Sequential Social Learning from Outcomes*

Tuval Danenberg † and Alexander Wolitzky ‡

July 30, 2026

Abstract

We study the classic sequential social learning model when agents observe signals of their predecessors' outcomes, such as their payoffs or output, rather than directly observing their actions. Two distinct forms of confounding—*match confounding* and *oppositional confounding*—can prevent learning. In both cases, informative private signals impede social learning by making other agents' actions less predictable, which makes it more difficult to draw unambiguous lessons from their outcomes. For this reason, learning succeeds when private signals are very noisy, but fails when they are very accurate. The model provides a rationale for why social learning may fail in societies where individuals rely on accurate, private information sources.

*We thank Daron Acemoglu, Navin Kartik, Stephen Morris, and Peter Sørensen for helpful comments and conversations.

†Department of Economics, MIT, tuvaldan@mit.edu

‡Department of Economics, MIT, wolitzky@mit.edu

1 Introduction

The sequential social learning model of Banerjee (1992) and Bikhchandani et al. (1992) ("BHW") posits that individuals observe private signals concerning an unknown state of the world and, in addition, learn from observing earlier agents' actions. This is a natural model of "herd behavior" (to use Banerjee's term) or "fads, fashion, [or] custom," (to use BHW's), where, by definition, agents' actions are publicly observed. However, in many social learning settings, agents' actions as well as their signals are private, and individuals observe only imperfect signals of others' outcomes, such as their payoffs or output. This situation arises whenever actions are "inputs" into a production process, where only the "output" is observed. For example, a farmer may observe her neighbor's crop yield but not her planting technique; a parent may observe whether his child's friends get sick but not whether they were vaccinated; or a pharmaceutical company CEO may observe the performance of a competitor's drug but not their research strategy. In all of these examples, the observed outcome (crop yield, health status, drug efficacy) is a stochastic consequence of both an agent's action and the state. This feature can lead to confounding between agents' actions and the state, which can impede social learning in a different manner from the information cascades emphasized by Banerjee and BHW. For instance, a parent who observes a certain disease prevalence in his community may not know whether this is caused by a highly infectious disease together with a high vaccination rate (in which case the parent should also vaccinate his own child), or a less infectious disease together with a lower vaccination rate (in which case perhaps he should not).[1]

This paper studies action-state confounding and the prospects for successful social learning in a variant of the classic sequential social learning model, where agents observe others' outcomes rather than their actions. We focus on a simple and canonical setting with a binary state $\theta \in \{L, H\}$, binary actions $a \in \{L, H\}$, and binary outcomes $y \in \{0, 1\}$. For example, a disease can be highly infectious, or not; a child can get vaccinated, or not; and she can get sick, or not. Agents observe private signals of the state as well as earlier agents' outcomes (but not actions). To focus on action-state confounding, we rule out infinite information cascades by assuming that, conditional on each action, the outcome statistically identifies the state. We ask when social learning succeeds, meaning that agents eventually take the correct action, $a = \theta$.

A preliminary result (Theorem 1) is that learning fails for some prior belief if and only if there exists a *confounding public belief* such that an agent who starts with this belief and

[1] We will use vaccination as a running example of our social learning problem, but we hasten to add that our model is one of pure social learning—i.e., there are only informational externalities, not direct payoff externalities—whereas, of course, in reality vaccination also provides payoff externalities.

then observes a private signal before choosing her action obtains each outcome with the same probability in both states, rendering her outcome uninformative for subsequent agents. This result follows from standard arguments (e.g., Smith and Sørensen, 2000).

Our first main result (Theorem 2) characterizes when a confounding belief exists. The result has three parts.

First, a confounding belief always exists if the outcome probabilities $\mathbb{P}_{a\theta}$—the probability of obtaining outcome 1 from action a in state θ—satisfy *match confounding*, meaning that $\mathbb{P}_{LL} > \mathbb{P}_{LH}$ and $\mathbb{P}_{HH} > \mathbb{P}_{HL}$. That is, under match confounding, the effect of the state θ on the probability $\mathbb{P}_{a\theta}$ of outcome 1 is determined by its match with the action a: for each action a, outcome 1 is more likely if $\theta = a$, so the action is correct. For example, in agriculture, the state can correspond to the suitability of the soil for different planting techniques, while the action corresponds to the chosen technique, and the outcome corresponds to whether the crop succeeds. In general, match confounding captures situations where the outcome is a signal of the current agent's payoff, so agents observe signals of others' performance or profit.

Second, a confounding belief exists under a condition on the private signal distribution if the outcome probabilities satisfy *oppositional confounding*, meaning that $\mathbb{P}_{HL} < \mathbb{P}_{HH} < \mathbb{P}_{LL} < \mathbb{P}_{LH}$. (The condition on private signals is that they are sufficiently informative, in a particular sense.) That is, under oppositional confounding, the effect of the state and the action on the probability of outcome 1 oppose each other—$\mathbb{P}_{a\theta}$ is increasing in θ (for each a) but decreasing in a (for each θ)—and the action "matters more," in that $\mathbb{P}_{HH} < \mathbb{P}_{LL}$. For instance, in the vaccine example, the state can correspond to the infectiousness of a disease, while the action is the decision to vaccinate (which is assumed to be optimal if and only if the disease is highly infectious), and the outcome is whether the child gets sick. In general, oppositional confounding describes settings where the action and the state have opposite effects on the outcome, and the action has a larger effect than the state.

Third, up to relabeling states and actions, these are the only possibilities. In the absence of match or oppositional confounding (or if there is oppositional confounding but private signals are insufficiently informative), learning succeeds for any prior belief.

These results characterize when learning succeeds for any prior, but they admit the possibility that learning can fail only for non-generic priors. Our remaining results characterize when learning succeeds for a generic set of priors, or, conversely, fails for an open and non-empty set of priors. We first provide general conditions for these properties based on a stability analysis of non-linear stochastic difference equations (Theorem 3), following Ellison and Fudenberg (1995) and Smith and Sørensen (2000). Building on these conditions, our second main result (Theorem 4) shows that, under either match or oppositional confound-

Table 1: Summary of Results

Outcome configuration	**Noisy private signals**	**Precise private signals**
Match confounding	Unstable confounding point *Learning generically succeeds*	Stable confounding point *Learning robustly fails*
Oppositional confounding	No confounding point *Learning succeeds*	Stable confounding point *Learning robustly fails*
Neither	*Learning succeeds*	*Learning succeeds*

ing (and assuming a regularity condition), learning generically *succeeds* if private signals are sufficiently *uninformative*, and learning robustly *fails* if private signals are sufficiently *informative*. Thus, informative private signals impede social learning, while uninformative ones facilitate it. The intuition for this result goes to the heart of the obstacle to social learning posed by action-state confounding. In sequential social learning with stochastic outcomes but no private signals, each agent can perfectly infer her predecessors' actions, except in non-generic cases of indifference. As we assume that the outcome identifies the state conditional on each action, this implies that social learning generically succeeds in the absence of private signals. Similarly, social learning generically succeeds if private signals are sufficiently uninformative that there is not enough uncertainty about predecessors' actions to confound learning from outcomes. Conversely, if private signals are very informative, there is a great deal of uncertainty about predecessors' actions, and we show that this leads learning to robustly fail under either match or oppositional confounding. In short, informative private signals create uncertainty about predecessors' actions, which makes it more difficult to draw unambiguous lessons from their outcomes.

A subtlety underlying these results is that, while informative private signals impede social learning under either match or oppositional confounding, they do so in a different manner in each case. With match confounding, there is always a confounding public belief, but it is unstable with uninformative private signals (and stable with informative ones). With oppositional confounding, there is no confounding public belief with uninformative private signals, but there is a stable confounding public belief with informative ones.[2] Thus, under match confounding, informative private signals stabilize the confounding public belief; while under oppositional confounding, they are also necessary for confounding. These points are summarized in Table 1, which organizes our main results.

While we find that informative private signals impede social learning, these signals of course also provide useful private information. Expected long-run payoffs are therefore non-

[2]With signals of intermediate informativeness, there can be an unstable confounding public belief.

monotone in private-signal precision: high when signals are noisy and social learning succeeds; low when signals are more precise and social learning fails; and then high again when signals are so precise that social learning is unnecessary.

Our results provide a novel mechanism through which improving the precision of private information sources (e.g., personalized news or social media feeds) can impede social learning when individual decisions are private. Consider vaccination. In a world where everyone receives the same public health information, it is relatively easy to estimate what share of other parents vaccinate their children, which makes it possible to draw inferences about the infectiousness of the disease or the efficacy of the vaccine from observing how many children get sick. If instead everyone has their own information source, a parent does not know what information other parents are getting and hence cannot infer the vaccination rate, so he cannot draw useful inferences from the prevalence of the disease. More generally, similar informational features may arise when individuals learn about the returns to many other personal choices, such as health-related behaviors (e.g., contraception, weight loss drugs, dietary supplements, diet or exercise regimes), or strategies for education (e.g., private tutoring or test prep), investment (e.g., leverage, cryptocurrency), or agriculture (e.g., planting technique, seed variety choice).[3] Given this, we believe that further investigating the relationship between the privacy of information sources and actions and the performance of social learning is a promising direction for future theoretical and empirical research.

Related Literature

Among the large literature on social learning following Banerjee and BHW (surveyed by Chamley, 2004 and Bikhchandani et al., 2024), the papers most related to the present one are Smith and Sørensen (2000) and Wolitzky (2018). Smith and Sørensen (2000) study a general sequential social learning model with observed actions and introduce the Markov-martingale methods we rely on.[4] Substantively, they also introduce the possibility of confounded social learning.[5] In their model, confounding occurs because agents have unobserved, heterogeneous preferences. Viewing an agent's choice as the cutoff preference type above which she takes action H, this is equivalent to allowing imperfectly observed choices, but where there are more than two options and observation "noise" is independent of the state. In contrast, we

[3]Empirical evidence on the role of social learning in these settings includes Allen et al. (2024) on vaccination, Munshi and Myaux (2006) on contraception, Aral and Nicolaides (2017) on exercise, and large literatures on agriculture and finance, surveyed by Foster and Rosenzweig (2010) and Kuchler and Stroebel (2021), respectively.

[4]Like us, Smith and Sørensen (2000) assume a binary state. Arieli and Mueller-Frank (2021) and Kartik et al. (2024) study sequential social learning with observed actions and multiple states.

[5]Antecedents include McLennan (1984), Bala and Goyal (1995), and Piketty (1995).

assume that, conditional on each action, outcomes identify the state. This makes our model and substantive results very different from Smith and Sørensen's. For instance, the match and oppositional confounding cases and the result that precise private signals hinder social learning do not arise in their model.

Wolitzky (2018) introduces outcome-based learning in a model of technology adoption similar to our oppositional confounding case. However, that paper assumes a continuum population with random sampling and deterministic adoption dynamics in each state, similar to Banerjee and Fudenberg (2004), rather than sequential social learning as in Banerjee (1992) and BHW. Thus, while the motivation of studying outcome-based learning is similar to the current paper, the analysis and results are very different. In particular, our key mechanism whereby informative private signals impede social learning relies on sequential social learning and is therefore absent in Wolitzky (2018): indeed, in that paper, social learning succeeds with informative private signals and fails with uninformative ones. We discuss this difference between outcome-based learning in these two settings in Section 5.2.[6]

Our result that precise private signals impede learning under match or oppositional confounding may be reminiscent of the familiar idea that social learning can be facilitated by "guinea pigs" or "sacrificial lambs" who lack social information and hence act only based on their private signals (e.g., Sgroi, 2002, Acemoglu et al., 2011, Song, 2016, Kremer et al., 2014, Che and Hörner, 2018). However, the logic is almost the opposite. The "guinea pigs" logic is that reducing an agent's social information makes their action more responsive to their private information, so others can learn from their action. Our logic is that reducing an agent's private information makes their action more responsive to social information—and hence more predictable—so others can learn from their outcome.

2 Preliminaries

2.1 Model

We consider a sequential social learning model where agents observe not their predecessors' actions, but their *outcomes*—noisy signals that depend on both the action and the state. The state, actions, and outcomes are all binary, and are denoted, respectively, by $\theta \in \{L, H\}$, $a \in \{L, H\}$, and $y \in \{0, 1\}$.

Payoffs are given by $u(a, \theta) = \mathbf{1}\{a = \theta\}$. Before choosing her action, each agent t observes the public history $h_{t-1} = (y_1, \ldots, y_{t-1})$ of her predecessors' outcomes and a private signal

[6]Fershtman (2024) studies sequential outcome-based social learning (where actions may or may not also be observed), focusing on the question of when equilibrium is utilitarian optimal.

of the state. The probability of outcome 1 when an agent takes action a and the state is θ is denoted $\mathbb{P}_{a\theta}$ and is independent across agents (conditional on actions and the state). We maintain the following assumption on the outcome probabilities.

Assumption 1.

1. $0 < \mathbb{P}_{a\theta} < 1$ for all a, θ.
2. $\mathbb{P}_{LL} \neq \mathbb{P}_{LH}$ and $\mathbb{P}_{HL} < \mathbb{P}_{HH}$.
3. $\mathbb{P}_{L\theta} \neq \mathbb{P}_{H\theta}$ for at least one of the two states θ.

Assumption 1.1 says that the outcome distribution has full support. Assumption 1.2 says that the state non-trivially affects the outcome, conditional on either action, and then without loss labels the actions and states so that, given action H, outcome 1 is more likely in state H. Without this assumption, either the model would have state-independent observation noise, which is equivalent to the "crazy types" allowed by Smith and Sørensen (2000); or it would fall in a mixed case where the outcome probabilities depend on the state conditional on one action but not the other.[7] Assumption 1.3 says that the action non-trivially affects the outcome. Without this assumption, the model would be one of individual rather than social learning. Finally, note that our model nests the standard observed-actions model as the case where $\mathbb{P}_{LL} = \mathbb{P}_{LH} = 0$ and $\mathbb{P}_{HL} = \mathbb{P}_{HH} = 1$: this case violates Assumptions 1.1 and 1.2, but it can be approximated by $(\mathbb{P}_{a\theta})$ satisfying these assumptions.

As in Smith and Sørensen (2000), we identify private signals with the beliefs they induce. Agent t's private belief that the state is H is denoted $p_t \in (0,1)$ and is conditionally i.i.d. with cdf F^θ. We assume that F^L and F^H admit continuous densities f^L and f^H supported on a common interval $[\underline{b}, \bar{b}] \subset [0,1]$, with $\underline{b} < \bar{b}$. Thus, private signals are continuously distributed and informative, but do not perfectly reveal the state.[8] Private beliefs are *bounded* if $0 < \underline{b} < \bar{b} < 1$ and *unbounded* if $\underline{b} = 0, \bar{b} = 1$. In addition, by the *no-introspection property* (Smith and Sørensen, 2000, Lemma A.1(a)), $f^L(p)/f^H(p) = (1-p)/p$ for all $p \in (\underline{b}, \bar{b})$.

Given an outcome history h_t, let q_t denote the *public belief* that the state is H, i.e., the posterior given history h_t and a neutral private belief $p = 1/2$. Let r denote the posterior belief of an agent with private belief p and public belief q, which by Bayes rule is given by

$$r = r(p,q) = \frac{pq}{pq + (1-p)(1-q)}. \tag{1}$$

[7] Xu (2025) analyzes a general model of state-independent observation noise. Wolitzky (2018) considers the mixed case in a continuum-agent model.

[8] We consider discrete signals, including the special case of completely uninformative signals, in Appendix D.

Note that $r \leqslant 1/2$—so action L is optimal—iff $p + q \leqslant 1$.

Finally, let $l = (1-q)/q$ denote the *public likelihood ratio* that the state is L rather than H. Let $\bar{p}(l) = l/(1+l)$, and note that $r \leqslant 1/2$ iff $p \leqslant \bar{p}(l)$. Denote the prior belief by $q_0 = \mathbb{P}(\theta = H)$ and the prior likelihood ratio by $l_0 = (1-q_0)/q_0$.

2.2 Belief Dynamics and Long-Run Learning

Since private beliefs are continuously distributed, agents almost surely have a strict preference for one action: the event $\{p = \bar{p}(l)\}$ has zero probability. The model thus has a unique Nash equilibrium, up to tie-breaking on zero-probability events. In equilibrium, in any period, the probability $\psi(y|\theta, l)$ that the outcome is y if the state is θ and the public likelihood ratio is l, and the next-period public likelihood ratio $\phi(y, l)$ following signal y, are given by

$$\begin{aligned} &\psi(1|\theta, l) = F^\theta(\bar{p}(l))\mathbb{P}_{L\theta} + (1 - F^\theta(\bar{p}(l)))\mathbb{P}_{H\theta}, \qquad \psi(0|\theta, l) = 1 - \psi(1|\theta, l), \qquad \text{and} \\ &\phi(y, l) = l\frac{\psi(y|L, l)}{\psi(y|H, l)}. \end{aligned} \tag{2}$$

As in Smith and Sørensen (2000), conditional on the state θ, the pair (y_t, l_t) follows a time-homogeneous Markov process on the state space $\{0, 1\} \times \mathbb{R}_+$: given (y_t, l_t), the next pair is $(y_{t+1}, \phi(y_{t+1}, l_t))$ with probability $\psi(y_{t+1}|\theta, l_t)$.

The following are standard observations (cf. Smith and Sørensen, 2000, Lemma 3). Note that it is without loss to condition on $\theta = H$; if instead $\theta = L$, the same analysis applies with l_t^{-1} in place of l_t.

Lemma 1. *The public likelihood ratio process (l_t) is a martingale conditional on $\theta = H$. Hence, conditional on $\theta = H$, (l_t) converges almost surely to a random variable $l_\infty = \lim_{t\to\infty} l_t$, with $\operatorname{supp}(l_\infty) \subset \mathbb{R}_+$.*

Still following Smith and Sørensen (2000), define the *cascade sets* $J_L = \{l : \operatorname{supp}(F) \subset [0, \bar{p}(l)]\}$, $J_H = \{l : \operatorname{supp}(F) \subset [\bar{p}(l), 1]\}$. If the public likelihood ratio l_t lies in the interior of a cascade set J_a, then agent t takes action a regardless of her private signal. However, whereas in Smith and Sørensen (2000) learning ceases if the public belief enters a cascade set, the exact opposite is true in our model: if the public belief is in a cascade set J_a, then the probability of outcome 1 in state θ is $\mathbb{P}_{a\theta}$, which differs across the states by Assumption 1.2. This implies that $\phi(0, l) \neq \phi(1, l)$ if l is in a cascade set—so, in our model, learning can only stop outside a cascade set.

We focus on the prospects for social learning in the long run. We say that *learning*

succeeds if the agents' expected payoffs converge to the first best:

$$\mathbb{P}\left(\mathbb{E}[u(a_{t+1},\theta)|l_t,\theta] \to 1\right) = 1 \qquad \text{for all } \theta \in \{L,H\}.$$

Conversely, *learning fails* if $\mathbb{P}\left(\mathbb{E}[u(a_{t+1},\theta)|l_t,\theta] \to 1\right) < 1$ for at least one state θ. A consequence of Lemma 1 is that learning succeeds if and only if the limit belief lies in the correct cascade set.

Lemma 2. $\mathbb{P}\left(\mathbb{E}[u(a_{t+1},H)|l_t,H] \to 1\right) = 1$ *if and only if* $\text{supp}(l_\infty) \subset J_H$.

Whether learning succeeds or fails may depend on the prior l_0.[9] Recall that a set is *meagre* if it is a countable union of nowhere dense sets. We call a set *generic* if its complement is meagre,[10] and *robust* if it is open and non-empty. Thus, learning *generically succeeds* if it succeeds outside of a meagre set of priors, and learning *robustly fails* if it fails for an open and non-empty set of priors.[11]

3 Identification and Confounding

3.1 Uniform Identification

A key determinant of learning is whether the outcome y statistically identifies the state θ. An important feature of our model is that this identification property depends endogenously on the public belief and agents' best responses.

To see this, let

$$\Delta(y,l) := \psi(y|H,l) - \psi(y|L,l),$$

the difference in the probability of outcome y between the two states when the public likelihood ratio is l. Note that $\Delta(y,l) = -\Delta(1-y,l)$ so we can define $\mathbf{\Delta}(l) := |\Delta(1,l)| = |\Delta(0,l)|$ as the magnitude of the difference in outcome probabilities.

Definition 1 (Identification)**.** The state is *identified* at public likelihood ratio l if $\mathbf{\Delta}(l) \neq 0$, and is *confounded* if $\mathbf{\Delta}(l) = 0$.

The state is *uniformly identified* (*UI*) if $\mathbf{\Delta}(l) \neq 0$ for all $l > 0$.

[9] That is, learning from one prior and learning from all priors differ in our model, as in Kartik et al. (2024).

[10] The topological term for what we call a generic set is *residual* or *comeagre*.

[11] This terminology is similar to Smith and Sørensen's, but they call a set generic if it is open and dense, so our notion of learning success is more permissive. For example, if learning succeeds for priors $q_0 \in [0,1]\backslash\mathbb{Q}$ and fails for priors $q_0 \in \mathbb{Q}$, then we would say that learning generically succeeds, while Smith and Sørensen would remain silent.

Uniform identification (UI) says that the outcome is informative of the state for any public likelihood ratio. By equation (2), this implies that $\phi(0,l) \neq \phi(1,l)$ for all l, so social learning never stops. Conversely, if UI fails, then there is some l—a *confounding point*, where $\mathbf{\Delta}(l) = 0$—such that if the public likelihood ratio reaches l then social learning ceases. Mathematically, the confounding points are the interior stationary points of the public likelihood ratio process (l_t).

By Assumption 1.2, a confounding point cannot lie in a cascade set. In other words, the only source of confounding in our model is confounding between the randomness of the current agent's action and the randomness of her outcome conditional on her action.

So far, the question of when UI holds in our model is entirely obscure, and we will investigate this momentarily. Nonetheless, a useful organizing result is that UI characterizes when learning succeeds for all priors.

Theorem 1. *Learning succeeds for all priors if and only if UI holds.*

The logic of Theorem 1 is standard. Clearly, if $\mathbf{\Delta}(l) = 0$ for some l, then learning fails when the prior equals l, because the public likelihood ratio is then fixed at l forever. Conversely, if UI holds then learning succeeds for any prior. To see this, note that the assumption that the signal distributions are continuous implies that $\psi(y|\theta,l)$ and $\phi(y,l)$ are continuous in l for each y. In turn, this implies (by Smith and Sørensen, 2000, Theorem B.2) that any limit belief $l \in \text{supp}(l_\infty)$ conditional on $\theta = H$ is a stationary point of (l_t). But UI implies that (l_t) cannot have an interior stationary point, so the only stationary point—and hence the only limit belief—is $l = 0$, so learning succeeds.

3.2 Match and Oppositional Confounding

In light of Theorem 1, two questions determine the prospects for learning: When does UI hold? And, when UI fails, does learning fail only for non-generic priors, or does it fail robustly? This section addresses the first question; we turn to the second in the next section.

We start with some terminology.

Definition 2. There is *match confounding* if $\mathbb{P}_{LH} < \mathbb{P}_{LL}$ and $\mathbb{P}_{HL} < \mathbb{P}_{HH}$.

There is *oppositional confounding* if $\mathbb{P}_{HL} < \mathbb{P}_{HH} < \mathbb{P}_{LL} < \mathbb{P}_{LH}$.

Under match confounding, for each action a, outcome 1 is more likely if $\theta = a$, so the action is correct. Match confounding thus captures situations where the outcome is a signal of the current agent's payoff, such as their performance or profit.[12]

[12] However, match confounding is broader than this, as it also covers the cases where $\mathbb{P}_{LH} < \mathbb{P}_{LL} < \mathbb{P}_{HL} < \mathbb{P}_{HH}$ or $\mathbb{P}_{HL} < \mathbb{P}_{HH} < \mathbb{P}_{LH} < \mathbb{P}_{LL}$, where $\Pr(y = 1)$ is non-monotone in the current agent's payoff.

Under oppositional confounding, $\mathbb{P}_{a\theta}$ is increasing in θ (for each a) but decreasing in a (for each θ), and the action "matters more," in that $P_{HH} < P_{LL}$. Oppositional confounding fits the vaccine example discussed in the introduction. For another example, suppose that the outcome is a signal of an agent's "output," with outcome 1 indicating high output. Suppose that high output is more likely when the state is H or the action is L, with the interpretation that a is an inverse measure of the agent's "effort"—for instance, because effort and θ are substitutes in producing output, so low effort ($a = H$) is optimal when $\theta = H$. Then, if $P_{HH} < P_{LL}$, there is oppositional confounding.

Oppositional confounding is similar to the "cost-saving innovation" case in Wolitzky (2018). There, action L is a high-cost, status-quo technology that successfully produces output with a known probability, and action H is a cost-saving innovation that always produces output with lower probability than the status-quo technology, but produces output with a high enough probability to be cost-effective in the "good state" where $\theta = H$.[13] In this setting, letting $\mathbb{P}_{a\theta}$ be the probability of producing output with action a in state θ, we have $\mathbb{P}_{HL} < \mathbb{P}_{HH} < \mathbb{P}_{LL} = \mathbb{P}_{LH}$, so the model fits the oppositional confounding case, with the difference that the state only affects the outcome probability for one action.[14]

We illustrate the match and oppositional confounding cases with a pair of examples.

Example 1. *Consider a symmetric instance of the match confounding case, where an agent obtains "success" ($y = 1$) with probability $3/4$ if she takes the correct action and with probability $1/4$ if she takes the incorrect action: $\mathbb{P}_{LL} = \mathbb{P}_{HH} = 3/4$, $\mathbb{P}_{LH} = \mathbb{P}_{HL} = 1/4$. Given a parameter $\alpha \in (0, 2]$ indexing private signal informativeness, assume that the private belief densities are given by:*

$$f^H(p) = \begin{cases} \frac{4p}{\alpha} & \text{if } p \in \left[\frac{1}{2} - \frac{\alpha}{4}, \frac{1}{2} + \frac{\alpha}{4}\right], \\ 0 & \text{otherwise,} \end{cases}$$

$$f^L(p) = \begin{cases} \frac{4(1-p)}{\alpha} & \text{if } p \in \left[\frac{1}{2} - \frac{\alpha}{4}, \frac{1}{2} + \frac{\alpha}{4}\right], \\ 0 & \text{otherwise.} \end{cases}$$

For p in the support, the corresponding cdfs are

$$F^H(p) = \frac{16p^2 - (2-\alpha)^2}{8\alpha}, \qquad F^L(p) = 1 - F^H(1-p),$$

[13] A running example of such a technology in Wolitzky (2018) is the introduction of "blue spoons" to discourage the inefficient over-use of fertilizer by farmers in Kenya, which was studied by Chandrasekhar et al. (2022).

[14] If Assumption 1.2 is relaxed to allow $\mathbb{P}_{LL} = \mathbb{P}_{LH}$, this introduces another possible type of learning failure, namely an infinite cascade on action L. We exclude this type of learning failure to focus on confounding between actions and state-dependent outcomes.

where the relationship between F^L and F^H follows from the fact that $f^L(p) = f^H(1-p)$ in this example. Thus, private signals are uninformative if $\alpha \approx 0$, but they become more informative as α increases. However, regardless of α, the neutral public likelihood ratio $l^* = 1$ is always a confounding point: since the signal distribution is symmetric, the probability of the correct action is the same in each state, and hence so is the probability of success.

While $l^* = 1$ is a confounding point for any α, in the next section we will see that it is stable if and only if α is above a certain cutoff, $\alpha^* \approx 0.7964$.

Example 2. *Consider an instance of the oppositional confounding case with $P_{HL} = 1/10$, $P_{HH} = 1/5$, $P_{LL} = 4/5$, $P_{LH} = 9/10$, and the same private belief densities as in Example 1 (again parameterized by $\alpha \in (0, 2]$).*

Given a public belief q, denote the probability that the current agent takes action H in each state by $\eta(q) = \Pr(a = H|\theta = H)$ and $\lambda(q) = \Pr(a = H|\theta = L)$. Suppressing q, we then have

$$\Pr(y = 1|\theta = H) = (0.2)\eta + (0.9)(1 - \eta), \qquad \Pr(y = 1|\theta = L) = (0.1)\lambda + (0.8)(1 - \lambda).$$

Confounding occurs at q if these two probabilities are equal, or equivalently if

$$\eta - \lambda = \frac{1}{7}.$$

Note that, if private signals are completely uninformative, then $\eta - \lambda = 0$ (because $\Pr(a = H)$ is the same in both states), and if private signals are perfectly informative, then $\eta - \lambda = 1$ (because $\eta = 1$ and $\lambda = 0$). Since, for any public belief q, $\eta(q)$ and $\lambda(q)$ are continuous in the private signal distribution, it is natural to expect that a confounding point exists if and only if private signals are sufficiently informative.

Indeed, since an agent takes action H if and only if her private belief satisfies $p > 1 - q$, we have $\eta(q) = 1 - F^H(1 - q)$ and $\lambda(q) = 1 - F^L(1 - q) = F^H(q)$, so

$$\eta(q) - \lambda(q) = 1 - F^H(q) - F^H(1 - q) = \frac{\alpha}{4} - \frac{1}{\alpha}(2q - 1)^2.$$

Thus, confounding occurs at q if and only if $(2q - 1)^2 = \alpha^2/4 - \alpha/7$. The solutions are

$$q = \frac{1}{2} \pm \frac{1}{2}\varrho, \quad \text{where} \quad \varrho = \sqrt{\frac{\alpha(7\alpha - 4)}{28}}, \tag{3}$$

if $\alpha \geqslant 4/7$, and there is no solution if $\alpha < 4/7$. When they exist, the corresponding confounding points are $l^* = (1 \mp \varrho)/(1 \pm \varrho)$.

Moreover, in the next section, we will see that both confounding points are stable whenever $\alpha > 4/7$.

The next theorem characterizes when UI holds.

Theorem 2. *UI fails if and only if one of the following holds:*

1. *Match confounding holds. In this case, there is a unique confounding point* l^**, which satisfies* $\Delta_l(1, l^*) < 0$.

2. *Oppositional confounding holds, and, for* $p_m := (\mathbb{P}_{LL} - \mathbb{P}_{HL})/(\mathbb{P}_{LL} - \mathbb{P}_{HL} + \mathbb{P}_{LH} - \mathbb{P}_{HH})$,

$$F^H(p_m)\mathbb{P}_{LH} + (1 - F^H(p_m))\mathbb{P}_{HH} - F^L(p_m)\mathbb{P}_{LL} - (1 - F^L(p_m))\mathbb{P}_{HL} \leqslant 0. \tag{4}$$

   In this case, if (4) *holds with equality then there is a unique confounding point* l^**, which satisfies* $\Delta_l(1, l^*) = 0$*; and if* (4) *holds strictly then there are exactly two confounding points* $l_1^* < l_2^*$*, which satisfy* $\Delta_l(1, l_1^*) < 0 < \Delta_l(1, l_2^*)$.

In particular, if there is neither match confounding nor oppositional confounding, then UI holds and learning succeeds for all priors.

Thus, under match confounding, a confounding point always exists, regardless of the private signal distributions. This follows because if $l = \infty$ then $\Delta(1, l) = \mathbb{P}_{LH} - \mathbb{P}_{LL} < 0$ (if the state is believed to be L, then outcome 1 is more likely when the true state is L), while if $l = 0$ then $\Delta(1, l) = \mathbb{P}_{HH} - \mathbb{P}_{HL} > 0$ (if the state is believed to be H, then outcome 1 is more likely when the true state is H), so by continuity there is some $l \in \mathbb{R}_+$ such that $\Delta(1, l) = 0$.

In contrast, under oppositional confounding, a confounding point exists iff (4) holds. This condition says that, at public belief $1 - p_m$ (where the corresponding threshold private signal is p_m), outcome 1 is more likely in state L. To see why this is the relevant condition, note that if the public belief is close to 0 or 1—so the current agent takes either $a = 0$ or $a = 1$ with high probability—then outcome 1 is more likely in state H, because $\mathbb{P}_{a\theta}$ is increasing in θ for each a. Thus, if there exists some interior public belief where outcome 1 is instead more likely in state L, then by continuity there must also exist a public belief where outcome 1 is equally likely in each state—that is, a confounding point. The proof of Theorem 2 shows that $1 - p_m$ is the public belief that maximizes the difference in the probability of outcome 1 between states L and H. Thus, a confounding point exists if and only if, at this belief, outcome 1 is more likely in state L.

Intuitively, (4) says that private signals are sufficiently informative, in a certain sense. When $\mathbb{P}_{HL} < \mathbb{P}_{HH} < \mathbb{P}_{LL} < \mathbb{P}_{LH}$, outcome 1 can be more likely in state L only if the

current agent takes action L with considerably higher probability in state L than in state H, conditional on the public belief. This holds if and only if private signals are sufficiently informative. This point can be formalized as follows, where we say that private signal distributions $(\tilde{F}^L, \tilde{F}^H)$ are *more precise evaluated at* p_m than (F^L, F^H) if $\tilde{F}^L(p_m) > F^L(p_m)$ and $\tilde{F}^H(p_m) < F^H(p_m)$; we say that sequences of private signal distributions (F_n^L, F_n^H) *converge to perfect information* if $F_n^L \to F_{\delta_0}$ and $F_n^H \to F_{\delta_1}$; and we say that (F_n^L, F_n^H) *converge to no information* if $F_n^L, F_n^H \to F_{\delta_{1/2}}$:[15]

Corollary 1. *Suppose that oppositional confounding holds. If UI holds for some signal distributions (F^L, F^H), it also holds for any signal distributions $(\tilde{F}^L, \tilde{F}^H)$ that are less precise evaluated at p_m. Conversely, if UI fails for some signal distributions (F^L, F^H), it also fails for any signal distributions $(\tilde{F}^L, \tilde{F}^H)$ that are more precise evaluated at p_m. Moreover, for any sequence (F_n^L, F_n^H) that converges to no information, for all sufficiently large n, (4) fails, so UI holds; and, for any sequence (F_n^L, F_n^H) that converges to perfect information, for all sufficiently large n, (4) holds strictly, so UI fails.*

To see the logic of the oppositional confounding case more concretely, recall the vaccine example from the introduction. If the public belief is either very confident that a disease is not infectious or that it is infectious, then everyone knows that parents who are currently deciding whether to vaccinate their children are either not vaccinating (in the first case) or vaccinating (in the second). Either way, the observed infection rate will be higher if the disease is actually infectious. In contrast, if the public belief is uncertain and parents obtain informative private signals, then parents will vaccinate much more when the disease is infectious, leading to a lower infection rate. Thus, by continuity, when private signals are informative there is some public belief that leads to the same infection rate whether or not the disease is infectious; whereas, when private signals are uninformative, the infection rate is always higher when the disease is infectious, for any public belief.

Finally, if there is neither match confounding nor oppositional confounding, then UI holds, and hence learning succeeds. This is easy to see in the case where the $\mathbb{P}_{a\theta}$'s are all distinct. In this case, if there is neither match confounding nor oppositional confounding (and $\mathbb{P}_{HL} < \mathbb{P}_{HH}$, by Assumption 1), then either $\max\{\mathbb{P}_{LL}, \mathbb{P}_{HL}\} < \min\{\mathbb{P}_{LH}, \mathbb{P}_{HH}\}$ or $\mathbb{P}_{LL} < \mathbb{P}_{LH} < \mathbb{P}_{HL} < \mathbb{P}_{HH}$. In the first case, outcome 1 is more likely in state H for any strategy profile. In the second case, outcome 1 is more likely in state H in any equilibrium, because in any equilibrium action H is played with higher probability in state H than in state L.

[15] Here, δ_p and F_{δ_p} denote the Dirac measure and the corresponding cdf on $p \in [0,1]$, and convergence of distributions is weak convergence.

4 Stability and Generic Learning

Whether the existence of confounding points implies that learning robustly fails depends on whether these points are stable. We thus investigate the stability of confounding points, which will yield our main conclusions regarding the prospects for successful social learning. We first derive general conditions for confounding points to be stable; then draw implications for the limits where private signal precision is either very high or very low; and, finally, illustrate these results in the case where private signals are drawn from a location-scale family.

4.1 General Conditions for Stability

We extend a result of Ellison and Fudenberg (1995) and Smith and Sørensen (2000) to obtain a general stability result for discrete Markov processes, and then apply it to the likelihood ratio process (l_t).

Lemma 3. *Let (x_t) be a Markov process over $\mathbb{R}_+$, with*

$$x_{t+1} = \begin{cases} \phi(0, x_t) & \text{with probability } p(x_t), \\ \phi(1, x_t) & \text{with probability } 1 - p(x_t). \end{cases}$$

Let $x^* \in \mathbb{R}_+$ satisfy $p(x^*) \in (0,1)$ and $\phi(0,x^*) = \phi(1,x^*) = x^*$, and suppose that there is a neighborhood of x^* on which $p(x)$ is continuous and $\phi(0,x)$ and $\phi(1,x)$ are C^1, with derivatives $\phi_x(0,x), \phi_x(1,x)$. Let $D \subset \mathbb{R}_+$ be the set of initial points x_0 from which there is positive probability of reaching x^* in finite time.

(a) If

$$|\phi_x(0,x^*)|^{p(x^*)} \cdot |\phi_x(1,x^*)|^{1-p(x^*)} > 1,$$

then, starting from any $x_0 \notin D$, $\mathbb{P}(x_t \to x^*) = 0$.

(b) If

$$|\phi_x(0,x^*)|^{p(x^*)} \cdot |\phi_x(1,x^*)|^{1-p(x^*)} < 1,$$

then there exists $\delta > 0$ such that, starting from any $x_0 \in (x^* - \delta, x^* + \delta)$, $\mathbb{P}(x_t \to x^*) > 0$.

Lemma 3 extends Lemma 1 of Ellison and Fudenberg (1995) to state-dependent transition probabilities and a more general process (x_t). Their lemma is the special case of Lemma 3 where $p(x)$ is constant in x, $x^* = 0$, and (x_t) is a process over $(0,1)$. As in their lemma, the logic of Lemma 3 is that (x_t) is approximately log-linear around x^* (as $\log(|x_t - x^*|)$ is

approximately a random walk with increments $\log(|\phi_x(0,x^*)|)$ and $\log(|\phi_x(1,x^*)|)$, so x^* is unstable if

$$p(x^*)\log\left(|\phi_x(0,x^*)|\right)+(1-p(x^*))\log\left(|\phi_x(1,x^*)|\right)>0,$$

and is stable if the opposite inequality holds. The extension of their lemma in part (b)—which shows that the process has positive probability of converging to a stable steady state—was already established by Smith and Sørensen (2000, Theorem C.1). We build on arguments from these papers to extend part (a), showing that the process has zero probability of converging to an unstable steady state. A caveat (which was ruled out by the assumptions of Ellison and Fudenberg, 1995) is that this result only applies to initial points from which the process cannot reach the unstable steady state in finite time.

We will focus on the implications of the instability condition in Lemma 3.a and the stability condition in Lemma 3.b for the public-belief process ϕ defined in (2). We wish to exclude the non-generic possibility that the confounding point is reached in finite time: i.e., that l_0 lies in the set D in Lemma 3. The following two lemmas verify that this possibility is indeed non-generic: for generic $(\mathbb{P}_{a\theta})\in[0,1]^4$ and a generic prior $l_0\in\mathbb{R}_+$, there is zero probability of reaching a confounding point in finite time.

Lemma 4. *There is a generic set $\mathcal{P}\subset[0,1]^4$ such that, for any $(\mathbb{P}_{a\theta})\in\mathcal{P}$, there is no nonempty open interval $I\subset\mathbb{R}_+$ over which $\phi(0,l)$ or $\phi(1,l)$ is constant in l.*

Intuitively, $\phi(y,l)$ is locally constant in l only if perturbing l shifts the cutoff private signal $\bar{p}(l)$ exactly so that the public posterior belief $\phi(y,l)$ remains constant. Lemma 4 shows that this cannot happen for generic outcome probabilities.

Henceforth, we maintain the conclusion of Lemma 4 as an assumption. In addition to holding for generic outcome probabilities $(\mathbb{P}_{a\theta})$ (by Lemma 4), this assumption also holds for arbitrary outcome probabilities and generic private signal distributions.[16]

Assumption 2. There is no nonempty open interval $I\subset\mathbb{R}_+$ over which $\phi(0,l)$ or $\phi(1,l)$ is constant.

Lemma 5. *Under Assumption 2, for any point l^*, the set of priors l_0 from which there is a positive probability of reaching l^* in finite time is meagre.*

We now translate the stability conditions in Lemma 3 into the public belief process defined in (2). For this, we first introduce two objects: the outcome likelihood ratio R and its elasticity ε.

[16]However, formalizing this last claim requires topologizing the space of signal distributions, which we omit.

For each outcome $y \in \{0,1\}$ and public likelihood ratio $l > 0$, define

$$R(y,l) := \frac{\psi(y|L,l)}{\psi(y|H,l)},$$

as the likelihood ratio for state L versus H derived from outcome y at belief l. This function is well defined and strictly positive by Assumption 1.1. The public belief dynamics in (2) are then given by

$$\phi(y,l) = l \cdot R(y,l), \tag{5}$$

and a confounding point l^* is characterized by $R(1,l^*) = 1$ (equivalently, $R(0,l^*) = 1$). The *elasticity of* R at (y,l) is

$$\varepsilon(y,l) := \frac{l\, R_l(y,l)}{R(y,l)} = \frac{d \log R(y,l)}{d \log l}.$$

From (5), $\phi_l(y,l) = R(y,l) + l \cdot R_l(y,l) = R(y,l)\,(1 + \varepsilon(y,l))$.

Let l^* be a confounding point—so there exists ψ^* such that $\psi(1|L,l^*) = \psi(1|H,l^*) = \psi^*$. Then $R(y,l^*) = 1$, so $\phi_l(y,l^*) = 1 + \varepsilon(y,l^*)$. Applying Lemma 3 to the process (l_t), the confounding point l^* is unstable if

$$|1 + \varepsilon(1,l^*)|^{\psi^*} \cdot |1 + \varepsilon(0,l^*)|^{1-\psi^*} > 1, \tag{Instability}$$

and it is stable if

$$|1 + \varepsilon(1,l^*)|^{\psi^*} \cdot |1 + \varepsilon(0,l^*)|^{1-\psi^*} < 1. \tag{Stability}$$

The following theorem is now a consequence of Lemmas 3 and 5.

Theorem 3.

1. *Learning generically succeeds if* (Instability) *holds at all confounding points.*

2. *Learning robustly fails if there exists a confounding point where* (Stability) *holds.*

Theorem 3 characterizes when learning generically succeeds (outside the knife-edge case where a confounding point lies on the boundary of the (in)stability condition). Its conditions are easy to check: there are at most two confounding points l^* (by Theorem 2), which are characterized by the equation $\mathbf{\Delta}(l^*) = 0$, and the (in)stability condition is a simple inequality that depends on the belief elasticities at a confounding point, $\varepsilon(y,l^*)$. In particular, a confounding point is stable if and only if these elasticities are small in absolute value, as the

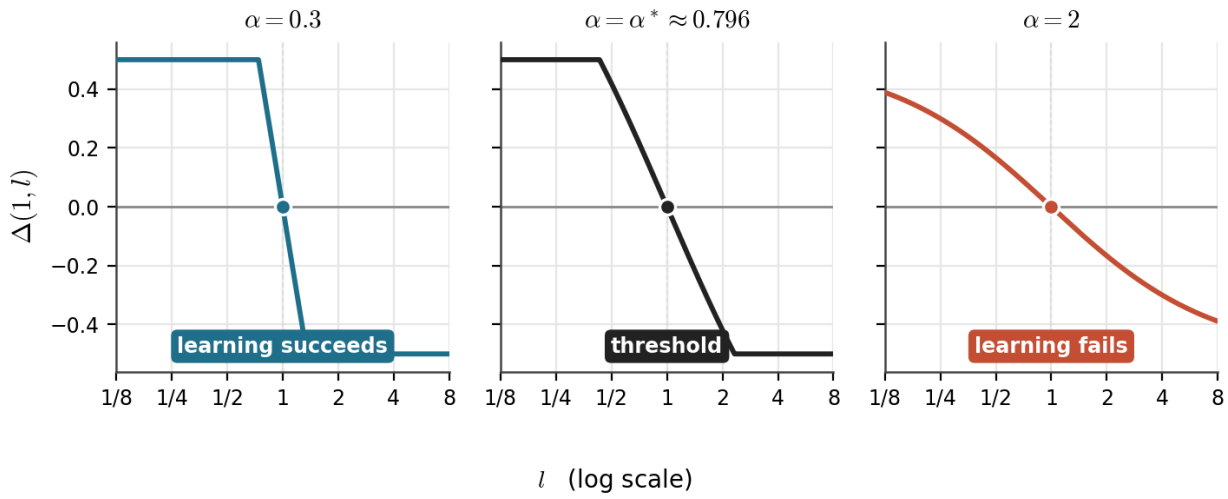

Figure 1: $\Delta(1, l)$ in Example 1 for three values of α, with l on a log scale. The confounding point is $l^* = 1$ for any α, and it is stable—so learning robustly fails—if $\alpha > \alpha^* \approx 0.7964$.

next result shows.[17]

Proposition 1. *Let l^* be a confounding point with $\varepsilon(1, l^*) \neq 0$. Define $\underline{y} := \arg\min_y \varepsilon(y, l^*) \in \{0, 1\}$ and $\underline{\psi} := \psi(\underline{y}|H, l^*) \in (0, 1)$. There exists a unique threshold $c(\underline{\psi}) \in (-2, -1.278)$, characterized by*

$$(-1 - c)^{\underline{\psi}} \left(1 - \tfrac{\underline{\psi}}{1-\underline{\psi}} c\right)^{1-\underline{\psi}} = 1, \tag{6}$$

such that (Stability) holds iff $\varepsilon(\underline{y}, l^*) > c(\underline{\psi})$ and (Instability) holds iff $\varepsilon(\underline{y}, l^*) < c(\underline{\psi})$.

Intuitively, a confounding point l^* is stable if the information content of outcomes (which is zero at l^*, by definition) is insensitive to shifts in the public belief around l^*. In this case, at a public belief near l^*, outcomes provide little information, so the public belief stays close to l^*. If instead the information content of outcomes is elastic, then perturbing the public belief away from l^* quickly makes outcomes more informative, precluding convergence to l^*.

We illustrate Theorem 3 in the context of Examples 1 and 2. In these examples, we use the formula

$$\varepsilon(1, l^*) = -\frac{l^* \delta^*}{\psi^*}, \qquad \varepsilon(0, l^*) = \frac{l^* \delta^*}{1 - \psi^*}, \tag{7}$$

where $\delta^* := \Delta_l(1, l^*) \neq 0$.[18]

[17] This result applies more generally to any model of learning from a binary outcome where the public likelihood ratio follows a Markov martingale and outcome probabilities have full support. The constant -1.278 approximates $-1 - W(1/e)$, where W is the principal branch of the Lambert W-function.

[18] To derive this, apply the quotient rule to $R(1, l) = \psi(1|L, l)/\psi(1|H, l)$ at l^* (where $\psi(1|L, l^*) = \psi(1|H, l^*) = \psi^*$) to get $R_l(1, l^*) = (\psi_l(1|L, l^*) - \psi_l(1|H, l^*))/\psi^* = -\Delta_l(1, l^*)/\psi^*$. Similarly $R_l(0, l^*) = \Delta_l(1, l^*)/(1 - \psi^*)$. Since $R(y, l^*) = 1$, the formulas follow from the definition of ε.

Example 3 (Example 1 cont.). *In Example 1, $l^* = 1$ is always a confounding point. The corresponding public belief is $q = 1/2$, and it is straightforward to calculate that $\eta = 1 - \lambda = 1/2 + \alpha/8$, $\psi^* = 1/2 + \alpha/16$, and $\delta^* = -1/(2\alpha)$. Since $\delta^* < 0$, we have $\varepsilon(0, l^*) < 0 < \varepsilon(1, l^*)$, so $\underline{y} = 0$, $\underline{\psi} = 1 - \psi^* = 1/2 - \alpha/16$, and $\varepsilon(\underline{y}, l^*) = l^*\delta^*/(1 - \psi^*) = -8/(\alpha(8 - \alpha))$. By Proposition 1, the point is stable iff $\varepsilon(\underline{y}, l^*) > c(\underline{\psi})$, which holds iff $\alpha > \alpha^* \approx 0.7964$. Thus the confounding point is stable (so learning robustly fails) if the range of private beliefs is larger than (approximately) $[0.3, 0.7]$, and it is unstable (so learning generically succeeds) if the range is smaller than this. Figure 1 illustrates. On the log axis, the slope of $\Delta(1, \cdot)$ at the confounding point is $l^*\delta^*$, the numerator of $\varepsilon(\underline{y}, l^*)$.*[19] *As α grows, the crossing flattens, which raises $\varepsilon(\underline{y}, l^*)$ until it clears the cutoff $c(\underline{\psi})$ at α^*.*

Example 4 (Example 2 cont.). *In Example 2, the confounding points $l^* = (1 \pm \varrho)/(1 \mp \varrho)$ exist if and only if $\alpha \geqslant 4/7$ (with ϱ defined in (3)). Consider the point $l^* = (1 - \varrho)/(1 + \varrho)$. It is straightforward to calculate that $\eta = 4/7 + \varrho/\alpha$, $\lambda = 3/7 + \varrho/\alpha$, $\psi^* = 1/2 - 7\varrho/10\alpha$, and $l^*\delta^* = -7\varrho(1 - \varrho^2)/10\alpha$. Since $l^*\delta^* < 0$, we have $\varepsilon(0, l^*) < 0 < \varepsilon(1, l^*)$, so $\underline{y} = 0$ and $\varepsilon(\underline{y}, l^*) = l^*\delta^*/(1 - \psi^*)$. As $-l^*\delta^* < 1 - \psi^*$ for any $\alpha > 4/7$, this gives $\varepsilon(\underline{y}, l^*) > -1 > c(\underline{\psi})$, so (Stability) holds by Proposition 1. The other point, $l^* = (1 + \varrho)/(1 - \varrho)$, is obtained by replacing ϱ with $-\varrho$, which swaps ψ^* with $1 - \psi^*$ and $\varepsilon(1, l^*)$ with $\varepsilon(0, l^*)$, leaving the left-hand side of (Stability) unchanged. Hence both confounding points are stable. In total, learning robustly fails if the range of private beliefs is larger than $[5/14, 9/14]$, and learning succeeds for any prior if the range is smaller than this. Figure 2 illustrates.*

4.2 The (Un)Informative Private Signal Limit

Our last set of results show that the (in)stability condition in Theorem 3 can be pinned down in the limits where private signals are either very informative or very noisy. This subsection considers general signal distributions that satisfy some regularity conditions; the next subsection shows that these conditions hold when signals come from a location-scale family with a strictly log-concave density and unbounded signal precision, such as a Gaussian distribution.

We rely on the following regularity conditions, which apply to sequences of signal distributions (F_n^L, F_n^H) that converge to either perfect information ($F_n^L \to F_{\delta_0}, F_n^H \to F_{\delta_1}$) or no information ($F_n^L, F_n^H \to F_{\delta_{1/2}}$).

[19] "Slope on the log axis" means the derivative with respect to $\ln l$, which at a confounding point equals $l^*\delta^*$. The figures use a base-2 log scale for readability. This multiplies every slope by the constant $\ln 2$ and so preserves all comparisons of steepness.

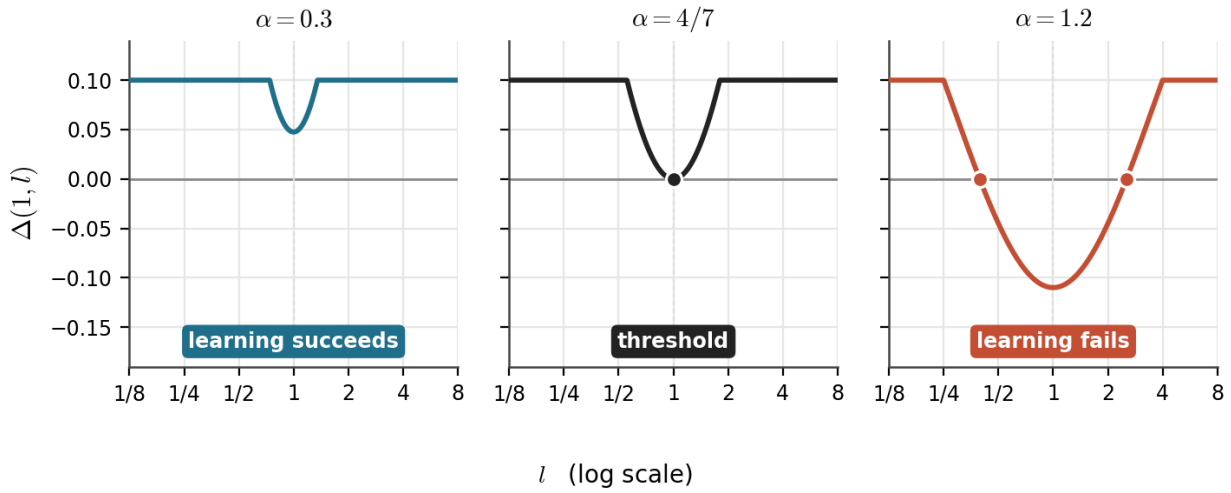

Figure 2: $\Delta(1, l)$ in Example 2 for three values of α, with l on a log scale. The confounding points are the zeros of $\Delta(1, \cdot)$; they exist if $\alpha \geqslant 4/7$ and are stable if $\alpha > 4/7$.

Definition 3. Signal distributions (F_n^L, F_n^H) that converge to perfect information are *tail-regular* if, for any sequence (p_n), the following hold:

1. If $\lim_{n\to\infty} F_n^L(p_n) \in (0, 1)$, then $\lim_{n\to\infty} f_n^H(p_n) = 0$.
2. If $\lim_{n\to\infty} F_n^H(p_n) \in (0, 1)$, then $\lim_{n\to\infty} f_n^L(p_n) = 0$.

Definition 4. Signal distributions (F_n^L, F_n^H) that converge to no information are *center-regular* if, for any sequence (p_n) satisfying $\lim_{n\to\infty} F_n^L(p_n) \in (0, 1)$, we have $\lim_{n\to\infty} f_n^L(p_n) = \infty$.[20]

Tail-regularity and center-regularity appear to be very mild conditions. Tail regularity is a weak kind of uniformity, which says that, as $(F_n^L, F_n^H) \to (F_{\delta_0}, F_{\delta_1})$, it is impossible to find a sequence $(p_n) \to 0$ where $F_n^L(p_n)$ remains bounded between 0 and 1, but $f_n^H(p_n) \not\to 0$ (or an analogous sequence $(p_n) \to 1$). Center regularity says that, as $F_n^L \to F_{\delta_{1/2}}$, it is impossible to find a sequence $(p_n) \to 1/2$ where $F_n^L(p_n)$ remains bounded between 0 and 1, but $f_n^H(p_n) \not\to \infty$. As we will see, these conditions hold whenever signals are drawn from a location-scale family with a strictly log-concave density and unbounded signal precision.

The following result also excludes the knife-edge case where $\mathbb{P}_{LL} = \mathbb{P}_{HH}$. As we will see, the same conclusion applies in this case, under a stronger regularity condition (which also holds when signals are drawn from a location-scale family with a strictly log-concave density and unbounded signal precision).

[20] By the no-introspection property, an equivalent condition is $\lim_{n\to\infty} F_n^H(p_n) \in (0, 1) \implies \lim_{n\to\infty} f_n^H(p_n) = \infty$.

Theorem 4. *Assume that there is either match confounding or oppositional confounding (otherwise, learning succeeds by Theorem 2), and that $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$. Then:*

1. *If the signal distributions (F_n^L, F_n^H) converge to perfect information and are tail-regular, then there exists $N > 0$ such that learning robustly fails for all $n > N$.*

2. *If the signal distributions (F_n^L, F_n^H) converge to no information and are center-regular, then there exists $N > 0$ such that learning generically succeeds for all $n > N$.*

The key observation behind Theorem 4 is that if private signals are very precise then the information content elasticities $\varepsilon(y, l^*)$ are small, because perturbing the public belief away from a confounding point l^* has little effect on the action distribution, and hence has little effect on the outcomes' information content. By Proposition 1, this implies that the confounding point is stable. Conversely, if private signals are very noisy then the information content elasticities are large, so any confounding point is unstable.

Theorem 4 follows from combining this stability analysis with the characterization of confounding points in Theorem 2. By Theorem 2, there is always a unique confounding point under match confounding; and under oppositional confounding there are no confounding points if private signals are very noisy, and there are two confounding points if private signals are very precise. Since confounding points are stable with precise private signals and unstable with noisy ones, it follows that, for either form of confounding, learning robustly fails with precise signals (because a stable confounding point exists), and learning generically succeeds with noisy signals (because the confounding point under match confounding is unstable, and there is no confounding point under oppositional confounding).

Remark 1. The assumption $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$ in Theorem 4 is needed only for Part 1 and only under match confounding (under oppositional confounding, $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$ holds by definition). In Appendix B we show that, under a stronger regularity condition on the private signal densities (*strong tail regularity*), Part 1 holds even when $\mathbb{P}_{LL} = \mathbb{P}_{HH}$. Moreover, strong tail regularity is also satisfied by location-scale families with unbounded scores.[21]

The reason for this complication is that the $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$ and $\mathbb{P}_{LL} = \mathbb{P}_{HH}$ cases behave differently with precise private signals. When $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$, the sequence of confounding points $l^*(n)$ converges to either 0 or ∞ as signals become perfect. In the limit, in one state agents always take the correct action, and in the other state both actions are taken with positive probability. For example, if $l^*(n) \to 0$, public information at the confounding point increasingly points to state H, and in the limit agents take the correct action in state H with

[21] In particular, Proposition 2 in the next subsection holds as stated, without requiring that $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$.

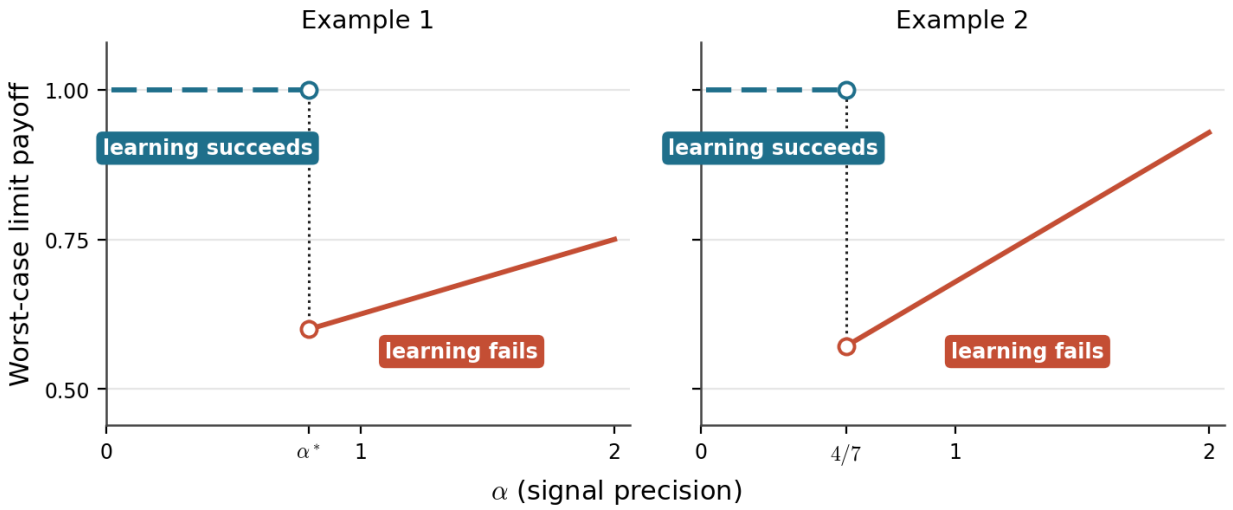

Figure 3: The (generic) worst-case limit payoff as a function of α in Examples 1 and 2. When learning fails, this payoff equals $1/2 + \alpha/8$ in Example 1 and $3/7 + \alpha/4$ in Example 2.

probability 1, while both actions are taken with positive probability in state L.[22] Instead, when $\mathbb{P}_{LL} = \mathbb{P}_{HH}$ under match confounding, the confounding belief threshold $\bar{p}(l^*(n))$ can lie above the region where private beliefs concentrate in state L and below the region where they concentrate in state H.[23] In the limit, agents effectively disregard the public information and take the correct action with probability converging to 1 in both states—which is why $\mathbb{P}_{LL} = \mathbb{P}_{HH}$ is necessary for confounding. Strong tail regularity is the required regularity condition in this regime.

Remark 2. Since we define learning success as expected payoffs converging to the first best, Theorem 4 implies that limit payoffs generically attain the first best when private signals are sufficiently noisy, but robustly fall short of it when they are sufficiently precise. However, precision also has a countervailing positive effect, since more precise private signals yield higher expected payoffs at any fixed public belief. Long-run payoffs are therefore non-monotone in signal precision under the premise of Theorem 4, as we now illustrate in Examples 1 and 2 (where a single precision cutoff separates learning success from failure).

Example 5 (Long-run payoffs in Examples 1 and 2)**.** *Given a prior l_0, consider the limit payoff $\lim_{t\to\infty} \mathbb{E}[u(a_{t+1}, \theta)]$, which averages the conditional expected payoffs $\mathbb{E}[u(a_{t+1}, \theta)|l_t, \theta]$ over the state and the realization of l_t. Define the (generic) "worst-case limit payoff" as its infimum over all priors $l_0 \in (0, \infty)$ from which no confounding point is reached in finite time*

[22] This is possible because the rate at which the confounding public belief moves toward H exactly offsets the increasingly accurate private information that the state is L.

[23] This can occur if either $l^*(n) \to l^* \in (0, \infty)$ or $l^*(n)$ converges to 0 or ∞ sufficiently slowly.

(by Lemma 5, this excludes only a meagre set of priors). When learning succeeds from a prior l_0, the limit payoff equals 1. When instead the public belief converges to a confounding point l^*, with corresponding public belief $q^* = 1/(1+l^*)$, the conditional expected payoffs converge to the probability of the correct action at l^*, which is $\eta(q^*)$ in state H and $1-\lambda(q^*)$ in state L. As $l_0 \to l^*$, the prior probability of state H approaches q^*, so when learning robustly fails the worst-case limit payoff equals $q^*\eta(q^*) + (1-q^*)(1-\lambda(q^*))$.[24]

In Example 1, the confounding point $l^* = 1$ has $q^* = 1/2$ and $\eta = 1-\lambda = 1/2+\alpha/8$, so the expected payoff at the confounding point is $1/2+\alpha/8$ in each state. The worst-case limit payoff is therefore 1 for $\alpha < \alpha^*$, where learning generically succeeds, and $1/2+\alpha/8$ for $\alpha > \alpha^*$, where learning robustly fails. In Example 2, there are no confounding points for $\alpha < 4/7$ and learning succeeds for every prior, so the worst-case limit payoff is 1. For $\alpha > 4/7$, the confounding point $l^* = (1-\varrho)/(1+\varrho)$ has $q^* = (1+\varrho)/2$, $\eta = 4/7+\varrho/\alpha$, and $\lambda = 3/7+\varrho/\alpha$, so the expected payoff at this confounding point is

$$q^*\eta + (1-q^*)(1-\lambda) = \frac{1+\varrho}{2}\left(\frac{4}{7}+\frac{\varrho}{\alpha}\right) + \frac{1-\varrho}{2}\left(\frac{4}{7}-\frac{\varrho}{\alpha}\right) = \frac{4}{7}+\frac{\varrho^2}{\alpha} = \frac{3}{7}+\frac{\alpha}{4},$$

where the last equality uses (3). Replacing ϱ with $-\varrho$ shows that the other confounding point yields the same payoff, so the worst-case limit payoff is $3/7+\alpha/4$ for $\alpha > 4/7$.

Figure 3 plots the worst-case limit payoff in these examples. In both examples, payoffs are U-shaped in α. In addition, since private signals in these examples do not converge to perfect information, payoffs do not return to the first best at $\alpha = 2$, where the worst-case payoff is 3/4 in Example 1 and 13/14 in Example 2.

4.3 Location-Scale Signal Distributions

Especially clean results are available when private signals are drawn from a location-scale family with a strictly log-concave density. In this case, center-regularity holds, and tail-regularity also holds if in addition private beliefs are unbounded.

Specifically, in this subsection we assume that in state $\theta \in \{L, H\}$, private signals are given by $s = \mathbf{1}\{\theta = H\} + \sigma_n\varepsilon$, where $\sigma_n > 0$ and ε is distributed according to a cdf G with a continuously differentiable, strictly positive, and strictly log-concave density g on $\mathbb{R}$. For each $\sigma > 0$, let $p_\sigma(s) := g((s-1)/\sigma)/[g((s-1)/\sigma) + g(s/\sigma)]$ denote the posterior belief that the state is H given signal s. Strict positivity and log-concavity of g ensure that $p_\sigma(s)$ is strictly increasing in s, and is therefore invertible (see Footnote 33 in Appendix A). Letting

[24] This relies on the fact that when l^* is stable, $\mathbb{P}(l_t \to l^*)$ tends to 1 as $l_0 \to l^*$. This follows from a similar argument as Lemma 3.

$s_\sigma(p)$ denote the inverse function, the signal distributions are given by

$$F_n^L(p) = G\left(\frac{s_{\sigma_n}(p)}{\sigma_n}\right), \quad F_n^H(p) = G\left(\frac{s_{\sigma_n}(p)-1}{\sigma_n}\right). \tag{8}$$

Let $r := g'/g$ denote the corresponding score function. The score function is decreasing by log-concavity, and is *unbounded* if $\lim_{x\to-\infty} r(x) = \infty$ and $\lim_{x\to\infty} r(x) = -\infty$. The score function is unbounded if and only if private beliefs are unbounded.

Proposition 2. *If $\sigma_n \to \infty$, then (F_n^L, F_n^H) converge to no information and are center-regular, and hence learning generically succeeds by Theorem 4.2. If the score function is unbounded and $\sigma_n \to 0$, then (F_n^L, F_n^H) converge to perfect information and are tail-regular, and hence learning robustly fails under match or oppositional confounding by Theorem 4.1.*

An unbounded score is sufficient for tail-regularity and hence for Theorem 4.1, but neither an unbounded score nor tail-regularity is necessary for the conclusion of Theorem 4.1 to hold. For example, in Appendix C, we show that if G is logistic (the best-known strictly log-concave distribution with a bounded score), then (F_n^L, F_n^H) do not satisfy tail-regularity as $\sigma_n \to 0$, yet the conclusion of Theorem 4.1 still holds.

Proposition 2 contrasts starkly with Smith and Sørensen's (2000) result that learning succeeds if and only if private beliefs are unbounded: in our model, for fixed outcome probabilities $\mathbb{P}_{a\theta}$, learning can succeed with bounded private beliefs but fail with unbounded ones. The logic is that precise private signals break incorrect cascades in Smith and Sørensen (2000), but cause (or stabilize) confounding in our model.

Finally, Theorem 4 and Proposition 2 pin down the two extreme cases—learning robustly fails when private signals are very precise and generically succeeds when they are very noisy—but these results do not establish that the impact of private signal precision is monotone, in that a single threshold σ^* separates the two regimes. Numerical results suggest that this monotonicity holds in the strictly log-concave location-scale model with an unbounded score. We have not been able to prove this result at this level of generality, because the calculations involved in verifying that there is a unique transition from (Stability) to (Instability) as the noise level σ increases, accounting for the dependence of the confounding point l^* on σ, are too complicated. However, the following result establishes monotonicity in a simpler special case: Gaussian noise and symmetric match confounding, where l^* is fixed independent of σ (as in Example 1).[25]

[25] Our proof involves a numerical calculation of the standard normal cdf, so strictly speaking it is not fully rigorous.

Proposition 3. *Consider the location-scale model with Gaussian noise: $g = \varphi$, $G = \Phi$. Assume symmetric match confounding: $\mathbb{P}_{LH} = \mathbb{P}_{HL} < \mathbb{P}_{LL} = \mathbb{P}_{HH}$. Then there exists $\sigma^* > 0$ such that learning robustly fails if $\sigma < \sigma^*$ and generically succeeds if $\sigma > \sigma^*$.*

5 Discussion

We conclude by discussing possible extensions beyond binary states, actions, and outcomes, and by relating our model to variants with unordered or random observations.

5.1 Beyond the Binary Setting

We have focused on a canonical setting with binary states, actions, and outcomes. This simple setting accommodates a variety of distinct noise structures—no confounding, match confounding, and oppositional confounding—and yields sharp predictions about long-run learning. We briefly discuss the prospects for extending the analysis beyond this binary setting.

With binary states and outcomes but more than two actions, our analysis mostly generalizes: the public belief remains a single likelihood ratio, and the $\boldsymbol{\Delta}(l)$ function retains a similar structure. We expect the main conclusions of Theorems 2 and 4 to extend to this setting, possibly under additional regularity assumptions on the outcome probabilities and the private signal distributions (e.g., an appropriate strengthening of tail and center regularity).[26]

When outcomes can take $M > 2$ values, learning generically succeeds, because uniform identification generically holds: the system of equations characterizing confounding requires that all M outcomes arise with equal probability in both states, which yields an overdetermined system with $M-1$ independent equations (the outcome probabilities) and 1 unknown (the public belief), so for generic outcome probabilities it has no solution.[27] However, an example of a (non-generic) information structure with multiple outcomes where learning fails under similar conditions as in the binary case arises when a binary "underlying outcome" $z \in \{0, 1\}$ is a stochastic function of (a, θ), and the observed outcome y (which can take any number of values) depends on (a, θ) only through z. For instance, perhaps the infectiousness of a strain of measles and whether a child is vaccinated determine whether the child contracts measles, but a child with (or without) measles can display a range of symptoms.

[26] Similarly, Smith and Sørensen's (2000) analysis applies with multiple actions.

[27] Similarly, Smith and Sørensen (2000) observe that confounding outcomes where rational agents take more than two actions with positive probability are non-generic in their model.

With more than two states, in general, the analysis becomes substantially more involved. The public belief becomes a vector of likelihood ratios; private signals induce distributions on a higher-dimensional simplex; and confounding becomes a condition at a point in a higher-dimensional space. One tractable case arises when $\Theta = [0, 1]$, the outcome probability $\mathbb{P}_{a\theta}$ and the signal distribution F^θ are both linear in θ (identifying state L with 0 and state H with 1), and there are either no private signals (so $\eta = \lambda$) or the effect of the action and the state on the probability of outcome 1 are additively separable (so $\mathbb{P}_{LL} + \mathbb{P}_{HH} - \mathbb{P}_{LH} - \mathbb{P}_{HL} = 0$). In this case, the probability of outcome 1 is linear in θ for any public belief, so confounding occurs for all states if and only if it occurs for states L and H, and hence Theorem 2 continues to characterize when UI holds. More generally, we leave the multiple states case as an open area for investigation, noting that, even with observed actions, the multiple states case has been studied only recently by Arieli and Mueller-Frank (2021) and Kartik et al. (2024).

In general, our point that informative private signals impede outcome-based learning by making actions less predictable does not seem specific to the binary setting, but the model does become considerably more complicated outside this setting, and there are many open questions about precisely how our results do or do not generalize.

5.2 Comparison to Unordered Observations

An important feature of our model is that, absent private signals, agents could perfectly infer each of their predecessor's actions. This feature stems from the assumption that agents move in sequence and observe the complete history of outcomes. It is therefore natural to compare our model with variants where agents observe incomplete or unordered samples of earlier agents' outcomes (as studied by Acemoglu et al. (2011), Smith and Sørensen (2020), and Callander and Hörner (2009) with observed actions), as well as a version of this model where a continuum of agents act each period, so the population dynamic is deterministic (as studied by Banerjee and Fudenberg (2004) with observed actions, and by Wolitzky (2018) with observed outcomes).

A version of our model with sequential (one-at-a-time) actions and observations of incomplete or unordered samples of outcomes seems difficult to analyze. The Markov-martingale structure of Smith and Sørensen (2000) and the current paper breaks down, because the sequence of beliefs conditional on social information is not a martingale. At the same time, the "improvement principle" of Banerjee and Fudenberg (2004), Acemoglu et al. (2011), and Smith and Sørensen (2020) also break down, because actions are not observed. It is thus unclear how to analyze this model, and, in particular, how to assess the impact of private signal precision.

A model similar to ours but with a continuum of agents acting each period and observations of unordered samples of outcomes was studied by Wolitzky (2018). That paper interprets the model as one of technology adoption and shows that social learning fails with a "cost-saving innovation"—which is similar to oppositional confounding—whenever private signals are sufficiently *uninformative*.[28] Thus, informative private signals facilitate social learning in Wolitzky (2018), but impede it in the present model.

The reason for this difference is as follows. In Wolitzky (2018), agents essentially observe the share of earlier agents obtaining outcome 1. The obstacle to social learning is that a greater frequency of outcome 1 can indicate that the state is either H or L, depending on the frequency of each action in each state; and the equilibrium population dynamic cannot pass from the region of state-contingent action frequencies where a greater frequency of outcome 1 is "good news" to the region where it is "bad news," because the locus of action frequencies where the outcome frequency is uninformative forms a reflecting barrier for the population dynamic. In this model, informative private signals facilitate social learning by ensuring that the action frequencies in the absence of social information are already close to efficient, so the population dynamic does not need to cross the reflecting barrier to converge to the efficient point. Specifically, if $\mathbb{P}_{HL} \leqslant \mathbb{P}_{HH} < \mathbb{P}_{LL} \leqslant \mathbb{P}_{LH}$ (a generalization of both oppositional confounding and the "cost-saving innovation" case), then the arguments in Wolitzky (2018) imply that social learning succeeds if and only if

$$(1 - F^H(1 - q_0))\mathbb{P}_{HH} + F^H(1 - q_0)\mathbb{P}_{LH} < (1 - F^L(1 - q_0))\mathbb{P}_{HL} + F^L(1 - q_0)\mathbb{P}_{LL},$$

so that outcome 1 is more likely in state L at the prior public belief q_0. This condition holds if private signals are very informative (as the LHS converges to $\mathbb{P}_{HH}$ and the RHS converges to $\mathbb{P}_{LL}$), and it fails if private signals are very noisy and $q_0 \neq 1/2$ (as, if $q_0 < 1/2$, the LHS converges to $\mathbb{P}_{LH}$ and the RHS converges to $\mathbb{P}_{LL}$; and, if $q_0 > 1/2$, the LHS converges to $\mathbb{P}_{HH}$ and the RHS converges to $\mathbb{P}_{HL}$).[29]

The distinction between the two models is easiest to appreciate in the case where private signals are entirely absent. In the current model, learning generically succeeds for any $(\mathbb{P}_{a\theta})$ satisfying Assumption 1, because agents can perfectly infer each of their predecessor's actions, and the outcome identifies the state, conditional on either action. In Wolitzky (2018), learning fails if $\mathbb{P}_{HL} \leqslant \mathbb{P}_{HH} < \mathbb{P}_{LL} \leqslant \mathbb{P}_{LH}$, because, in the absence of social information, $\eta = \lambda$, and hence outcome 1 is more likely in state H; whereas successful learning requires

[28] As discussed in Section 3.2, the cost-saving innovation case corresponds to $\mathbb{P}_{HL} < \mathbb{P}_{HH} < \mathbb{P}_{LL} = \mathbb{P}_{LH}$.

[29] Wolitzky (2018) does not consider match confounding. In that model, match confounding would lead to a failure of social learning for some private signal distributions, but the set of such distributions would shrink to a non-generic set as agents' samples grow large, unlike in the oppositional confounding case.

$\eta = 1$ and $\lambda = 0$, implying that outcome 1 is more likely in state L; and the equilibrium population dynamic cannot converge from an (η, λ)-pair where outcome 1 is more likely in state H to one where it is more likely in state L. Sufficiently informative private signals overcome the obstacle to social learning in Wolitzky (2018), whereas they are the very source of the obstacle in the present paper.

The impact of private signal precision is thus very different in deterministic, continuum agent settings and in stochastic, sequential-move ones. This should not be too surprising, as the models are quite different: e.g., Smith and Sørensen (2020) describe the former class of models as capturing information transmission rather than social learning, because in principle "society" has enough information to take the correct action after a single period. In our view, these results call for care in matching models to applications, as well as further research on the role of private information and imperfect observation in different social learning settings.

Appendix

A Proofs

Proof of Lemma 1. We have

$$\mathbb{E}[l_{t+1}|l_t, H] = \sum_{y \text{ s.t. } \psi(y|l_t,H)>0} \psi(y|l_t, H) l_t \frac{\psi(y|l_t, L)}{\psi(y|l_t, H)} = l_t \sum_{y \text{ s.t. } \psi(y|l_t,H)>0} \psi(y|l_t, L) = l_t.$$

The last equality follows from mutual absolute continuity of the $\mathbb{P}_{a\theta}$ and of F^L, F^H, which implies that $\psi(y|l_t, L) > 0 \iff \psi(y|l_t, H) > 0$ for any y, l_t, and hence $\sum_{y \text{ s.t. } \psi(y|l_t,H)>0} \psi(y|l_t, L) = 1$. Hence, (l_t) is a nonnegative martingale, which converges by Doob's theorem. ∎

Proof of Lemma 2. By Lemma 1, it suffices to consider convergent sequences of likelihood ratios l_t. If $l_t \to l^* \notin J_H$ then $\bar{p}(l^*) > \underline{b}$, so by continuity of F^H and $\bar{p}$,

$$\mathbb{P}(a_{t+1} = L|l_t, H) = F^H(\bar{p}(l_t)) \to F^H(\bar{p}(l^*)) > 0,$$

and hence $\lim_{t\to\infty} \mathbb{E}[u(a_{t+1}, \theta)|l_t, H] < 1$. Similarly, if $l_t \to l^* \in J_H$ then $\mathbb{P}(a_{t+1} = L|l_t, H) \to 0$, so $\lim_{t\to\infty} \mathbb{E}[u(a_{t+1}, \theta)|l_t, H] = 1$. Thus, learning succeeds iff $l_\infty \in J_H$ almost surely, or equivalently (since $\text{supp}(l_\infty)$ and J_H are closed) $\text{supp}(l_\infty) \subset J_H$. ∎

Proof of Theorem 1. We show that UI implies that learning succeeds for any prior (the converse being immediate). Assume w.l.o.g. that $\theta = H$. Continuity of the signal distributions F^θ implies that $\psi(y|\theta, l)$ and $\phi(y, l)$ are continuous in l, for each y. Hence, Theorem B.2 in Smith and Sørensen (2000) implies that any $l^* \in \text{supp}(l_\infty)$ is stationary: i.e., $\phi(y, l^*) = l^*$ for all y such that $\psi(y|H, l^*) > 0$. By definition of ϕ, l^* satisfies this condition only if (i) $l^* = 0$ or (ii) $l^* > 0$ and $\psi(y|L, l^*) = \psi(y|H, l^*)$ for all y such that $\psi(y|H, l^*) > 0$. But the second scenario is ruled out by UI, so we have $\text{supp}(l_\infty) = \{0\} \in J_H$, and thus learning succeeds by Lemma 2. ∎

Proof of Theorem 2. Let $g(p) := \Delta(1, \bar{p}^{-1}(p))$. Since $\bar{p}$ is a strictly increasing bijection from $(0, \infty)$ onto $(0, 1)$, confounding points correspond to roots of g in $(0, 1)$, and

$\operatorname{sign} \Delta_l(1, l^*) = \operatorname{sign} g'(\bar{p}(l^*))$. Expanding,

$$g(p) = F^H(p)\mathbb{P}_{LH} + (1 - F^H(p))\mathbb{P}_{HH} - F^L(p)\mathbb{P}_{LL} - (1 - F^L(p))\mathbb{P}_{HL}.$$

Note that $g(\underline{b}) = \mathbb{P}_{HH} - \mathbb{P}_{HL} > 0$ and $g(\bar{b}) = \mathbb{P}_{LH} - \mathbb{P}_{LL} \neq 0$. Outside $[\underline{b}, \bar{b}]$, g is constant and nonzero, so any zero of g lies in $(\underline{b}, \bar{b})$, where

$$g'(p) = f^H(p)(\mathbb{P}_{LH} - \mathbb{P}_{HH}) - f^L(p)(\mathbb{P}_{LL} - \mathbb{P}_{HL}).$$

By the no-introspection property $f^H(p)/f^L(p) = p/(1-p)$. Thus, $\operatorname{sign} g'(p)$ equals the sign of a monotone function of p, so g has at most one critical point on $(\underline{b}, \bar{b})$. Finally, simple algebra shows that when one exists it equals p_m.

If match confounding holds, then $g(\underline{b}) > 0 > g(\bar{b})$, so by continuity g has a root p^*. This root is unique since g has at most one critical point, and $g'(p^*) < 0$ since g transitions from positive to negative at p^*.

Now assume match confounding fails. Then $g(\underline{b}), g(\bar{b}) > 0$, so UI fails iff g attains a nonpositive value on $(\underline{b}, \bar{b})$. We show that this occurs iff there is oppositional confounding and (4) holds.

If oppositional confounding also fails, then $\mathbb{P}_{HH} \geqslant \mathbb{P}_{LL}$. Since F^H first-order stochastically dominates F^L, for every $p \in (\underline{b}, \bar{b})$ we can use a standard monotone coupling to construct coupled random variables (actions) A_H, A_L taking values in $\{L, H\}$ such that $A_H \geqslant A_L$ almost surely and $\Pr(A_\theta = L) = F^\theta(p)$. For any realization, $\mathbb{P}_{A_H H} \geqslant \mathbb{P}_{A_L L}$: if $A_H = A_L$, this follows from $\mathbb{P}_{aH} > \mathbb{P}_{aL}$ (by the absence of match confounding for $a = L$; by Assumption 1.2 for $a = H$); if $A_H = H$ and $A_L = L$, it follows from $\mathbb{P}_{HH} \geqslant \mathbb{P}_{LL}$. Moreover, since $p \in (\underline{b}, \bar{b})$, the event $\{A_H = A_L\}$ has positive probability, so the inequality is strict in expectation, giving $g(p) = \mathbb{E}[\mathbb{P}_{A_H H} - \mathbb{P}_{A_L L}] > 0$. Hence, UI holds.

If oppositional confounding holds, both $\mathbb{P}_{LH} - \mathbb{P}_{HH}$ and $\mathbb{P}_{LL} - \mathbb{P}_{HL}$ are strictly positive, so $p_m \in (0, 1)$ is well-defined. Since p_m is the only candidate critical point of g, g attains a nonpositive value on $(\underline{b}, \bar{b})$ iff (4) holds (noting that if (4) holds then $p_m \in (\underline{b}, \bar{b})$). When (4) holds with equality, p_m is the unique root of g and $g'(p_m) = 0$; when it holds strictly, g has exactly two roots $p_1^* < p_m < p_2^*$, with $g'(p_1^*) < 0 < g'(p_2^*)$. ∎

Proof of Corollary 1. Under oppositional confounding, $d_L := \mathbb{P}_{LL} - \mathbb{P}_{HL} > 0$ and $d_H := \mathbb{P}_{LH} - \mathbb{P}_{HH} > 0$, so $p_m = d_L/(d_L + d_H) \in (0, 1)$. Collecting terms, the left-hand side of (4) equals

$$\mathbb{P}_{HH} - \mathbb{P}_{HL} + d_H F^H(p_m) - d_L F^L(p_m), \tag{9}$$

which is strictly increasing in $F^H(p_m)$ and strictly decreasing in $F^L(p_m)$. By Theorem 2, under oppositional confounding UI fails if and only if (4) holds: that is, if and only if (9) is nonpositive.

The first two claims are immediate. If UI holds for (F^L, F^H) then (9) is positive, and any $(\tilde{F}^L, \tilde{F}^H)$ that is less precise at p_m has $\tilde{F}^L(p_m) < F^L(p_m)$ and $\tilde{F}^H(p_m) > F^H(p_m)$, which only raises (9), so UI still holds. If UI fails for (F^L, F^H) then (9) is nonpositive, and any $(\tilde{F}^L, \tilde{F}^H)$ that is more precise at p_m has $\tilde{F}^L(p_m) > F^L(p_m)$ and $\tilde{F}^H(p_m) < F^H(p_m)$, which only lowers (9), so UI still fails.

For the limiting claims we evaluate (9) along each limit, using $p_m \in (0,1)$. If (F_n^L, F_n^H) converges to perfect information, then p_m is a continuity point of both F_{δ_0} and F_{δ_1}, so $F_n^L(p_m) \to 1$ and $F_n^H(p_m) \to 0$, and (9) converges to $\mathbb{P}_{HH} - \mathbb{P}_{HL} - d_L = \mathbb{P}_{HH} - \mathbb{P}_{LL} < 0$. Hence, (4) holds strictly for all sufficiently large n, and UI fails.

If (F_n^L, F_n^H) converges to no information and $p_m \neq 1/2$, then p_m is a continuity point of $F_{\delta_{1/2}}$, so $F_n^L(p_m)$ and $F_n^H(p_m)$ both tend to 0 when $p_m < 1/2$, and both to 1 when $p_m > 1/2$. In the former case (9) tends to $\mathbb{P}_{HH} - \mathbb{P}_{HL} > 0$, in the latter to $\mathbb{P}_{HH} - \mathbb{P}_{HL} + d_H - d_L = \mathbb{P}_{LH} - \mathbb{P}_{LL} > 0$. If instead $p_m = 1/2$, then $d_L = d_H$, so (9) equals $\mathbb{P}_{HH} - \mathbb{P}_{HL} + d_H(F_n^H(1/2) - F_n^L(1/2))$. By the no-introspection property,

$$F_n^L(1/2) - F_n^H(1/2) = \int_{p \leqslant 1/2} \left(\frac{f_n^L(p)}{f_n^H(p)} - 1 \right) dF_n^H(p) = \int_{p \leqslant 1/2} \frac{1-2p}{p} dF_n^H(p), \tag{10}$$

which tends to 0 as $F_n^L, F_n^H \to \delta_{1/2}$,[30] so (9) again tends to $\mathbb{P}_{HH} - \mathbb{P}_{HL} > 0$. In every case (9) tends to a strictly positive limit, so (4) fails for all sufficiently large n, and UI holds. ■

Proof of Lemma 3. We prove part (a). Part (b) is implied by Theorem C.1 of Smith and Sørensen (2000).

Define the shifted process (z_t) by $z_t := x_t - x^*$. Assume towards contradiction that starting from some initial condition $x_0 \notin D$, there is positive probability that $x_t \to x^*$ and therefore $z_t \to 0$. We proceed in three steps. First, we use a coupling argument and a Taylor approximation to approximate $(|z_t|)$ by a nonnegative linear process (w_t). Second, we show that once z_t is sufficiently close to 0, we have $w_t \leqslant |z_t|$. Thus, if $z_t \to 0$ then so does w_t. Finally, we use the strong law of large numbers to show that $\mathbb{P}(w_t \to 0) = 0$, yielding a contradiction. This structure follows Ellison and Fudenberg (1995), but each step is adjusted to allow more general continuation functions and state-dependent transition probabilities.

[30] By no-introspection $\frac{1}{p} dF_n^H(p) = \frac{1}{1-p} dF_n^L(p)$, and $\frac{1-2p}{p} \leqslant \frac{1}{p}$ with $\frac{1}{1-p} \leqslant 2$ for $p \leqslant 1/2$, so for any $\gamma \in (0, 1/2)$ the integral in (10) over $(0, 1/2-\gamma]$ is at most $2F_n^L(1/2-\gamma)$, while on $(1/2-\gamma, 1/2]$ the integrand is at most $2\gamma/(1/2-\gamma)$. Letting $n \to \infty$ then $\gamma \to 0$ gives the claim.

In particular, the definition of the coupled process is more involved and resembles the one used by Smith and Sørensen (2000) to extend part (b) of the lemma.

Step 1. Define $\gamma_i := \phi_x(i, x^*)$. Assume w.l.o.g that $|\gamma_0| \geqslant |\gamma_1|$ (relabeling otherwise). By hypothesis, $|\gamma_0|^{p(x^*)}|\gamma_1|^{1-p(x^*)} > 1$, so we must also have $|\gamma_1| > 0$. Since $\phi(i, x)$ are C^1 in x in a neighborhood of x^*, a Taylor approximation of $\phi(i, x)$ around x^* implies that

$$\phi(i, x_t) - x^* = \gamma_i(x_t - x^*) + o(|x_t - x^*|).$$

Let $\tilde{\phi}(i, z_t)$ and $\tilde{p}(z_t)$ be the continuation functions and transition probabilities for (z_t), respectively. These satisfy $\tilde{\phi}(i, z_t) = \gamma_i z_t + o(|z_t|)$ for $i = 0, 1$ and $\tilde{p}(z) = p(z + x^*)$ for all z. Denote $p^* := p(x^*) = \tilde{p}(0)$.

For any $w > 0$ and any $\varepsilon > 0$ such that $\varepsilon < \min\{|\gamma_1|, p^*\}$, define $\tilde{\phi}_\varepsilon(i, w) = (|\gamma_i| - \varepsilon)w$, and choose $\delta > 0$ such that both of the following hold:

$$\begin{aligned} &\text{(i)} \quad \tilde{\phi}_\varepsilon(i, w) \leqslant \min\{|\tilde{\phi}(i, w)|, |\tilde{\phi}(i, -w)|\} \text{ for all } w \in [0, \delta), \\ &\text{(ii)} \quad \tilde{p}(z) \geqslant p^* - \varepsilon \text{ for all } z \in (-\delta, \delta). \end{aligned} \tag{11}$$

Since there is positive probability that $z_t \to 0$, there exists $T \in \mathbb{N}$ such that the event $A := \{z_t \to 0 \text{ and } |z_t| < \delta \text{ for all } t \geqslant T\}$ also has positive probability. We define a stochastic process (w_t) coupled to (z_t) using the standard monotone coupling construction. Let (ζ_t) be a sequence of i.i.d uniform-(0,1) random variables, and recast (z_t) by setting $y_t = 0$ if and only if $\zeta_t \leqslant \tilde{p}(z_{t-1})$ and $y_t = 1$ otherwise, and defining $z_t = \tilde{\phi}(y_t, z_{t-1})$. Then, define (w_t) on $[0, \infty)$ by

$$w_t = \begin{cases} |z_t| & \text{if } t \leqslant T, \\ \tilde{\phi}_\varepsilon(0, w_{t-1}) & \text{if } t > T \text{ and } \zeta_t \leqslant p^* - \varepsilon, \\ \tilde{\phi}_\varepsilon(1, w_{t-1}) & \text{if } t > T \text{ and } \zeta_t > p^* - \varepsilon. \end{cases}$$

Step 2. We claim that if event A has positive probability then so does the event $B := \{w_t \to 0\}$. It suffices to show that $A \subset B$, which we establish by proving inductively that on the event A, for every nonnegative integer s we have $w_{T+s} \leqslant |z_{T+s}|$. The base case $s = 0$ holds by definition of (w_t), since on event A we have $w_T = |z_T|$. Assume that the induction hypothesis holds for $s \geqslant 0$ and consider $s + 1$. Since $\tilde{\phi}_\varepsilon(i, w)$ are increasing in w and $\tilde{\phi}_\varepsilon(0, w) \geqslant \tilde{\phi}_\varepsilon(1, w)$ for all $w > 0$, it follows that on the event A,

$$w_{T+s+1} \leqslant \tilde{\phi}_\varepsilon(y_{T+s+1}, w_{T+s}) \leqslant \tilde{\phi}_\varepsilon(y_{T+s+1}, |z_{T+s}|) \leqslant |\tilde{\phi}(y_{T+s+1}, z_{T+s})| = |z_{T+s+1}|.$$

The first inequality holds because $|z_{T+s}| < \delta$ and condition (ii) in (11) imply $\tilde{p}(z_{T+s}) \geqslant p^* - \varepsilon$, which implies that the process (w_t) takes the larger 0-transition only if $y_{T+s+1} = 0$. The second inequality follows from the induction hypothesis and the monotonicity of $\tilde{\phi}_\varepsilon$ in its second argument. The third inequality follows from $|z_{T+s}| < \delta$ and condition (i) in (11).

Step 3: We complete the proof by showing that $\Pr(w_t \to 0) = 0$. If $w_t \to 0$, then the process $(\log(w_t))$ must converge to $-\infty$. Moreover, we have $\log(w_{T+s}) = \log(w_T) + \sum_{\tau=1}^{s} u_\tau$, where the u_τ are i.i.d Bernoulli random variables taking values $\log(|\gamma_0| - \varepsilon)$ and $\log(|\gamma_1| - \varepsilon)$ with probabilities $p^* - \varepsilon$ and $1 - p^* + \varepsilon$, respectively.[31] Thus, the strong law of large numbers implies that the probability that $\log(w_t)$ converges to $-\infty$ must be 0 if $\mathbb{E}[u_\tau] > 0$, i.e., if

$$(p^* - \varepsilon)\log(|\gamma_0| - \varepsilon) + (1 - p^* + \varepsilon)\log(|\gamma_1| - \varepsilon) > 0.$$

This inequality holds for all sufficiently small $\varepsilon > 0$ by continuity and the hypothesis that $|\gamma_0|^{p^*}|\gamma_1|^{1-p^*} > 1$, completing the proof. ∎

Proof of Lemma 4. Fix $y \in \{0, 1\}$. By (2), for any $l \neq l'$, we have $\phi(y, l) = \phi(y, l')$ iff

$$l\psi(y|L, l)\psi(y|H, l') - l'\psi(y|L, l')\psi(y|H, l) = 0.$$

Since $\psi(y|\theta, l)$ is an affine function of $\mathbb{P}_{L,\theta}$ and $\mathbb{P}_{H,\theta}$, this is a polynomial equation in $(\mathbb{P}_{a\theta})$. Moreover, the equation is not an identity: for example, it fails if all four $\mathbb{P}_{a\theta}$'s are equal. Therefore, the set of solutions $\mathcal{Z}_{l,l'} \subset [0, 1]^4$ has empty interior by the identity theorem for analytic functions, and is therefore nowhere dense as a closed set with empty interior. Now, if $\phi(y, l)$ is constant over a nonempty open interval then $(\mathbb{P}_{a\theta})$ must lie in $\mathcal{Z}_{l,l'}$ for some distinct $l, l' \in \mathbb{Q}$, and hence $(\mathbb{P}_{a\theta}) \in \mathcal{Z} := \bigcup_{l,l' \in \mathbb{Q}: l \neq l'} \mathcal{Z}_{l,l'}$. Finally, $\mathcal{Z}$ is meagre as a countable union of nowhere dense sets, so its complement is generic. ∎

Proof of Lemma 5. Let $\phi^{-t}(l^*)$ be the set of priors l_0 from which there is a positive probability of reaching l^* in t periods: this can be defined inductively by $\phi^{-1}(l^*) = \{l : l^* \in \{\phi(0, l), \phi(1, l)\}\}$, $\phi^{-t}(l^*) = \{l : \{\phi(0, l), \phi(1, l)\} \cap \phi^{-(t-1)}(l^*) \neq \varnothing\}$. We show that $\phi^{-t}(l^*)$ is closed and has empty interior, and hence is nowhere dense. This implies that $\bigcup_{t \geqslant 1} \phi^{-t}(l^*)$, the set of priors from which there is a positive probability of reaching l^* in finite time, is meagre.

First, $\phi^{-1}(l^*)$ is closed, because $\{l^*\}$ is closed and $\phi(y, l)$ is continuous. Inductively, $\phi^{-t}(l^*)$ is closed.

[31] Here we use $x_0 \notin D$, which ensures that $z_t \neq 0$ for all t, and therefore $w_t \neq 0$.

Second, $\phi^{-1}(l^*)$ has empty interior. Suppose for contradiction it contained a nonempty open interval I. Then $I = A_0 \cup A_1$, where $A_y = \{l \in I : \phi(y, l) = l^*\}$ is closed in I as the preimage of $\{l^*\}$ under the continuous map $\phi(y, \cdot)$. Since I cannot be written as the union of two closed sets both with empty interior (such a union is nowhere dense in I), some A_y contains a nonempty open sub-interval I', on which $\phi(y, \cdot)$ is constant (equal to l^*), contradicting Assumption 2.

Finally, suppose for contradiction that $\phi^{-(t-1)}(l^*)$ has empty interior, but $\phi^{-t}(l^*)$ contains a nonempty open interval I. By definition of $\phi^{-t}(l^*)$, for each $l \in I$ at least one of $\phi(0, l), \phi(1, l)$ lies in $\phi^{-(t-1)}(l^*)$, so $I = B_0 \cup B_1$ where $B_y = \{l \in I : \phi(y, l) \in \phi^{-(t-1)}(l^*)\}$. Each B_y is closed in I, as the preimage of the closed set $\phi^{-(t-1)}(l^*)$ under the continuous map $\phi(y, \cdot)$. Since I cannot be written as the union of two closed sets both with empty interior, some B_y contains a nonempty open sub-interval I'. Then $\phi(y, I') \subset \phi^{-(t-1)}(l^*)$ is the continuous image of an interval, hence an interval (possibly degenerate). If $\phi(y, I')$ is a single point, then $\phi(y, \cdot)$ is constant on I', contradicting Assumption 2. Otherwise, $\phi(y, I')$ contains a nonempty open interval, contradicting the hypothesis that $\phi^{-(t-1)}(l^*)$ has empty interior. ∎

Proof of Theorem 3. In each state θ, the process (l_t) is a Markov process over $\mathbb{R}_+$ with

$$l_{t+1} = \begin{cases} \phi(0, l_t) & \text{with probability } \psi(0|\theta, l_t), \\ \phi(1, l_t) & \text{with probability } 1 - \psi(0|\theta, l_t). \end{cases}$$

Lemma 3 applies to this process at any confounding point l^*. First, $\psi(y|\theta, l)$ is continuous in l and satisfies $\psi(y|\theta, l) \in (0, 1)$, since it is a convex combination of $\mathbb{P}_{L\theta}$ and $\mathbb{P}_{H\theta}$ and $\mathbb{P}_{a\theta} \in (0, 1)$ by Assumption 1.1. Second, since a confounding point cannot lie in a cascade set, $\bar{p}(l^*) \in (\underline{b}, \bar{b})$, and F^L and F^H are C^1 on $(\underline{b}, \bar{b})$ because f^L and f^H are continuous there. Hence $\phi(0, \cdot)$ and $\phi(1, \cdot)$ are C^1 in a neighborhood of l^*. Moreover, for any confounding point l^*, the set D_{l^*} of initial points l_0 from which there is positive probability of reaching l^* in finite time is meagre, by Lemma 5. Since there are at most two confounding points by Theorem 2, the union of the sets D_{l^*} over all confounding points is also meagre. Hence, if the instability condition holds at all confounding points, then learning generically succeeds. Conversely, if the stability condition holds at a confounding point l^*, then learning fails from any prior in a neighborhood of l^*. ∎

Proof of Proposition 1. By (7), $\varepsilon(1, l^*)$ and $\varepsilon(0, l^*)$ have opposite signs, so $\varepsilon(1, l^*) \neq 0$

implies that exactly one of them is negative, namely $\varepsilon(\underline{y}, l^*)$. Let $u := \varepsilon(\underline{y}, l^*) < 0$. By (7), $\varepsilon(1-\underline{y}, l^*) = -\frac{\underline{\psi}}{1-\underline{\psi}}u > 0$, so we can write the left hand side of (Stability) and (Instability) as,

$$|1+u|^{\underline{\psi}}\left(1-\frac{\underline{\psi}}{1-\underline{\psi}}u\right)^{1-\underline{\psi}} =: T(u,\underline{\psi}),$$

where the second factor is positive since $u < 0$. It remains to study the sign of $T(u,\underline{\psi}) - 1$ on $u < 0$.

For $u \in (-1, 0)$, both factors of T are positive. By the weighted AM-GM inequality,

$$T(u,\underline{\psi}) \leqslant \underline{\psi}(1+u) + (1-\underline{\psi})\left(1-\frac{\underline{\psi}}{1-\underline{\psi}}u\right) = 1,$$

with equality iff $1+u = 1-\frac{\underline{\psi}}{1-\underline{\psi}}u$, i.e., $u = 0$. Hence $T(u,\underline{\psi}) < 1$ on $(-1,0)$, and $T(-1,\underline{\psi}) = 0 < 1$.

For $u < -1$, $T(u,\underline{\psi}) = (-(1+u))^{\underline{\psi}}(1-\frac{\underline{\psi}}{1-\underline{\psi}}u)^{1-\underline{\psi}}$, and

$$\frac{d\log T(u,\underline{\psi})}{du} = \frac{\underline{\psi}}{1+u} - \frac{\underline{\psi}}{1-\frac{\underline{\psi}}{1-\underline{\psi}}u}.$$

Since $1+u < 0$ and $1-\frac{\underline{\psi}}{1-\underline{\psi}}u > 0$, both terms on the right-hand side are negative. Thus $T(\cdot,\underline{\psi})$ is strictly decreasing on $(-\infty,-1)$. Since $T(-1^-,\underline{\psi}) = 0$ and $T(u,\underline{\psi}) \to \infty$ as $u \to -\infty$, there is a unique $c(\underline{\psi}) \in (-\infty,-1)$ with $T(c(\underline{\psi}),\underline{\psi}) = 1$; moreover $T < 1$ on $(c(\underline{\psi}),-1)$ and $T > 1$ on $(-\infty, c(\underline{\psi}))$. Combined with the previous paragraph, this proves the cutoff characterization.

Finally, at $u = -2$ the first factor of T is 1 and the second is $(\frac{1+\underline{\psi}}{1-\underline{\psi}})^{1-\underline{\psi}} > 1$, so $T(-2,\underline{\psi}) > 1$ and $c(\underline{\psi}) > -2$. At $u = -1 - W(1/e) \approx -1.278$ a short computation gives $T(-1-W(1/e),\underline{\psi}) < 1$, so $c(\underline{\psi}) < -1 - W(1/e)$.[32] ∎

Proof of Theorem 4. Let $d_L := \mathbb{P}_{LL} - \mathbb{P}_{HL}$ and $d_H := \mathbb{P}_{LH} - \mathbb{P}_{HH}$. If (F_n^L, F_n^H) converges to perfect information and n is sufficiently large, we say signals are *near perfect*. If (F_n^L, F_n^H) converges to no information and n is sufficiently large, we say signals are *near noise*.

Step 1 (characterization of confounding points). By Theorem 2.1, under match confounding there is always a unique confounding point l^*, which satisfies $\Delta_l(1, l^*) < 0$. By Corollary 1, under oppositional confounding with near-noise signals UI holds, so there is no confounding

[32] Write $w := W(1/e)$, so $\log w = -(1+w)$. At $u = -1-w$ the factors of T are $w^{\underline{\psi}}$ and $(\frac{1+\underline{\psi}w}{1-\underline{\psi}})^{1-\underline{\psi}}$, so $\log x < x-1$ for $x \neq 1$ gives $\log T(-1-w,\underline{\psi}) = -\underline{\psi}(1+w) + (1-\underline{\psi})\log\frac{1+\underline{\psi}w}{1-\underline{\psi}} < -\underline{\psi}(1+w) + \underline{\psi}(1+w) = 0$. Hence $T(-1-w,\underline{\psi}) < 1$.

point, and with near-perfect signals (4) holds strictly, so by Theorem 2.2 there are two confounding points $l_1^* < l_2^*$ with $\Delta_l(1, l_1^*) < 0 < \Delta_l(1, l_2^*)$. In sum, with near-perfect signals there is, under either form of confounding, exactly one confounding point with $\Delta_l(1, l^*) < 0$, and with near-noise signals there is a unique confounding point under match confounding and none under oppositional confounding.

Step 2 (sufficient conditions for (in)stability). Fix a confounding point l^* where $\delta^* := \Delta_l(1, l^*) < 0$. Let $p^* = \bar{p}(l^*)$ and $\psi^* = \psi(1|L, l^*) = \psi(1|H, l^*)$. We derive sufficient conditions for l^* to satisfy (Stability) and (Instability). Since $\psi(1|\theta, l) = F^\theta(\bar{p}(l))\mathbb{P}_{L\theta} + (1 - F^\theta(\bar{p}(l)))\mathbb{P}_{H\theta}$ for $\theta \in \{L, H\}$, and $\bar{p}(l) = l/(1 + l)$, we have

$$\psi_l(1|\theta, l^*) = f^\theta(\bar{p}(l^*))\bar{p}'(l^*)d_\theta = f^\theta(p^*)\frac{d_\theta}{(1 + l^*)^2}.$$

Thus,

$$\delta^* = \psi_l(1|H, l^*) - \psi_l(1|L, l^*) = \frac{1}{(1 + l^*)^2}\left(f^H(p^*)d_H - f^L(p^*)d_L\right).$$

Multiplying by l^* and using the identity $\frac{l^*}{(1+l^*)^2} = p^*(1 - p^*)$, we get

$$l^*\delta^* = p^*(1 - p^*)\left(f^H(p^*)d_H - f^L(p^*)d_L\right).$$

Because $\delta^* < 0$, the term $1 - \frac{l^*\delta^*}{\psi^*}$ in (Stability) is strictly greater than 1. For the term inside the other absolute value, if $-l^*\delta^* < 1 - \psi^*$, then $1 + \frac{l^*\delta^*}{1-\psi^*} > 0$. In this case, we can drop the absolute values and apply the weighted AM-GM inequality:

$$\left(1 + \frac{l^*\delta^*}{1 - \psi^*}\right)^{(1-\psi^*)}\left(1 - \frac{l^*\delta^*}{\psi^*}\right)^{\psi^*} < (1 - \psi^*)\left(1 + \frac{l^*\delta^*}{1 - \psi^*}\right) + \psi^*\left(1 - \frac{l^*\delta^*}{\psi^*}\right) = 1.$$

On the other hand, if $-l^*\delta^* > 2(1 - \psi^*)$, then $1 + \frac{l^*\delta^*}{1-\psi^*} < -1$, meaning its absolute value is strictly greater than 1. Since $\left|1 - \frac{l^*\delta^*}{\psi^*}\right| > 1$ as well, their product is strictly greater than 1. In sum, for a confounding point l^* where $\Delta_l(1, l^*) < 0$, a sufficient condition for stability is $-l^*\delta^* < 1 - \psi^*$, and a sufficient condition for instability is $-l^*\delta^* > 2(1 - \psi^*)$.

We will apply these conditions to sequences of confounding points $l^*(n)$ such that $\Delta_l(1, l^*(n)) < 0$. For such a sequence, let $p^*(n) = \bar{p}(l^*(n))$, $\psi^*(n) = \psi(1|L, l^*(n))$, and $\delta^*(n) = \Delta_l(1, l^*(n))$, and define the sequence $h(n)$ as the value of $-l^*(n)\delta^*(n)$ at these points:

$$h(n) := -p^*(n)(1 - p^*(n))\left(f_n^H(p^*(n))d_H - f_n^L(p^*(n))d_L\right) \tag{12}$$

Since $\psi^*(n)$ is bounded away from 0 and 1 (by Assumption 1.1), to establish part 1 of the

theorem, it suffices to show that $h(n) \to 0$ with near-perfect signals. Conversely, to establish part 2 of the theorem, it suffices to show that $h(n) \to \infty$ with near-noise signals and match confounding (as, by Step 1, UI holds with near-noise signals and oppositional confounding).

Step 3 (possible limits of private belief cdfs at confounding points). Let (F_n^L, F_n^H) be a sequence of private belief distributions, and suppose that for all sufficiently large n there exists a confounding point $l^*(n)$, with associated private belief threshold $p^*(n) = \bar{p}(l^*(n))$. Further assume that the sequence $(F_n^L(p^*(n)), F_n^H(p^*(n)))$ converges to some limit $(F_\infty^L, F_\infty^H) \in [0,1]^2$. Since $\Delta(1, l^*(n)) = 0$, we have

$$F_n^H(p^*(n))(\mathbb{P}_{HH} - \mathbb{P}_{LH}) - F_n^L(p^*(n))(\mathbb{P}_{HL} - \mathbb{P}_{LL}) = \mathbb{P}_{HH} - \mathbb{P}_{HL}. \tag{13}$$

We claim that at least one of F_∞^L, F_∞^H lies in $(0,1)$. Suppose for contradiction that both are in $\{0,1\}$. By stochastic dominance, $F_n^L \geqslant F_n^H$ (Smith and Sørensen (2000) Lemma A.1(c)), so $F_\infty^L \geqslant F_\infty^H$. This leaves three possible cases: $(0,0)$, $(1,0)$, and $(1,1)$. We show that none of these are possible.

If $F_\infty^L = F_\infty^H = 0$, then the left-hand side of (13) converges to 0, while the right-hand side is $\mathbb{P}_{HH} - \mathbb{P}_{HL} > 0$ by Assumption 1.2.

If $F_\infty^L = F_\infty^H = 1$, then the left-hand side of (13) converges to $\mathbb{P}_{HH} - \mathbb{P}_{LH} - \mathbb{P}_{HL} + \mathbb{P}_{LL}$, while the right-hand side is $\mathbb{P}_{HH} - \mathbb{P}_{HL}$. This is equal to the right-hand side only if $\mathbb{P}_{LL} = \mathbb{P}_{LH}$, which is ruled out by Assumption 1.2.

If $F_\infty^L = 1$ and $F_\infty^H = 0$, then the left-hand side of (13) converges to $\mathbb{P}_{LL} - \mathbb{P}_{HL}$. This is equal to the right-hand side only if $\mathbb{P}_{LL} = \mathbb{P}_{HH}$, which is ruled out by assumption.

Step 4 (proof of part 1). By Step 1, for either type of confounding, with near-perfect signals there is a unique confounding point $l^*(n)$ such that $\Delta_l(1, l^*(n)) < 0$. We show that $h(n) \to 0$, which implies that learning robustly fails by Step 2 and Theorem 3.

By the no-introspection property, $f_n^L(p)/f_n^H(p) = (1-p)/p$ for all $p \in (\underline{b}, \bar{b})$. Noting that $p^*(n) \in (\underline{b}, \bar{b})$, we can rewrite (12) as

$$\begin{aligned} h(n) &= (1 - p^*(n)) f_n^H(p^*(n)) \left[(1 - p^*(n)) d_L - p^*(n) d_H\right] \\ &= p^*(n) f_n^L(p^*(n)) \left[(1 - p^*(n)) d_L - p^*(n) d_H\right]. \end{aligned} \tag{14}$$

Since $p^*(n) \in [0,1]$ for all n, and the expression in square brackets is bounded, to prove that $h(n) \to 0$ it suffices to prove that along every subsequence as $n \to \infty$, there exists a sub-subsequence where either $f_n^H(p^*(n)) \to 0$ or $f_n^L(p^*(n)) \to 0$. Given an arbitrary subsequence $n \to \infty$, we can choose a sub-subsequence for which $(F_n^L(p^*(n)), F_n^H(p^*(n))) \to (F_\infty^L, F_\infty^H)$. By

Step 3, either $F_\infty^L \in (0,1)$ or $F_\infty^H \in (0,1)$. In the first case, tail-regularity (Condition 1) implies $f_n^H(p^*(n)) \to 0$; in the second, tail-regularity (Condition 2) implies $f_n^L(p^*(n)) \to 0$ (noting in both cases that tail-regularity of (F_n^L, F_n^H) implies tail-regularity of any subsequence).

Step 5 (proof of part 2). It suffices to show that $h(n) \to \infty$ with near-noise signals and match confounding. As in Step 4, it suffices to prove this for a sub-subsequence along which $(F_n^L(p^*(n)), F_n^H(p^*(n))) \to (F_\infty^L, F_\infty^H)$. By Step 3, at least one of F_∞^L, F_∞^H lies in $(0,1)$. This implies that $p^*(n) \to 1/2$, as otherwise $F_n^L, F_n^H \to \delta_{1/2}$ implies that $F_n^L(p^*(n)), F_n^H(p^*(n))$ both converge to either 0 or 1. Applying the no-introspection property as in the proof of Corollary 1, convergence to complete noise further implies that $F_\infty^L = F_\infty^H$. By center regularity (of the subsequence), this implies that $f_n^L(p^*(n)), f_n^H(p^*(n)) \to \infty$. Since

$$(1-p^*(n))\left[(1-p^*(n))d_L - p^*(n)d_H\right] \to \frac{1}{4}\left(\mathbb{P}_{LL} - \mathbb{P}_{LH} + \mathbb{P}_{HH} - \mathbb{P}_{HL}\right) > 0,$$

this implies that

$$h(n) = (1-p^*(n))f_n^H(p^*(n))\left[(1-p^*(n))d_L - p^*(n)d_H\right] \to \infty,$$

completing the proof. ∎

Proof of Proposition 2. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), G)$, and let $\varepsilon : \Omega \to \mathbb{R}$ denote the identity map, so that $\varepsilon \sim G$. Define the random variables

$$P_n^L(\varepsilon) := p_{\sigma_n}(\sigma_n \varepsilon) = \frac{g(\varepsilon - 1/\sigma_n)}{g(\varepsilon - 1/\sigma_n) + g(\varepsilon)},$$
$$P_n^H(\varepsilon) := p_{\sigma_n}(1 + \sigma_n \varepsilon) = \frac{g(\varepsilon)}{g(\varepsilon) + g(\varepsilon + 1/\sigma_n)}.$$

Then P_n^L has cdf F_n^L and P_n^H has cdf F_n^H. Since g is a log-concave density on $\mathbb{R}$, we have $g(x) \to 0$ as $|x| \to \infty$. Hence for every fixed $\varepsilon \in \mathbb{R}$, if $\sigma_n \to 0$ then $\lim_{n\to\infty} P_n^L(\varepsilon) = 0$ and $\lim_{n\to\infty} P_n^H(\varepsilon) = 1$. And if $\sigma_n \to \infty$ then $\lim_{n\to\infty} P_n^L(\varepsilon) = \lim_{n\to\infty} P_n^H(\varepsilon) = 1/2$. Since almost sure convergence implies weak convergence, this gives $F_n^L \to F_{\delta_0}$ and $F_n^H \to F_{\delta_1}$ as $\sigma_n \to 0$, and $F_n^L, F_n^H \to F_{\delta_{1/2}}$ as $\sigma_n \to \infty$. Thus, (F_n^L, F_n^H) converges to perfect information if $\sigma_n \to 0$ and to no information if $\sigma_n \to \infty$.

Turning to the regularity conditions, note that[33]

$$\frac{dp_\sigma(s)}{ds} = \frac{p_\sigma(s)\left(1-p_\sigma(s)\right)}{\sigma}\left[r\left(\frac{s-1}{\sigma}\right) - r\left(\frac{s}{\sigma}\right)\right].$$

By the inverse function theorem, $s'_\sigma(p) = 1/p'_\sigma(s_\sigma(p))$, so by (8) private belief densities are given by

$$\begin{aligned} f_n^H(p) &= \frac{d}{dp}F_n^H(p) = \frac{g\left(\frac{s_{\sigma_n}(p)-1}{\sigma_n}\right)}{p(1-p)\left[r\left(\frac{s_{\sigma_n}(p)-1}{\sigma_n}\right) - r\left(\frac{s_{\sigma_n}(p)}{\sigma_n}\right)\right]}, \\ f_n^L(p) &= \frac{d}{dp}F_n^L(p) = \frac{g\left(\frac{s_{\sigma_n}(p)}{\sigma_n}\right)}{p(1-p)\left[r\left(\frac{s_{\sigma_n}(p)-1}{\sigma_n}\right) - r\left(\frac{s_{\sigma_n}(p)}{\sigma_n}\right)\right]}. \end{aligned} \tag{15}$$

Center regularity: Assume $\sigma_n \to \infty$, and consider a sequence (p_n) with $\lim_{n\to\infty} F_n^H(p_n) = q \in (0,1)$. Since $F_n^H \to F_{\delta_{1/2}}$, it must be the case that $p_n \to 1/2$. For each n, let $s_n := s_{\sigma_n}(p_n)$ denote the signal that induces belief p_n, and define $x_n := (s_n - 1)/\sigma_n$. By (8), $F_n^H(p_n) = G(x_n)$, so $G(x_n) \to q$, which implies $x_n \to G^{-1}(q) \in \mathbb{R}$. Furthermore, $s_n/\sigma_n = x_n + 1/\sigma_n \to G^{-1}(q)$.

By continuity of the score function $r = g'/g$, $\lim_{n\to\infty}[r(x_n) - r(x_n + 1/\sigma_n)] = 0$. Then by (15),

$$\lim_{n\to\infty} f_n^H(p_n) = \lim_{n\to\infty} \frac{g(x_n)}{p_n(1-p_n)[r(x_n) - r(x_n+1/\sigma_n)]} = \frac{4g(G^{-1}(q))}{\lim_{n\to\infty}[r(x_n) - r(x_n+1/\sigma_n)]} = \infty.$$

Tail regularity: Assume that the score function is unbounded above and below and that $\sigma_n \to 0$. We verify the two conditions for tail-regularity. First, consider a sequence (p_n) such that $\lim_{n\to\infty} F_n^L(p_n) = q \in (0,1)$. Since $F_n^L \to F_{\delta_0}$ and $F_n^H \to F_{\delta_1}$, $q < 1$ implies $p_n \to 0$, and hence $\lim_{n\to\infty} F_n^H(p_n) = 0$. As before, let $s_n = s_{\sigma_n}(p_n)$ and $x_n = (s_n - 1)/\sigma_n$. Then $F_n^H(p_n) = G(x_n) \to 0$ implies $x_n \to -\infty$, and $F_n^L(p_n) = G(x_n + 1/\sigma_n) \to q$ implies $x_n + 1/\sigma_n \to G^{-1}(q) \in \mathbb{R}$. Using the identity $p_n(1-p_n) = \frac{g(x_n)g(x_n+1/\sigma_n)}{(g(x_n)+g(x_n+1/\sigma_n))^2}$, we can rewrite $f_n^H(p_n)$ from (15) as:

$$f_n^H(p_n) = \frac{\left(g(x_n) + g(x_n+1/\sigma_n)\right)^2}{g(x_n+1/\sigma_n)[r(x_n) - r(x_n+1/\sigma_n)]}. \tag{16}$$

As $n \to \infty$, $g(x_n) \to 0$ and $g(x_n + 1/\sigma_n) \to g(G^{-1}(q)) > 0$. Thus, the numerator converges

[33] Since g is log concave, $r(x) = (\log g)'(x)$ is decreasing so the expression in square brackets is positive, and therefore $p_\sigma(s)$ is strictly increasing as claimed in Remark 4.3.

to $g(G^{-1}(q))^2$ and the $g(\cdot)$ term in the denominator converges to $g(G^{-1}(q))$. Because the score function is unbounded above, $r(x_n) \to \infty$ as $x_n \to -\infty$. Meanwhile, $r(x_n + 1/\sigma_n) \to r(G^{-1}(q))$, a finite constant. Therefore, the term in brackets diverges to ∞, which yields $\lim_{n\to\infty} f_n^H(p_n) = 0$.

Second, consider a sequence (p_n) such that $\lim_{n\to\infty} F_n^H(p_n) = q \in (0,1)$. Since $F_n^H \to F_{\delta_1}$ and $F_n^L \to F_{\delta_0}$, $q > 0$ implies $p_n \to 1$, and hence $\lim_{n\to\infty} F_n^L(p_n) = 1$. Here, $F_n^H(p_n) = G(x_n) \to q$ implies $x_n \to G^{-1}(q) \in \mathbb{R}$, and $F_n^L(p_n) = G(x_n+1/\sigma_n) \to 1$ implies $x_n+1/\sigma_n \to \infty$. Similarly rewriting $f_n^L(p_n)$ gives:

$$f_n^L(p_n) = \frac{(g(x_n) + g(x_n + 1/\sigma_n))^2}{g(x_n)[r(x_n) - r(x_n + 1/\sigma_n)]}. \tag{17}$$

As $n \to \infty$, $g(x_n + 1/\sigma_n) \to 0$ and $g(x_n) \to g(G^{-1}(q)) > 0$. Because the score function is unbounded below, $r(x_n + 1/\sigma_n) \to -\infty$ as $x_n + 1/\sigma_n \to \infty$, while $r(x_n)$ converges to a finite constant. The term in brackets therefore diverges to ∞, yielding $\lim_{n\to\infty} f_n^L(p_n) = 0$. ∎

Proof of Proposition 3. Write $\bar{P} := \mathbb{P}_{LL} = \mathbb{P}_{HH}$, $\underline{P} := \mathbb{P}_{LH} = \mathbb{P}_{HL}$, and $d := \bar{P} - \underline{P} > 0$. Note that in the notation of the proof of Theorem 4, $d = d_L = -d_H$. For convenience, reparametrize the scale by $\omega := 1/(2\sigma) \in (0, \infty)$.

By symmetry, the confounding point for any ω is $l^* = 1$, which is unique by Theorem 2. Thus, the private-belief threshold is $p^* := \bar{p}(l^*) = 1/2$, and, again by symmetry, the signal that induces it is $s^* = 1/2$ for every ω. By (8), $F^H(p^*) = \Phi(-\omega)$ and $F^L(p^*) = \Phi(\omega)$.

For private signal distributions with scale ω let $\psi^*(\omega) := \psi(1|L, l^*) = \psi(1|H, l^*)$ and $h(\omega) := -l^*\Delta_l(1, l^*)$ (the quantity $h(n)$ of (12) taking ω rather than n as an argument). Explicitly,

$$\psi^*(\omega) = \underline{P} + d \cdot \Phi(\omega) \qquad \text{and} \qquad h(\omega) = d \cdot \varphi(\omega)/\omega > 0. \tag{18}$$

The first formula follows from (2). For the second, since $p^* = 1/2$ the bracketed term in (14) equals d, so $h = f^H(p^*)d/2$, while (15) with the Gaussian score $r(u) = -u$ gives $f^H(p^*) = 2\varphi(\omega)/\omega$.

Using (7), define

$$m_1 := \phi_l(1, l^*) = 1 + \varepsilon(1, l^*) = 1 + \frac{h}{\psi^*} > 1, \qquad m_0 := \phi_l(0, l^*) = 1 + \varepsilon(0, l^*) = 1 - \frac{h}{1 - \psi^*}.$$

By Theorem 3 (which applies with Gaussian noise, as $\phi(y, \cdot)$ is real-analytic and non-constant, so Assumption 2 holds), l^* is stable, hence learning robustly fails, when the *stability index* $S(\omega) := \psi^* \log m_1 + (1 - \psi^*) \log |m_0|$ is negative, and unstable, hence learning generically

succeeds, when $S(\omega) > 0$. Note that S is C^1 around any ω for which $m_0 \neq 0$.

Since $\varepsilon(1, l^*) = h/\psi^* > 0 > -h/(1-\psi^*) = \varepsilon(0, l^*)$, the smaller elasticity is at $\underline{y} = 0$, with $\underline{\psi} = 1 - \psi^*$. Proposition 1 then implies that $S(\omega) > 0$ whenever $h/(1-\psi^*) > 2$ and $S(\omega) < 0$ whenever $h/(1-\psi^*) < 1.278$. As $\omega \to 0$, $h \to \infty$ while $1 - \psi^* \to 1 - \underline{P} - d/2 > 0$; as $\omega \to \infty$, $h \to 0$ while $1 - \psi^* \to 1 - \bar{P} > 0$. Hence $S > 0$ for sufficiently small ω and $S < 0$ for sufficiently large ω, so S has a zero (though undefined where $m_0 = 0$, S tends to $-\infty$ there, so the intermediate value theorem still yields a zero).

We now establish that this zero is unique by proving that every crossing is a down-crossing, i.e., $S(\omega) = 0$ implies $S'(\omega) < 0$. This implies that there exists ω^* such that $S > 0$ for all $\omega < \omega^*$ and $S < 0$ for all $\omega > \omega^*$, completing the proof.

Let ω_0 be a zero of S. Since ω_0 lies on the boundary between (Stability) and (Instability), Proposition 1 gives $\varepsilon(0, l^*) = c(1-\psi^*) < -1$, so $m_0(\omega_0) < 0$. By continuity, $m_0 < 0$ on an open neighborhood of ω_0, on which

$$S' = \psi^{*\prime} \log \frac{m_1}{|m_0|} + \psi^* \frac{m_1'}{m_1} + (1-\psi^*)\frac{m_0'}{m_0}.$$

The definitions of m_1 and m_0 give $\psi^* m_1 = \psi^* + h$ and $(1-\psi^*)m_0 = 1 - \psi^* - h$. Differentiating these identities and dividing by m_1 and m_0, respectively, gives

$$\psi^* \frac{m_1'}{m_1} = \frac{\psi^{*\prime} + h'}{m_1} - \psi^{*\prime} \qquad \text{and} \qquad (1-\psi^*)\frac{m_0'}{m_0} = -\frac{\psi^{*\prime} + h'}{m_0} + \psi^{*\prime},$$

so $S' = \psi^{*\prime} \log(m_1/|m_0|) + (\psi^{*\prime} + h')(1/m_1 + 1/|m_0|)$, using $-1/m_0 = 1/|m_0|$. From (18), $\psi^{*\prime}(\omega) = d \cdot \varphi(\omega) = h(\omega)\omega$ and, since $\varphi'(\omega) = -\omega\varphi(\omega)$, $h'(\omega) = -h(\omega)(\omega^2+1)/\omega$, so $\psi^{*\prime} + h' = -h/\omega$. Moreover, $S(\omega_0) = 0$ gives $\psi^* \log m_1 = (1-\psi^*)\log(1/|m_0|)$, so $\log(m_1/|m_0|) = (1/\psi^*)\log(1/|m_0|)$ at ω_0. Substituting, we have

$$\begin{aligned} S'(\omega_0) &= \frac{h}{\omega_0}\left((\omega_0^2/\psi^*)\log(1/|m_0|) - 1/m_1 - 1/|m_0|\right) \\ &\leqslant \frac{h}{\omega_0}\left(\omega_0^2/(e\psi^*|m_0|) - 1/m_1 - 1/|m_0|\right) \\ &< \frac{h}{\omega_0}\left(\omega_0^2/(e\psi^*|m_0|) - 1/|m_0|\right) = \frac{h}{\omega_0|m_0|}\left(\omega_0^2/(e\psi^*) - 1\right) < 0. \end{aligned}$$

The first inequality applies $x \log(1/x) \leqslant 1/e$, valid for every $x > 0$, at $x = |m_0|$. The second drops the term $-1/m_1 < 0$. The final inequality is implied by the following claim, which completes the proof. This last step relies on a numerical calculation involving Φ.

Claim. $\omega_0^2 < e\psi^*(\omega_0)$.

Proof. Throughout, ψ^* and h are evaluated at ω_0. As established above, the smaller

elasticity is $\varepsilon(0, l^*) = c(1-\psi^*)$, where $c(\underline{\psi})$ is the cutoff in Proposition 1. As $c(\underline{\psi}) < -1 - W(1/e)$ for every $\underline{\psi}$, writing $\rho_0 := 1 + W(1/e) \approx 1.278$, the ratio $\rho := h/(1-\psi^*) = -\varepsilon(0, l^*)$ satisfies $\rho > \rho_0$.

We wish to relate ω_0 to ψ^*. To this end, note that (18) splits each outcome probability into a belief-independent term plus a belief-dependent term: $\psi^* = \underline{P} + d \cdot \Phi(\omega_0)$ and $1 - \psi^* = (1 - \bar{P}) + d \cdot \Phi(-\omega_0)$. By Assumption 1, $\underline{P} > 0$ and $1 - \bar{P} > 0$, so each outcome probability strictly exceeds its belief-dependent term, i.e., $d \cdot \Phi(-\omega_0) < 1 - \psi^*$ and $d \cdot \Phi(\omega_0) < \psi^*$. Substituting $d = \rho(1-\psi^*)\omega_0/\varphi(\omega_0)$, which follows from $h = d \cdot \varphi(\omega_0)/\omega_0$, and using $\rho > \rho_0$, we obtain

$$M(\omega_0) := \frac{\omega_0 \Phi(-\omega_0)}{\varphi(\omega_0)} < \frac{1}{\rho} < \frac{1}{\rho_0}, \qquad \frac{1-\psi^*}{\psi^*} < \frac{\varphi(\omega_0)}{\rho_0 \omega_0 \Phi(\omega_0)}. \tag{19}$$

Step 1: $\omega_0 < \omega_\star$, where $\omega_\star \approx 1.549$ is the unique solution of $M(\omega_\star) = 1/\rho_0$. The map M is strictly increasing on $(0, \infty)$: its derivative equals $(1+\omega^2)\Phi(-\omega)/\varphi(\omega) - \omega$, which is positive by the Mills bound $\Phi(-\omega) > \omega\varphi(\omega)/(1+\omega^2)$. Rewriting the Mills bound as $M(\omega) > \omega^2/(1+\omega^2)$, we get $M(\omega) > 1/\rho_0$ whenever $\omega^2/(1+\omega^2) \geqslant 1/\rho_0$, i.e., whenever $\omega^2 \geqslant 1/(\rho_0 - 1)$. Since $M(0) = 0$, the solution $\omega_\star$ exists, is unique, and satisfies $\omega_\star^2 < 1/(\rho_0 - 1) < 4$, a bound we use in Step 2. The inequality $M(\omega_0) < 1/\rho_0$ then gives $\omega_0 < \omega_\star$.

Step 2: $\rho_0 \Phi(\omega)(e - \omega^2) \geqslant \omega\varphi(\omega)$ for every $\omega \in (0, \omega_\star)$. Write $v(\omega) := \Phi(\omega)(e - \omega^2)$. On $(0, 2)$, v is concave: $v''(\omega) = -2\Phi(\omega) - \omega\varphi(\omega)(e + 4 - \omega^2)$, and both terms are negative there since $\omega^2 < 4 < e + 4$. As $\omega_\star < 2$ by Step 1, v is concave on $(0, \omega_\star)$, so it is bounded below on this interval by the smaller of its endpoint values $v(0) = e/2$ and $v(\omega_\star)$. Thus, for every $\omega \in (0, \omega_\star)$, $\rho_0 v(\omega) \geqslant \min\{\rho_0 e/2, \rho_0 v(\omega_\star)\}$. Both values exceed $(2\pi e)^{-1/2} \approx 0.242$: indeed $\rho_0 e/2 > 1$, while $\rho_0 v(\omega_\star) \approx 0.382$. Finally, $\omega\varphi(\omega) \leqslant \varphi(1) = (2\pi e)^{-1/2}$ for every $\omega > 0$, as the derivative of $\omega\varphi(\omega)$ equals $\varphi(\omega)(1 - \omega^2)$.

Step 3: $\omega_0^2 < e\psi^*$. By the second inequality of (19) and then Step 2, applicable since $\omega_0 < \omega_\star$ by Step 1,

$$\frac{\omega_0^2}{\psi^*} = \omega_0^2 + \omega_0^2 \cdot \frac{1-\psi^*}{\psi^*} < \omega_0^2 + \frac{\omega_0 \varphi(\omega_0)}{\rho_0 \Phi(\omega_0)} \leqslant \omega_0^2 + (e - \omega_0^2) = e.$$

■

References

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

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on facebook. *Science*, *384*, eadk3451.

Aral, S., & Nicolaides, C. (2017). Exercise contagion in a global social network. *Nature Communications*, *8*, 14753.

Arieli, I., & Mueller-Frank, M. (2021). A general analysis of sequential social learning. *Mathematics of Operations Research*, *46*, 1235–1249.

Bala, V., & Goyal, S. (1995). A theory of learning with heterogeneous agents. *International Economic Review*, *36*, 303–323.

Banerjee, A. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, *107*, 797–817.

Banerjee, A., & Fudenberg, D. (2004). Word-of-mouth learning. *Games and Economic Behavior*, *46*, 1–22.

Bikhchandani, S., Hirshleifer, D., Tamuz, O., & Welch, I. (2024). Information cascades and social learning. *Journal of Economic Literature*, *62*, 1040–1093.

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

Callander, S., & Hörner, J. (2009). The wisdom of the minority. *Journal of Economic Theory*, *144*, 1421–1439.

Chamley, C. (2004). *Rational herds: Economic models of social learning*. Cambridge University Press.

Chandrasekhar, A. G., Duflo, E., Kremer, M., Pugliese, J. F., Robinson, J., & Schilbach, F. (2022). *Blue spoons: Sparking communication about appropriate technology use* (tech. rep.). National Bureau of Economic Research.

Che, Y.-K., & Hörner, J. (2018). Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, *133*, 871–925.

Ellison, G., & Fudenberg, D. (1995). Word-of-mouth communication and social learning. *The Quarterly Journal of Economics*, *110*, 93–125.

Fershtman, D. (2024). A robust wisdom of the crowd. *Available at SSRN 4982712*.

Foster, A. D., & Rosenzweig, M. R. (2010). Microeconomics of technology adoption. *Annual Review of Economics*, *2*, 395–424.

Kartik, N., Lee, S., Liu, T., & Rappoport, D. (2024). Beyond unbounded beliefs: How preferences and information interplay in social learning. *Econometrica*, *92*, 1033–1062.

Kremer, I., Mansour, Y., & Perry, M. (2014). Implementing the "wisdom of the crowd". *Journal of Political Economy*, *122*, 988–1012.

Kuchler, T., & Stroebel, J. (2021). Social finance. *Annual Review of Financial Economics*, *13*, 37–55.

McLennan, A. (1984). Price dispersion and incomplete learning in the long run. *Journal of Economic Dynamics and Control*, *7*, 331–347.

Munshi, K., & Myaux, J. (2006). Social norms and the fertility transition. *Journal of Development Economics*, *80*, 1–38.

Piketty, T. (1995). Social mobility and redistributive politics. *The Quarterly Journal of Economics*, *110*, 551–584.

Sgroi, D. (2002). Optimizing information in the herd: Guinea pigs, profits, and welfare. *Games and Economic Behavior*, *39*, 137–166.

Smith, L., & Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*, *68*, 371–398.

Smith, L., & Sørensen, P. (2020). Rational social learning with random sampling.

Song, Y. (2016). Social learning with endogenous observation. *Journal of Economic Theory*, *166*, 324–333.

Wolitzky, A. (2018). Learning from others' outcomes. *American Economic Review*, *108*, 2763–2801.

Xu, W. (2025). Social learning through coarse signals of others' actions. *Journal of Economic Theory*, *229*, 106066.

Online Appendix

B Allowing $\mathbb{P}_{LL} = \mathbb{P}_{HH}$ in Theorem 4

We introduce an additional regularity condition that allows $\mathbb{P}_{LL} = \mathbb{P}_{HH}$ in Theorem 4.

Definition 5. Signal distributions (F_n^L, F_n^H) that converge to perfect information are *strongly tail-regular* if they are tail-regular, and, for every sequence (p_n) with $\lim_{n\to\infty} F_n^L(p_n) = 1$ and $\lim_{n\to\infty} F_n^H(p_n) = 0$,

$$\lim_{n\to\infty} (1 - p_n) f_n^H(p_n) = 0. \tag{20}$$

By the no-introspection property $f_n^L(p)/f_n^H(p) = (1-p)/p$, (20) is equivalent to $\lim_{n\to\infty} p_n f_n^L(p_n) = 0$. Strong tail-regularity still appears to be a mild condition. It says that, if $F_n^H(p)$ is small, then $f_n^H(p)$ is much smaller than $1/(1-p)$.

Corollary 2. *In Theorem 4, the condition $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$ can be dropped if tail-regularity is strengthened to strong tail regularity.*

Proof. In the notation of the proof of Theorem 4, the assumption $\mathbb{P}_{LL} \neq \mathbb{P}_{HH}$ is used only in Step 3, and only to rule out the limit $(F_\infty^L, F_\infty^H) = (1, 0)$. Under convergence to no information, Step 5 shows that $F_\infty^L = F_\infty^H$, so $(F_\infty^L, F_\infty^H) = (1, 0)$ is impossible regardless, and Part 2 follows from Step 5 unchanged.

For Part 1, Steps 1 and 2 go through unchanged, and it suffices to show that $h(n) \to 0$ along any sub-subsequence where $(F_n^L(p^*(n)), F_n^H(p^*(n))) \to (F_\infty^L, F_\infty^H)$. If at least one of F_∞^L, F_∞^H lies in $(0, 1)$, Step 4 gives $h(n) \to 0$ via tail-regularity. In the remaining case, $(F_\infty^L, F_\infty^H) = (1, 0)$, strong tail-regularity gives $(1 - p^*(n)) f_n^H(p^*(n)) \to 0$, and substituting into (14) yields $h(n) \to 0$ since the bracketed term is bounded. ∎

We now verify that strong tail-regularity holds for location-scale families with strictly log-concave densities and unbounded score. We retain the notation of Section 4.3.

Proposition 4. *Assume the score function is unbounded and $\sigma_n \to 0$. Then (F_n^L, F_n^H) are strongly tail-regular.*

Proof. Convergence to perfect information and tail-regularity were established in Proposition 2. We verify (20).

Let (p_n) satisfy $F_n^L(p_n) \to 1$ and $F_n^H(p_n) \to 0$. Set $s_n := s_{\sigma_n}(p_n)$, $x_n := (s_n - 1)/\sigma_n$, and $y_n := s_n/\sigma_n$, so that $y_n - x_n = 1/\sigma_n$. From (8), $F_n^L(p_n) = G(y_n)$ and $F_n^H(p_n) = G(x_n)$.

Since G is a continuous cdf that is strictly increasing on $\mathbb{R}$, $G(y_n) \to 1$ and $G(x_n) \to 0$ imply $y_n \to \infty$ and $x_n \to -\infty$. As in the proof of Proposition 2, the identity $p_n(1-p_n) = g(x_n)g(y_n)/(g(x_n)+g(y_n))^2$ combined with (15) yields

$$(1-p_n)f_n^H(p_n) = \frac{g(x_n)+g(y_n)}{r(x_n)-r(y_n)}. \tag{21}$$

Since g is a log-concave density on $\mathbb{R}$, we have $g(u) \to 0$ as $|u| \to \infty$. Hence $g(x_n) \to 0$ and $g(y_n) \to 0$, so the numerator of (21) tends to 0. Since the score is unbounded above and below, $r(x_n) \to \infty$ and $r(y_n) \to -\infty$, so the denominator tends to $+\infty$. Therefore (21) tends to 0, establishing (20). ∎

C Logistic Noise

This appendix shows that the location-scale model with logistic noise is an example where private beliefs are bounded and tail-regularity fails, but the conclusion of Theorem 4 still holds. Thus, within the location scale model, unbounded beliefs and tail-regularity are sufficient for the conclusion of Theorem 4, but not necessary.

Consider the location-scale model and assume that noise is *standard logistic*,

$$g(x) = \frac{e^{-x}}{(1+e^{-x})^2}, \qquad G(x) = \frac{1}{1+e^{-x}}.$$

The score is $r(x) := g'(x)/g(x) = 1 - 2G(x)$, so $r'(x) = -2g(x) < 0$ and g is strictly log-concave. The score is bounded: $r(x) \in (-1,1)$ for all $x \in \mathbb{R}$. Since g is symmetric around 0 and satisfies $g(x) = G(x)(1-G(x))$, we have, for any $s \in \mathbb{R}$ and $\sigma > 0$,

$$\frac{g(s/\sigma)}{g((s-1)/\sigma)} = e^{-1/\sigma}\left(\frac{1+e^{-(s-1)/\sigma}}{1+e^{-s/\sigma}}\right)^2 \in (e^{-1/\sigma}, e^{1/\sigma}),$$

where the bounds follow from taking the limits $s \to \pm\infty$. Hence

$$p_\sigma(s) = \left(1 + \frac{g(s/\sigma)}{g((s-1)/\sigma)}\right)^{-1} \in \left(\frac{1}{1+e^{1/\sigma}}, 1 - \frac{1}{1+e^{1/\sigma}}\right).$$

That is, private beliefs are bounded for every $\sigma > 0$, though the support converges to $[0,1]$ as $\sigma \to 0$. In addition, tail-regularity also fails, as can be seen via the sequence $p^*(n)$ in Case A in the subsequent proof.

Proposition 5. *Assume that there is either match confounding or oppositional confounding. In the location-scale model with standard logistic noise, there exist* $0 < \sigma_1 \leqslant \sigma_2$ *such that learning robustly fails if* $\sigma < \sigma_1$ *and learning generically succeeds if* $\sigma > \sigma_2$*.*

Proof. That learning generically succeeds for large σ follows from Proposition 2: the standard logistic density is strictly log-concave, so (F_n^L, F_n^H) are center-regular when $\sigma_n \to \infty$. We focus on the small-σ direction, following the structure of the proof of Theorem 4.1 but replacing the appeal to tail-regularity by a direct computation in the logistic model. As there, write $d_L := \mathbb{P}_{LL} - \mathbb{P}_{HL}$ and $d_H := \mathbb{P}_{LH} - \mathbb{P}_{HH}$.

Let (σ_n) be any sequence with $\sigma_n \to 0$. We show that learning robustly fails for n sufficiently large. By Theorem 3 and Step 2 of the proof of Theorem 4, it suffices to exhibit a confounding point $l^*(n)$ with $\delta^*(n) := \Delta_l(1, l^*(n)) < 0$ satisfying the sufficient condition for stability $h(n) < 1 - \psi^*(n)$, where $h(n) := -l^*(n)\delta^*(n)$ and $\psi^*(n) := \psi(1|H, l^*(n))$.

Step 1 (existence of a confounding point with $\delta^* < 0$*).* By the argument in the proof of Proposition 2, (F_n^L, F_n^H) converges to perfect information. Step 1 of the proof of Theorem 4 relies only on convergence to perfect information, not on tail-regularity, and hence applies: under match confounding, for every n there is a unique confounding point, which satisfies $\delta^*(n) < 0$; under oppositional confounding, for n large (4) holds strictly, so there are two confounding points, exactly one of which satisfies $\delta^*(n) < 0$. In either case, fix a confounding point $l^*(n)$ with $\delta^*(n) < 0$ and let $p^*(n) = \bar{p}(l^*(n))$.

Step 2 (the possible limits). Since $(F_n^L(p^*(n)), F_n^H(p^*(n))) \in [0,1]^2$ for every n, any subsequence admits a sub-subsequence along which this pair converges to some $(F_\infty^L, F_\infty^H) \in [0,1]^2$. Passing to a further sub-subsequence if necessary, assume also that $p^*(n)$ converges. As in Step 4 of the proof of Theorem 4, it suffices to verify $h(n) < 1 - \psi^*(n)$ along every such sub-subsequence.

Note that (F_∞^L, F_∞^H) cannot have both coordinates in $(0,1)$, as this would simultaneously imply both $p^*(n) \to 0$ and $p^*(n) \to 1$. Combined with $F_\infty^L \geqslant F_\infty^H$ (stochastic dominance) and Step 3 of the proof of Theorem 4—where excluding $(0,0)$ and $(1,1)$ uses only Assumption 1—this leaves three cases:

- *Case A:* $F_\infty^L \in (0,1)$, $F_\infty^H = 0$, and $p^*(n) \to 0$.
- *Case B:* $F_\infty^L = 1$, $F_\infty^H \in (0,1)$, and $p^*(n) \to 1$.
- *Case C:* $(F_\infty^L, F_\infty^H) = (1,0)$. By Step 3 of the proof of Theorem 4, this case arises only if $\mathbb{P}_{LL} = \mathbb{P}_{HH}$, and hence only under match confounding.

Step 3 (computing the limits of $h(n)$ and $\psi^*(n)$). Let $s_n := s_{\sigma_n}(p^*(n))$ and $x_n := (s_n - 1)/\sigma_n$, so $F_n^H(p^*(n)) = G(x_n)$ and $F_n^L(p^*(n)) = G(x_n + 1/\sigma_n)$. Throughout, we use the logistic identities $g(u) = G(u)(1 - G(u))$ and $r(u) = 1 - 2G(u)$.

Case A. Here $F_n^H(p^*(n)) \to 0$ and $F_n^L(p^*(n)) \to F_\infty^L \in (0,1)$, so $x_n \to -\infty$ and $x_n + 1/\sigma_n \to \xi_A := G^{-1}(F_\infty^L) \in \mathbb{R}$. Thus, $g(x_n) \to 0$, $g(x_n + 1/\sigma_n) \to g(\xi_A) > 0$, $r(x_n) \to 1$, and $r(x_n + 1/\sigma_n) \to r(\xi_A)$, so by (16),

$$f_n^H(p^*(n)) = \frac{(g(x_n) + g(x_n + 1/\sigma_n))^2}{g(x_n + 1/\sigma_n)[r(x_n) - r(x_n + 1/\sigma_n)]} \to \frac{g(\xi_A)}{1 - r(\xi_A)} = \frac{1 - F_\infty^L}{2},$$

where the last equality uses the logistic identities $g(\xi_A) = F_\infty^L(1 - F_\infty^L)$ and $1 - r(\xi_A) = 2F_\infty^L$.

Passing to the limit in (13) and using $F_\infty^H = 0$ pins down $F_\infty^L d_L = \mathbb{P}_{HH} - \mathbb{P}_{HL}$. Then, by (14) and using $p^*(n) \to 0$,

$$h(n) = (1 - p^*(n)) f_n^H(p^*(n)) \left[(1 - p^*(n))d_L - p^*(n)d_H\right] \to \frac{1 - F_\infty^L}{2} \cdot d_L = \frac{\mathbb{P}_{LL} - \mathbb{P}_{HH}}{2},$$

where the last equality uses $(1 - F_\infty^L)d_L = d_L - (\mathbb{P}_{HH} - \mathbb{P}_{HL}) = \mathbb{P}_{LL} - \mathbb{P}_{HH}$. Also,

$$\psi^*(n) = F_n^L(p^*(n))\mathbb{P}_{LL} + (1 - F_n^L(p^*(n)))\mathbb{P}_{HL} \to \mathbb{P}_{HL} + F_\infty^L d_L = \mathbb{P}_{HH}.$$

Therefore

$$\lim_{n\to\infty} [1 - \psi^*(n) - h(n)] = 1 - \mathbb{P}_{HH} - \frac{\mathbb{P}_{LL} - \mathbb{P}_{HH}}{2} = 1 - \frac{\mathbb{P}_{LL} + \mathbb{P}_{HH}}{2} > 0,$$

since $\mathbb{P}_{LL}, \mathbb{P}_{HH} < 1$ by Assumption 1.1.

Case B. Here $F_n^L(p^*(n)) \to 1$ and $F_n^H(p^*(n)) \to F_\infty^H \in (0,1)$, so $x_n + 1/\sigma_n \to \infty$ and $x_n \to \xi_B := G^{-1}(F_\infty^H) \in \mathbb{R}$. Thus, $g(x_n) \to g(\xi_B) > 0$, $g(x_n + 1/\sigma_n) \to 0$, $r(x_n) \to r(\xi_B)$, and $r(x_n + 1/\sigma_n) \to -1$, so by (17),

$$f_n^L(p^*(n)) = \frac{(g(x_n) + g(x_n + 1/\sigma_n))^2}{g(x_n)[r(x_n) - r(x_n + 1/\sigma_n)]} \to \frac{g(\xi_B)}{r(\xi_B) + 1} = \frac{F_\infty^H}{2},$$

where the last equality uses the logistic identities $g(\xi_B) = F_\infty^H(1 - F_\infty^H)$ and $r(\xi_B) + 1 = 2(1 - F_\infty^H)$. Passing to the limit in (13) and using $F_\infty^L = 1$ pins down $F_\infty^H d_H = \mathbb{P}_{LL} - \mathbb{P}_{HH}$. Then, by the second line of (14) and using $p^*(n) \to 1$,

$$h(n) = p^*(n) f_n^L(p^*(n)) \left[(1 - p^*(n))d_L - p^*(n)d_H\right] \to \frac{F_\infty^H}{2} \cdot (-d_H) = \frac{\mathbb{P}_{HH} - \mathbb{P}_{LL}}{2}.$$

Also,

$$\psi^*(n) = F_n^H(p^*(n))\mathbb{P}_{LH} + (1 - F_n^H(p^*(n)))\mathbb{P}_{HH} \to \mathbb{P}_{HH} + F_\infty^H d_H = \mathbb{P}_{LL}.$$

Therefore

$$\lim_{n\to\infty} [1 - \psi^*(n) - h(n)] = 1 - \mathbb{P}_{LL} - \frac{\mathbb{P}_{HH} - \mathbb{P}_{LL}}{2} = 1 - \frac{\mathbb{P}_{LL} + \mathbb{P}_{HH}}{2} > 0.$$

Case C. Here $F_n^L(p^*(n)) \to 1$ and $F_n^H(p^*(n)) \to 0$, so $x_n \to -\infty$, $x_n + 1/\sigma_n \to \infty$, $r(x_n) \to 1$, and $r(x_n + 1/\sigma_n) \to -1$. Under $\mathbb{P}_{LL} = \mathbb{P}_{HH}$, the right-hand side of (13) equals d_L, so (13) rearranges to $(1 - F_n^L(p^*(n)))d_L = -F_n^H(p^*(n))d_H$, i.e., $1 - G(x_n + 1/\sigma_n) = \kappa G(x_n)$, where $\kappa := -d_H/d_L$. Recall that Case C can only arise under match confounding, which together with $\mathbb{P}_{LL} = \mathbb{P}_{HH}$ gives $d_L > 0$, $d_H < 0$, and hence $\kappa > 0$.

Set $a_n := g(x_n)$ and $b_n := g(x_n + 1/\sigma_n)$; both tend to 0. Using $g(u) = G(u)(1 - G(u))$,

$$\frac{a_n}{b_n} = \frac{G(x_n)(1 - G(x_n))}{G(x_n + 1/\sigma_n)(1 - G(x_n + 1/\sigma_n))} = \frac{G(x_n)(1 - G(x_n))}{G(x_n + 1/\sigma_n)\,\kappa\,G(x_n)} = \frac{1 - G(x_n)}{\kappa\,G(x_n + 1/\sigma_n)} \to \frac{1}{\kappa}.$$

Hence by (16),

$$f_n^H(p^*(n)) = \frac{(a_n + b_n)^2}{b_n[r(x_n) - r(x_n + 1/\sigma_n)]} = \frac{b_n(a_n/b_n + 1)^2}{r(x_n) - r(x_n + 1/\sigma_n)} \to 0,$$

since $b_n \to 0$ while $(a_n/b_n + 1)^2 \to (1/\kappa + 1)^2$ and $r(x_n) - r(x_n + 1/\sigma_n) \to 2$. Then by (14), $h(n) \to 0$ since $f_n^H(p^*(n)) \to 0$ and the other factors are bounded. Also,

$$\psi^*(n) = F_n^L(p^*(n))\mathbb{P}_{LL} + (1 - F_n^L(p^*(n)))\mathbb{P}_{HL} \to \mathbb{P}_{LL}.$$

Therefore $1 - \psi^*(n) - h(n) \to 1 - \mathbb{P}_{LL} > 0$.[34] ∎

D Discrete Signals

In this appendix, we replace the assumption that private beliefs are continuously distributed with the assumption that they take finitely many values. We show that learning generically succeeds in this discrete signal setting.

Let $K \geqslant 1$ and assume that private beliefs are supported on $\{p_1, \ldots, p_K\}$ with $0 < p_1 < \cdots < p_K < 1$ and state-contingent signal probabilities $\pi_k^\theta = \mathbb{P}(p = p_k|\theta)$. Since each p_k is the

[34] Since $\mathbb{P}_{LL} = \mathbb{P}_{HH}$, the limit equals $1 - (\mathbb{P}_{LL} + \mathbb{P}_{HH})/2$, matching Cases A and B. This is because in all three cases $\psi^*(n) \to \min\{\mathbb{P}_{LL}, \mathbb{P}_{HH}\}$ and $h(n) \to |\mathbb{P}_{HH} - \mathbb{P}_{LL}|/2$, so $\psi^*(n) + h(n) \to (\mathbb{P}_{LL} + \mathbb{P}_{HH})/2$.

posterior probability that the state is H given the corresponding signal, identifying signals with these beliefs implies the no-introspection property: $\pi_k^H/\pi_k^L = l_k := p_k/(1-p_k)$ for each k. We maintain Assumption 1.

In contrast to the continuous signal case, with discrete signals there are public beliefs where agents are indifferent between actions with positive probability: an agent facing public likelihood ratio l is indifferent with positive probability whenever $l = l_k$ for some k. Let $T = \{l_k\}_{k=1}^K$ denote the set of likelihood ratios where agents are indifferent with positive probability. There can now be multiple equilibria, depending on how agents break ties on T. However, $\psi(y|\theta, l)$ and $\phi(y, l)$ are still uniquely defined as in (2), for all $l \notin T$.

Theorem 5. *There exists an open and dense set $\mathcal{S} \subset [0,1]^4$ such that, for any $(\mathbb{P}_{a\theta}) \in \mathcal{S}$, there is a generic set of priors from which learning succeeds in any equilibrium.*

Proof. We first define the set $\mathcal{S}$. Any public likelihood ratio $l \notin T$ induces a strict best response, where signal p_k leads to action L iff $l_k < l$. Since $p_1 < \cdots < p_K$ implies $l_1 < \cdots < l_K$, the signals leading to L are $p_1, \ldots, p_{j(l)}$, for $j(l) = \max\{k : l > l_k\}$ (or $j(l) = 0$ if $l < l_1$). Hence,

$$\psi(1|\theta, l) = \sum_{i \leqslant j(l)} \pi_i^\theta \mathbb{P}_{L\theta} + \sum_{i > j(l)} \pi_i^\theta \mathbb{P}_{H\theta},$$

and the confounding condition $\mathbf{\Delta}(l) = 0$ is

$$\sum_{i \leqslant j} \pi_i^H \mathbb{P}_{LH} + \sum_{i > j} \pi_i^H \mathbb{P}_{HH} = \sum_{i \leqslant j} \pi_i^L \mathbb{P}_{LL} + \sum_{i > j} \pi_i^L \mathbb{P}_{HL}, \tag{22}$$

with $j = j(l)$. (For $j = 0$, (22) reduces to $\mathbb{P}_{HH} = \mathbb{P}_{HL}$, and for $j = K$ it reduces to $\mathbb{P}_{LH} = \mathbb{P}_{LL}$, both of which are ruled out by Assumption 1.) This is a linear equation in $(\mathbb{P}_{a\theta})$. Define $\mathcal{S} \subset [0,1]^4$ as the set of $(\mathbb{P}_{a\theta})$ for which (22) fails for every $j \in \{1, \ldots, K-1\}$. The complement of $\mathcal{S}$ is a finite union of hyperplanes in $[0,1]^4$, so $\mathcal{S}$ is open and dense.

Now fix any $(\mathbb{P}_{a\theta}) \in \mathcal{S}$. We show that, for a generic set of priors, the public likelihood ratio l_t cannot hit T in finite time in any equilibrium. To see this, on the set $\mathbb{R}_+ \backslash T$ where ψ and ϕ are well-defined independent of the choice of equilibrium, $\psi(y|\theta, l)$ is piecewise constant in l, so $\phi(y, l) = l \cdot \psi(y|L, l)/\psi(y|H, l)$ is piecewise linear and strictly increasing in l on each piece. In particular, for each $y \in \{0, 1\}$ and each $l' \in \mathbb{R}_+$, the set $\{l \in \mathbb{R}_+ \backslash T : \phi(y, l) = l'\}$ is finite, and hence so is the set $T \cup \{l \in \mathbb{R}_+ \backslash T : \phi(y, l) = l' \text{ for some } y \in \{0, 1\}\}$, which contains all likelihood ratios from which l' can be reached in one period in any equilibrium. Now, define the sets $T_t \subset \mathbb{R}_+$ by $T_0 = T$ and, inductively, $T_t = T \cup \{l \in \mathbb{R}_+ \backslash T : \phi(y, l) \in T_{t-1} \text{ for some } y \in \{0, 1\}\}$. Clearly, T_t contains all likelihood ratios from which T can be reached in t periods in any equilibrium. Moreover, since T is finite, each T_t is finite as a finite union of finite sets,

so $\bigcup_t T_t$ is countable. Hence, its complement is a generic set of priors from which l_t cannot hit T in finite time in any equilibrium.

Finally, learning succeeds in any equilibrium for any such prior, by a standard argument. Indeed, fix a prior $l_0 \in \mathbb{R}_+ \setminus \bigcup_t T_t$, assume without loss that $\theta = H$, and let $L \subset \mathbb{R}_+ \setminus T$ be the countable set of public likelihood ratios that can be reached in finite time starting from l_0. Then $\bigcup_{l \in L} \{\mathbf{\Delta}(l)\}$ is a finite set of strictly positive numbers (since $\mathbf{\Delta}(l)$ takes at most $K+1$ possible values on $\mathbb{R}_+ \setminus T \supseteq L$, which are all strictly positive since $(\mathbb{P}_{a\theta}) \in \mathcal{S}$), so there exists $m > 0$ such that $\mathbf{\Delta}(l) \geqslant m$ for all $l \in L$. Now fix any $l^* > 0$ and let $\delta = l^*/2$. For any $l \in (l^* - \delta, l^* + \delta) \cap L$ and any $y \in \{0, 1\}$,

$$|\phi(y, l) - l| = l \cdot \frac{\mathbf{\Delta}(l)}{\psi(y|H, l)} \geqslant (l^*/2) \cdot m.$$

Moreover, $\psi(y|H, l) \in (0, 1)$ also takes finitely many values on L, so it is bounded away from 0 and from 1. Hence, the argument in Theorem B.1 in Smith and Sørensen (2000) applies with $I = (l^* - \delta, l^* + \delta)$, giving $l^* \notin \text{supp}(l_\infty)$.[35] Since $l^* > 0$ was arbitrary, $\text{supp}(l_\infty) = \{0\}$, so $l_t \to 0$ almost surely and learning succeeds. ■

Remark 3. A special case of discrete signals arises when private signals are completely uninformative. This is the case $K = 1$, where the no-introspection property combined with $\pi_1^H = \pi_1^L = 1$ forces $p_1 = 1/2$, so the single signal is uninformative. Theorem 5 applies, and in this case requires no further restriction on $(\mathbb{P}_{a\theta})$: as the proof shows, $\mathcal{S}$ is defined as the complement of a finite union of hyperplanes indexed by $\{1, \ldots, K-1\}$, which is empty when $K = 1$, so $\mathcal{S} = [0, 1]^4$. Hence learning generically succeeds in any equilibrium. This is consistent with Theorem 4.2, which shows that learning generically succeeds as private signals become uninformative; the same holds in the limiting case where private signals carry no information at all.

[35] The theorem does not apply directly, as it requires that $|\phi(y, l) - l|$ is bounded away from zero for every $l \in I$, whereas we establish this only for $l \in I \cap L$. But since the process remains in L in every period, their proof applies verbatim.