individuals find it possible that predecessors may have an arbitrarily informative DGP,

6.1 Sufficient Conditions for Herding

The following theorem provides a sufficient condition for herding.

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 $x^p = o(F^0(x))$ as $x \to 0$ for some $p \in (0, \alpha)$, herding occurs $\mathbb{P}^*$-almost surely, and incorrect herding occurs with a $\mathbb{P}^*$-strictly positive probability.*

Theorem 3 is a parallel statement of Theorem 2 when signals are unbounded. The restriction $\overline{F}_i^0(x) \leq ax^\alpha$ says that the true DGP is bounded by some power function. This condition is relatively weak and can cover many interesting DGPs.[12] The condition $x^p = o(F^0(x))$ means that the tail of $F^0(x)$ is sufficiently fat, so it is parallel to the conditions in Theorem 2. Thus, Theorem 3 says that when individuals consider a highly informative DGP, herding occurs almost surely and the herding can be incorrect. The intuition is similar to that behind Theorem 2—whenever individuals perceive a highly informative DGP, it creates a strong herding force that cannot be offset by any other model, so an incorrect herding can emerge. Also, Theorem 3 only requires $\mathcal{F}_0$ to contain a specific model without imposing restrictions on other structures, so it is easy to hold in many interesting cases as implied by the following corollary.

Corollary 1. *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 incorrect herding occurs with a $\mathbb{P}^*$-strictly positive probability.*

Corollary 1 says that if the true DGP has a "power tail," arbitrarily small ambiguity in the power of $\overline{F}^0$ is sufficient to trigger an incorrect herding. This suggests that within this class of models, complete learning is not robust in a sense. Below is a concrete example.

Example 4. For better exposition, this example focuses on the actual signal s_i (instead of normalized signals). Consider the signal space $\mathcal{S} = (0, 1)$; signals are i.i.d., and the DGP takes the form of $g_m = (g_m^0, g_m^1)$, where

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

The true DGP is $g_{m_0}^\theta$ where $m_0 > 0$. It is easy to see that signals are unbounded—i.e., $g_{m_0}^0(s)/g_{m_0}^1(s)$ is unbounded, so complete learning will occur if individuals precisely perceive the true DGP. Suppose that individuals are ambiguous and perceive a set $M_\varepsilon = [m_0 - \varepsilon, m_0 + \varepsilon] \subset \mathbb{R}_+$. Corollary 1 implies that for *all* $\varepsilon > 0$, complete learning no longer holds, and the society will settle on an incorrect action with a strictly positive probability.

[12] The condition was also adopted by Rosenberg and Vieille (2019). As they pointed out, one class of DGPs that violates the condition is $\overline{F}_i^0(x) \sim 1/|\log(x)|$ as $x \to 0$.

6.2 Discussion: Conditions for Complete Learning

Previous discussion shows that complete learning will collapse under ambiguity. It is natural to ask when complete learning occurs. Section 1 of the Supplementary Material provides a necessary and sufficient condition for complete learning for DGPs with power tails. The condition can be intuitively explained as follows. To achieve complete learning, we need to exclude two sources of incomplete learning: (i) incorrect herding and (ii) action non-convergence. This implies that $\mathcal{F}_0$ must be restricted in two ways: (i) to exclude incorrect herding, $\mathcal{F}_0$ cannot contain any highly informative DGP in the sense of Theorem 1, and (ii) to exclude action non-convergence, $\mathcal{F}_0$ must contain some DGP that is adequately informative, such that the society will reach a consensus in the end. As can be seen, complete learning requires more complicated conditions than incorrect herding, in that we need to restrict multiple aspects of the model set. This provides another perspective into why incorrect herding seems to be more robust than complete learning under ambiguity.

7 Other Ambiguity Preferences

The key results of this paper can be extended to a wider class of ambiguity preferences. Below are two important examples: the α-max-min EU preference and the smooth ambiguity preference.

7.1 α-MEU Model

Consider first the case in which individuals have α-maxmin expected utility ($\alpha-$MEU) preferences (Hurwicz, 1951; Ghirardato et al., 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 max-min EU model, and $\alpha = 0$ corresponds to the max-max EU model. We have the following fact.

Fact 1. *Under $\alpha-$MEU preferences, previous results hold for **all** $\alpha \in [0, 1]$.*

Thus an information cascade can arise for all ambiguity attitudes captured by the α-MEU model. The result may seem surprising, but it comes from the simple fact that

$$\begin{aligned}\min_{\pi} \mathbb{E}_{\pi} u(a,1) > \min_{\pi} \mathbb{E}_{\pi} u(a,0) &\Leftrightarrow \min_{\pi} [1-\mathbb{E}_{\pi} u(a,0)] > \min_{\pi} [1-\mathbb{E}_{\pi} u(a,1)] \\ &\Leftrightarrow \max_{\pi} \mathbb{E}_{\pi} u(a,1) > \max_{\pi} \mathbb{E}_{\pi} u(a,0), \end{aligned} \tag{2}$$

so the decision under the max-min model is the same as that under the max-max model, and hence under any $\alpha-$MEU model. Notice that this fact relies on the binary and symmetric action space; with general action space, the ambiguity attitude usually affects decisions, but the cascade result is still valid.[13] Fundamentally, the cascade result is driven by the asymmetry between herding and contrarian fueled by ambiguity, but the asymmetric effect is not unique to ambiguity aversion. Suppose that individuals have max-max preferences—i.e., their decisions are determined by the best-case utility. Under high ambiguity, every action in a herd can be interpreted as containing a lot of information, so the best-case utility of herding can be very high; in contrast, the best-case utility of breaking a herd cannot exceed the case where previous actions contain no information, so it is more limited. Therefore, the asymmetric effect still emerges.

7.2 Smooth Ambiguity Model

The $\alpha-$MEU model makes a restrictive assumption whereby decisions only depend on the best and the worst cases. To address this concern, I consider the extension in which individuals have the smooth ambiguity preference, as axiomatized by Klibanoff et al. (2005):

$$V_i(a) = \phi^{-1}\left(\int_{\mathcal{F}_0^{i-1}} \phi\left[\mathbb{E}_{F_1,...,F_{i-1}}\left(U(a,\theta)\,|h_i\right)\right] d\mu_i\left(F_1,...,F_{i-1}|h_i\right)\right),$$

where (i) $\mu_i(F_1,...,F_{i-1}|h_i)$ stands for the second-order belief on $(F_1,...,F_{i-1})$ conditional on history h_i and (ii) ϕ denotes the second-order utility function, whose curvature describes the ambiguity attitude. It turns out that an information cascade can still occur when individuals are sufficiently *ambiguity-sensitive*; i.e., when ϕ is sufficiently concave or convex. Below is an example.

Example 5. The model set consists of parametric models,

$$\mathcal{F}_0 = \left\{F^{\theta}(x,\alpha) : \alpha \in \mathcal{A}\right\}, \text{ where } \mathcal{A} \subset \mathbb{R}^n,$$

[13] The difference is that with different attitudes, society may settle on different actions, as discussed in Section 8.

where every DGP has the same support $[1/\gamma, \gamma]$ with $\gamma < \infty$.[14] Every individual's DGP can be represented by a parameter $\alpha_i \in \mathcal{A}$. Suppose that α_is are i.i.d. drawn according to g, where g has full support on $\mathcal{A}$. All probabilities are calculated according to g. Further suppose that $\mathcal{F}_0$ features sufficient ambiguity; e.g., it satisfies the conditions in Theorem 2. Individuals have the following preference

$$V_i(a) = \left[\int_{\mathcal{A}^{i-1}} \left[\mathbb{E}_{\alpha_1,...,\alpha_{i-1}} u(a,\theta)\right]^{1-\sigma} d\mu_i(\alpha_1, ..., \alpha_{i-1}|h_i)\right]^{\frac{1}{1-\sigma}},$$

where $\mu_i : A^{i-1} \times \mathcal{A}^{i-1} \to \mathbb{R}_{++}$ denotes an updating rule that delivers a full-support posterior after finite history.[15] The parameter σ represents the ambiguity attitude. When $\sigma = 0$, it corresponds to the EU case. In this case, the learning outcome depends on the true model g and the updating rule μ_i. As $|\sigma| \to \infty$, it approaches the max-min or max-max case. It turns out that for ***any*** updating rule μ_i that leads to a full-support posterior, any g, and any $\varepsilon > 0$, there exists some $\overline{\sigma} < \infty$ such that an information cascade can occur with a probability greater than $1 - \varepsilon$ whenever $|\sigma| > \overline{\sigma}$. The proof is in the Appendix.

Example 5 comes from two facts: (i) the smooth model approaches the max-min or max-max model as ϕ's curvature goes to infinity, and (ii) an information cascade occurs in finite time. By making the curvature of ϕ sufficiently large, we can ensure that the belief paths under the smooth model and the max-min (or max-max) model are very close up to any finite time, so a cascade can also occur with an arbitrarily large probability. Example 5 focuses on a specific ϕ for illustration, but the same result holds for general ϕs.

In summary, with general ambiguity preferences, an information cascade still occurs when (i) there is sufficient ambiguity—i.e., there are sufficiently many models, and (ii) individuals are sufficiently ambiguity-sensitive—i.e., their decisions are adequately influenced by the best or the worst outcomes. Therefore, the information cascade result in the benchmark model does not only represent an extreme case.

8 Discussion and Extension

Multiple Actions and States. When there are multiple actions or states, an information cascade can still arise under sufficient ambiguity. The difference is that the ambiguity

[14] A similar result also holds for unbounded signals, which is discussed in Section 3.3 of the Supplementary Material.

[15] This is a weak assumption and covers many interesting updating rules; e.g., (i) Bayes' rule with a full-support prior and (ii) the dynamically consistent updating rule of the smooth ambiguity model with regular ϕs (Hanany and Klibanoff, 2008).

attitude can affect which actions will be adopted by the society. For example, if individuals are ambiguity-averse, the society may settle on safe actions, whereas if individuals are ambiguity-loving, the society will select risky actions. It can be further shown that with sufficient informational and prior ambiguity, individuals will only choose from the safest or the riskiest actions in the limit, which contain at most two actions when the state space is binary. In this case, assuming a binary action space is without loss of generality for the paper's purpose. This extension is discussed in Sections 4-5 of the Supplementary Material.

Ambiguity and Model Misspecification. The benchmark model imposes no restriction on $\mathcal{F}_0$, so individuals may perceive some incorrect models. This raises the question of how important incorrect beliefs are in producing the result. This can be answered by considering Example 5. Now suppose that individuals correctly specify g, so their beliefs μ_is are objectively correct. Example 5 shows that for all $\varepsilon > 0$, an information cascade still occurs with a probability greater than $1 - \varepsilon$, when individuals are sufficiently ambiguity-sensitive. The result takes a sharper form with max-min preferences, in which an information cascade occurs almost surely (Section 3 of the Supplementary Material). In this example, individuals are correctly specified, but the cascade result is still valid. It suggests that the main result is produced by ambiguity and ambiguity preference and does not require individuals to be misspecified.

The similarity between misspecification and ambiguity is that both create distortions to correct Bayesian learning. However, misspecification distorts correct Bayesian learning with a focus on the "correct" part, whereas ambiguity distorts correct Bayesian learning with a focus on the "Bayesian" part. The former allows individuals to perceive an incorrect model but still maintains expected utility preferences and Bayes' rule; however, the latter assumes non-expected utility preferences and employs learning rules under ambiguous information. As can be seen, the distortions emerge from different mechanisms, so they often generate different learning dynamics. Also, the distortions in the paper always point in the direction of producing a cascade when there is sufficient ambiguity, whereas the distortions from incorrect beliefs depend on the model's perception and can produce many results. The discussion of the misspecified learning literature can be found in Section 9.

Ambiguity and Bayesian Model Uncertainty. The uncertainty in this paper is represented by a set of models. It is natural to consider an alternative setup in which individuals are Bayesian and hold a prior over models. In this setup, it turns out that the social learning outcome would be prior-dependent. For example, if signals are unbounded and the prior assigns a strictly positive probability to the true DGPs, complete learning will occur (Kalai and Lehrer, 1993); however, for other priors, complete learning may not occur. Essentially, by assigning a specific prior, we are requiring that individuals follow a particular rule to interpret information, which is in the same spirit as assigning a particular model in the standard case. Different from the Bayesian paradigm, this paper allows individuals to consider multiple interpretations simultaneously, so we can discuss which outcome is more robust with respect to alternative models. Admittedly, this paper's learning outcome would also depend on (i) the model set, as it depends on the true model in the standard case, and (ii) the ambiguity preference, as it depends on the prior in the Bayesian uncertainty case. However, the ambiguity framework allows us to interpret the dependence in an interesting way, such that an information cascade always occurs when (i) there is sufficient ambiguity, i.e., the model set is sufficiently large, and (ii) when individuals are sufficiently ambiguity sensitive. It therefore provides an alternative interpretation of herding behavior from the perspective of ambiguity.

Ambiguous Network Structures. A cascade can also occur when individuals face ambiguity about the network structure—i.e., they do not know what their predecessors can actually observe. Essentially, network structures play the same role as DGPs, in that they both imply particular interpretations of past actions. As such, an information cascade can also emerge from a similar mechanism. This is discussed in Section 8 of the Supplementary Material.

9 Related Literature

This paper contributes to the growing literature on learning under ambiguity. Most works in this thread of literature focus on individual learning. Marinacci (2002) and Marinacci and Massari (2019) study in an individual learning problem in terms of whether ambiguity will fade away asymptotically. Epstein and Schneider (2007) introduce an α-maximum likelihood learning rule and investigate a dynamic portfolio choice problem. Battigalli et al. (2019) study a learning problem in which data are endogenously generated from an experimentation process. Fryer Jr et al. (2019) and Chen (2022) study learning problems where individuals are biased in interpreting ambiguous information. This paper complements the literature by investigating a social learning problem, where informational ambiguity seems to occur very naturally. A relevant paper is by Ford et al. (2013), who study a sequential trading model in which traders face ambiguity and have neo-additive capacity expected utility (CEU) preferences (Chateauneuf et al., 2007). They show that ambiguity can produce both herding and contrarian. This arises from the property that the CEU is bounded away from 0 and 1, but the ask-bid prices can fully adjust to 0 and 1, so the discrepancy provides room for herding and contrarian. In contrast, this paper's result comes from a different mechanism that employs the asymmetry between cascade and non-cascade under ambiguous information; also, the ambiguity in the paper mainly produces herding but not contrarian. In addition to the aforementioned applications, learning under ambiguity is also examined in recent works in decision theory—e.g., Cheng (2022); Kovach (2021); and Tang (2022);—and experimental economics; e.g., De Filippis et al. (2022) and Epstein et al. (2019).

This paper is closely related to the literature on social learning with misspecified models. Bohren (2016) and Bohren and Hauser (2021) examine a sequential social learning problem in which individuals misspecify the true model. Bohren (2016) finds that different model specifications can lead to different learning outcomes—e.g., complete learning, incomplete learning, and cyclical actions. Bohren and Hauser (2021) incorporate these results in a more general framework. They find that complete learning is robust with respect to small misspecifications, which stands in contrast to this paper's finding that complete learning may be non-robust. The difference is driven by their assumption that the society has a positive fraction of "autarkic agents" who only act according to their private signals. (Frick et al., 2020a,b) also find that complete learning is not robust but in different settings. Specifically, Frick et al. (2020a) consider a social learning problem in which the state space is continuous and individuals with different preferences randomly meet with each other. Frick et al. (2020b) propose a local martingale-based approach and show the fragility of sequential social learning in an environment in which signals are bounded and individuals have heterogeneous risk preferences. Compared with previous papers on misspecified social learning, this paper assumes that individuals consider a set of models instead of a specific model.[16] In the paper, the learning outcome does not rely on whether a particular model is perceived—under high ambiguity, the inclusion of any model perception will not change the almost sure occurrence of a cascade. As such, my paper complements the literature by suggesting a way to think about which outcome is more robust when individuals can consider multiple model specifications.

This paper also connects to the literature on social learning with non-Bayesian agents. The literature shows that incorrect learning can emerge if individuals follow some naive learning rules—for example, when they do not fully account for predecessors' inferences (Eyster and Rabin, 2010), when they follow a coarse inference rule (Guarino and Jehiel, 2013), or when they follow some average rule to aggregate the opinions from others (DeMarzo et al., 2003; Molavi et al., 2018; Dasaratha and He, 2020). In this paper, individuals are not naive, and they understand how others make inferences. The paper's deviation from the Bayesian paradigm is mainly created by ambiguity and ambiguity preferences.

[16] See Hansen and Marinacci (2016) for a survey on ambiguity and model misspecification.

10 Conclusion

This paper studies a sequential social learning problem in which individuals face ambiguity about other people's DGPs. I find that under sufficient ambiguity, an information cascade almost surely occurs without regard to many details of the learning environment. Interestingly, some results that feature non-cascades are fragile with respect to a small perturbation in ambiguity. The paper focuses on the sequential social learning model, so it would be interesting to investigate how individuals learn in other environments—e.g., general networks, repeated interactions, heterogeneous preferences, etc.

Bibliography

Anderson, Lisa R and Charles A Holt (1997) "Information cascades in the laboratory," *American Economic Review*, 847–862. 9

Banerjee, Abhijit V (1992) "A simple model of herd behavior," *Quarterly Journal of economics*, 107 (3), 797–817. 2, 9

Battigalli, Pierpaolo, Alejandro Francetich, Giacomo Lanzani, and Massimo Marinacci (2019) "Learning and self-confirming long-run biases," *Journal of Economic Theory*, 183, 740–785. 21

Bikhchandani, Sushil, David Hirshleifer, Omer Tamuz, and Ivo Welch (2021) "Information cascades and social learning,"Technical report, National Bureau of Economic Research. 2

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992) "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of Political Economy*, 100 (5), 992–1026. 2, 9

Bohren, J Aislinn (2016) "Informational herding with model misspecification," *Journal of Economic Theory*, 163, 222–247. 2, 22

Bohren, J Aislinn and Daniel N Hauser (2021) "Learning with heterogeneous misspecified models: Characterization and robustness," *Econometrica*, 89 (6), 3025–3077. 2, 22

Çelen, Boğaçhan and Shachar Kariv (2004) "Distinguishing informational cascades from herd behavior in the laboratory," *American Economic Review*, 94 (3), 484–498. 9

Chateauneuf, Alain, Jürgen Eichberger, and Simon Grant (2007) "Choice under uncertainty with the best and worst in mind: Neo-additive capacities," *Journal of Economic Theory*, 137 (1), 538–567. 21

Chen, Jaden Yang (2022) "Biased learning under ambiguous information," *Journal of Economic Theory*, 105492. 21

Cheng, Xiaoyu (2022) "Relative maximum likelihood updating of ambiguous beliefs," *Journal of Mathematical Economics*, 99, 102587. 22

Dasaratha, Krishna and Kevin He (2020) "Network structure and naive sequential learning," *Theoretical Economics*, 15 (2), 415–444. 22

De Filippis, Roberta, Antonio Guarino, Philippe Jehiel, and Toru Kitagawa (2022) "Non-Bayesian updating in a social learning experiment," *Journal of Economic Theory*, 199, 105188. 22

DeMarzo, Peter M, Dimitri Vayanos, and Jeffrey Zwiebel (2003) "Persuasion bias, social influence, and unidimensional opinions," *Quarterly Journal of Economics*, 118 (3), 909–968. 22

Eichberger, Jürgen, Simon Grant, and David Kelsey (2016) "Randomization and dynamic consistency," *Economic Theory*, 62 (3), 547–566. 7

Ellsberg, Daniel (1961) "Risk, ambiguity, and the Savage axioms," *Quarterly Journal of Economics*, 643–669. 3

Epstein, Larry G, Yoram Halevy et al. (2019) *Hard-to-interpret signals*: University of Toronto, Department of Economics. 22

Epstein, Larry G and Martin Schneider (2007) "Learning under ambiguity," *Review of Economic Studies*, 74 (4), 1275–1303. 21

Eyster, Erik and Matthew Rabin (2010) "Naive herding in rich-information settings," *American Economic Journal: Microeconomics*, 2 (4), 221–43. 22

Ford, JL, David Kelsey, and Wei Pang (2013) "Information and ambiguity: herd and contrarian behaviour in financial markets," *Theory and Decision*, 75 (1), 1–15. 21

Frick, Mira, Ryota Iijima, and Yuhta Ishii (2020a) "Misinterpreting others and the fragility of social learning," *Econometrica*, 88 (6), 2281–2328. 22

——— (2020b) "Stability and robustness in misspecified learning models." 22, 30

Fryer Jr, Roland G, Philipp Harms, and Matthew O Jackson (2019) "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization," *Journal of the European Economic Association*, 17 (5), 1470–1501. 21

Ghirardato, Paolo, Fabio Maccheroni, and Massimo Marinacci (2004) "Differentiating ambiguity and ambiguity attitude," *Journal of Economic Theory*, 118 (2), 133–173. 17

Gilboa, Itzhak and David Schmeidler (1989) "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18, pp–141. 3, 7

Golub, Benjamin and Evan Sadler (2017) "Learning in social networks," *Available at SSRN 2919146*. 2

Guarino, Antonio and Philippe Jehiel (2013) "Social learning with coarse inference," *American Economic Journal: Microeconomics*, 5 (1), 147–74. 22

Hanany, Eran and Peter Klibanoff (2008) "Updating ambiguity averse preferences,"Technical report, Discussion Paper. 19

Hansen, Lars Peter et al. (2014) "Uncertainty outside and inside economic models," *Journal of Political Economy*, 122 (5), 945–87. 2

Hansen, Lars Peter and Massimo Marinacci (2016) "Ambiguity aversion and model misspecification: An economic perspective," *Statistical Science*, 31 (4), 511–515. 2, 22

Herrera, Helios and Johannes Hörner (2012) "A necessary and sufficient condition for information cascades." 2

Hurwicz, Leonid (1951) "Some specification problems and applications to econometric models," *Econometrica*, 19 (3), 343–344. 17

Kalai, Ehud and Ehud Lehrer (1993) "Rational learning leads to Nash equilibrium," *Econometrica*, 1019–1045. 20

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji (2005) "A smooth model of decision making under ambiguity," *Econometrica*, 73 (6), 1849–1892. 7, 18

Kovach, Matthew (2021) "Ambiguity and partial Bayesian updating," *arXiv preprint arXiv:2102.11429*. 22

Marinacci, Massimo (2002) "Learning from ambiguous urns," *Statistical Papers*, 43 (1), 143. 21

——— (2015) "Model uncertainty," *Journal of the European Economic Association*, 13 (6), 1022–1100. 2

Marinacci, Massimo and Filippo Massari (2019) “Learning from ambiguous and misspecified models,” *Journal of Mathematical Economics*, 84, 144–149. 21

Molavi, Pooya, Alireza Tahbaz-Salehi, and Ali Jadbabaie (2018) “A theory of non-Bayesian social learning,” *Econometrica*, 86 (2), 445–490. 22

Pires, Cesaltina (2002) “A rule for updating ambiguous beliefs,” *Theory and Decision*, 53 (2), 137–152. 6

Rosenberg, Dinah and Nicolas Vieille (2019) “On the efficiency of social learning,” *Econometrica*, 87 (6), 2141–2168. 16, 31

Saito, Kota (2012) “Subjective timing of randomization and ambiguity,”Technical report, mimeo. 7

Seo, Kyoungwon (2009) “Ambiguity and second-order belief,” *Econometrica*, 77 (5), 1575–1605. 7

Smith, Lones and Peter Sørensen (2000) “Pathological outcomes of observational learning,” *Econometrica*, 68 (2), 371–398. 2, 5, 9, 10, 15, 27, 30

Smith, Lones, Peter Norman Sørensen, and Jianrong Tian (2021) “Informational herding, optimal experimentation, and contrarianism,” *Review of Economic Studies*, 88 (5), 2527–2554. 2

Tang, Rui (2022) “A Theory of Contraction Updating,” *Available at SSRN*. 22

Wald, Abraham (1950) “Statistical decision functions..” 3

A Proofs

A.1 Proof of Theorem 1

Lemma 2. *For all normalized DGP, F, 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 $\operatorname{supp}(F)$).

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

Lemma 3. *When $\mathcal{F}_0 = \mathcal{F}$, 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. 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 DGP such that $\operatorname{supp}(F_\gamma) = \left\{\gamma, \frac{1}{\gamma}\right\}$, i.e., the "most informative" DGP 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

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

From Lemma 2 (1), we know that for all $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$,

$$\inf_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)} \geq 1. \tag{4}$$

Combining (3) and (4), we obtain $r_{i+1} \geq \sqrt{\gamma} \times r_i$ for all $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$, when $a_i = 1$. The discussion of $a_i = 0$ is symmetric. □

Theorem. *When $\mathcal{F}_0 = \mathcal{F}$, an information cascade occurs $\mathbb{P}^*$-almost surely.*

Proof. If signals are unbounded, i.e., $\gamma = \infty$, 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. If signals are bounded, Lemma 3 implies that there exists some $K < \infty$ such that for all $r_i \in \left(\frac{1}{\gamma}, \gamma\right)$, K consecutive action θs will bring r_i into the cascade set C_θ, so an information cascade of action θ is triggered. When $r_i \geq 1$, 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 lead to a cascade of action 1. Lemma 2 implies that

$$\frac{\mathbb{P}^* (\lambda_i > 1)}{1 - \mathbb{P}^* (\lambda_i > 1)} = \frac{1 - \overline{F}^0 (1)}{\overline{F}^0 (1)} = \frac{\overline{F}^1 (1)}{\overline{F}^0 (1)} \geq \frac{1}{\gamma} \Rightarrow \mathbb{P}^* (\lambda_i > 1) \geq \frac{1}{1+\gamma}. \tag{5}$$

As a result,

$$\mathbb{P}^* (\text{Cascade}|r_i \geq 1) \geq \mathbb{P}^* (\lambda_i, \lambda_{i+1}, ..., \lambda_{i+K-1} > 1|r_i \geq 1) \geq \left(\frac{1}{1+\gamma}\right)^K > 0, \tag{6}$$

and similarly,

$$\mathbb{P}^* (\text{Cascade}|r_i < 1) \geq \left(\frac{\gamma}{1+\gamma}\right)^K > 0. \tag{7}$$

Levy's 0-1 Law shows that as $i \to \infty$, we have

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

(6) and (7) imply that $\mathbb{P}^* (\text{Cascade}|h_i) > \left(\frac{1}{1+\gamma}\right)^K > 0$ for all i, so we must have $1_{\text{Cascade}} = 1$ $\mathbb{P}^*$-almost surely—i.e., a cascade almost surely happens. □

A.2 Proof of Theorem 2

Proof. **Proof of Theorem 2 (1):** Suppose that there is $F \in \mathcal{F}_0$, which is discrete at γ. Denote $p = F^0 \left(\frac{1}{\gamma}\right) > 0$, 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:

$$\bar{l}_{i+1} = \bar{l}_i \times \sup_{F_i \in \mathcal{F}_0} \frac{1 - F_i^1\left(\frac{1}{r_i}\right)}{1 - F_i^0\left(\frac{1}{r_i}\right)} \geq \bar{l}_i \times \frac{1 - F^1\left(\frac{1}{r_i}\right)}{1 - F^0\left(\frac{1}{r_i}\right)} \geq \bar{l}_i \cdot \left[\lim_{r \to \gamma} \frac{1 - F^1\left(\frac{1}{r}\right)}{1 - F^0\left(\frac{1}{r}\right)}\right] = \bar{l}_i \cdot \frac{1 - \frac{1}{\gamma} \cdot p}{1 - p}, \tag{8}$$

where the inequality line comes from Property (3) in Lemma 2, and the last equality comes from the discreteness of signals. Also, 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.

Proof of Theorem 2 (2): Suppose that there exists some $F \in \mathcal{F}_0$ such that F^1 is continuously differentiable on $(\gamma - \varepsilon, \gamma)$ with $f^1(\gamma) > \frac{2}{\gamma - 1}$. When F is discrete at γ, an information cascade occurs almost surely, as implied by condition (1). I thus focus on the case in which F 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 $I'(\gamma) \equiv \lim_{\delta \to 0} I'(\gamma - \delta)$; then we have

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

where the second 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 Theorem 3

A.3.1 Local Stability under Ambiguity

I first introduce the notion of local stability under ambiguity, which is parallel to that in Bayesian learning—e.g., Smith and Sørensen (2000) and Frick et al. (2020b).

Definition 5. [Local Stability under Ambiguity] State 0 (or state 1) is ***locally stable*** if there exists some $r \in \mathbb{R}_{++}$ 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 set Π_0 with $r_0 < r$ (or $r_0 > R$).

Here $\mathbb{P}^*_{r_0}$ denotes the true probability measure conditional on the average prior likelihood ratio's being r_0. Roughly, state θ is locally stable if posteriors will converge to δ_θ with a strictly positive probability when priors are close to δ_θ. We have the following results.

Lemma 4. *Suppose that $\mathcal{F}_0$ contains a DGP with unbounded signals. Then, a herd 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 the herd of action 1. First, suppose that $r_i \to \infty$; then we must have a herd of action 1, because if an action 0 is taken by an individual i, then

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

which contradicts $r_i \to \infty$. Second, suppose that a herd of action 1 occurs, then

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

so $\{r_i\}$ is an increasing sequence and has a limit in $\mathbb{R} \cup \{+\infty\}$. If r_i does not diverge to infinity, it must converge to some $R < \infty$. Let F be the unbounded DGP that $\mathcal{F}_0$ contains; then

$$r_{i+1} \geq r_i \times \sqrt{\frac{1 - F^1(1/r_i)}{1 - F^0(1/r_i)}}. \tag{9}$$

Taking the limit on both sides of (9), we obtain $R \geq \sqrt{\frac{1-F^1(1/R)}{1-F^0(1/R)}} \times R$, so $\sqrt{\frac{1-F^1(1/R)}{1-F^0(1/R)}} \leq 1$. Since F has unbounded signals, Lemma 2 (1) implies that $\sqrt{\frac{1-F^1(1/R)}{1-F^0(1/R)}} > 1$, which is a contradiction. As a consequence, $r_i \to \infty$. □

Lemma 5. *Suppose that $\mathcal{F}_0$ contains a DGP with unbounded signals. If both* 0 *and* 1 *are locally stable, then (i) both correct and incorrect herding occur with a $\mathbb{P}^*$-strictly positive probability and (ii) herding occurs $\mathbb{P}^*$-almost surely.*

Proof. (i) From the definition of local stability, Lemma 4, and the fact that $\{r_i\}$ is a Markov process, we know that both correct and incorrect herding occur with a strictly positive probability when r_i is sufficiently large or small—that is, $r_i \in C = \{r_i < r\} \cup \{r_i > R\}$ for some $r, R \in (0, +\infty)$. Outside of C, r_i is bounded away from 0 and $+\infty$, so there exists some $K < \infty$ such that K identical actions can bring r_i into C.[17] This further implies that $\{r_i \to 0\}$ and $\{r_i \to \infty\}$ both occur with a strictly positive probability—i.e., both types of herding occur with a positive probability. (ii) Denote by $H = \{r_i \to 0\} \cup \{r_i \to \infty\}$, which denotes the event of herding by Lemma 4. Levy's 0-1 Law implies that $\mathbb{P}^*(H|h_i) \to \mathbb{P}^*(H|h_\infty) = 1_H \in \{0, 1\}$. The arguments in (i) imply that we can find a constant $\delta > 0$ such that for all possible history h_i, $\mathbb{P}^*(H|h_i) > \delta$, so $1_H = 1$ almost surely—i.e., herding almost surely occurs. □

A.3.2 Formal Proof of Theorem 3

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. Rosenberg and Vieille (2019) show 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 $\sqrt{\frac{1-F^1(x)}{1-F^0(x)}} \sim \sqrt{1+F^0(x)} = 1 + \frac{1}{2}F^0(x) + o(F^0(x))$, which proves the lemma. □

Lemma 7. *Under the conditions of Theorem 3, state* 1 *is locally stable.*

Proof. We want to show that there exists some $R < \infty$ such that for all $r_0 \geq R$, the probability of an action-1 herd is greater than some $\varepsilon > 0$. Let H_θ denote the event in which $a_i = \theta$ for all i, i.e., an action-θ herd. We have

$$\mathbb{P}^*_{r_0}(H_1) = \lim_{i\to\infty} \mathbb{P}^0_{r_0}(a_1 = a_2 = ...a_i = 1) = \prod_{i=1}^{\infty}\left[1 - \overline{F}^0_i\left(\frac{1}{r_i}\right)\right] \geq \prod_{i=1}^{\infty}\left[1 - a \times \left(\frac{1}{r_i}\right)^{\alpha}\right], \tag{10}$$

[17] Suppose that $r_i \in [r, R]$. Let F be any unbounded DGP contained in $\mathcal{F}_0$. Then when $a_i = 0$, we have $r_{i+1} \leq r_i \times \sqrt{\frac{F^1(1/r_i)}{F^0(1/r_i)}} \leq r_i \times \sqrt{\frac{F^1(1/r)}{F^0(1/r)}}$. Because $r \in (0, \infty)$, we have $r_{i+1}/r_i \leq \sqrt{\frac{F^1(1/r)}{F^0(1/r)}} \equiv \beta < 1$, so after $K = \left\lceil \log_\beta^{r/R} \right\rceil + 1$ consecutive action 0s, we $r_{i+K} < r$. Similarly, K consecutive action 1s will result in $r_{i+K} > R$.

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

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

where F denotes the DGP in $\mathcal{F}_0$ such that $x^p = o(F^0(x))$. Let $q \in (p, \alpha)$, then we have

$$\begin{aligned}
\lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1(1/r)}{1-F^0(1/r)}} - 1}{\left(1 + \frac{1}{r^q}\right)^{1/q} - 1} &= \lim_{r\to\infty} \frac{\sqrt{\frac{1-F^1(1/r)}{1-F^0(1/r)}} - 1}{\frac{1}{r^q}} \times \lim_{r\to\infty} \frac{\frac{1}{r^q}}{\left(1 + \frac{1}{r^q}\right)^{1/q} - 1} \\
&= \lim_{r\to\infty} \frac{\frac{1}{2}F^0(1/r)}{\frac{1}{r^q}} \times \lim_{r\to\infty} \frac{\frac{1}{r^q}}{\left(1 + \frac{1}{r^q}\right)^{1/q} - 1} \qquad (11) \\
&> \lim_{r\to\infty} \frac{\frac{1}{2}(1/r)^p}{\frac{1}{r^q}} \times q = \infty,
\end{aligned}$$

where (11) 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

$$r_i \geq \left(r_0^q + i\right)^{1/q}, \qquad \forall i \geq 1. \qquad (12)$$

After substituting (12) into (10), we know that for all $r_0 \geq R$,

$$\mathbb{P}_{r_0}^*(H_1) \geq \prod_{i=1}^{\infty} \left[1 - a \times \left(\frac{1}{r_i}\right)^{\alpha}\right] \geq \prod_{i=1}^{\infty} \left[1 - a \times \frac{1}{\left(r_0^q + i\right)^{\alpha/q}}\right] \geq \prod_{i=1}^{\infty} \left[1 - a \times \frac{1}{\left(R^q + i\right)^{\alpha/q}}\right].$$

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 $i \geq 1$. 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 $\sum a \times \frac{1}{(R^q+i)^{\alpha/q}} < \infty$, so

$$\mathbb{P}^*_{r_0}(H_1) \geq \prod_{i=1}^{\infty} \left[1 - a \times \frac{1}{(R^q+i)^{\alpha/q}}\right] \equiv \varepsilon > 0,$$

which establishes the local stability of state 1. □

Lemma 8. *Under the conditions of Theorem 3, state* 0 *is locally stable.*

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

$$\mathbb{P}^*_{r_0}(H_0) = \prod_{i=1}^{\infty} F^0\left(\frac{1}{r_i}\right) = \prod_{i=1}^{\infty} \left[1 - F^1(r_i)\right] \geq \prod_{i=1}^{\infty} \left[1 - F^0(r_i)\right] = \mathbb{P}^*_{1/r_0}(H_1),$$

which says that the probability of a correct herd is higher than that of an incorrect herd. From Step Lemma 7, there exists R such that $\mathbb{P}^*_{1/r_0}(H_1) \geq \varepsilon > 0$ for all $1/r_0 > R$. Let $r = 1/R$, so we also have $\mathbb{P}^*_{r_0}(H_0) \geq \varepsilon > 0$ for all $r_0 < r$, which establishes the local stability of state 0. □

Combining Lemmas 5 to 8, we know that herding occurs almost surely, and an incorrect herd occurs with a strictly positive probability, so Theorem 3 is proved.

A.4 Proof of the Claim in Example (5)

Proof. For simplicity, suppose that $\mathcal{F}_0$ satisfies the conditions in Theorem 2 by containing a discrete model.[18] Under the max-min (and max-max) model, we have $r_{i+1}/r_i \begin{cases} \geq \beta & a_i = 1 \\ \leq 1/\beta & a_i = 0 \end{cases}$ for all $r_i \in (1/\gamma, \gamma)$ and for some $\beta > 1$. Let R_i denote the (essential) average likelihood ratio under the smooth ambiguity model—i.e., individual i chooses action 1 if $\lambda_i \times R_i > 1$ and action 0 otherwise (except for the tie case). Fixing an $\epsilon < \beta - 1$, the continuity of ϕ implies that for all $I < \infty$, there exists some $\overline{\sigma} < \infty$ such that whenever $|\sigma| \geq \overline{\sigma}$, we have

$$R_{i+1}/R_i \begin{cases} \geq \beta - \epsilon & a_i = 1 \\ \leq \frac{1}{\beta - \epsilon} & a_i = 0 \end{cases} \text{ for all } i \leq I.$$

Therefore, when $i \leq I$, an information cascade will be triggered after at most $K \equiv \lceil \log^{\gamma}_{\beta-\epsilon} \rceil + 1$ consecutive actions. Let N_i denote the event that $R_i \in \left(\frac{1}{\gamma}, \gamma\right)$; i.e., an information cascade

[18]The case for continuous model can be discussed analogously as in the proof of Theorem 2.

does not occur when i is making decision. Let $p \equiv \min \left\{\mathbb{P}_g^* (\lambda_i \geq 1), \mathbb{P}_g^* (\lambda_i < 1)\right\}$, where $\mathbb{P}_g^*$ denotes the probability measure induced by g; i.e, the objective probability measure in the problem. We have

$$\frac{\mathbb{P}_g^* (N_{i+K})}{\mathbb{P}_g^* (N_i)} = \frac{\mathbb{P}_g^* (N_{i+K} \cap N_i)}{\mathbb{P}_g^* (N_i)} = \mathbb{P}_g^* (N_{i+K} | N_i) \leq (1-p)^K < 1,$$

where the first equality comes from the fact that $N_{i+K} \subset N_i$, and the last inequality comes from $p \geq \frac{1}{1+\gamma}$ (from the Proof of Theorem 1). Therefore, the expected number of individuals before I who do not face a cascade is

$$\mathbb{E}_g^* \left(\sum_{i \leq I} 1_{N_i} \right) = \sum_{i \leq I} \mathbb{P}_g^* (N_i) < \frac{K-1}{1-(1-p)^K} \equiv M < \infty.$$

As a consequence,

$$\mathbb{P}_g^* (N_1 = ... = N_I = 1) \times I \leq \mathbb{E}_g^* \left(\sum_{i \leq I} 1_{N_i} \right) < M \Rightarrow \mathbb{P}_g^* (N_1 = ... = N_I = 1) \leq M/I.$$

It means that the probability of no cascade before individual I is less than M/I, which implies that the probability of an information cascade is greater than $1 - M/I$. Note that M is a constant, so when I becomes arbitrarily large, the probability of a cascade can be arbitrarily close to 1. □