Misspecified Beliefs about Time Lags[*]

Yingkai Li   Harry Pei

December 14, 2020

Abstract

We examine the long-term behavior of a Bayesian agent who has a misspecified belief about the time lag between actions and feedback, and learns about the payoff consequences of his actions over time. Misspecified beliefs about time lags result in attribution errors, which have no long-term effect when the agent's action converges, but can lead to arbitrarily large long-term inefficiencies when his action cycles. Our proof uses concentration inequalities to bound the frequency of action switches, which are useful to study learning problems with history dependence. We apply our methods to study a policy choice game between a policy-maker who has a correctly specified belief about the time lag and the public who has a misspecified belief.

Keywords: time lag, misspecified belief, Bayesian learning, action cycles, history dependence, concentration inequality.

1 Introduction

We study learning problems faced by Bayesian decision makers who have misspecified beliefs about the time lag between decisions and feedback. We examine the long-term consequences of such belief misspecifications both in single-agent decision-making problems and in games of collective decision-making.

Misperception about time lags is prevalent among decision makers at various levels, ranging from leaders in organizations to ordinary citizens. For example, a manager decides how much resource to allocate to R&D. Unlike efforts on production and sales, investments in R&D are unlikely to pay off in the short run, and moreover, it is usually unclear when and whether they will pay off. Repenning and Sterman (2002) show that these time lags hinder an organization's learning about the optimal resource allocation by "*complicating the attribution of causality between actions and results*". Rahmandad, Repenning, and Sterman (2009) point out that what slows down organizational learning is not the delay per se, but instead, people's misperceptions about the delay. Consequences of such misperceptions include the so-called *capability traps* (Repenning and Sterman, 2002), in which members of an organization work hard on production at the expense of cutting back on the time allocated to R&D and maintenance, that ultimately results in low productivity.

Similarly, fans of football clubs tend to credit or blame their current managers for their team's performances while ignoring the effects of previous managers' decisions. Many people believe that reopening the economy is safe amidst the COVID-19 pandemic when the number of cases and hospitalizations in Georgia, Florida, and Arizona went down three weeks after these states' reopenings.[1] However, the number of cases and hospitalizations started to surge six to eight weeks after these states' reopennings, forcing some of them to partially return to lockdown.

We propose a model that incorporates such misperceptions. In every period, an agent chooses an action and observes an outcome that determines his payoff. The agent faces uncertainty about the *state*, i.e., the mapping from his actions to the outcome distributions. He observes the history of actions and outcomes and updates his belief according to Bayes rule. We assume that the true state belongs to the support of the agent's prior belief and that the outcome is informative about the state regardless of the agent's action.

The outcome distribution in period t depends only on the agent's action in period $t - k^*$ while the agent believes that it depends on his action in period $t - k'$, where k' is different from k^*.[2] Our formulation can

---

[1] The state of Georgia reopened in late April, and on May 23rd, Governor Brian Kemp shared the news that hospitalizations are down by 30% since the state reopened. Similar patterns arise after the reopening of Florida, Texas and Arizona in early May. However, the number of cases and hospitalizations in these states started to surge from late June to July. See https://www.latimes.com/world-nation/story/2020-05-23/georgia-reopened-first-the-data-say-whatever-you-want-them-to and https://www.cnbc.com/2020/06/29/more-states-reverse-or-slow-reopening-plans-as-coronavirus-cases-climb.html

[2] Section 5 extends our result to situations where (1) the outcome distribution depends on a weighted average of the agent's current and past actions, or (2) the agent faces uncertainty about the time lag and learns about it over time.

capture, for example, an individual underestimates the time it takes for workouts to have effects on fitness, a policy-maker underestimates or overestimates the time it takes for a curriculum reform to have effects on students' academic achievements, and so on.

This novel form of belief misspecification interferes learning through an *attribution error*, which has no long-term effect when the agent's action converges but can lead to mislearning when the agent's action changes over time. Theorem 1 shows that the mislearning caused by attribution errors can lead to arbitrarily large long-term inefficiencies in the sense that for an open set of states, there exist prior beliefs that include the true state in their support such that the asymptotic frequency with which the agent takes his optimal action is arbitrarily close to zero. This stands in contrast to the benchmark scenario with a correctly specified belief about the time lag, in which the agent chooses his optimal action almost surely in the long run.

The first challenge in establishing this result stems from the fact that our learning problem exhibits nontrivial history-dependence. This is because the agent's current-period action directly affects his future observations. The second challenge arises from the observation that the agent's action cannot converge to anything suboptimal and inefficiencies can only arise when the agent's action cycles in the long run. As a result, one needs to bound the frequency of action switches in order to quantify the amount of mislearning, which is a key step toward showing that the posterior probability of the true state is low in the long run.

We develop a new technique using concentration inequalities. First, we examine an auxiliary problem in which the true state is excluded from the agent's prior belief. We use the Chernoff-Hoeffding inequality to show that in expectation, the agent switches actions within a finite number of periods. We then establish a concentration inequality on unbounded random variables in order to bound the frequency of action switches. Next, we study situations in which the true state occurs with small but positive probability. We use the Azuma-Hoeffding inequality to show that due to the mislearning caused by frequent action switches, the true state occurs with low probability in the agent's posterior for all periods. This explains why the agent's actions cycle over time even when the true state belongs to the support of his prior belief.

We apply our framework to study a dynamic policy choice game between a policy-maker who has a correctly specified belief about the time lag and the public who has a misspecified belief. The policy-maker wants to implement a socially beneficial reform but cannot do so without the public's support. The public prefers the reform to the status quo in one state and prefers the status quo in the other state. This conflict of interest can arise when the reform has positive externalities or the status quo has negative externalities on marginalized groups (e.g., massive gatherings during a pandemic has negative externalities on the immuno-compromised) which the policy-maker cares about but the majority of citizens fail to internalize. Therefore, a reform can be both optimal for the benevolent policy-maker and suboptimal for the majority of citizens.

We characterize the maximal frequency that the policy-maker can implement the reform when he has no private information about the state. We show that the policy-maker's optimal payoff equals the maximal frequency of reform in an auxiliary game where he knows the state but the public is naive in the sense that they fail to recognize the informational content of the policy-maker's behaviors. Intuitively, this is because the policy-maker can asymptotically learn the true state, and when the reform is optimal for the public, he can implement the reform in almost every period regardless of the public's prior.

We also construct a class of strategies under which the policy-maker can approximately achieve his optimal payoff, according to which he proposes the reform with frequency close to a half when the public entertains a pessimistic belief about the reform, and proposes the reform with frequency strictly greater than a half when the public entertains an optimistic belief about the reform. The former maximizes the amount of mislearning and the latter maximizes the frequency of reform subject to a constraint that the expected amount of mislearning is non-negative. The key step is to use the Wald inequality and show that conditional on the reform being suboptimal for the public, the policy-maker's future proposals are accepted with probability close to 1 when he started to propose the reform with frequency greater than a half.

Our work contributes to a growing literature on misspecified learning by studying environments with history dependence. The agent in our model has a misspecified belief about the dynamic structure of the problem and his past actions can affect future outcomes. This stands in contrast to most of the existing works such as Berk (1966), Nyarko (1991), Esponda and Pouzo (2016), Fudenberg, Romanyuk, and Strack (2017), Bohren and Hauser (2020), Frick, Iijima, and Ishii (2020), Esponda, Pouzo, and Yamamoto (2020), and Fudenberg, Lanzani, and Strack (2020) that exclude history dependence.

Several recent papers study misspecifed learning models with history dependence and provide conditions for the steady states. Shalizi (2009) provides sufficient conditions for the convergence of posterior belief when there is no endogenous action choice and the signals in different periods can be correlated. He (2020) examines misspecified learning in two-period optimal stopping problems in which an agent mistakenly believes that the outcome in the second period is negatively correlated with that in the first period. Esponda and Pouzo (2020) study a single-agent Markov decision problem with misspecified beliefs about the state transition function. Molavi (2020) examines a dynamic general equilibrium model in which an agent's choice in the current period affects the constraints he face in the future. By contrast, we focus on the dynamics of an agent's behavior in history-dependent learning problems instead of the steady states. We show that the long-run outcome can be inefficient by bounding the frequency of action switches.[3]

---

[3]Esponda, Pouzo, and Yamamoto (2020) introduce stochastic approximation techniques and characterize the frequency of the agent's actions in misspecified learning problems without history-dependence.

The attribution error in our model is related to Eliaz and Spiegler (2020), who study an agent's long-term behavior when he updates his belief according to a misspecified causal model. They propose a solution concept that characterizes the steady states of the above learning process, rather than examining the dynamics of actions and beliefs. Spiegler (2013) examines the dynamic interaction between an agent and a sequence of principals, each of them acts only once and chooses whether to intervene. The agent attributes changes of a state variable to the latest intervention, which is applicable when some of the principal's actions (intervention) are more salient than others (no intervention). Jehiel and Samuelson (2012) characterize an informed long-run player's payoff and behavior when he faces a sequence of short-run players who mistakenly believe that all types of the long-run player use stationary strategies. By contrast, we study a different type of attribution error, where the agent has wrong beliefs about the delay between actions and feedback.

2   Model

Time is discrete, indexed by $t = 1, 2, \ldots$ In period t, a Bayesian agent chooses an action $a_t \in A$, and then observes an outcome $y_t \in Y$. We assume that both A and Y are finite sets.

Our modeling innovation is to introduce time lags between decisions and feedback as well as the agent's misperception about the time lag. Formally, there exist two non-negative integers $k^*, k' \in \mathbb{N}$ with $k^* \neq k'$, such that the distribution of y_t depends only on a_{t-k^*}, while the agent believes that it depends on $a_{t-k'}$.

The agent faces uncertainty about the distribution over outcomes (which we call the *state*) and learns about it over time by observing the history of actions and outcomes. A typical state is denoted by $F \equiv \{F(\cdot|a)\}_{a \in A}$, with $F(\cdot|a) \in \Delta(Y)$. Let $F^* \equiv \{F^*(\cdot|a)\}_{a \in A}$ be the *true state*, namely, y_t is distributed according to $F^*(\cdot|a_{t-k^*})$. The agent's prior belief about the state is $\pi_0 \in \Delta(\mathscr{Y})$, with $\mathscr{Y} \equiv \left(\Delta(Y)\right)^A$ and $\text{supp}(\pi_0)$ is finite.[4] After the agent learns that the state is F, he believes that y_t is distributed according to $F(\cdot|a_{t-k'})$. The agent observes $h^t \equiv \{\ldots, a_{-1}, a_0, a_1, \ldots, a_{t-1}, y_1, \ldots, y_{t-1}\}$ in period t and his posterior belief is denoted by $\pi_t \in \Delta(\mathscr{Y})$. All the actions before period 1 are exogenously given. In order to focus on misspecified belief about the time lag, we focus on prior beliefs that are *regular*:

Regular Prior Belief. π_0 *is regular with respect to F^* if*

1. $F^* \in \text{supp}(\pi_0)$, *and for every $F \in \text{supp}(\pi_0)$ and $a \in A$, $F(\cdot|a)$ has full support.*

2. *for every $F, F' \in \text{supp}(\pi_0)$ and $a \in A$, we have $F(\cdot|a) \neq F'(\cdot|a)$.*

---

[4]When there are infinitely many states, Diaconis and Freedman (1986) and Shalizi (2009) show that the agent's posterior belief may not converge to the true state even when the true state belongs to the support of his prior belief. We abstract away from this complication in order to focus on the economic implications of misspecified beliefs about time lags.

The first part requires that the true state F^* belongs to the support of the agent's prior belief and that the agent cannot rule out any state no matter which action he takes and which outcome he observes. This rules out canonical forms of belief misspecifications studied by Berk (1966), Nyarko (1991), and Esponda and Pouzo (2016) in which F^* is excluded from the agent's prior belief. The second part requires that the observed outcome is informative about the state regardless of the agent's action, which is satisfied for generic finite subsets of $\mathscr{Y}$. It rules out lack-of-identification problems, such as safe-arms in bandit models.

The agent's stage-game payoff is $v(y_t)$. We assume that $\arg\max_{a\in A}\{\sum_{y\in Y} v(y)F^*(y|a)\}$ is a singleton, and its unique element is denoted by a^*, i.e., the agent has a unique optimal action under the true state. This is satisfied for generic $F^* \in \mathscr{Y}$ and $v : Y \to \mathbb{R}$ given that A and Y are finite sets. The agent's strategy is $\sigma : \mathscr{H} \to \Delta(A)$, where $\mathscr{H}$ is the set of histories. Strategy σ is optimal if $\sigma(h^t)$ maximizes the expected value of $\sum_{s=0}^{+\infty} \delta^s v(y_{t+s})$ at every h^t, where $\delta \in [0,1)$ is the agent's discount factor. We focus on settings such that either $\delta \in (0,1)$ or $(\delta, k') = (0,0)$. This is because the agent is indifferent between all actions when $\delta = 0$ and $k' \geq 1$. Let $\Sigma^*(\pi_0)$ be the set of strategies that are optimal for the agent when his prior is π_0.

For some useful benchmarks, the agent chooses a^* in every period after he learns that the true state is F^* even if he entertains a misspecified belief about the time lag. If there is no belief misspecification, i.e., $k^* = k'$, then according to Berk's Theorem (Berk, 1966), the agent's action converges to a^* almost surely.

Remark: Our baseline model focuses on situations in which the outcome in every period is affected only by one of the agent's actions. Section 5 discusses extensions where the outcome in period t depends on a convex combination of the agent's past and current-period actions, and the agent has a wrong belief about the weights of different actions. We also consider settings in which the agent faces uncertainty about the time lag and learns about it over time, but the support of his prior belief excludes the true time lag.

The agent's payoff in our baseline model depends only on the observed outcome. Our results extend when the agent's payoff also depends on the state. When the agent's payoff depends directly on his action (e.g., different actions have different costs), his action can be suboptimal even when he learns the true state. This is because when $k^* \neq k'$, the agent either overestimates or underestimates the time it takes for his action to have an effect, which can lead to suboptimal decisions since the agent discounts future payoffs. Our results extend to settings where the agent's payoff is $v(y_t) - c(a_t)$, as long as the absolute value of $c(\cdot)$ is small enough such that the agent has a strict incentive to choose a^* after he learns the true state.

3   Result

First, we show that if the agent's action converges, then it can only converge to his optimal action. Moreover, the asymptotic frequency of his optimal action must be strictly positive when his prior belief is regular.

Lemma 3.1. *Suppose either* $\delta \in (0,1)$ *or* $(\delta, k') = (0,0)$*. If* π_0 *is regular with respect to* F^* *and* a_t *converges to* a *with positive probability, then* $a = a^*$*. Furthermore,*

$$\liminf_{t \to +\infty} \mathbb{E}^\sigma \left[\frac{1}{t} \sum_{s=1}^{t} \mathbf{1}\{a_s = a^*\} \right] > 0 \text{ for every } \sigma \in \Sigma^*(\pi_0). \tag{3.1}$$

The proof is in Appendix B. Intuitively, the only way in which misspecified beliefs about time lags can interfere learning is through an *attribution error*, namely, the agent attributes the effects of a_{t-k^*} to $a_{t-k'}$. When the agent's action converges, a_{t-k^*} and $a_{t-k'}$ are the same so the attribution error does not affect his learning. Since $F^* \in \text{supp}(\pi_0)$ and there is no lack-of identification problem, the agent will learn the true state almost surely. This implies that the agent's actions are asymptotically efficient, which contradicts the presumption that his action converges to something other than a^*. The agent takes his optimal action with positive asymptotic frequency since for every $\varepsilon > 0$, the following event occurs with positive probability:

$$\mathscr{E}^\varepsilon \equiv \{\text{there exists } T \in \mathbb{N} \text{ such that } \pi_t(F^*) > 1 - \varepsilon \text{ for every } t \geq T\}. \tag{3.2}$$

Intuitively, this is because y is informative about the state, so there always exists a signal realization that increases the posterior probability of F^*. The probability that the agent chooses a^* in all future periods is strictly positive when the posterior probability of F^* is close to 1, which implies (3.1).

Despite the agent's action cannot converge to anything other than a^*, the attribution errors caused by misspecified beliefs can lead to arbitrarily large long-term inefficiencies. Theorem 1 shows that the frequency with which the agent takes his optimal action can be arbitrarily close to 0.

Theorem 1. *Suppose either* $\delta \in (0,1)$ *or* $(\delta, k') = (0,0)$*. For every* $\gamma > 0$*, there exists an open set* $\mathscr{Y}^o \subset \mathscr{Y}$ *such that for every* $F^* \in \mathscr{Y}^o$*, there is a prior belief* π_0 *that is regular with respect to* F^* *under which*

$$\limsup_{t \to +\infty} \mathbb{E}^\sigma \left[\frac{1}{t} \sum_{s=1}^{t} \mathbf{1}\{a_s = a^*\} \right] < \gamma \quad \text{for every} \quad \sigma \in \Sigma^*(\pi_0). \tag{3.3}$$

Theorem 1 implies that Bayesian agents fail to take their optimal action even when the true state belongs to the support of their prior belief and the observed outcome can statistically identify the state. Intuitively, this is because attribution errors lead to *mislearning* when the agent switches actions. In particular, when the

agent's action changes over time, the probability of the true state can decrease in expectation. When action switches are frequent enough, the amount of mislearning outweighs the what the agent learns when he takes the same action in consecutive periods. As a result, his posterior belief may attach a low probability to F^* in all periods. Under some F^* and π_0 that is regular with respect to F^*, such an event occurs with probability arbitrarily close to 1, which leads to arbitrarily large asymptotic inefficiencies.

The proof is in Appendix C. We explain the logic behind our argument using an example, which illustrates how attribution errors lead to action cycles in the long run and how to bound the frequency of action switches using concentration inequalities.

Illustrative Example: Suppose $A \equiv \{0,1\}$, $\delta = 0$, $k^* = 1$, and $k' = 0$. That is, the agent is myopic, the distribution of y_t depends only on the agent's action in period $t-1$ while the agent believes that y_t is affected by his action in period t. Let $Y \equiv \{y_0, y_1, y_2\}$, $v(y_0) = v(y_1) = 0$, $v(y_2) = 1$, and the support of π_0 is $\{F^*, F_0, F_1\}$, with

$$F^*(y_i|a=0) = \begin{cases} \varepsilon & i=0 \\ 1-3\varepsilon & i=1 \\ 2\varepsilon & i=2 \end{cases} \qquad F^*(y_i|a=1) = \begin{cases} 1-2\varepsilon & i=0 \\ \varepsilon & i=1 \\ \varepsilon & i=2 \end{cases}$$

$$F_0(y_i|a=0) = \begin{cases} 2/3-\varepsilon & i=0 \\ 1/3-\varepsilon & i=1 \\ 2\varepsilon & i=2 \end{cases} \qquad F_0(y_i|a=1) = \begin{cases} 2/3-\varepsilon/2 & i=0 \\ 1/3-\varepsilon/2 & i=1 \\ \varepsilon & i=2 \end{cases}$$

$$F_1(y_i|a=0) = \begin{cases} 1/3-\varepsilon/2 & i=0 \\ 2/3-\varepsilon/2 & i=1 \\ \varepsilon & i=2 \end{cases} \qquad F_1(y_i|a=1) = \begin{cases} 1/3-\varepsilon & i=0 \\ 2/3-\varepsilon & i=1 \\ 2\varepsilon & i=2 \end{cases}$$

These distributions are depicted in Figure 1. One can verify that the optimal actions in states F^* and F_0 are both 0, and the optimal action in state F_1 is 1.

Since the true state F^* belongs to the support of π_0 and y can statistically identify the state, the agent's action converges to 0 almost surely when he has a correctly specified belief about the time lag. However,
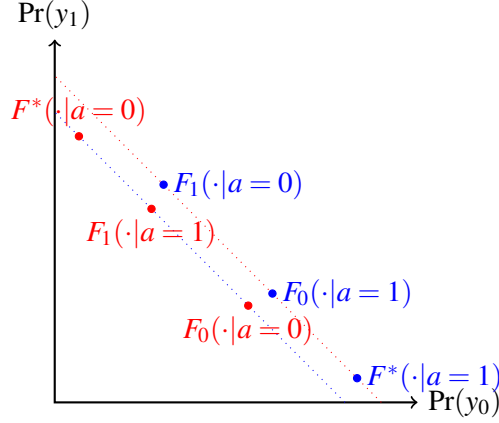
Figure 1: Example with Action Cycles: Signal Distributions under F^*, F_0 and F_1

when the agent has a misspecified belief about the time lag, we sketch an argument which shows that the asymptotic frequency of the suboptimal action can be arbitrarily close to $1/2$.

Claim 1. *For every $\eta > 0$, there exists $\overline{\varepsilon} > 0$ such that when $\varepsilon < \overline{\varepsilon}$, there exists a full support π_0 under which*

$$\limsup_{t \to +\infty} \frac{1}{t} \mathbb{E}^\sigma \left[\sum_{s=1}^t \mathbf{1}\{a_s = a^*\} \right] < \frac{1}{2} + \eta \quad \text{for every} \quad \sigma \in \Sigma^*(\pi_0).$$

Our argument proceeds in two steps. First, we examine an auxiliary learning problem where F^* occurs with zero probability. Let $l_t \equiv \log \frac{\pi_t(F_0)}{\pi_t(F_1)}$. By definition, the agent has a strict incentive to take action 0 when $l_t > 0$, and has a strict incentive to take action 1 when $l_t < 0$.

A useful observation from Figure 1 is that when the agent takes action 0, F_1 is closer to F^* compared to F_0. As a result, the log likelihood ratio l_t decreases in expectation when the agent chooses action 0 in two consecutive periods. Let τ_0 be the number of periods with which the agent's action switches back to action 1 when $a_{t-2} = 1$ and $a_{t-1} = a_t = 0$. The Chernoff-Hoeffding inequality implies that:

$$\Pr(\tau_0 \geq s) \leq \Pr(l_{t+s} \geq 0) \leq \exp\left(-2s \left(\frac{l_t}{s} + \underbrace{\mathbb{E}[l_s - l_{s-1}]}_{<0} \right)^2 \right), \tag{3.4}$$

from which we know that the distribution of τ_0 is first order stochastically dominated by an exponential distribution and therefore, has bounded first and second moments.

Similarly, when the agent takes action 1, F_0 is closer to F^* compared to F_1. As a result, l_t increases in expectation when the agent takes action 1 in two consecutive periods. Let τ_1 be the number of periods with which the agent's action switches back to action 0 when $a_{t-2} = 0$ and $a_{t-1} = a_t = 1$. A similar argument

based on the Chernoff-Hoeffding inequality implies that τ_1 has bounded first and second moments.

In order to bound the frequency of action switches from below using the above conclusions on τ_0 and τ_1, we establish a concentration inequality that applies to unbounded random variables (Lemma A.3).[5] This inequality implies that for every $\varepsilon > 0$, there exists a large enough $T \in \mathbb{N}$ such that

$$\mathbb{E}\left[\frac{\#\{t \leq T | a_{t-1} = a_t = 1\}}{\#\{t \leq T | a_{t-1} = 1, a_t = 0\}}\right] = \mathbb{E}\left[\frac{\#\{t \leq T | a_{t-1} = a_t = 1\}}{\#\{t \leq T | a_{t-1} = 0, a_t = 1\}}\right] \leq \mathbb{E}[\tau_1] + \varepsilon,$$

and

$$\mathbb{E}\left[\frac{\#\{t \leq T | a_{t-1} = a_t = 0\}}{\#\{t \leq T | a_{t-1} = 1, a_t = 0\}}\right] = \mathbb{E}\left[\frac{\#\{t \leq T | a_{t-1} = a_t = 0\}}{\#\{t \leq T | a_{t-1} = 0, a_t = 1\}}\right] \leq \mathbb{E}[\tau_0] + \varepsilon.$$

When $\pi_0(F_1)$ and $\pi_0(F_0)$ are close, the expectations of τ_0 and τ_1 are close, and therefore, the asymptotic frequencies of both actions are close to $1/2$. The above inequalities imply that the asymptotic frequency of action switches is strictly positive and is close to $\frac{1}{1+\mathbb{E}[\tau_0]}$ and $\frac{1}{1+\mathbb{E}[\tau_1]}$.

Next, we consider the case in which the true state F^* belongs to the support of π_0 but occurs with low probability. We show that with probability close to 1, the agent's posterior belief attaches a low probability to F^* in all periods. Formally, for every $\eta > 0$, there exists $\overline{\pi} > 0$ such that when the prior probability of F^* is less than $\overline{\pi}$, the probability of the event that

$$\max\left\{\frac{\pi_t(F^*)}{\pi_t(F_0)}, \frac{\pi_t(F^*)}{\pi_t(F_1)}\right\} < \eta \text{ for all } t \in \mathbb{N} \tag{3.5}$$

is at least $1 - \eta$. Intuitively, both $\log \frac{\pi_t(F^*)}{\pi_t(F_0)}$ and $\log \frac{\pi_t(F^*)}{\pi_t(F_1)}$ increase in expectation when $a_t = a_{t-1}$ since F^* is the true state. However, as can be seen from Figure 1, $F^*(\cdot | a = 0)$ is further away from $F^*(\cdot | a = 1)$ compared to both $F_0(\cdot | a = 0)$ and $F_1(\cdot | a = 0)$, and $F^*(\cdot | a = 1)$ is further away from $F^*(\cdot | a = 0)$ compared to both $F_0(\cdot | a = 1)$ and $F_1(\cdot | a = 1)$. Due to the attribution errors caused by misspecified beliefs about the time lag, both log likelihood ratios *decrease* in expectation when $a_t \neq a_{t-1}$.

When $\pi_t(F^*)$ is low, the agent's best reply problem is similar to the one he faces in the auxiliary scenario where F^* is excluded from the support of his prior, in which case he frequently switches actions. Those action switches together with the attribution error lead to mislearning. When $\mathbb{E}[\tau_0]$ and $\mathbb{E}[\tau_1]$ are sufficiently small, action switches are frequent, so the mislearning caused by attribution errors outweighs what the agent can learn when he takes the same action in adjacent periods.

In order to formalize this intuition, we provide a lower bound on the probability of event (3.5) using

---

[5]The Chernoff-Hoeffding inequality only applies to bounded random variables. Corollary 5.5 in Lattimore and Szepesvári (2020) and Jin et al. (2019) establish concentration inequalities for random variables with sub-Gaussian distributions. By contrast, our result is more general since it only requires the random variable to have bounded first and second moments.

concentration inequalities. The Chernoff-Hoeffding inequality does not apply since the agent's belief affects his actions, so the log likelihood ratios between F^* and F_0, and between F^* and F_1 can exhibit serial correlations. We overcome this challenge by constructing a martingale process with bounded increments from the log likelihood ratios and then applying the Azuma-Hoeffding inequality. We show that with probability close to 1, the agent's belief attaches low probability to F^* in all periods. This together with the frequent action switches explains why he takes the inefficient action with positive asymptotic frequency and his actions cycle over time.

Remark: In our example, the asymptotic frequency with which the agent takes the inefficient action is close to $1/2$. In Appendix B, we construct F^* as well as regular prior beliefs with respect to F^* such that the frequency of the inefficient action is close to 1. In this example, one can simply modify the outcome distributions such that both τ_1 and τ_0 have low expectations, but the expectation of τ_1 is significantly greater than the expectation of τ_0.

4 Application: Dynamic Policy Choice Game

In order to demonstrate the applicability of our techniques to bound the frequency of action switches, we analyze a dynamic policy choice game between

- a principal who strategically makes policy proposals, learns about the state over time, and has a correctly specified belief about the time lag between the chosen policy and the observed feedback,

- a Bayesian agent who can veto the principal's proposals, and learn about the outcome distribution under a misspecified belief about the time lag.[6]

In every period, a society needs to make a collective choice between two policies $a_t \in A \equiv \{0, 1\}$. In period t, the principal makes a proposal $\widetilde{a}_t \in A$. If $\widetilde{a}_t = 0$, then action 0 is automatically implemented, i.e., $a_t = 0$. If $\widetilde{a}_t = 1$, then the agent chooses whether to accept ($a_t = 1$) or veto ($a_t = 0$) the principal's proposal.

Both the principal and the agent face uncertainty about the *state*, which is contained in $\mathscr{F} \equiv \{F_0, F_1\}$. Their common prior belief is $\pi_0 \in \Delta(\mathscr{F})$, which we assume has full support.[7]

The principal and the agent agree to disagree in terms of the time lag between decisions and feedback. The principal has a correctly specified belief about the time lag and knows that the distribution of y_t depends

---

[6]Our analysis also applies to a sequence of myopic agents, each plays the game only once.

[7]Extensions to environments with more than two states are available upon request. Our results also apply when the principal and the agent agree to disagree about the state distribution. For example, when the principal's prior belief is p_0 that is different from π_0, one needs to replace π_0 by p_0 in RHS of (4.6).

only on a_{t-k^*}, with $k^* \geq 1$. The agent believes that the distribution of y_t depends on a_t, i.e., $k' = 0$. In period t, both players observe $h^t \equiv \{\widetilde{a}_s, a_s, y_s\}_{s=0}^{t-1}$ and update their beliefs about the state according to Bayes rule. Let $\mathscr{H}$ be the set of histories. Let $\sigma_p : \mathscr{H} \to [0,1]$ be the principal's strategy, which maps the histories to the probability that he proposes action 1, with $\sigma_p \in \Sigma_p$. Let $\sigma_a : \mathscr{H} \to [0,1]$ be the agent's strategy, which maps the histories to the probability with which he approves action 1, with $\sigma_a \in \Sigma_a$.

The principal is patient and maximizes the frequency of action 1.[8] The agent is myopic and his stage-game payoff in period t is $v(y_t)$.[9] Since the principal has no private information about the state, under a *no signaling what you don't know* condition (Fudenberg and Tirole, 1991), neither the agent's belief nor his best reply depends on the principal's proposals $\{\widetilde{a}_0, ..., \widetilde{a}_t\}$ or the principal's strategy σ_p. Without loss of generality, we assume that $\sum_{y \in Y} v(y) F_1(y|1) > \sum_{y \in Y} v(y) F_1(y|0)$ and $\sum_{y \in Y} v(y) F_0(y|0) > \sum_{y \in Y} v(y) F_0(y|1)$, that is, action a is optimal for the agent in state F_a for every $a \in A$.[10]

This game fits applications where a benevolent policy-maker (i.e., the principal) wants to persuade the public (i.e., the agents) to stop taking actions that have negative externalities on marginalized groups (action 0, for example, massive gatherings during a pandemic have negative externalities on people who are immunocompromised), or to adopt reforms that have positive externalities (action 1, for example, reducing greenhouse gas emission has positive externalities on future generations). Action 0 is interpreted as a *status quo action*, which the policy-maker has the ability to implement by himself. By contrast, the public's co-operation is crucial for the implementation of the socially beneficial action. For example, the government can issue a mask mandate for the purpose of slowing down the spread of a virus, but this mandate won't be effective unless the majority of citizens cooperate. However, taking the socially beneficial action is against the agent's private interest in state F_0 and he learns about which action is optimal over time.

Our result characterizes the maximal frequency that the principal can implement the socially beneficial action by taking advantage of the agent's misspecified beliefs. We also describe the qualitative features of the principal's strategy from which he approximately attains his optimal payoff. We assume that the agent is not indifferent between action 0 and action 1 at any history.

Assumption 1. π_0 *is such that the agent is not indifferent between action* 0 *and action* 1 *at every* h^t.

Assumption 1 is satisfied for generic prior belief π_0 given that A and Y are finite sets. This assumption implies that the agent's optimal strategy is unique, which we denote by σ_a^*. The principal's asymptotic

---

[8]We evaluate the patient principal's payoff using the long-run averages. This is a common practice in undiscounted games, see for example, Hart (1985) and Forges (1992).

[9]We comment on the case in which $k' \neq 0$ and the agent's discount factor is strictly positive by the end of this section.

[10]If action 1 is optimal for the agent in both states, then the principal can implement action 1 with frequency 1 regardless of the agent's prior belief and belief misspecification. If action 0 is optimal for the agent in both states, then the frequency of action 1 is zero regardless of the principal's strategy.

payoff from strategy σ_p is between

$$\underline{V}(\sigma_p) \equiv \liminf_{t \to +\infty} \frac{1}{t} \mathbb{E}^{(\sigma_p, \sigma_a^*)} \left[\sum_{s=1}^{t} a_s \right] \quad \text{and} \quad \overline{V}(\sigma_p) \equiv \limsup_{t \to +\infty} \frac{1}{t} \mathbb{E}^{(\sigma_p, \sigma_a^*)} \left[\sum_{s=1}^{t} a_s \right], \tag{4.1}$$

where $\mathbb{E}^{(\sigma_p, \sigma_a^*)}[\cdot]$ is the expectation under (σ_p, σ_a^*). The principal's payoff when he optimally chooses his strategy is bounded between $\underline{V} \equiv \sup_{\sigma_p \in \Sigma_p} \underline{V}(\sigma_p)$ and $\overline{V} \equiv \sup_{\sigma_p \in \Sigma_p} \overline{V}(\sigma_p)$.

We introduce some notation to characterize the principal's optimal payoff. Let

$$\mathscr{E}^* \equiv \left\{ \text{there exists } t \in \mathbb{N} \text{ such that for every } s \geq t, \ \sum_{F \in \mathscr{F}} \pi_s(F) \sum_{y \in Y} v(y) F(y|1) > \sum_{F \in \mathscr{F}} \pi_s(F) \sum_{y \in Y} v(y) F(y|0) \right\}$$

be the event that action 1 is strictly optimal for the agent starting from some period. Let

$$q^* \equiv \sup_{\sigma_p \in \Sigma_p} \Pr(\mathscr{E}^* | \sigma_p, \sigma_a^*, F_0). \tag{4.2}$$

Intuitively, q^* is the maximal probability of event $\mathscr{E}^*$ when the state is F_0 and the agent plays according to his optimal strategy. Let $X_{a \to a'}$ be a random variable such that

$$X_{a \to a'} = \log \frac{F_1(y|a')}{F_0(y|a')} \text{ with probability } F_0(y|a) \text{ for every } y \in Y. \tag{4.3}$$

Intuitively, $X_{a \to a'}$ is the change in the log likelihood ratio between F_0 and F_1 when the true state is F_0, the previous period action was a, and the current period action is a'. Let

$$\lambda \equiv \sup_{\widehat{\lambda} \geq 0} \frac{\widehat{\lambda} + 1}{\widehat{\lambda} + 2} \tag{4.4}$$

subject to

$$\widehat{\lambda} \mathbb{E}[X_{1 \to 1}] + \mathbb{E}[X_{1 \to 0} + X_{0 \to 1}] > 0. \tag{4.5}$$

Moreover, we define $\lambda = 0$ if the above inequality is never satisfied. One can verify that $\lambda \in (\frac{1}{2}, 1) \cup \{0\}$. Intuitively, λ is the maximal frequency of action 1 such that the log likelihood ratio between F_0 and F_1 does not increase in expectation, or in another word, the amount of mislearning in state F_0 is non-negative.

Theorem 2. *If π_0 satisfies Assumption 1, then*

$$\overline{V} = \underline{V} = \pi_0(F_1) + \pi_0(F_0) q^* \lambda. \tag{4.6}$$

Theorem 2 implies that the principal's asymptotic payoff exists (i.e., $\overline{V} = \underline{V}$) and characterizes its value. At the optimum, the asymptotic frequency of action 1 is 1 in state F_1 and is $q^* \lambda$ in state F_0.

The proof is in Appendix D. We provide an intuitive explanation in three steps, using an example in which there are two outcomes $Y = \{y_g, y_b\}$, the outcome distributions are given by:

$$F_1(y|a) \equiv \begin{cases} r & \text{if } (a,y) = (0,y_b) \text{ or } (1,y_g) \\ 1-r & \text{if } (a,y) = (1,y_b) \text{ or } (0,y_g), \end{cases}$$

$$F_0(y|a) \equiv \begin{cases} r & \text{if } (a,y) = (1,y_b) \text{ or } (0,y_g) \\ 1-r & \text{if } (a,y) = (0,y_b) \text{ or } (1,y_g), \end{cases}$$

where $r \in (1/2, 1)$ is a parameter, and the agent's payoff is 1 when the outcome is y_g and is 0 otherwise.

In this example, the agent strictly prefers action 0 if and only if $\log \frac{\pi_t(F_0)}{\pi_t(F_1)} > 0$. One can verify that $X_{1\to1}$ first order stochastically dominates both $X_{1\to0}$ and $X_{0\to1}$. Therefore, the maximum that defines q^* is attained when the principal proposes the opposite action to what was implemented k^* periods ago, that is, $\tilde{a}_t = 1 - a_{t-k^*}$ for every $t \in \mathbb{N}$. According to the maximization problem that defines λ,

$$\lambda = \max_{\widehat{\lambda}} \frac{\widehat{\lambda}+1}{\widehat{\lambda}+2} \quad \text{subject to} \quad \widehat{\lambda}\mathbb{E}[X_{1\to1}] + \mathbb{E}[X_{1\to0} + X_{0\to1}] \geq 0. \tag{4.7}$$

Since $X_{1\to1}$ first order stochastically dominates both $X_{1\to0}$ and $X_{0\to1}$, the constraint is binding and the maximum in (4.7) is attained when the ratio between taking the same action in consecutive periods and action switches is $\widehat{\lambda}$.

Step 1:   We consider an *auxiliary game* in which the principal knows the true state but the agent is naive in the sense that he fails to extract information from the principal's proposals. We show that when the true state belongs to $\mathscr{F}_1$, the principal's asymptotic payoff in the auxiliary game is 1 regardless of the agent's prior belief. Moreover, the principal can attain this payoff by proposing action 1 in every period.

Let $l_t \equiv \log \frac{\pi_t(F_0)}{\pi_t(F_1)}$. Agent t strictly prefers action 1 when $l_t < 0$ and strictly prefers action 0 when $l_t > 0$. Since l_t decreases in expectation when the agent takes the same action in two consecutive periods and F_1 is the true state, the Wald's inequality (Lemma A.1) implies that

1. for every $l_t < 0$, the probability of the event $\{l_\tau < 0 \text{ for every } \tau \geq t\}$ is bounded away from 0,

2. for every $l_t > 0$, the probability of the event $\{l_\tau > 0 \text{ for every } \tau \geq t\}$ is 0.

Since the log likelihood ratio process is absorbed with positive probability at negative values and is absorbed

with zero probability at positive values, we know that with probability 1, there exists $T \in \mathbb{N}$ such that $l_t < 0$ for all $t \geq T$. Therefore, the principal's asymptotic payoff is 1 regardless of the agent's prior belief.

Step 2:   We show that principal's optimal payoff in the auxiliary game where the state is F_0 is $q^* \lambda$, that is,

$$U(F_0, \pi_0) = q^* \lambda. \tag{4.8}$$

Recall that (1) q^* is the maximal probability that the agent eventually has an incentive to approve action 1, which is attained when the principal proposes action 1 with frequency approximately $1/2$ the agent knows event $\mathscr{E}^*$, and (2) according to (4.7), λ is the maximal frequency that the principal can propose action 1 subject to the constraint that the log likelihood ratio between F_0 and F_1 does not increase in expectation.

The principal faces a tradeoff between increasing the frequency that he proposes action 1 and increasing the probability that the agent is willing to approve action 1. The former allows him to propose action 1 with frequency as high as λ, but in order to maximize the probability with which the agent approves action 1, he needs to propose it with frequency close to $1/2$.

Equation (4.8) suggests that such a tradeoff has no impact on the principal's asymptotic payoff since in the auxiliary game where the state is F_0, the principal can attain an expected payoff as if (1) the agent eventually approves action 1 for all periods with its maximal probability q^*, and (2) the principal can propose action 1 with its maximal frequency subject to the mislearning constraint, which equals λ.

The definitions of q^* and λ imply that $U(F_0, \pi_0) \leq q^* \lambda$. We show that $U(F_0, \pi_0) \geq q^* \lambda$ by constructing a family of strategies under which the principal's asymptotic payoff is arbitrarily close to $q^* \lambda$. Each strategy in this class is characterized by a cutoff log likelihood ratio $l_\varepsilon^* < 0$ such that the principal proposes $\widetilde{a}_t = 1 - a_{t-k^*}$ if the log likelihood ratio is above l_ε^* in all previous period, and proposes action 1 with frequency close to but less than λ when the log likelihood ratio has fall below l_ε^* in at least one period.

The key step is to show that under the proposed strategy (1) the probability with which the log likelihood ratio falls below l_ε^* in at least one period is arbitrarily close to q^*, and (2) conditional on the log likelihood ratio falls below l_ε^*, the probability that it is strictly negative in all future periods is close to 1.

We establish these two claims using concentration inequalities. The intuition behind the first claim is that conditional on $l_t > 0$, the probability that the log likelihood ratio is positive in all future periods is strictly positive, so the log likelihood ratio will eventually escape any bounded interval with probability 1. As a result, the probability that $l_t < 0$ for all t large enough equals the probability that $l_t < l_\varepsilon^*$ for all t large enough. The intuition behind the second statement is that according to (4.8), one can construct strategies

under which the frequency of proposing action 1 is close to λ, yet the log likelihood ratio is non-increasing in expectation. For an example of such a strategy, let $T_1, T_2 \in \mathbb{N}$ be such that T_1 is even and $T_2/T_1 \in (\lambda - \varepsilon, \lambda)$. The principal's strategy is divided into $T \equiv T_1 + T_2$ period blocks such that he proposes action 0 in period 1, 3, ... $T_1 - 1$ within each block, and proposes action 1 otherwise. The Wald's inequality implies that the probability with which the log likelihood ratio exceeds 0 in some period after t is small when $l_t < l_\varepsilon^*$ and l_ε^* is small enough.

To conclude, when the principal uses this class of strategies, he can ensure that with probability close to q^*, the agent is willing to approve action 1 in all future periods, and conditional on this event, he can propose policy 1 with frequency arbitrarily close to λ. This explains why the tradeoff he faces between inducing mislearning and increasing the frequency of proposing action 1 diminishes in the long run.

Step 3:   We show that the principal's payoff in our dynamic policy choice game with symmetric uncertainty equals his expected payoff in the auxiliary game. Formally, let $U(F, \pi_0)$ be the principal's payoff in the auxiliary game when the state is F and the agent's prior belief is π_0. We show that:

$$\underline{V} = \overline{V} = \pi_0(F_0)U(F_0, \pi_0) + \pi_0(F_1)U(F_1, \pi_0). \tag{4.9}$$

This is implied by the following two inequalities:

$$\overline{V} \leq \pi_0(F_0)U(F_0, \pi_0) + \pi_0(F_1)U(F_1, \pi_0) \tag{4.10}$$

and

$$\underline{V} \geq \pi_0(F_0)U(F_0, \pi_0) + \pi_0(F_1)U(F_1, \pi_0). \tag{4.11}$$

Inequality (4.10) is straightforward since the principal's payoff is weakly greater in the auxiliary game given that he has more information and the agent does not extract information from his proposals.

In order to establish inequality (4.11), let σ_p^ε be the principal's strategy such that his asymptotic payoff is more than $U(F_0, \pi_0) - \varepsilon$ in the auxiliary game where the state is F_0. Since y is informative about the state, for every $\varepsilon > 0$, there exists $T \in \mathbb{N}$ such that for each of the principal's strategy and every $a \in A$, the principal's posterior belief in period T attaches probability greater than $1 - \varepsilon$ to F_a when F_a is the true state.

Consider the principal's asymptotic payoff by using the following strategy in the original game with symmetric uncertainty:

1. he plays according to σ_p^ε in the first T periods,

2. if his period T posterior belief attaches probability greater than $1 - \varepsilon$ to state F_1, then he proposes action 1 in all future periods,

3. if his period T posterior belief attaches probability less than $1 - \varepsilon$ to state F_1, he continues to use strategy σ_p^{ε}.

Under the above strategy, the probability with which the principal proposes 1 in every period after T is at least $1 - \varepsilon$ conditional on the state being F_1. Since $U(F_1, \pi_0) = 1$ for all π_0, the principal's asymptotic payoff conditional on state F_1 is at least $1 - \varepsilon$. Conditional on the true state being F_0, the probability with which he uses σ_p^{ε} in every period is at least $1 - \varepsilon$, so his payoff is at least $(1 - \varepsilon)(U(F_0, \pi_0) - \varepsilon)$. Therefore, his expected asymptotic payoff converges to $\pi_0(F_0)U(F_0, \pi_0) + \pi_0(F_1)$ as ε goes to 0.

Remarks: Our formula for the principal's optimal payoff is reminiscent of the repeated zero sum games of Aumann and Maschler (1995) and the Bayesian persuasion games of Kamenica and Gentzkow (2011), where an informed player's payoff in a binary-state setting is a piece-wise linear and concave function of the uninformed player's prior belief. Our formula for the principal's highest equilibrium payoff (4.6) is not continuous since q^* depends on the agent's prior belief and exhibits discontinuity in general.

When $\delta > 0$, the agent may have incentives to experiment, which depend on the principal's strategy. As a result, the agent's optimal strategy is not unique when the log likelihood ratio between F_0 and F_1 is close to the cutoff at which a myopic agent is indifferent. In general, for every $\delta \in (0,1)$ and k', there exist two cutoffs l^* and l^{**} with $0 < l^* < l^{**} < +\infty$ such that regardless of the principal's strategy, the agent has a strict incentive to approve action 1 when $l_t < l^*$ and has a strict incentive to veto action 1 when $l_t > l^{**}$. When $l_t \in [l^*, l^{**}]$, the agent's incentive depends on the principal's strategy. Nevertheless, when the discount factor is positive but small enough, the set of agent-optimal strategy is small and our approach provides lower and upper bounds on the principal's payoff. The two bounds coincide as the agent's discount factor converges to 0, in which case our approach can exactly characterize the principal's asymptotic payoff.

5   Discussions

We discuss extensions and generalizations of our main result.

Uncertainty about the time lag: In our baseline model, the agent faces uncertainty about the outcome distribution but has a degenerate prior about the time lag. In general, the agent may also learn about the time lag under a misspecified model.

Our main result extends when the agent faces uncertainty both about the outcome distribution and the time lag.[11] Formally, there is a finite set of states $\mathscr{F} \subset \mathscr{Y}$ and a finite set of possible time lags $K \subset \mathbb{N}$. The agent has a full support prior belief $\pi_0 \in \Delta(\mathscr{F} \times K)$. In order to focus on the effects of misspecified belief about the time lag, we assume that $F^* \in \mathscr{F}$ and $k^* \notin K$.

Similar to the baseline model, the agent chooses $a^* \equiv \arg\max_{a \in A} \sum_{y \in Y} v(y) F^*(y|a)$ in every period when he learns the true state regardless of his belief about the time lag. As a result, the agent's action cannot converge to actions other than a^* and a^* occurs with positive asymptotic frequency. When $\delta \in (0, 1)$ or $(\delta, k') = (0, 0)$, there exists $F^* \in \mathscr{Y}$ and a prior belief π_0 that is regular with respect to F^* under which the agent takes his optimal action with frequency arbitrarily close to 0.

General forms of belief misspecifications: In our baseline model, the distribution of y_t depends only on one of the agent's actions. In practice, the outcome distribution can be affected by multiple actions.

We extend our results when the distribution of y_t depends on a convex combination of the agent's current-period action and his actions in the last $k \in \mathbb{N}$ periods where k is an exogenous parameter. In particular, when the state is $F \equiv \{F(\cdot|a)\}_{a \in A}$, y_t is distributed according to $\sum_{a \in A} \alpha_t(a) F(\cdot|a)$ where $\alpha_t(a) \equiv \sum_{j=0}^{k} \beta_j \mathbf{1}\{a_{t-j} = a\}$, $\beta \equiv (\beta_0, ..., \beta_k) \in \mathbb{R}_+^{k+1}$, and $\sum_{j=0}^{k} \beta_j = 1$. The agent has a wrong belief about the convex weights of different actions, and believes that when the state is F, y_t is distributed according to $\sum_{a \in A} \widehat{\alpha}_t(a) F(\cdot|a)$, where $\widehat{\alpha}_t(a) \equiv \sum_{j=0}^{k} \widehat{\beta}_j \mathbf{1}\{a_{t-j} = a\}$ with $\widehat{\beta} \equiv (\widehat{\beta}_0, ..., \widehat{\beta}_k) \in \mathbb{R}_+^{k+1}$, $\sum_{j=0}^{k} \widehat{\beta}_j = 1$, and $\widehat{\beta} \neq \beta$. This general formulation captures for example, when the agent overestimates or underestimates the effects of his current-period action on the current-period outcome, i.e., when $\widehat{\beta}_0 \neq \beta_0$. Theorem 1 extends under a stronger identification condition that for every $F, F' \in \mathscr{F}$ and $\alpha \in \Delta(A)$, we have $F(\cdot|\alpha) \neq F'(\cdot|\alpha)$.

---

[11] If the agent only faces uncertainty about the time lag but knows the outcome distribution, then he takes his optimal action in every period and misspecified belief about the time lag is irrelevant for his behavior and payoff.

A   Probability Tools

We state three results in probability theory, which will be used in our subsequent proofs. The first result is the Wald nequality, which bounds the probability of the union of tail events from above.

Lemma A.1 (Wald, 1944)**.** *Let $\{Z_t\}_{t\in\mathbb{N}}$ be a sequence of i.i.d. random variables with finite support, strictly negative mean, and takes a positive value with positive probability. Let $r^* > 0$ be the unique real number that satisfies $\mathbb{E}_{z\sim Z_1}[\exp(r^*z)] = 1$. We have*

$$\Pr\left[\bigcup_{n=1}^{\infty}\left\{\sum_{t=1}^{n}Z_t \geq c\right\}\right] \leq \exp(-r^* \cdot c) \text{ for every } c > 0.$$

The second result is the Azuma-Hoeffding inequality, that applies to martingales with bounded increments.

Lemma A.2 (Azuma-Hoeffding inequality)**.** *Let $\{Z_0, Z_1, \cdots\}$ be a martingale such that $|Z_k - Z_{k-1}| \leq c_k$. Then for every $N \in \mathbb{N}$ and $\varepsilon_1 > 0$, we have*

$$\Pr[Z_N - Z_0 \geq \varepsilon_1] \leq \exp\left(-\frac{\varepsilon_1^2}{2\sum_{k=1}^{N}c_k^2}\right).$$

The third result extends the Chernoff-Heoffding inequality to random variables with unbounded support and finite first and second moments.

Lemma A.3. *For any $\lambda > 1$ and any sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$ with finite mean $\mu > 0$ and finite variance, we have*

$$\Pr\left[\sum_{i=1}^{n}X_i \geq \lambda\mu n\right] \leq \exp(-cn)$$

where $c \equiv \max_{t>0}\{\lambda\mu t - \ln\mathbb{E}\left[e^{tX_1}\right]\}$. For any $\lambda \in (0,1)$, we have

$$\Pr\left[\sum_{i=1}^{n}X_i \leq \lambda\mu n\right] \leq \exp(-c'n)$$

where $c' \equiv \max_{t>0}\{-\lambda\mu t - \ln\mathbb{E}\left[e^{-tX_1}\right]\}$

Proof. The proof is similar to that of the Chernoff-Hoeffding inequality. If $\lambda > 1$, then

$$\Pr\left[\sum_{i=1}^{n}X_i \geq \lambda n\mu\right] = \Pr\left[e^{t\sum_{i=1}^{n}X_i} \geq e^{t\lambda n\mu}\right] \leq \frac{1}{e^{t\lambda n\mu}}\cdot\prod_{i=1}^{n}\mathbb{E}\left[e^{tX_i}\right] \text{ for every } t \in \mathbb{N},$$

where the last inequality holds by the Markov inequality. Set $t \in \mathbb{N}$ in order to maximize $\lambda \mu t - \ln \mathbb{E}\left[e^{tX_1}\right]$, we have

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \lambda n \mu\right] \leq \exp\{-cn\}.$$

If $\lambda < 1$, then

$$\Pr\left[\sum_{i=1}^{n} X_i \leq \lambda n \mu\right] = \Pr\left[e^{-t\sum_{i=1}^{n} X_i} \geq e^{-t\lambda n \mu}\right] \leq \frac{1}{e^{-t\lambda n \mu}} \cdot \prod_{i=1}^{n} \mathbb{E}\left[e^{-tX_i}\right] \text{ for every } t \in \mathbb{N},$$

Set $t \in \mathbb{N}$ in order to maximize $-\lambda \mu t - \ln \mathbb{E}\left[e^{-tX_1}\right]$, we have

$$\Pr\left[\sum_{i=1}^{n} X_i \leq \lambda n \mu\right] \leq \exp\{-c'n\}.$$

$\square$

Remark: In Lemma A.3, let $c(t) = \lambda \mu t - \ln \mathbb{E}\left[e^{tX_1}\right]$. One can verify that $c(0) = 0$ and $c'(0) = (\lambda - 1)\mu > 0$ for $\lambda > 1$ and $\mu > 0$. Moreover, $c''(0) = \text{Var}[X_1]$ is finite. Therefore, $c = \max_{t>0}\{\lambda \mu t - \ln \mathbb{E}\left[e^{tX_1}\right]\}$ is strictly positive for $\lambda > 1$ and $\mu > 0$. Similarly, we can also show that $c' > 0$ for $\lambda < 1$ and $\mu > 0$.

B   Proof of Lemma 3.1

The conclusion of Lemma 3.1 is implied by the following two claims.

Claim 2. *For $\varepsilon \in (0,1)$ and π_0 that has finite support, there exists $\eta > 0$ and a sufficiently large T such that the $\pi_T(F^*) > 1 - \varepsilon$ with probability at least η.*

Claim 3. *Suppose $\delta \in (0,1)$ or $(\delta, k') = (0,0)$. For any finite set of states $\mathscr{F} \subseteq \Delta(\mathscr{Y})$, there exist $\varepsilon, \eta \in (0,1)$ such that if $\pi_0(F^*) > 1 - \varepsilon$ and π_0 is supported on $\mathscr{F}$, then a^* is chosen for all periods with probability at least η.*

Combining those two claims, we have that for any prior π with finite support, there exists $\eta > 0$ and $T > 0$ such that with probability η, action a^* is chosen for all periods after T, which implies that the limiting frequency of a^* is at least $\eta > 0$.

Proof Claim 2. Let $\mathscr{F}$ be the support of π_0. For every pair of states $F_0, F_1 \in \mathscr{F}$, let

$$X_{a \to a'}(F_0, F_1) \equiv \log \frac{F_0(y|a')}{F_1(y|a')} \text{ with probability } F^*(y|a) \text{ for every } y \in Y.$$

Let $\bar{l}$ be the largest realization of $|X_{a \to a'}(F_0, F_1)|$ for any $a, a' \in A$ and $F_0, F_1 \in \mathscr{F}$. One can verify that the increment of $\log \frac{\pi_t(F_0)}{\pi_t(F_1)}$ follows the same distribution as random variable $X_{a_{t-k^*} \to a_{t-k'}}(F_0, F_1)$. Therefore,

$$\log \frac{\pi_T(F_0)}{\pi_T(F_1)} - \log \frac{\pi_0(F_0)}{\pi_0(F_1)} - \sum_{t=1}^{T} \mathbb{E}\left[X_{a_{t-k^*} \to a_{t-k'}}(F_0, F_1)\right]$$

is a martingale with bounded increments. According to Lemma A.2, for every $\varepsilon > 0$, $T > 0$, π, and F_0, F_1, and any sequence of actions $a_0, \ldots, a_T$, we have

$$\Pr\left[\log \frac{\pi_T(F_0)}{\pi_T(F_1)} - \log \frac{\pi_0(F_0)}{\pi_0(F_1)} - \sum_{t=1}^{T} \mathbb{E}\left[X_{a_{t-k^*} \to a_{t-k'}}(F_0, F_1)\right] \geq \varepsilon\right] \leq \exp\left(-\frac{\varepsilon_1^2}{2T\bar{l}^2}\right). \tag{B.1}$$

Now consider an auxiliary scenario where $k^* = k'$. One can verify that $X_{a_{t-k'} \to a_{t-k'}}(F, F^*) < 0$ for every $F_1 \neq F^*$. For every $\varepsilon > 0$, let $T \in \mathbb{N}$ be large enough such that

1. $|\mathscr{F}| \cdot \exp\left(-\frac{\varepsilon_1^2}{2T\bar{l}^2}\right) < 1$.

2. For any $F \in \mathscr{F} \setminus F^*$, $-\log \frac{\pi_0(F)}{\pi_0(F^*)} - \sum_{t=1}^{T} \mathbb{E}[X a_{t-k'} \to a_{t-k'}(F, F^*)] - \varepsilon > \log \frac{|\mathscr{F}|(1-\varepsilon)}{\varepsilon}$.

By applying union bound to inequality (B.1) for all $F \in \mathscr{F} \setminus F^*$, we know that with strictly positive probability, $\log \frac{\pi_T(F)}{\pi_T(F^*)} \geq \log \frac{|\mathscr{F}|(1-\varepsilon)}{\varepsilon}$ for any $F \in \mathscr{F} \setminus F^*$. This implies that $\pi_T(F^*) \geq 1 - \varepsilon$. Note that by Condition 2, this event also occurs with strictly positive probability when the true state is F^* with time lag k^*. This concludes the proof of Claim 2. $\square$

Proof of Claim 3. For every finite set $\mathscr{F} \subseteq \Delta(\mathscr{Y})$ such that $F^* \in \mathscr{F}$, let $\pi_\varepsilon \in (0,1)$ be the lowest probability such that for every $\pi \in \Delta(\mathscr{F})$, if π attaches probability at least π_ε to F^*, then the agent has a strict incentive to choose a^*. For simplicity, let $X(F) \equiv X_{a^* \to a^*}(F, F^*)$ for any $F \in \mathscr{F} \setminus F^*$. By definition, we have $\mathbb{E}[X(F)] < 0$. Let $r_F^* > 0$ be defined via $\mathbb{E}_{x \sim X(F)}[\exp(r_F^* \cdot x)] = 1$. According to Lemma A.1, for a sequence of i.i.d. random variables $X_1, \ldots, X_t$ that is distributed according to $X(F)$, we have

$$\Pr\left[\bigcup_{n=1}^{\infty}\left\{\sum_{t=1}^{n} X_t \geq c\right\}\right] \leq \exp(-r_F^* \cdot c) \text{ for every } c > 0.$$

Let $r^* = \min_{F \in \mathscr{F} \setminus F^*} r_F^*$. Let c be such that $|\mathscr{F}| \cdot \exp(-r^* \cdot c) < 1$, and let $\varepsilon > 0$ be such that $c + \bar{l} \cdot \max\{k^*, k'\} + \log \frac{\varepsilon}{1-\varepsilon} < \log \frac{\pi_\varepsilon}{|\mathscr{F}| \cdot (1-\pi_\varepsilon)}$, we know that for the first $\max\{k^*, k'\}$ periods, the agent always chooses action a^* and the log likelihood ratio between F and F^* increases by at most $\bar{l} \cdot \max\{k^*, k'\}$. More-

over, for any $t > \bar{l} \cdot \max\{k^*, k'\}$, with probability $1 - |\mathscr{F}| \cdot \exp(-r^* \cdot c) > 0$, we have that

$$\log \frac{\pi_t(F)}{\pi_t(F^*)} < c + \bar{l} \cdot \max\{k^*, k'\} + \log \frac{\pi_0(F)}{\pi_0(F^*)} < c + \bar{l} \cdot \max\{k^*, k'\} + \log \frac{\varepsilon}{1 - \varepsilon} < \log \frac{\pi_\varepsilon}{|\mathscr{F}| \cdot (1 - \pi_\varepsilon)}$$

for any $F \in \mathscr{F} \setminus F^*$. This implies that $\pi_t(F^*) > \pi_\varepsilon$ for any $t > 0$ and hence the agent chooses action a^* in all future periods. $\qquad\square$

C   Proof of Theorem 1

We establish inequality (3.3) when $|Y| = 2$, and later adjust the proof to environments where $|Y| \geq 3$. For simplicity, we first consider the case in which $k^* = 1$ and $k' = 0$. Our argument straightforwardly generalizes to other values of k^* and k' as long as $k^* \neq k'$.

Since $|Y| = 2$, the agent chooses the action that induces the high-payoff outcome with higher probability. Without loss of generality, let $Y \equiv \{y_0, y_1\}$, and let $v(y_0) = 0$ and $v(y_1) = 1$. Let $\zeta, \zeta' \in (0, 1/4)$, and let

$$F^*(y_1 | a = 0) \equiv \frac{1}{2} \qquad\qquad F^*(y_1 | a = 1) \equiv \zeta';$$

$$F_0(y_1 | a = 0) \equiv 2\zeta \qquad\qquad F_0(y_1 | a = 1) \equiv \zeta; \qquad\qquad \text{(C.1)}$$

$$F_1(y_1 | a = 0) \equiv 3\zeta \qquad\qquad F_1(y_1 | a = 1) \equiv 4\zeta.$$

The optimal action under F^* is 0. Let $\pi_0(F_0) \equiv \frac{1-\zeta'}{2}$, $\pi_0(F_1) \equiv \frac{1-\zeta'}{2}$, and $\pi(F) \equiv \zeta'$. First, we show that for any $\gamma > 0$, there exist ζ and ζ' such that the asymptotic frequency of action 0 is less than γ. By the end of this section, we identify the crucial components in this construction to show that there exists an open set of distributions such that the agent can have arbitrarily small frequency of choosing action a^*.

First we bound the expected number of times with which the agent chooses action 0 and action 1 when the true distribution is F^* and the agent's prior belief attach small but positive probability to F^*. Let

$$X_{a \to a'}(F_0, F_1) \equiv \log \frac{F_0(y | a')}{F_1(y | a')} \text{ with probability } F^*(y | a) \text{ for every } y \in Y. \qquad\qquad \text{(C.2)}$$

Intuitively, this is the change in the log likelihood ratio between F_0 and F_1 when the previous period action was a and the current period action is a'. Let $\bar{l}(F_0, F_1)$ be the largest realization of $X_{1 \to 0}(F_0, F_1)$, and let $\underline{l}(F_0, F_1)$ be the smallest realization of $X_{0 \to 1}(F_0, F_1)$. By construction we have $\bar{l}(F_0, F_1) > 0$ and $\underline{l}(F_0, F_1) < 0$. In what follows, we omit the dependence of (F_0, F_1) and write $\bar{l}$ and $\underline{l}$ instead.

Consider a hypothetical scenario in which the support of the agent's prior belief is $\{F_0, F_1\}$. For any

discount factor $\delta \in [0,1)$, there exists $l^* \in \mathbb{R}$ depending on δ such that the agent is indifferent between actions 0 and 1 when $\log \frac{\pi(F_0)}{\pi(F_1)} = l^*$, and strictly prefers action 0 if and only if $\log \frac{\pi(F_0)}{\pi(F_1)} \geq l^*$. For every $l > l^*$, let random variable $\tau_0(l)$ be the number of consecutive periods with which the agent takes action 0 when the initial value of $\log \frac{\pi(F_0)}{\pi(F_1)}$ is l. For every $l < l^*$, let random variable $\tau_1(l)$ be the number of consecutive periods with which the agent takes action 1 when the initial value of $\log \frac{\pi(F_0)}{\pi(F_1)}$ is l.

Claim 4. *Random variables $\tau_0(l)$ and $\tau_1(l)$ have finite mean and variance for every $l \in \mathbb{R}$.*

Proof. According to Bayes rule, $l_t = l_{t-1} + X_{0 \to 0}(F_0, F_1)$ for every $t \in \mathbb{N}$. Let H be the maximal difference in the realization of random variable $X_0(F_0, F_1)$. The Chernoff-Hoeffding inequality implies that

$$\Pr[l_t > l^*] \leq \exp\left(-\frac{2(-t \cdot \mathbb{E}[X_{0 \to 0}(F_0, F_1)] - l_0 + l^*)^2}{tH^2}\right).$$

Since $\Pr[l_t > l^*]$ vanishes exponentially as $t \to +\infty$, $\tau_0(l)$ has finite mean and variance for every $l \geq l^*$. Similarly, one can also show that $\tau_1(l)$ has a finite mean and variance for every $l \leq l^*$. $\square$

Next, suppose F^* belongs to the support of agent's prior belief. Let $\tau_0^\varepsilon(l)$ be the number of consecutive periods with which the agent takes action 0 when the initial value of $\log \frac{\pi(F_0)}{\pi(F_1)}$ is $l - \varepsilon$ and switches to action 1 if the log likelihood ratio is below $l^* + \varepsilon$ after the first period. Let $\tau_1^\varepsilon(l)$ be the number of consecutive periods with which the agent takes action 1 when the initial value of $\log \frac{\pi(F_0)}{\pi(F_1)}$ is $l + \varepsilon$ and switches to action 0 if the log likelihood ratio is above $l^* - \varepsilon$ after the first period.

For every $\varepsilon > 0$, there exists $\pi_\varepsilon > 0$ such that when the prior belief $\pi \in \Delta(\mathscr{F})$ satisfies $\pi(F^*) < \pi_\varepsilon$, the agent strictly prefers action 0 when $l > l^* + \varepsilon$ and strictly prefers action 1 when $l < l^* - \varepsilon$. Let $X_{\tau_0}(\varepsilon)$ be the random variable that has the same distribution as $\tau_0(\bar{l} + l^* + 2\varepsilon)$ and $X'_{\tau_0}(\varepsilon)$ be the random variable that has the same distribution as $\tau_0^\varepsilon(l^*)$. Let $X_{\tau_1}(\varepsilon)$ be the random variable that has the same distribution as $\tau_1(\underline{l} + l^* - 2\varepsilon)$ and $X'_{\tau_1}(\varepsilon)$ be the random variable that has the same distribution as $\tau_1^\varepsilon(l^*)$. For every $\varepsilon > 0$ and $\eta > 0$, let

$$\hat{c}_0(\varepsilon, \eta) \equiv \max_{t > 0} \left\{(1 + \eta)t\mathbb{E}[X_{\tau_0}(\varepsilon)] - \ln \mathbb{E}\left[e^{tX_{\tau_0}(\varepsilon)}\right]\right\}$$

$$c'_0(\varepsilon, \eta) \equiv \max_{t > 0} \left\{-(1 - \eta)t\mathbb{E}[X'_{\tau_0}(\varepsilon)] - \ln \mathbb{E}\left[e^{-tX'_{\tau_0}(\varepsilon)}\right]\right\}$$

$$c_0(\varepsilon, \eta) \equiv \min\{\hat{c}_0(\varepsilon, \eta), c'_0(\varepsilon, \eta)\}$$

and

$$\hat{c}_1(\varepsilon, \eta) \equiv \max_{t>0} \left\{ (1+\eta) t \mathbb{E}[X_{\tau_1}(\varepsilon)] - \ln \mathbb{E}\left[e^{t X_{\tau_1}(\varepsilon)} \right] \right\}$$

$$c_1'(\varepsilon, \eta) \equiv \max_{t>0} \left\{ -(1-\eta) t \mathbb{E}[X_{\tau_1}'(\varepsilon)] - \ln \mathbb{E}\left[e^{-t X_{\tau_1}'(\varepsilon)} \right] \right\}$$

$$c_1(\varepsilon, \eta) \equiv \min\{\hat{c}_1(\varepsilon, \eta), c_1'(\varepsilon, \eta)\}.$$

Let

$$K_0(\varepsilon, \eta) \equiv \frac{1}{c_0(\varepsilon, \eta)} \cdot \log \frac{e^{c_0(\varepsilon, \eta)}}{\varepsilon(e^{c_0(\varepsilon, \eta)} - 1)}$$

and

$$K_1(\varepsilon, \eta) \equiv \frac{1}{c_1(\varepsilon, \eta)} \cdot \log \frac{e^{c_1(\varepsilon, \eta)}}{\varepsilon(e^{c_1(\varepsilon, \eta)} - 1)}.$$

Let $\overline{\eta}_0(\varepsilon, \eta) \in \mathbb{R}_+$ be such that

$$c_0(\varepsilon, \bar{\eta}_0(\varepsilon, \eta)) = \frac{\log(K_0(\varepsilon, \eta)/\varepsilon)}{K_0(\varepsilon, \eta)},$$

and $\overline{\eta}_1(\varepsilon, \eta) \in \mathbb{R}_+$ be such that

$$c_1(\varepsilon, \bar{\eta}_1(\varepsilon, \eta)) = \frac{\log(K_1(\varepsilon, \eta)/\varepsilon)}{K_1(\varepsilon, \eta)}.$$

For every $k \in \mathbb{N}$, let $t_k^0 \in \mathbb{N}$ be the kth time such that $a_{t_k} = 0$ and $a_{t_k-1} = 1$ and $t_k^1 \in \mathbb{N}$ be the kth time such that $a_{t_k} = 1$ and $a_{t_k-1} = 0$. Let S_t^0 be the total number of adjacent periods until t in which the agent chooses action 0. Let S_t^1 be the total number of adjacent periods until t in which the agent chooses action 1.

Claim 5. *For every $\varepsilon > 0$ and $\eta > 0$, if $\pi(F^*) < \pi_\varepsilon$, then the following event happens with probability at least $1 - 6\varepsilon$:*

$$S_{t_k^0}^0 \leq [k(1+\eta) + K_0(\varepsilon, \eta)(1 + \bar{\eta}_0(\varepsilon, \eta))] \mathbb{E}[\tau_0(\bar{l} + l^* + 2\varepsilon)]$$

$$S_{t_k^0}^0 \geq (k - K_0(\varepsilon, \eta)) \cdot (1-\eta) \mathbb{E}[\tau_0^\varepsilon(l^*)]$$

$$S_{t_k^1}^1 \leq [k(1+\eta) + K_1(\varepsilon, \eta)(1 + \bar{\eta}_1(\varepsilon, \eta))] \mathbb{E}[\tau_1(\underline{l} + l^* - 2\varepsilon)]$$

$$S_{t_k^1}^1 \geq (k - K_1(\varepsilon, \eta)) \cdot (1-\eta) \mathbb{E}[\tau_1^\varepsilon(l^*)]$$

for every $k \in \mathbb{N}$.

Proof. We establish the upper and lower bounds for $S_{t_k^0}^0$. The ones for $S_{t_k^1}^1$ can be derived using a similar

argument. Claim 4 implies that $\mathbb{E}[X_{\tau_0}(\varepsilon)]$ is finite. Suppose π is such that $\pi(F^*) < \pi_\varepsilon$. Since $S^0_{t^0_k}$ is first order stochastically dominated by $\sum_{i=1}^k x_i$ with $x_i \sim X_{\tau_0}(\varepsilon)$, Lemma A.3 implies that

$$\Pr[S^0_{t^0_k} > k(1+\eta) \cdot \mathbb{E}[\tau_0(\bar{l}+l^*+2\varepsilon)]] \leq \exp(-k \cdot c_0(\varepsilon,\eta)).$$

The union bound implies that

$$\Pr\left[\bigcup_{k \geq K_0(\varepsilon,\eta)} \{S^0_{t^0_k} > k(1+\eta) \cdot \mathbb{E}[\tau_0(\bar{l}+l^*+2\varepsilon)]\}\right] \leq \sum_{k \geq K_0(\varepsilon,\eta)} \exp(-k \cdot c_0(\varepsilon,\eta)) \leq \varepsilon.$$

Moreover, for every $k < K_0(\varepsilon,\eta)$, we have

$$\Pr\left[S^0_{t^0_k} > k(1+\bar{\eta}_0(\varepsilon,\eta)) \cdot \mathbb{E}[\tau_0(\bar{l}+l^*+2\varepsilon)])\right] \leq \frac{\varepsilon}{K_0(\varepsilon,\eta)}.$$

Take the union of these events, we have

$$\Pr\left[\bigcup_{k \geq 1} \{S^0_{t^0_k} > k(1+\eta) \cdot \mathbb{E}[\tau_0(\bar{l}+l^*+2\varepsilon)] + K_0(\varepsilon,\eta)(1+\bar{\eta}(\varepsilon,\eta)) \cdot \mathbb{E}[\tau_0(\bar{l}+l^*+2\varepsilon)]\}\right]$$

$$\leq \Pr\left[\bigcup_{k \geq K_0(\varepsilon,\eta)} \{S^0_{t^0_k} > k(1+\eta) \cdot \mathbb{E}[\tau_0(\bar{l}+l^*+2\varepsilon)]\}\right] + \Pr\left[\bigcup_{k < K_0(\varepsilon,\eta)} \{S^0_{t^0_k} > k(1+\bar{\eta}_0(\varepsilon,\eta))\}\right] \leq 2\varepsilon.$$

Moreover, $S^0_{t^0_k}$ first order stochastically dominates $\sum_{i=1}^k x_i$ with $x_i \sim X'_{\tau_0}(\varepsilon)$. Lemma A.3 implies that

$$\Pr[S^0_{t^0_k} < k(1-\eta) \cdot \mathbb{E}[\tau^\varepsilon_1(l^*)]] \leq \exp(-k \cdot c_0(\varepsilon,\eta)).$$

By union bound, we have

$$\Pr\left[\bigcup_{k \geq K_0(\varepsilon,\eta)} \{S^0_{t^0_k} < k(1-\eta) \cdot \mathbb{E}[\tau^\varepsilon_1(l^*)]\}\right] \leq \sum_{k \geq K_0(\varepsilon,\eta)} \exp(-k \cdot c_0(\varepsilon,\eta)) \leq \varepsilon. \qquad \square$$

Claim 6. *For every $\varepsilon > 0$, there exists a prior belief $\pi_0 \in \Delta(\mathscr{F})$ such that event $\{\pi(F^*) < \pi_\varepsilon\}$ occurs with probability at least $1 - 6\varepsilon$.*

Proof. Let $l_t(F, F') \equiv \log \frac{\pi_t(F)}{\pi_t(F')}$. Let $X_t(F^*) \equiv l_t(F^*, F_0) - l_{t-1}(F^*, F_0)$, $Z_0(F^*) \equiv l_0(F^*, F_0)$ and $Z_t(F^*) = Z_{t-1}(F^*) + l_t(F^*, F_0) - l_{t-1}(F^*, F_0) - \mathbb{E}\left[X_t(F^*) \middle| h^{t-1}\right]$ for every $t \geq 1$. One can verify that $\{Z_t(F^*)\}_{t \in \mathbb{N}}$ is a martingale. By definition, $l_t(F^*, F_0) = Z_t(F^*) + \sum_{t' \leq t} \mathbb{E}\left[X_{t'}(F^*) \middle| h^{t'-1}\right]$. Let H be the difference between the

maximal realization of $X_0(F^*, F_0)$ and the minimal realization of $X_0(F^*, F_0)$. We have $|Z_t(F^*) - Z_{t-1}(F^*)| \leq H$ for every $t \in \mathbb{N}$. According to Lemma A.2,

$$\Pr[Z_t(F^*) - Z_0(F^*) \geq t\eta] \leq \exp\left(-\frac{t\eta^2}{2H^2}\right) \text{ for every } \eta \in \mathbb{R}_+.$$

Let $T \equiv \frac{2H^2}{\eta^2} \log \frac{e^{\eta^2/(2H^2)}}{\varepsilon(e^{\eta^2/(2H^2)}-1)}$. Take the union of these events, we obtain the following upper bound:

$$\Pr\left[\bigcup_{t \geq T} \{Z_t(F^*) - Z_0(F^*) \geq t\eta\}\right] \leq \sum_{t \geq T} \exp\left(-\frac{t\eta^2}{2H^2}\right) \leq \varepsilon.$$

Moreover, for every $t < T$, we have

$$\Pr\left[Z_t(F^*) - Z_0(F^*) \geq \frac{2H^2 \log(T/\varepsilon)}{\eta^2}\right] \leq \frac{\varepsilon}{T}.$$

Take the union bound, we obtain

$$\Pr\left[\bigcup_{t \geq 1} \left\{l_t(F^*, F_0) - l_0(F^*, F_0) \geq t\eta + \frac{2H^2 \log(T/\varepsilon)}{\eta^2} + \sum_{t' \leq t} \mathbb{E}\left[X_{t'}(F^*) \mid h^{t'-1}\right]\right\}\right] \leq 2\varepsilon. \qquad \text{(C.3)}$$

Similarly, let $X'_t(F^*) \equiv l_t(F^*, F_1) - l_{t-1}(F^*, F_1)$, we have

$$\Pr\left[\bigcup_{t \geq 1} \left\{l_t(F^*, F_1) - l_0(F^*, F_1) \geq t\eta + \frac{2H^2 \log(T/\varepsilon)}{\eta^2} + \sum_{t' \leq t} \mathbb{E}\left[X'_{t'}(F^*) \mid h^{t'-1}\right]\right\}\right] \leq 2\varepsilon. \qquad \text{(C.4)}$$

Let

$$T'_0 = K_0(\varepsilon_1, \eta)(1 + \bar{\eta}_0(\varepsilon_1, \eta))\mathbb{E}[\tau_0(\bar{l} + l^* + 2\varepsilon_1)] + K_0(\varepsilon_1, \eta)\mathbb{E}[\tau_0^\varepsilon(l^*)]$$

$$T'_1 = K_1(\varepsilon_1, \eta)(1 + \bar{\eta}_1(\varepsilon_1, \eta))\mathbb{E}[\tau_1(\underline{l} + l^* - 2\varepsilon_1)] + K_1(\varepsilon_1, \eta)\mathbb{E}[\tau_1^\varepsilon(l^*)].$$

Note that the expected log likelihood of F^* when switching from action 0 to action 1 is $\sum_{y \in Y} F^*(y|a) \log F(y|a')$, which diverges to $-\infty$ as ζ' goes to 0, while the the expected log likelihood of $F \in \{F_0, F_1\}$ remains bounded. Thus, according to Claim 5, for sufficiently small ζ',

$$\sum_{t' \leq t} \mathbb{E}\left[X_{t'}(F^*) \mid h^{t'-1}\right] \leq T'_0 \cdot (\bar{Z} + \eta) - t \cdot \eta \text{ for every } t \in \mathbb{N}$$

occurs with probability at least $1 - 2\varepsilon_1$. Similar bound holds for $\sum_{t' \leq t} \mathbb{E}\left[X'_{t'}(F^*) \mid h^{t'-1}\right]$. Combining these

with inequalities (C.3) and (C.4), for sufficiently small $l_0(F^*, F_0), l_0(F^*, F_1)$, we have

$$\Pr\left[\bigcup_{t\geq 1}\{l_t(F^*, F_0) \geq \log \pi_\varepsilon \text{ or } l_t(F^*, F_1) \geq \log \pi_\varepsilon\}\right] \leq 6\varepsilon.$$

Therefore, the probability of the event that $\pi_t(F^*) < \pi_\varepsilon$ for every $t \in \mathbb{N}$ is at least $1 - 6\varepsilon$.   □

According to Claim 6, there exists a positive probability event under which the probability that the agent's posterior belief attaches to F^* is sufficiently small in all periods. Conditional on this event, Claim 5 implies that the agent's action cycles between 0 and 1. In the last step, we bound the asymptotic frequency of action 0:

$$\sup_{\sigma \in \Sigma^*(\pi_0)} \left\{ \limsup_{t \to +\infty} \mathbb{E}^\sigma\left[\frac{1}{t}\sum_{s=1}^{t}\mathbf{1}\{a_s = a^*\}\right]\right\} = \limsup_{k\to+\infty} \mathbb{E}\left[\frac{S^0_{t^0_k}}{S^1_{t^1_k} + S^0_{t^0_k}}\right]$$

$$\leq \limsup_{k\to+\infty} \frac{[k(1+\eta) + K_0(\varepsilon,\eta)(1 + \bar{\eta}_0(\varepsilon,\eta))]\mathbb{E}[\tau_0(\bar{l} + l^* + 2\varepsilon)]}{(k - K_1(\varepsilon,\eta))\cdot(1-\eta)\mathbb{E}[\tau_1^\varepsilon(l^*)]} + 8\varepsilon$$

$$= \frac{\mathbb{E}[\tau_0(\bar{l} + l^* + 2\varepsilon)]}{(1-\eta)\mathbb{E}[\tau_1^\varepsilon(l^*)]} + 8\varepsilon \leq \gamma.$$

The first inequality holds by directly applying Claims 5 and 6 and note that when the events in the claims fails, the expected frequency is at most 1. Moreover, for any $\gamma' < \gamma$, there exists $\zeta > 0$ in the construction of F_0, F_1 such that $\frac{\mathbb{E}[\tau_0(\bar{l}+l^*+2\varepsilon)]}{(1-\eta)\mathbb{E}[\tau_1^\varepsilon(l^*)]} < \gamma$ due to the fact that in our construction, $\mathbb{E}[\tau_0(\bar{l} + l^* + 2\varepsilon)]$ is bounded from above for any $\zeta > 0$ while $\mathbb{E}[\tau_1^\varepsilon(l^*)] \geq \frac{1}{2\log\frac{1-\zeta}{1-4\zeta}}$ approaches infinity as $\zeta \to 0$. Thus, the last inequality holds by simply setting ε, η, ζ to be small enough constants.

Remark: Note that proof of Theorem 1 does not hinge on the parameters in the design of the instance in (C.1). We summarize the important features of the construction of F^* and π_0:

1. The KL-divergence between $F^*(\cdot|0)$ and $F^*(\cdot|1)$ is sufficiently large.[12] This is sufficient to establish that the posterior probability of F^* converges to 0.

2. $F_1(\cdot|0)$ is closer to $F^*(\cdot|0)$ compared to $F_0(\cdot|0)$, and $F_0(\cdot|1)$ is closer to $F^*(\cdot|1)$ compared to $F_1(\cdot|1)$. This is sufficient to established that the action cycles between 0 and 1 for infinite periods if the agent does not believe F^* happens with high probability.

3. The expected log likelihood ratio $\mathbb{E}[X_{1\to 1}(F_0, F_1)]$ is sufficiently close to 0. This is sufficient to established that the number of time periods required for the agent to switch action from 1 to 0 is sufficiently

---

[12]KL-divergence is not symmetric, and it is sufficient to have either $D(F^*(\cdot|0)\,\|\,F^*(\cdot|1))$ or $D(F^*(\cdot|1)\,\|\,F^*(\cdot|0))$ is large.

large, which implies the limit frequency of action 0 is sufficiently small.

As is evident from Claim 5 and 6, essentially any instance satisfying those three properties is sufficient to show that the limit frequency of $a*$ is sufficiently small, and the example in (C.1) is an illustration that satisfies all three properties. Next we discuss the generalization the inefficiency result to broader settings.

- When $|Y| \geq 3$, there exists a subset $Y' \subseteq Y$ and $|Y'| = 2$ such that each distribution in the support of the agent's belief coincides for outcomes $y \in Y \backslash Y'$. For any $\gamma > 0$, by setting distributions F^*, F_0, F_1 such that (1) the probability that the realized outcome $y \in Y \backslash Y'$ is sufficiently small , and (2) the conditional distribution on Y' is the same as what we constructed for the case there are only two outcomes, we can show that the expected average frequency of choosing action a^* can be smaller than γ.

- When $|A| \geq 3$, there exists a subset $A' \subseteq A$ and $|A'| = 2$. For any $\gamma > 0$, by setting distributions F^*, F_0, F_1 such that (1) it is always suboptimal to choose any action $a \in A \backslash A'$ for any distribution π, and (2) the distribution when choosing action $a \in A'$ is the same as what we constructed for the case there are only two actions, we can show that the expected average frequency of choosing action a^* can be smaller than γ.

- For general time lag $k^* \neq k'$, we need to have an additional step to show that the number of periods before the agent switches the action is not always $|k^* - k' - 1|$. This is obvious when $k^* = 1$ and $k' = 0$. For the general case, we can show that the probability of such event happens is strictly between $(0, 1)$, and when the KL-divergence between $F^*(\cdot|0)$ and $F^*(\cdot|1)$ is sufficiently large, the attribution error is sufficiently large, and the posterior belief on F^* still converges to 0.

D   Proof of Theorem 2

Appendix D.1 characterizes the principal's payoff in the auxiliary game. Appendix D.2 establishes the connections between the principal's payoff in the auxiliary game and his payoff in the original game with symmetric uncertainty.

D.1   Payoff in the Auxiliary Game

This section examines the principal's asymptotic payoff in an auxiliary game in which he knows the true state but the agent is naive in the sense that she ignores the informational content of the principal's proposals

and updates her belief based only on the chosen policies and observed signals. For every $F \in \mathscr{F}$, let

$$\underline{u}(\sigma_p, F) \equiv \liminf_{t \to +\infty} \frac{1}{t} \mathbb{E}^{\sigma_p} \left[\sum_{s=1}^{t} a_s \middle| F \right] \tag{D.1}$$

and

$$\overline{u}(\sigma_p, F) \equiv \limsup_{t \to +\infty} \frac{1}{t} \mathbb{E}^{\sigma_p} \left[\sum_{s=1}^{t} a_s \middle| F \right] \tag{D.2}$$

be the lower and upper bounds on the principal's asymptotic payoffs when he uses strategy σ_p and the true state is F. Let

$$\underline{U}(F) \equiv \sup_{\sigma_p \in \Sigma_p} \underline{u}(\sigma_p, F) \quad \text{and} \quad \overline{U}(F) \equiv \sup_{\sigma_p \in \Sigma_p} \overline{u}(\sigma_p, F) \tag{D.3}$$

We establish two lemmas.

Lemma D.1. *For every $\pi_0 \in \Delta(\mathscr{F})$ that has full support, we have $\underline{U}(F_1) = \overline{U}(F_1) = 1$.*

Proof. Let $\pi_{t,i}$ be the posterior probability of distribution F_i according to the agent's belief in period t. Let $l_t \equiv \log \frac{\pi_{t,1}}{\pi_{t,0}}$. There exist a threshold l^* such that the agent chooses action $a_t = 0$ if and only if $l_t > l^*$.

We show that for any $\varepsilon > 0$, the following strategy for the principal achieves payoff at least $1 - \varepsilon$. The strategy of the principal is to always propose action 0 until the log likelihood satisfies $l_t > l^* + c$, where c is defined later in the analysis. The principal switches to always proposing action 1 if the above condition is satisfied. Note that when the principal propose action 0 for all periods, there is no attribution error, and the agent learns the correct distribution. By inequality (B.1), for any $\varepsilon_1 > 0$, any prior π_0, and any parameter c, there exists $T > 0$ such that with probability at least $1 - \varepsilon_1$, $l_T > l^* + c$. Thus with probability at least $1 - \varepsilon_1$, the principal switches to proposing action 1 before time T. Moreover, by Claim 3, for any $\varepsilon_2 > 0$, there exists $c > 0$ such that with probability at least $1 - \varepsilon_2$, $l_{t+T} > l_T - c$ for all $t > 0$. By setting $\varepsilon_1 = \varepsilon_2 = \frac{\varepsilon}{2}$ and apply the union bound, with probability at least $1 - \varepsilon$, we have $l_t > l^*$ for any $t > T$. Thus the payoff of the principal is at least $1 - \varepsilon$ with the given strategy. Taking $\varepsilon \to 1$ gives the desired bound. □

Lemma D.2. *For every $\pi_0 \in \Delta(\mathscr{F})$ that has full support,*

$$\underline{U}(F_0) = \overline{U}(F_0) = q^* \lambda. \tag{D.4}$$

The proof consists of two parts. In Section D.1.1, we show that

$$\sup_{\sigma_p \in \Sigma_p} \overline{u}(\sigma_p, F_0) \leq q^* \lambda \text{ for every } F_0 \in \mathscr{F}_0. \tag{D.5}$$

In Section D.1.2, we show that

$$\sup_{\sigma_p \in \Sigma_p} \underline{u}(\sigma_p, F_0) \geq q^* \lambda \text{ for every } F_0 \in \mathscr{F}_0. \tag{D.6}$$

D.1.1   Proof of Lemma D.2: Establish the Payoff Upper Bound

Let

$$\Pi_1 \equiv \left\{ \pi \in \Delta(\mathscr{F}) \,\Big|\, \text{argmax}_{i \in \{0,1\}} \Big\{ \sum_{F \in \mathscr{F}} \pi(F) \sum_{y \in Y} F(y|i) v(y) \Big\} = \{1\} \right\}$$

be the set of beliefs under which the agent strictly prefers action 1. Claim 3 implies that there exists $\underline{p} > 0$ such that for every $\pi_t \notin \Pi_1$ and $\sigma_p \in \Sigma_p$, we have

$$\Pr\left(\pi_s \notin \Pi_1 \text{ for every } s \geq t \,\Big|\, F_0, \sigma_p \right) > \underline{p}. \tag{D.7}$$

For every $k \in \mathbb{N}$, we say that π_t crosses Π_1 in period k (or equivalently, there is a *crossing* in period k) if $\pi_{k-1} \in \Pi_1$ and $\pi_k \notin \Pi_1$, or $\pi_{k-1} \notin \Pi_1$ and $\pi_k \in \Pi_1$. The uniform lower bound in (D.7) implies that for every $\sigma_p \in \Sigma_p$, the expected number of crossings is finite almost surely. Therefore,

$$\Pr\left(\exists t \in \mathbb{N} \text{ s.t. } \pi_s \notin \Pi_1 \text{ for every } s \geq t \,\Big|\, F_0, \sigma_p \right) + \Pr\left(\underbrace{\exists t \in \mathbb{N} \text{ s.t. } \pi_s \in \Pi_1 \text{ for every } s \geq t}_{\equiv \text{event } \mathscr{E}_{\sigma_p}} \,\Big|\, F_0, \sigma_p \right) = 1 \tag{D.8}$$

Let $\mathscr{E}_{\sigma_p}$ be the event that there exists $t \in \mathbb{N}$ such that $\pi_s \in \Pi_1$ for every $s \geq t$. Let σ_p^* be the strategy of the principal for maximizing probability of event $\mathscr{E}_{\sigma_p^*}$ given prior π_0. By definition, q^* is the probability of event $\mathscr{E}_{\sigma_p^*}$ when the principal uses strategy σ_p^*.

The principal's asymptotic payoff conditional on event $\{\exists t \in \mathbb{N} \text{ s.t. } \pi_s \notin \Pi_1 \text{ for every } s \geq t\}$ is zero. We conclude the proof by showing that his asymptotic payoff conditional on event $\mathscr{E}_{\sigma_p}$ is at most λ for every $\sigma_p \in \Sigma_p$ satisfying $\Pr[\mathscr{E}_{\sigma_p}|F_0, \sigma_p] > 0$.

Suppose toward a contradiction that there exists $\varepsilon > 0$ such that conditional on $\mathscr{E}_{\sigma_p}$, the asymptotic frequency of policy 1 is more than $\lambda + \varepsilon$ when the true state is F_0. First, we observe that the asymptotic frequency of $(a_{t-1}, a_t) = (1, 0)$ equals that of $(a_{t-1}, a_t) = (0, 1)$ regardless of the principal's strategy σ_p. The definition of λ then implies that

$$\lim_{t \to +\infty} \mathbb{E}\left[\log \frac{\pi_t(F_0)}{\pi_t(F_1)} \,\Big|\, \mathscr{E}_{\sigma_p}, \sigma_p \right] = +\infty. \tag{D.9}$$

As a result, the agent strictly prefers action 0 asymptotically when the principal uses strategy σ_p conditional

on event $\mathscr{E}_{\sigma_p}$. This contradicts the definition of event $\mathscr{E}_{\sigma_p}$ under which the agent strictly prefers policy 1.

D.1.2 Proof of Lemma D.2: Attain the Payoff Upper Bound

We construct $\sigma_p^\varepsilon \in \Sigma_p$ for every $\varepsilon > 0$ such that

$$\underline{u}(\sigma_p^\varepsilon, F_0) \geq q^* \lambda - \varepsilon. \tag{D.10}$$

For every $\sigma_p \in \Sigma_p$ and $l \in \mathbb{R}$, let $\mathscr{E}_{\sigma_p, l}$ be the following event when the principal uses strategy σ_p,

$$\log \frac{\pi_s(F_1)}{\pi_s(F_0)} \geq l \text{ for every } s \geq t \text{ for some } t \in \mathbb{N}. \tag{D.11}$$

If $\Pi(l) \subset \Delta(\mathscr{F})$ be the set of beliefs that satisfy (D.11). Recall the definition of $\mathscr{E}_{\sigma_p}$ in (D.8), which implies the existence of $l^* \in \mathbb{R}_+$ such that $\mathscr{E}_{\sigma_p, l} \subset \mathscr{E}_{\sigma_p}$ and $\Pi(l) \subset \Pi_1$ for every $l \geq l^*$. Recall that σ_p^* is the strategy that maximizes the probability of event $\mathscr{E}_{\sigma_p^*}$ given prior π_0..

Lemma D.3. *For every $\pi_0 \in \Delta(\mathscr{F})$ that has full support and $l \in \mathbb{R}$, we have $\Pr[\mathscr{E}_{\sigma_p^*, l} | F_0, \sigma_p^*] = q^*$.*

Proof of Lemma D.3: As shown before, there exists $\underline{p} > 0$ such that for every $\pi_t \notin \Pi_1$, the probability with which $\pi_s \notin \Pi_1$ for every $s \geq t$ is at least $\underline{p}$ when the true state is F_0. As a result, for every $l \in \mathbb{R}_+$, the probability of the following event is zero under any strategy in Σ_p:

- $\pi_s \in \Pi_1 \backslash \Pi(l)$ for every $s \geq t$.

This implies that for every $l \in \mathbb{R}$ that satisfies $\Pi(l) \subset \Pi_1$, we have

$$\Pr\left(\exists t \in \mathbb{N} \text{ s.t. } \pi_t \in \Pi(l) \text{ for all } s \geq t \Big| F_0, \sigma_p^*\right) = \Pr\left(\exists t \in \mathbb{N} \text{ s.t. } \pi_t \in \Pi_1 \text{ for all } s \geq t \Big| F_0, \sigma_p^*\right). \tag{D.12}$$

and moreover,

$$\Pr\left(\exists t \in \mathbb{N} \text{ s.t. } \pi_t \in \Pi_1 \text{ for all } s \geq t \Big| F_0, \sigma_p^*\right) + \Pr\left(\exists t \in \mathbb{N} \text{ s.t. } \pi_t \notin \Pi_1 \text{ for all } s \geq t \Big| F_0, \sigma_p^*\right) = 1. \tag{D.13}$$

Equations (D.12) and (D.13) together imply that $\Pr[\mathscr{E}_{\sigma_p^*, l} | F_0, \sigma_p^*] = \Pr[\mathscr{E}_{\sigma_p^*} | F_0, \sigma_p^*]$, while the latter equals q^*. $\square$

Next we focus on the case when $\lambda > 0$ since the case $\lambda = 0$ is trivial. In this case, we know that $\mathbb{E}[X_{1\to 0} + X_{0\to 1}] > 0$ since $\mathbb{E}[X_{1\to 1}] < 0$.

For small enough $\varepsilon > 0$, let $T_1, T_2 \in \mathbb{N}$ be two positive integers such that T_1 is even and $\frac{\frac{T_1}{2}+T_2}{T_1+T_2} \in \left(\lambda - \varepsilon, \lambda\right)$. Let $T \equiv T_1 + T_2$. Let $\overline{\sigma}_p \in \Sigma_p$ be defined as:

- $\overline{\sigma}_p(h^t) = 0$ if there exists $k \in \mathbb{N}$ such that $t \in \{kT+2, kT+4, ..., kT+T_1\}$,

- $\overline{\sigma}_p(h^t) = 1$ otherwise.

According to $\overline{\sigma}_p$, the frequency with which the principal proposes policy 1 belongs to the interval $(\lambda - \varepsilon, \lambda)$.

Let $l_t \equiv \log \frac{\pi_t(F_1)}{\pi_t(F_0)}$, and let X_T be the increment of l_t from period t to $t+T$ when policy 0 is chosen in period $t+2, t+4, ..., t+T_1$ and policy 1 is chosen in other periods. Let $\overline{H}$ be the maximal realization of X_T.

Let $r^* > 0, \eta > 0$ be such that $\mathbb{E}_{z \sim X_T}[\exp(r^* z)] = 1$ and $\exp(-r^* \cdot \eta) < \varepsilon$. Let $\overline{l} \in \mathbb{R}$ be large enough such that $\Pi(\overline{l}) \subset \Pi_1$. Recall the definition of σ_p^*. Let $\sigma_p^\varepsilon \in \Sigma_p$ be defined as:

- $\sigma_p^\varepsilon(h^t) = \sigma_p^*(h^t)$ if $\pi_t \in \Pi(\overline{l}+\eta+\overline{H})$ for all $t' < t$;

- $\sigma_p^\varepsilon(h^t) = \overline{\sigma}_p(h^t)$ otherwise.

Conditional on π_t reaches $\Pi(\overline{l}+\eta+\overline{H})$, the Wald's inequality in Lemma A.1 implies that the probability with which $\pi_s \in \Pi(\overline{l})$ for every $s \geq t$ is at least $1 - \varepsilon$, which implies that the principal's asymptotic payoff is at least $q^*(\lambda - \varepsilon)$ when the true state is F_0.

D.2   Connections between Auxiliary Game & Original Game

We show that the principal's payoff in the original game equals his expected payoff in the auxiliary game studied in Appendix D.1. First, we show that the principal learns the true state asymptotically regardless of the chosen policies.

Lemma D.4. *For every* $\sigma_p \in \Sigma_p$, $F \in \mathscr{F}$, *and* $\varepsilon > 0$, *there exists* $\tau \in \mathbb{N}$ *such that*

$$\Pr\left(\pi_\tau(F) > 1 - \varepsilon \,\middle|\, F\right) > 1 - \varepsilon. \tag{D.14}$$

Proof of Lemma D.4: Let Q_F be the probability measure over $\mathscr{H}$ induced by distribution F and let Q_p be the probability measured over $\mathscr{H}$ induced by the principal's prior belief $\pi_0 \in \Delta(\mathscr{F})$. For every history h^t, let $q_{F|h^t} \in \Delta(A \times Y)$ be the principal's belief about (a_t, y_t) conditional on the true state being F, and let $q_{\pi_t|h^t} \in \Delta(A \times Y)$ be the principal's belief about (a_t, y_t) when his belief about the state is $\pi_t \in \Delta(\mathscr{F})$. The chain rule for relative entropy implies that

$$-\log \pi_0(F) \geq d\left(Q_F \,\middle\|\, Q_p\right) = \sum_{t=0}^{\infty} \mathbb{E}_{Q_F}\left[d\left(q_{F|h^t} \,\middle\|\, q_{\pi_t|h^t}\right)\right]. \tag{D.15}$$

Conditional 2 implies that $d(q_{F|h^t}\|q_{F'|h^t}) > 0$ for every $F \neq F'$. Since F is finite, for every $\varepsilon > 0$, there exists $\eta > 0$ such that $d(q_{F|h^t}\|q_{p|h^t}) > \eta$ for every $\pi_0 \in \Delta(\mathscr{F})$ satisfying $\pi_0(F) \leq 1 - \varepsilon$. Inequality (D.15) implies the existence of $\tau \in \mathbb{N}$ such that

$$\sum_{t=\tau}^{\infty} \mathbb{E}_{Q_F}\left[d\left(q_{F|h^t}\|q_{\pi_t|h^t}\right)\right] \leq \eta\varepsilon. \tag{D.16}$$

The Markov's inequality implies that the probability with which $d(q_{F|h^t}\|q_{\pi_t|h^t}) > \eta$ is strictly less than ε for every $t \geq \tau$, or equivalently, the probability with which $\pi_t(F) \leq 1 - \varepsilon$ is less than ε for every $t \geq \tau$. $\square$

Lemma D.5. *We have* $\overline{V} = \underline{V} = \sum_{F \in \mathscr{F}} \pi_0(F)\overline{U}(F) = \sum_{F \in \mathscr{F}} \pi_0(F)\underline{U}(F)$.

Proof of Lemma D.5: Since $\overline{U}(F) = \underline{U}(F)$ for every $F \in \mathscr{F}$, we have $\overline{V} \leq \sum_{F \in \mathscr{F}} \pi_0(F)\overline{U}(F)$. We show $\underline{V} \geq \sum_{F \in \mathscr{F}} \pi_0(F)\overline{U}(F)$ by constructing a strategy σ_p^{ε} for every $\varepsilon > 0$ under which $\underline{V}(\sigma_p^{\varepsilon}) \geq \sum_{F \in \mathscr{F}} \pi_0(F)\overline{U}(F) - \varepsilon$.

For every $\varepsilon > 0$, let $\tau \in \mathbb{N}$ be such that $\Pr\left(\pi_\tau(F) > 1 - \varepsilon \middle| F\right) > 1 - \varepsilon$, and let $\sigma_p^{F_0}(\varepsilon) \in \Sigma_p$ be the strategy under which the principal obtains utility $\underline{u}(\sigma_p^{F_0}(\varepsilon), F_0) \geq \overline{U}(F_0) - \varepsilon$ if the true state is F_0. Such a strategy exists according to Lemma D.2. Similarly, let $\sigma_p^{F_1}(\varepsilon)$ be the strategy under which the principal obtains utility $\underline{u}(\sigma_p^{F_1}(\varepsilon), F_1) \geq 1 - \varepsilon$. Let $\sigma_p^{\varepsilon} \in \Sigma_p$ be a strategy under which

- The principal follows $\sigma_p^{F_0}(\varepsilon)$ for every $t \leq \tau$.

- If $\pi_\tau(F_0) \geq 1 - \varepsilon$, then the principal follows $\sigma_p^{F_0}(\varepsilon)$ starting from period τ.

- Otherwise, he follows $\sigma_p^{F_1}(\varepsilon)$ starting from period τ.

Next, we establish a lower bound on the principal's asymptotic payoff from strategy σ_p^{ε}. Conditional on the true state is F_0, the probability with which the principal plays $\sigma_p^F(\varepsilon, \tau)$ is at least $1 - \varepsilon$. Conditional on the true state is F_1, there exists $T > 0$ such that the probability with which the principal proposes 1 in every period after $\tau + T$ is greater than $1 - \varepsilon$ according to Lemma D.1. As a result, the principal's asymptotic payoff from σ_p^{ε} is at least $(1 - \varepsilon)\sum_{F \in \mathscr{F}} \pi_0(F)\overline{U}(F)$. Since the principal's stage-game payoff is between 0 and 1, we have $(1 - \varepsilon)\sum_{F \in \mathscr{F}} \pi_0(F)\overline{U}(F) \geq \sum_{F \in \mathscr{F}} \pi_0(F)\overline{U}(F) - \varepsilon$. $\square$

Generalizations.   Finally, we discuss the generalization of our result in broader settings.

1. All results in this section does not hinge on the fact that $k^* = 1$ and $k' = 0$. In fact, all the lemmas and claims hold directly for general time lags.

2. When $|\mathscr{F}| > 2$, we denote $\mathscr{F}_0$ as the set of distributions with optimal action 0 for the agent and $\mathscr{F}_1$ as the set of distributions with optimal action 1 for the agent. The results directly generalize when $|\mathscr{F}_1| > 1$. As we observe from Lemma D.1, the payoff of the principal in the auxiliary game does not depend on the prior when the true state is in $\mathscr{F}_1$. Thus the principal can simply learn the true distribution with high probability as described in Lemma D.5. However, things are more complicated when $|\mathscr{F}_0| > 1$. The main reason is that the payoff of the principal depends on the prior π_0 in the auxiliary game when the true state is in $\mathscr{F}_1$. When the principal faces uncertainly over $\mathscr{F}_0$, if there does not exist a strategy σ_p that maximizes the probability of the event $\mathscr{E}_{\sigma_p}$ simultaneously for all $F_0 \in \mathscr{F}_0$, the principal suffers a non-negligible utility loss in the process of learning the true state.

References

Robert J Aumann and Michael Maschler. *Repeated games with incomplete information*. MIT press, 1995.

Robert Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, pages 51–58, 1966.

Aislinn Bohren and Daniel N Hauser. Learning with model misspecification: Characterization and robustness. *working paper*, 2020.

Persi Diaconis and David Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

Kfir Eliaz and Ran Spiegler. A model of competing narratives. *American Economic Review*, 110(12): 3786–3816, 2020.

Ignacio Esponda and Demian Pouzo. Berk-nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3):1093–1130, 2016.

Ignacio Esponda and Demian Pouzo. Equilibrium in misspecified markov decision processes. *Theoretical Economics*, forthcoming, 2020.

Ignacio Esponda, Demian Pouzo, and Yuichi Yamamoto. Asymptotic behavior of Bayesian learners with misspecified models. *arXiv preprint arXiv:1904.08551*, 2020.

Françoise Forges. Repeated games of incomplete information: non-zero-sum. *Handbook of game theory with economic applications*, 1:155–177, 1992.

Mira Frick, Ryota Iijima, and Yuhta Ishii. Stability and robustness in misspecified learning models. *working paper*, 2020.

Drew Fudenberg and Jean Tirole. Perfect Bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory*, 53(2):236–260, 1991.

Drew Fudenberg, Gleb Romanyuk, and Philipp Strack. Active learning with a misspecified prior. *Theoretical Economics*, 12(3):1155–1189, 2017.

Drew Fudenberg, Giacomo Lanzani, and Philipp Strack. Limits points of endogenous misspecified learning. *working paper*, 2020.

Sergiu Hart. Nonzero-sum two-person repeated games with incomplete information. *Mathematics of Operations Research*, 10(1):117–153, 1985.

Kevin He. Mislearning from censored data: The gambler's fallacy in optimal-stopping problems. *Working Paper*, 2020.

Philippe Jehiel and Larry Samuelson. Reputation with analogical reasoning. *The Quarterly Journal of Economics*, 127(4):1927–1969, 2012.

Yaonan Jin, Yingkai Li, Yining Wang, and Yuan Zhou. On asymptotically tight tail bounds for sums of geometric and exponential random variables. *arXiv preprint arXiv:1902.02852*, 2019.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Pooya Molavi. Macroeconomics with learning and misspecification: A general theory and applications. *Working Paper*, 2020.

Yaw Nyarko. Learning in mis-specified models and the possibility of cycles. *Journal of Economic Theory*, 55(2):416–427, 1991.

Hazhir Rahmandad, Nelson Repenning, and John Sterman. Effects of feedback delay on learning. *System Dynamics Review*, 25(4):309–338, 2009.

Nelson Repenning and John Sterman. Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly*, 47:265–295, 2002.

Cosma Rohilla Shalizi. Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1039–1074, 2009.

Ran Spiegler. Placebo reforms. *American Economic Review*, 103(4):1490–1506, 2013.

Abraham Wald. On cumulative sums of random variables. *The Annals of Mathematical Statistics*, 15(3):283–296, 1944.