Esponda, Ignacio; Pouzo, Demian

Article

Equilibrium in misspecified Markov decision processes

Theoretical Economics

Provided in Cooperation with:
The Econometric Society

Suggested Citation: Esponda, Ignacio; Pouzo, Demian (2021) : Equilibrium in misspecified Markov decision processes, Theoretical Economics, ISSN 1555-7561, The Econometric Society, New Haven, CT, Vol. 16, Iss. 2, pp. 717-757, https://doi.org/10.3982/TE3843

This Version is available at:
https://hdl.handle.net/10419/253497

Theoretical Economics 16 (2021), 717–757 1555-7561/20210717

Equilibrium in misspecified Markov decision processes

Ignacio Esponda
Department of Economics, University of California

Demian Pouzo
Department of Economics, University of California

We provide an equilibrium framework for modeling the behavior of an agent who holds a simplified view of a dynamic optimization problem. The agent faces a Markov decision process, where a transition probability function determines the evolution of a state variable as a function of the previous state and the agent's action. The agent is uncertain about the true transition function and has a prior over a set of possible transition functions; this set reflects the agent's (possibly simplified) view of her environment and may not contain the true function. We define an equilibrium concept and provide conditions under which it characterizes steady-state behavior when the agent updates her beliefs using Bayes' rule.

Keywords. Misspecified model, Markov decision process, equilibrium.

JEL classification. C61, D83.

1. Introduction

Early interest in studying the behavior of agents who hold misspecified views of the world (e.g., Arrow and Green 1973, Kirman 1975, Sobel 1984, Kagel and Levin 1986, Nyarko 1991, Sargent 1999) has recently been renewed by the work of Piccione and Rubinstein (2003), Jehiel (2005), Eyster and Rabin (2005), Jehiel and Koessler (2008), Esponda (2008), Esponda and Pouzo (2016, 2017, 2019), Eyster and Piccione (2013), Spiegler (2013, 2016, 2017), Fudenberg et al. (2017), Heidhues et al. (2018, 2021) and Eliaz and Spiegler (2020), among others. There are least two reasons for this interest. First, it is natural for agents to be uncertain about their complex environment and to represent this uncertainty with parsimonious parametric models that are likely to be misspecified. Second, endowing agents with misspecified models can explain how certain biases in behavior arise endogenously as a function of the primitives.

The previously cited papers focus on problems that are intrinsically "static" in the sense that they can be viewed as repetitions of static problems where the only link between periods arises because the agent is learning the parameters of the model. Yet dynamic decision problems, where an agent chooses an action that affects a state variable

Ignacio Esponda: iesponda@ucsb.edu
Demian Pouzo: dpouzo@econ.berkeley.edu
We thank Vladimir Asriyan, Hector Chade, Xiaohong Chen, Emilio Espino, Drew Fudenberg, Bruce Hansen, Philippe Jehiel, Bart Lipman, Jack Porter, Philippe Rigollet, Tom Sargent, Ran Spiegler, Philipp Strack, Iván Werning, and several seminar participants for helpful comments.

 https://doi.org/10.3982/TE3843

(other than a belief), are ubiquitous in economics. The goal of this paper is to provide a tractable framework to study dynamic settings where the agent has a possibly misspecified model.

We study a Markov decision process where a single agent chooses actions at discrete time intervals. A transition probability function describes how the agent's action and the current state affects the next period's state. The current payoff is a function of states and actions. As is well known, this problem can be represented recursively via the Bellman equation

$$V(s) = \max_{x \in \Gamma(s)} \pi(s,x) + \delta \int_{\mathbb{S}} V\big(s'\big) Q\big(ds' \mid s, x\big), \tag{1}$$

where s is the current state, x is the agent's choice variable from a feasible set $\Gamma(s)$, π is the payoff function, Q is the transition probability function, and δ is the discount factor.

In realistic environments, the agent often has to deal with two difficult issues: a potentially large state space (i.e., the curse of dimensionality) and uncertainty about the transition probability function. For example, (1) may represent a dynamic savings problem where the agent decides every period what fraction x of her wealth to save. The state variable s is a vector that includes wealth as well as any variable that helps predict returns to savings, such as previous interest rates and other macroeconomic indicators. The function Q represents the return function, and, naturally, the agent may not even be sure which indicators are relevant in predicting returns. In such a complex environment, it is reasonable to expect the agent to simplify the problem and focus only on certain variables by solving a version of (1) where Q is replaced by a "simpler" transition function.

The main objective of this paper is to provide a framework for modeling the behavior of an agent who holds a simplified view of the dynamic optimization problem represented by (1). Our approach is to postulate that the agent is endowed with a family of transition probability functions, $\{Q_\theta : \theta \in \Theta\}$, indexed by a parameter space Θ. This family captures both the uncertainty of the agent as well as the way in which she simplifies the problem. In particular, the agent's model is misspecified whenever the true model Q is not in $\{Q_\theta : \theta \in \Theta\}$. For example, the agent may incorrectly believe that certain macroeconomic indicators are irrelevant for predicting returns, but she may still be uncertain as to the predictive value of the remaining indicators. Each period, the agent observes the current state, chooses an action, and then updates her belief using Bayes' rule when the new state is realized.

Our main contribution is to introduce an equilibrium concept to describe the steady states of the agent's learning dynamics when the agent is a Bayesian learner with a misspecified model. To characterize the agent's steady-state behavior, the modeler simply solves problem (1), except that the true transition function Q is replaced by the agent's perception of this transition, $\bar{Q}_{\mu^*} = \int_\Theta Q_\theta \mu^*(d\theta)$, where μ^* is interpreted as the agent's equilibrium belief over all models in Θ. As any other equilibrium object, the equilibrium belief μ^* is determined endogenously. In addition to gaining tractability, we focus on equilibrium behavior because it is standard in economics and allows us to relate our

findings to previous work, and also because we are interested in the long-run implications of model misspecification and not necessarily on mistakes that arise from limited opportunities to learn.

We say that a probability distribution over state–action pairs is a Berk–Nash equilibrium if it satisfies two requirements. First, there exists a belief over Θ such that, for any state–action pair in the support of the equilibrium distribution, the agent's action given the state is optimal given the belief, and, moreover, the belief puts probability 1 on the set of parameter values that are "closest" to the true transition probability function over state–action pairs. The notion of closest is formalized by a weighted version of Kullback–Leibler divergence, where the weights in turn depend on the equilibrium distribution. Second, the agent's equilibrium behavior gives rise to a particular Markov process over states and actions, and we require the equilibrium distribution to be a stationary distribution of this process.

We then illustrate how our equilibrium concept can help analyze environments that seemed previously intractable using three examples. First, we consider the problem of an agent facing a dynamic effort task who fails to take into account that his performance today is affected by his performance yesterday. Second, we consider a stochastic growth model where the agent incorrectly assumes that productivity and preference shocks are independent. Finally, we consider a production problem with Markov shocks and uncertain cost, where the decision maker has an incorrect parametric specification of the cost function.

We conclude by investigating one possible foundation for our equilibrium concept. Consider the case where the agent has a prior belief μ over Θ that is updated using Bayes' rule based on the current state, the agent's decision, and the state observed next period, $\mu' = B(s, x, s', \mu)$, where B denotes the Bayesian operator and μ' is the posterior belief. One convenience of Bayesian updating is that we can represent this problem recursively via the following Bellman equation, where the state variable now also includes the agent's belief:

$$W(s,\mu) = \max_{x\in\Gamma(s)} \pi(s,x) + \delta \int\int W\big(s',\mu'\big) Q_\theta\big(ds' \mid s,x\big)\mu(d\theta), \tag{2}$$

where $\mu' = B(s, x, s', \mu)$ is the updated belief.

In this environment, a natural question is whether the limiting distribution of state–action pairs corresponds to a Berk–Nash equilibrium. In the static case, where there is no state variable s, the answer has been shown to be yes under fairly mild assumptions (see Esponda and Pouzo 2016). A remarkable feature of this result, which is shared by other equilibrium foundations, such as the foundation for Nash and self-confirming equilibrium (e.g., Fudenberg and Kreps 1993, 1995), is that the modeler does not need to tackle the problem of belief updating so as to characterize limiting behavior, but rather applies a fixed equilibrium belief.

In the dynamic environments that we study in this paper, the answer to our question is more nuanced. We show that the answer is positive if one of three conditions is satisfied. The first condition is that the environment is subjectively static, in the sense that the the agent believes (possibly incorrectly) that the current state does not affect

the future state. The second condition is that the environment is identified, a condition that essentially requires that the agent's belief is uniquely determined irrespective of the agent's action.[1] The third condition is that all states are visited with positive probability in the steady state. At least one of these three conditions is typically satisfied in applications. We show by example that if neither of these conditions is satisfied, then steady states cannot generally be characterized by an equilibrium approach where the agent holds a fixed, equilibrium belief, and this is true even if the agent's model is correctly specified. In contrast, the modeler is forced to consider the more complicated problem with belief updating, as represented by (2). As we explain in Section 5, the difference in results between the static and the dynamic settings arises from the fact that updating a belief can never decrease the agent's continuation value in the static case (because of a nonnegative value of experimentation), but it may decrease it when both the belief and another state variable change.

A few other people have also studied the problem of misspecified learning by economic agents outside the traditional static setting where one agent repeatedly faces the same problem every period. Blume and Easley (1998, Section 5) study a competitive economy. Bohren and Hauser (2017) and Frick et al. (2020b) study social learning environments. Rabin and Vayanos (2010) and Ortoleva and Snowberg (2015) study environments with misspecification in nonindependent and nonidentically distributed settings where own actions do not affect beliefs (i.e., passive learning). He (2018) studies misspecification in an optimal stopping problem. Molavi (2019) considers a recursive general-equilibrium framework that nests a class of macroeconomics models in which agents learn with misspecified models.[2] With the exception of some stochastic growth problems (e.g., Koulovatianos et al. 2009), there are very few applications of the types of misspecified, active learning Markovian decision environments we consider in this paper. By proposing a tractable equilibrium approach, we hope to stimulate applications in this area.

More generally, the paper is related to the literature that provides learning foundations for equilibrium concepts, such as Nash or self-confirming equilibrium (see Fudenberg and Levine 1998 for a survey). In contrast to this literature, we consider Markov decision problems and allow for misspecified models. Particular types of misspecifications have been studied in extensive form games. Jéhiel (1995) considers the class of repeated alternating-move games and assumes that players forecast only a limited number of time periods into the future; see Jéhiel (1998) for a learning foundation.[3] We share the feature that the learning process takes place within the play of the game and that beliefs are those that provide the best fit given the data. As with much of this literature, our learning foundation for the equilibrium concept does not guarantee that

[1] Identification rules out situations where beliefs are incorrect due to lack of experimentation, which is a hallmark of the bandit (e.g., Rothschild 1974, McLennan 1984, Easley and Kiefer 1988) and the self-confirming equilibrium (e.g., Battigalli 1987, Fudenberg and Levine 1993, Dekel et al. 2004, Fershtman and Pakes 2012) literatures.

[2] In macroeconomics, there are several models where agents make forecasts using statistical models that are misspecified (e.g., Evans and Honkapohja 2001, Chapter 13, Sargent 1999, Chapter 6).

[3] Jehiel and Samet (2007) consider the general class of extensive form games with perfect information and assume that players simplify the game by partitioning the nodes into similarity classes.

behavior converges to the equilibrium, but only that if it converges, it must converge to an equilibrium; see Section 5.2 for further discussion.

Finally, a particular class of examples that fit our framework involve a typical coarseness misspecification or a type of correlation neglect that has been studied in previous frameworks, such as analogy-based expectation equilibrium (Jehiel 2005, Jehiel and Koessler 2008) and Bayesian networks (Spiegler 2016, 2017).

The framework and equilibrium notion are presented in Sections 2 and 3. In Section 4, we work through several examples, and in Section 5 we provide a foundation for the equilibrium notion.

2. Markov decision processes

We begin by describing the environment faced by the agent.

Definition 1. A *Markov decision process* (MDP) is a tuple $\langle \mathbb{S}, \mathbb{X}, q_0, Q, \pi, \delta \rangle$, where

- $\mathbb{S}$ is a nonempty and finite set of states
- $\mathbb{X}$ is a nonempty and finite set of actions
- $q_0 \in \Delta(\mathbb{S})$ is a probability distribution on the initial state
- $Q : \mathbb{S} \times \mathbb{X} \to \Delta(\mathbb{S})$ is a transition probability function
- $\pi : \mathbb{S} \times \mathbb{X} \times \mathbb{S} \to \mathbb{R}$ is a per-period payoff function
- $\delta \in [0, 1)$ is a discount factor.

We sometimes use MDP(Q) to denote an MDP with transition probability function Q and exclude the remaining primitives.

The timing is as follows. At the beginning of each period $t = 0, 1, 2, \ldots$, the agent observes state $s_t \in \mathbb{S}$ and chooses an action $x_t \in \mathbb{X}$. (It is straightforward to incorporate a feasible set of actions that depends on the state.) Then a new state s_{t+1} is drawn according to the probability distribution $Q(\cdot \mid s_t, x_t)$ and the agent receives payoff $\pi(s_t, x_t, s_{t+1})$ in period t. The initial state s_0 is drawn according to the probability distribution q_0. As usual, the objective of the agent is to choose a feasible policy rule to maximize expected discounted utility, $\sum_{t=0}^{\infty} \delta^t \pi(s_t, x_t, s_{t+1})$.

By the principle of optimality, the agent's problem can be cast recursively as

$$V(s) = \max_{x \in \mathbb{X}} \int_{\mathbb{S}} \{\pi(s, x, s') + \delta V(s')\} Q(ds' | s, x), \tag{3}$$

where $V : \mathbb{S} \to \mathbb{R}$ is the (unique) solution to the Bellman equation (3).

Definition 2. An action x *is optimal given* s in the MDP(Q) if

$$x \in \arg\max_{\hat{x} \in \mathbb{X}} \int_{\mathbb{S}} \{\pi(s, \hat{x}, s') + \delta V(s')\} Q(ds' | s, \hat{x}).$$

3. Subjective Markov decision processes

Our main objective is to study the behavior of an agent who faces an MDP but is uncertain about the transition probability function. We begin by introducing a new object to model the problem with uncertainty, which we call the subjective Markov decision process (SMDP). We then define the notion of a Berk–Nash equilibrium of an SMDP.

3.1 *Setup*

Definition 3. A *subjective Markov decision process* (SMDP) is an MDP, $\langle \mathbb{S}, \mathbb{X}, q_0, Q, \pi, \delta \rangle$, and a nonempty family of transition probability functions, $\mathcal{Q}_\Theta = \{Q_\theta : \theta \in \Theta\}$, where each transition probability function $Q_\theta : \mathbb{S} \times \mathbb{X} \to \Delta(\mathbb{S})$ is indexed by a parameter value $\theta \in \Theta$.

We interpret the set $\mathcal{Q}_\Theta$ as the different transition probability functions (or models of the world) that the agent considers possible. We sometimes use SMDP$(Q, \mathcal{Q}_\Theta)$ to denote an SMDP with true transition probability function Q and a family of transition probability functions $\mathcal{Q}_\Theta$.

Definition 4. An SMDP$(Q, \mathcal{Q}_\Theta)$ is *misspecified* if $Q \notin \mathcal{Q}_\Theta$; otherwise, it is *correctly specified*. It is *subjectively static* if π and all elements in $\mathcal{Q}_\Theta$ do not depend on the current state. It is *static* if, in addition to being subjectively static, the true transition probability function Q does not depend on the current state.

An SMDP describes the agent's subjective perception of the environment. In particular, the agent has a correct perception of the state space, the action space, and the payoff function, but she is uncertain about the transition probability function. The static case was previously studied by Esponda and Pouzo (2016). An SMDP is subjectively static if the agent believes it is static, even though it might not actually be a static environment. This property plays an important role in one of our main results.

Definition 5. A *regular subjective Markov decision process* (regular-SMDP) is an SMDP that satisfies the following conditions:

- The set Θ is a compact subset of an Euclidean space.
- The function $Q_\theta(s' \mid s, x)$ is continuous as a function of $\theta \in \Theta$ for all $(s, x, s') \in \mathbb{S} \times \mathbb{X} \times \mathbb{S}$.
- There is a dense set $\hat{\Theta} \subseteq \Theta$ such that, for all $\theta \in \hat{\Theta}$, $Q_\theta(s' \mid s, x) > 0$ for all $(s, x, s') \in \mathbb{S} \times \mathbb{X} \times \mathbb{S}$ such that $Q(s' \mid s, x) > 0$.

The first two conditions in Definition 5 place parametric and continuity assumptions on the subjective models.[4] The last condition plays two roles. First, it rules out

[4]Without the assumption of a finite-dimensional parameter space, Bayesian updating need not converge to the truth for most priors and parameter values even in correctly specified statistical settings (Freedman 1963, Diaconis and Freedman 1986). Note that the parametric assumption is only a restriction if the set of states or actions is nonfinite, a case we consider in some of the examples.

a stark form of misspecification by guaranteeing that there exists at least one parameter value that can rationalize every feasible observation. Second, it implies that the correspondence of parameters that are a closest fit to the true model, to be defined in the next section, is upper hemicontinuous, which, in particular, implies the existence of equilibrium.

3.2 *Equilibrium*

The goal of this section is to ne the notion of Berk–Nash equilibrium of an SMDP. The goal of the solution concept is to predict a distribution over outcomes (meaning state–action pairs), $m \in \Delta(\mathbb{S} \times \mathbb{X})$, as a function of the primitives of the environment. In Section 5, we interpret an equilibrium distribution over state–action pairs as the limiting frequency of state–action pairs in an environment where the agent is Bayesian and updates her belief about the transition probability function in each period.

NOTATION. For a given probability distribution over state–action pairs, $m \in \Delta(\mathbb{S} \times \mathbb{X})$, we denote the marginal over $\mathbb{S}$ by $m_{\mathbb{S}}$, the marginal over $\mathbb{X}$ by $m_{\mathbb{X}}$, and the two conditional probability distributions by $m_{\mathbb{X}|\mathbb{S}}$ and $m_{\mathbb{S}|\mathbb{X}}$. We sometimes abuse notation and eliminate the subscripts when referring to marginals and conditional distributions if there is no room for confusion.

The next definition is used to place constraints on the agent's equilibrium belief $\mu \in \Delta(\Theta)$ when the equilibrium distribution over state–action pairs is m.

DEFINITION 6. The *weighted Kullback–Leibler divergence* (wKLD) is a mapping $K_Q\colon \Delta(\mathbb{S} \times \mathbb{X}) \times \Theta \to \bar{\mathbb{R}}_+$ such that for any $m \in \Delta(\mathbb{S} \times \mathbb{X})$ and $\theta \in \Theta$,[5]

$$K_Q(m, \theta) = \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|s,x)}\left[\ln\left(\frac{Q(S'|s,x)}{Q_\theta(S'|s,x)}\right)\right] m(s,x).$$

The *set of closest parameter values given* $m \in \Delta(\mathbb{S} \times \mathbb{X})$ is the set

$$\Theta_Q(m) \equiv \arg\min_{\theta\in\Theta} K_Q(m, \theta).$$

The set $\Theta_Q(m)$ can be interpreted as the set of parameter values that constitute the best fit with the true transition probability function Q when outcomes are drawn from the distribution m.

LEMMA 1. *(i) For every* $m \in \Delta(\mathbb{S} \times \mathbb{X})$ *and* $\theta \in \Theta$, $K_Q(m, \theta) \geq 0$, *with equality holding if and only if* $Q_\theta(\cdot \mid s, x) = Q(\cdot \mid s, x)$ *for all* (s, x) *such that* $m(s, x) > 0$. *(ii) For any regular* SMDP$(Q, \mathcal{Q}_\Theta)$, $m \mapsto \Theta_Q(m)$ *is nonempty, compact-valued, and upper hemicontinuous (uhc).*

[5]We follow the standard convention that $\ln(0) \cdot 0 = 0$.

Most proofs are provided in the Appendix.
We now define equilibrium.

DEFINITION 7. A probability distribution over state–action pairs, $m \in \Delta(\mathbb{S} \times \mathbb{X})$, is a *Berk–Nash equilibrium* of the SMDP$(Q, \mathcal{Q}_\Theta)$ if there exists a belief $\mu \in \Delta(\Theta)$ such that (i) and (ii) below hold.

(i) *Optimality*. For all $(s, x) \in \mathbb{S} \times \mathbb{X}$ such that $m(s, x) > 0$, x is optimal given s in the MDP$(\bar{Q}_\mu)$, where $\bar{Q}_\mu = \int_\Theta Q_\theta \mu(d\theta)$.

(ii) *Belief Restriction*. We have $\mu \in \Delta(\Theta_Q(m))$.

Moreover, the following condition holds:

(iii) *Stationarity*. For all $s' \in \mathbb{S}$, $m_\mathbb{S}(s') = \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} Q(s' \mid s, x) m(s, x)$.

Condition (i) in the definition of Berk–Nash equilibrium requires actions to be optimal in the MDP where the transition probability function is $\int_\Theta Q_\theta \mu(d\theta)$. Condition (ii) requires that the agent puts positive probability only on the set of closest parameter values given m, $\Theta_Q(m)$. Finally, to interpret condition (iii), note that, for states that occur with positive probability, we can replace $m(s, x)$ with $m_{\mathbb{X}|\mathbb{S}}(x \mid s) m_\mathbb{S}(s)$ in the right-hand side of the expression. In particular, we can think of the agent as following the strategy of choosing actions according to the probability distribution $m_{\mathbb{X}|\mathbb{S}}(\cdot \mid s) \in \Delta(\mathbb{X})$ in state s. Thus, the equilibrium transition probability function over states is given by $s \mapsto Q(\cdot|s, x) m_{\mathbb{X}|\mathbb{S}}(x \mid s)$, and condition (iii) simply says that $m_\mathbb{S}$ is an invariant distribution for this equilibrium transition probability function. In the special case of a static environment, our definition collapses to the single-agent definition in Esponda and Pouzo (2016).

The next result establishes the existence of equilibrium in any regular SMDP.

THEOREM 1. *For any regular SMDP, a Berk–Nash equilibrium exists.*

3.3 *Identification*

Identification plays an important role in the results that follow. In statistics, identification refers to the capacity to infer a unique data generating process from the observed, exogenous data. In our environment, the notion of identification is a bit more nuanced, because the data observed by the agent are endogenous, in the sense that they depend on the agent's actions. Thus, following Esponda and Pouzo (2016), it is natural to consider two notions of identification. These notions distinguish between outcomes on and off the equilibrium path.

DEFINITION 8. An SMDP is *weakly identified given* $m \in \Delta(\mathbb{S} \times \mathbb{X})$ if $\theta, \theta' \in \Theta_Q(m)$ implies that $Q_\theta(\cdot \mid s, x) = Q_{\theta'}(\cdot \mid s, x)$ for all $(s, x) \in \mathbb{S} \times \mathbb{X}$ such that $m(s, x) > 0$; if the condition is satisfied for all $(s, x) \in \mathbb{S} \times \mathbb{X}$, we say that the *SMDP is identified given* m. An SMDP is *(weakly) identified* if it is (weakly) identified for all $m \in \Delta(\mathbb{S} \times \mathbb{X})$.

Weak identification implies that, for any equilibrium distribution m, the agent has a unique belief *along the equilibrium path*, i.e., for states and actions that occur with positive probability. But there could be many beliefs consistent with what happens for those state–action pairs that have zero probability. Thus, weak identification allows one to capture bandit situations, where the agent settles for an action, but may have incorrect beliefs about the benefits she would have obtained with a different action. Weak identification is a fairly weak condition and its failure is often associated with knife-edge cases (see, for example, the coin example by Berk (1966)).

Identification strengthens the definition of weak identification by requiring that beliefs are also unique off the equilibrium path. Under identification, it is as if the agent can eventually learn (possibly incorrectly) the primitives of the environment irrespective of her choice of actions.

PROPOSITION 1. *Consider a correctly specified and identified SMDP with corresponding* MDP(Q). *If m is a Berk–Nash equilibrium of the SMDP, then, for all (s, x) in the support of m, x is optimal given s in the* MDP(Q).

PROOF. Suppose m is a Berk–Nash equilibrium. Then there exists $\mu \in \Delta(\Theta_Q(m))$ such that, for all (s, x) in the support of m, x is optimal given s. Because the SMDP is correctly specified, there exists θ^* such that $Q_{\theta^*} = Q$ and, therefore, by Lemma 1(i), $\theta^* \in \Delta(\Theta_Q(m))$. Then, by identification, any $\hat{\theta} \in \Theta_Q(m)$ satisfies $Q_{\hat{\theta}} = Q_{\theta^*} = Q$, implying that, for all (s, x) in the support of m, x is also optimal given s in the MDP(Q). □

Proposition 1 says that in environments where the agent is uncertain about the transition probability function but her subjective model is both correctly specified and identified, then Berk–Nash equilibrium corresponds to the solution of the MDP under correct beliefs about the transition probability function.

4. EXAMPLES

Applications in the literature on agents with misspecified models have for the most part concentrated on static environments. We hope that the equilibrium concept developed in this paper encourages researchers to explore misspecification in the types of dynamic environments that are central to many economic applications. For this purpose, we pick three standard dynamic environments, and, for each case, introduce a novel misspecification and show how the equilibrium concept can be used to derive concrete predictions. Overall, we hope to convey that Berk–Nash equilibrium can help expand the scope of the classical dynamic programming approach in economics.

Some of the examples in this section assume, for convenience, a nonfinite set of actions and states. While the equilibrium concept extends in a straightforward manner to nonfinite settings, the proofs of the results we provide in the next section rely on finiteness assumptions; we leave the extension to nonfinite settings for further work.

4.1 *Dynamic effort task*

We use the following stylized version of a dynamic effort task to illustrate the steps required to find a Berk–Nash equilibrium.

MDP. In each period t, the agent chooses whether to put high or low effort in a task, $x_t \in \mathbb{X} = \{H, L\}$, where H represents high effort and L represents low effort. The task then fails or succeeds, $s_{t+1} \in \mathbb{S} = \{0, 1\}$, where 0 denotes failure and 1 denotes success. The payoff is $\pi(L, s_{t+1}) = s_{t+1}$ under low effort and $\pi(H, s_{t+1}) = s_{t+1} - c$ under high effort, where c is the cost of high effort. The probability of a success is 1 if the agent puts high effort: $Q(1 \mid s, H) = 1$ for all $s \in \{0, 1\}$. The probability of success if the agent puts low effort depends on the state: The probability of success is $q_0 \equiv Q(1 \mid 0, L)$ if the last task resulted in a failure and is $q_1 \equiv Q(1 \mid 1, L)$ if it resulted in a success. This simple setup captures several problems where the agent's success depends not only on her action, but also on a previous success or failure. For example, a firm that sells a product today may increase its chances of selling a product tomorrow due to word-of-mouth advertising. Alternatively, an agent who succeeds on a task today may feel motivated and find it easier to succeed on the task tomorrow for the same level of effort.

For concreteness, we assume that

$$0 < q_0 < 1 - c < q_1 < 1. \tag{4}$$

In particular, the probability of a success under low effort is higher if the past task was a success compared to a failure. A myopic agent who knows the primitives will find it optimal to choose H in state $s = 0$ (because q_0, the expected payoff from L, is lower than $1 - c$, the payoff from H) and choose L in state $s = 1$ (because $1 - c < q_1$). It is also relatively easy to see that this strategy is optimal irrespective of the discount factor of the agent.

SMDP. The agent believes, incorrectly, that the effort task is not dynamic. Formally, $\mathcal{Q}_\Theta = \{Q_\theta : \theta \in \Theta\}$, where $\Theta = [0, 1]$, and, for all $\theta \in \Theta$, $Q_\theta(1 \mid s, H) = 1$ and $Q_\theta(1 \mid s, L) = \theta$ for all $s \in \{0, 1\}$. In particular, the agent knows that the probability of success is 1 if she puts high effort, but the agent does not know the probability of success if she puts low effort. Moreover, the agent believes that the probability of success under low effort is independent of the current state. For example, the firm might be unaware that word-of-mouth advertising is important or the agent may fail to take into account how performance today affects her motivation tomorrow. This is an example of a subjectively static SMDP because the contemporaneous payoff function π and the perceived transitions do not depend on the current state.

Equilibrium. For simplicity, we restrict attention to equilibria satisfying the natural refinement that the agent's action does not depend on the state: $m_L \equiv m_{\mathbb{X}|\mathbb{S}}(L \mid 0) = m_{\mathbb{X}|\mathbb{S}}(L \mid 1)$ and $1 - m_L = m_{\mathbb{X}|\mathbb{S}}(H \mid 0) = m_{\mathbb{X}|\mathbb{S}}(H \mid 1)$. This is a natural refinement because the agent does not think the current state matters, but it potentially leaves out mixed-strategy equilibria where the agent is indifferent between the two actions and for some reason decides to use a tie-breaking rule that depends on the state.

Stationarity. Condition (iii) in the definition of Berk–Nash equilibrium requires

$$\begin{aligned} m_{\mathbb{S}}(1) &= \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} Q(1 \mid s,x) m_{\mathbb{X}|\mathbb{S}}(x \mid s) m_{\mathbb{S}}(s) \\ &= (1-m_L) + m_L\big(q_0 m_{\mathbb{S}}(0) + q_1 m_{\mathbb{S}}(1)\big). \end{aligned}$$

Solving this equation for $m_{\mathbb{S}}(1)$, we obtain the stationary probability of $s=1$ as a function of the agent's behavior, m_L:

$$m_{\mathbb{S}}(1) = \frac{1-m_L(1-q_0)}{1-m_L(q_1-q_0)}. \tag{5}$$

Beliefs. The wKLD is given by

$$\begin{aligned} K_Q(m,\theta) &= \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} m_{\mathbb{X}|\mathbb{S}}(x \mid s) m_{\mathbb{S}}(s) \sum_{s'\in\mathbb{S}} Q\big(s' \mid s,x\big) \ln \frac{Q\big(s' \mid s,x\big)}{Q_\theta\big(s' \mid s,x\big)} \\ &= -m_L\big\{m_{\mathbb{S}}(0)\big(q_0 \ln\theta + (1-q_0)\ln(1-\theta)\big) \\ &\quad + m_{\mathbb{S}}(1)\big(q_1 \ln\theta + (1-q_1)\ln(1-\theta)\big)\big\} + \text{Const}, \end{aligned}$$

where Const is a term that does not depend on θ.

If $m_L > 0$, then

$$\theta_Q(m) = \big(1-m_{\mathbb{S}}(1)\big)q_0 + m_{\mathbb{S}}(1)q_1 \tag{6}$$

is the unique parameter value that minimizes the wKLD function. Intuitively, (6) is a weighted average of the probabilities that low effort yields a success in each state, q_0 and q_1, where the weights are given by the stationary probabilities of each state. If, however, $m_L = 0$, the wKLD is constant in the parameter and any $\theta \in \Theta$ minimizes wKLD.

We make a second refinement and restrict attention to equilibria where (6) is the unique minimizer even if L is chosen with probability 0: $m_L = 0$. One rationale is that the agent has a small but vanishing probability of trembling and, consistent with the first restriction, this probability does not depend on the state.

Optimality. Because the agent believes that the problem is static, the optimal strategy is to choose the action that maximizes the current period's payoff. Let

$$D(\theta) \equiv \theta - (1-c) \tag{7}$$

denote the perceived expected payoff *difference* of choosing L versus H under the belief that the parameter value is θ with probability 1. If $D(\theta) > 0$, then L is the unique optimal strategy: $m_L = 1$. If, alternatively, $D(\theta) < 0$, then H is the unique optimal strategy: $m_L = 0$. Finally, if $D(\theta) = 0$, there is no restriction on m_L.

Equilibrium. By (5) and assumption (4), $m_{\mathbb{S}}(1)$ is continuous and decreasing as a function of m_L. Intuitively, the higher is the probability of low effort, the lower is the stationary probability of being in the state $s=1$, where the task is successful. Also by (6) and assumption (4), $\theta_Q(m)$ is continuous and increasing as a function of $m_{\mathbb{S}}(1)$. Thus, we can combine (5) and (6) to produce a mapping that, in a slight abuse of notation,

we denote by $m_L \mapsto \theta_Q(m_L)$ that is continuous and decreasing: As m_L increases, the probability of state $s=1$, $m_{\mathbb{S}}(1)$, decreases, which in turn yields a decrease in θ_Q.

Finally, we take the mapping $m_L \mapsto \theta_Q(m_L)$ together with (7) to form the mapping that, in a slight abuse of notation, we denote by $m_L \mapsto D(m_L)$, where $D(m_L) = \theta_Q(m_L) - (1-c)$ is the agent's perceived expected payoff difference of choosing L versus H under the belief that minimizes KLD when the agent chooses no effort with probability m_L. Simple algebra (combining (5), (6), and (7)) shows that

$$D(m_L) = \big(q_1 - m_L(q_1 - q_0)\big)/\big(1 - m_L(q_1 - q_0)\big) - (1-c). \tag{8}$$

The mapping $m_L \mapsto D(m_L)$ is decreasing because, as explained earlier, $m_L \mapsto \theta_Q(m_L)$ is decreasing. To find the equilibria, it is convenient first to compute $D(0)$ and $D(1)$. Simple algebra yields $D(0) = q_1 - (1-c) > 0$. Intuitively, if $m_L = 0$, then the agent is spending all the time in state $s=1$, and so a small tremble resulting in action L occurs in a state where the probability of success is q_1. Thus, a small tremble leads the agent to believe that the probability of success under L is q_1. Since $q_1 > 1-c$, the agent would then like to deviate and choose L with positive probability. As m_L increases, however, state $s=0$ becomes more likely and the agent becomes more pessimistic about the probability of a success under L.

The most pessimistic belief for the agent is at $m_L = 1$. Simple algebra yields $D(1) = q_0/(1-(q_1-q_0)) - (1-c)$. If the primitives (q_0, c, q_1) are such that $D(1) \geq 0$, then there is a unique equilibrium where $m_L^* = 1$. If, however, $D(1) < 0$, then there is a unique equilibrium and it given by the mixed action $m_L^* \in (0,1)$ that solves $D(m_L^*) = 0$. Using the expression in (8), it is easy to see that the mixed equilibrium action is given by $m_L^* = (q_1 - (1-c))/(c(q_1 - q_0))$.

Figure 1 shows an example where the equilibrium action is mixed. In addition to demonstrating the mechanics underlying the equilibrium concept, this example illustrates the importance of allowing the agent to take mixed actions, a feature that is not needed in standard dynamic optimization settings.

4.2 *Stochastic growth with correlated shocks*

Stochastic growth models have been central to studying optimal intertemporal allocation of capital and consumption since the work of Brock and Mirman (1972). Freixas (1981) and Koulovatianos et al. (2009) assume that agents learn the distribution over productivity shocks with correctly specified models. We follow Hall (1997) and subsequent literature in incorporating shocks to both preferences and productivity. We show that there is underinvestment in equilibrium whenever shocks are positively correlated but agents fail to account for this correlation.

MDP. In each period t, an agent observes $s_t = (y_t, z_t) \in \mathbb{S} = \mathbb{R}_+ \times \{L, H\}$, where y_t is wealth and z_t is an independent and identically distributed (i.i.d.) utility shock, and chooses how much wealth to save, $x_t \in [0, y_t] \subseteq \mathbb{X} = \mathbb{R}_+$, consuming the rest. Current period utility is $\pi(y_t, z_t, x_t) = z_t \ln(y_t - x_t)$. Wealth the next period, y_{t+1}, is given by

$$\ln y_{t+1} = \alpha^* + \beta^* \ln x_t + \varepsilon_t,$$

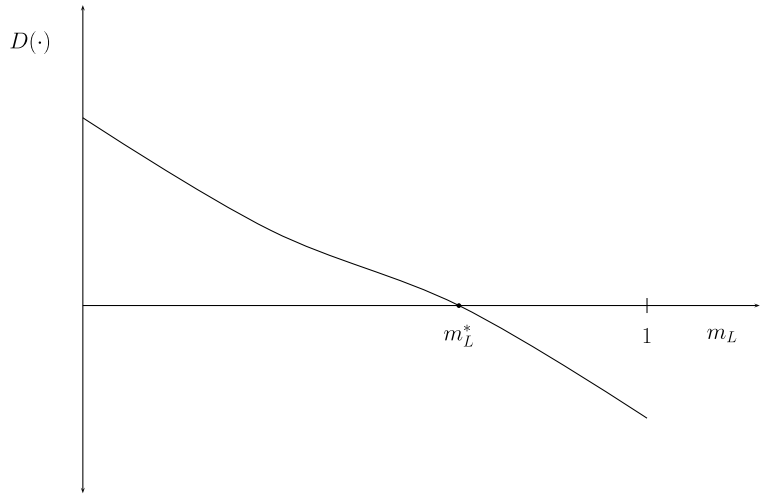

FIGURE 1. Equilibrium of the dynamic effort environment.

where $\varepsilon_t = \gamma^* z_t + \xi_t$ is an unobserved i.i.d. productivity shock, $\xi_t \sim N(0,1)$, and $0 < \delta\beta^* < 1$, where $\delta \in [0,1)$ is the discount factor. The utility shock can be interpreted as a shock to home or nonmarket production technologies (e.g., Bencivenga 1992). We assume that $\gamma^* > 0$, so that the utility and productivity shocks are positively correlated. For example, technological advances increase productivity of both market and nonmarket activities. Let $0 < L < H$ and let $q \in (0,1)$ be the probability that the shock is H. Formally, $Q(y', z' \mid y, z, x)$ is such that y' and z' are independent, y' has a log-normal distribution with mean $\alpha^* + \beta^* \ln x + \gamma^* z$ and unit variance, and $z' = H$ with probability q.

SMDP. The agent believes that

$$\ln y_{t+1} = \alpha + \beta \ln x_t + \varepsilon_t, \tag{9}$$

where $\varepsilon_t \sim N(0,1)$ and is *independent* of the utility shock. For simplicity, we assume that the agent knows the distribution of the utility shock and is uncertain about $\theta = (\alpha, \beta) \in \Theta = \mathbb{R}^2$. The subjective transition probability function $Q_\theta(y', z' \mid y, z, x)$ is such that y' and z' are independent, y' has a log-normal distribution with mean $\alpha + \beta \ln x$ and unit variance, and $z' = H$ with probability q. The agent has a misspecified model because she believes that the productivity and utility shocks are independent, when in fact $\gamma^* \neq 0$.

Equilibrium. *Optimality*. The Bellman equation for the agent is

$$V(y,z) = \max_{0 \le x \le y} z \ln(y-x) + \delta E\big[V\big(Y', Z'\big) \mid x\big],$$

and it is straightforward to verify that the optimal strategy is to invest a fraction of wealth that depends on the utility shock and the unknown parameter β, i.e., $x = A_z(\beta) \cdot y$, where $A_L(\beta) = \frac{\delta\beta((1-q)L+qH)}{(1-\delta\beta(1-q))H+\delta\beta(1-q)L}$ and $A_H(\beta) = \frac{\delta\beta((1-q)L+qH)}{\delta\beta qH+(1-\delta\beta q)L} < A_L(\beta)$, provided that $\beta\delta < 1$, which will be true in equilibrium. For the agent who knows the primitives, the optimal strategy is to invest fractions $A_L(\beta^*)$ and $A_H(\beta^*)$ in the low and high

state, respectively. Since $\beta \mapsto A_z(\beta)$ is increasing, the equilibrium strategy of a misspecified agent can be compared to the optimal strategy by comparing the equilibrium belief about β with the true β^*.

Beliefs and stationarity. Let $A = (A_L, A_H)$, with $A_H < A_L$, represent a strategy, where A_z is the proportion of wealth invested given utility shock z. Because the agent believes that ε_t is independent of the utility shock and normally distributed, the minimizers of the wKLD function are the estimands of a linear regression model, which are unique, and, therefore, this SMDP is identified provided the agent invests more than zero with positive probability.[6] In particular, for a strategy represented by $A = (A_L, A_H)$, the parameter value $\hat{\beta}(A)$ that minimizes wKLD is

$$\begin{aligned}\hat{\beta}(A) &= \frac{\operatorname{Cov}\big(\ln Y', \ln X\big)}{\operatorname{Var}(\ln X)} \\ &= \frac{\operatorname{Cov}\big(\ln Y', \ln(A_Z Y)\big)}{\operatorname{Var}\big(\ln(A_Z Y)\big)} \\ &= \beta^* + \gamma^* \frac{\operatorname{Cov}(Z, \ln A_Z)}{\operatorname{Var}(\ln A_Z) + \operatorname{Var}(\ln Y)},\end{aligned}$$

where Cov and Var are taken with respect to the (true) distribution of (Y, Z). Since $A_H < A_L$, then $\operatorname{Cov}(Z, \ln A_Z) < 0$. Therefore, the assumption that $\gamma^* > 0$ implies that the bias $\hat{\beta}(A) - \beta^*$ is negative and its magnitude depends on the strategy A. Intuitively, the agent invests a larger fraction of wealth when z is low, which happens to be during times when ε is also low.

Equilibrium. We establish that there exists at least one equilibrium with positive investment by showing that there is at least one fixed point of the mapping $\beta \mapsto \hat{\beta}(A_L(\beta), A_H(\beta))$. This mapping is continuous and satisfies $\hat{\beta}(A_L(0), A_H(0)) = \hat{\beta}(A_L(1/\delta), A_H(1/\delta)) = \beta^*$ and $\hat{\beta}(A_L(\beta), A_H(\beta)) < \beta^*$ for all $\beta \in (0, 1/\delta)$. Then, since $\delta\beta^* < 1$, there is at least one fixed point β^M, and any fixed point satisfies $\beta^M \in (0, \beta^*)$. Thus, the misspecified agent underinvests in equilibrium compared to the optimal strategy.[7] The conclusion is reversed if $\gamma^* < 0$, illustrating how the framework provides predictions about beliefs and behavior that depend on the primitives (as opposed to simply postulating that the agent is over- or underconfident about productivity).

4.3 *Production with uncertain cost*

Finally, we consider an agent who produces with uncertain costs. This example illustrates two features of the framework. First, unlike the previous examples, the agent

[6] From (9) and Gaussianity of the residuals, the wKLD is proportional to the expected (under the true measure) square of the residual in expression (9). Thus, the minimizers of the wKLD coincide with the values of (α, β) that provide the best fit under this loss when the data are distributed according to the true probability measure.

[7] It is also an equilibrium not to invest, $A = (0, 0)$, supported by the belief $\beta^* = 0$, which cannot be disconfirmed since investment does not take place. But this equilibrium is not robust to experimentation (e.g., it does not survive a refinement where the belief when not investing is required to be the limit of the belief as the fraction invested goes to zero).

knows the dynamics governing the state variable. Instead, the agent has uncertainty about the per-period payoff. The example shows how to incorporate this kind of uncertainty into the framework. Second, in contrast to the previous examples, where the agent directly omitted a variable or neglected a correlation, we consider a case where the agent incorporates all relevant variables into her model but uses an incorrect functional form.

MDP. Each period t, an agent observes a productivity shock $z \in \mathbb{Z} = \{z_1, \dots, z_K\} \subset \mathbb{R}_+$ and chooses an input $x \in \mathbb{X} \subset \mathbb{R}_+$. As a result, the agent obtains a payoff of $z \ln x - c(x)$ in that period, where $c(x) = \phi(x)\epsilon$ is the cost of choosing x, and ϵ is a random, independent cost shock distributed according to the distribution p^*, which has support equal to $[0, \infty)$. Let $Q(z' \mid z)$ be the probability that tomorrow's productivity shock is z' given the current shock z. We assume that there is a unique stationary distribution over these productivity shocks, denoted by $q = (q_1, \dots, q_K)$.

SMDP. The agent knows all the primitives except the cost function $c(\cdot)$. The agent believes that $c_\theta(x) = x\epsilon$ and $\epsilon \sim p_\theta$, where p_θ has support equal to $[0, \infty)$. For concreteness, we assume that ϵ follows an exponential distribution, $p_\theta(\epsilon) = (1/\theta)e^{-(1/\theta)\epsilon}$. In particular, the agent's model is misspecified if either cost is nonlinear, i.e., $\phi(\cdot)$ is nonlinear, or the true distribution over cost shocks, p^*, does not belong to the exponential family.

The framework presented in this paper assumes that the agent knows the per-period payoff function and may be uncertain about the transition function. To fit this example into the framework, we simply let the cost c be part of the state as follows:

$$V(z, c) = \max_x \int \big(zf(x) - c' + \delta V\big(z', c'\big)\big) Q\big(dz' \mid z\big) Q^C\big(dc' \mid x\big).$$

The variable c' is the unknown cost of production at the time the agent has to choose x. Its distribution is given by $Q^C(dc' \mid x)$, which is the distribution of $c' = c(x)$ as described above. The agent knows Q, but does not know Q^C. In particular, the agent has a parametric family of transitions, where $Q^C_\theta(dc' \mid x)$ is the distribution of $c' = c_\theta(x)$.

Equilibrium. *Optimality*. Suppose the agent has a degenerate belief on some θ. Because the transition of c' does not depend on c and the transition of z' does not depend on x, the agent's optimization problem reduces to the simple static optimization problem $\max_x z \ln x - xE_\theta[\epsilon]$. Noting that $E_\theta[\epsilon] = \theta$, it follows that the optimal input choice in state z_j is

$$x_j = z_j/\theta \tag{10}$$

for $j \in \{1, \dots, K\}$.

Stationarity. The stationarity condition implies that the marginal of m over $\mathbb{Z}$ is equal to the stationary distribution over z, which is given by $q = (q_1, \dots, q_K)$. Therefore, the stationary distribution over $\mathbb{X}$, denoted by $m_\mathbb{X}$, is given by $m_\mathbb{X}(x_j) = q_j$, where x_j satisfies (10), and it is equal to zero otherwise.

Beliefs. The part of the wKLD function that depends on θ is given by

$$\begin{aligned}\sum_x E_{Q(\cdot|x)}\big[\log Q_\theta^C\big(c' \mid x\big)\big]m_{\mathbb{X}}(x) &= \sum_j E_{Q(\cdot|x_j)}\big[\log p_\theta\big(c'/x_j\big)\big]q_j \\ &= \sum_j E_{Q(\cdot|x_j)}\left[-\frac{1}{\theta}\big(c'/x_j\big) - \ln\theta\right]q_j \\ &= -\frac{1}{\theta}E_{p^*}[\epsilon]\sum_j\big(\phi(x_j)/x_j\big)q_j - \ln\theta.\end{aligned}$$

There is a unique parameter value θ that maximizes this expression, and so this SMDP is identified. This unique minimizer is given by

$$\theta = E_{p^*}[\epsilon]\sum_j\big(\phi(x_j)/x_j\big)q_j. \tag{11}$$

The right-hand side of this expression is a weighted average of the expected average costs. This expression depends on the assumption that ϵ follows an exponential distribution, and it would differ for different families of distributions. For example, for the case of the log-normal distribution, the average cost should be replaced by the logarithm of the average cost.

Equilibrium. To solve for equilibrium, we first combine (10) and (11) to obtain

$$\theta^* = E_{p^*}[\epsilon]\sum_j\big(\theta^*\phi\big(z_j/\theta^*\big)/z_j\big)q_j. \tag{12}$$

A solution θ^* to (12) corresponds to an equilibrium belief. To find the equilibrium action as a function of the shock, we simply replace the equilibrium belief θ^* into the optimality condition (10). To illustrate, suppose that the true cost function is quadratic, i.e., $\phi(x) = x^2$. Then there is a unique solution to (12), and, therefore, a unique equilibrium belief $\theta^* = (E_{p^*}[\epsilon]E_q[z])^{1/2}$ and action

$$x_j^* = z_j/\big(E_{p^*}[\epsilon]E_q[z]\big)^{1/2}. \tag{13}$$

We can contrast this expression with the optimal action of an agent who knows the correct primitives and solves $\max_x z\ln x - x^2E_{p^*}[\epsilon]$, thus obtaining the optimal action

$$x_j^{\text{opt}} = \big(z_j/\big(2E_{p^*}[\epsilon]\big)\big)^{1/2}. \tag{14}$$

The optimal action depends on the productivity shock, while the optimal action for the misspecified agent depends on both the shock and the average shock. The reason is that the agent incorrectly believes the marginal cost is constant and learns this marginal cost by averaging over the marginal costs experienced in equilibrium, and the distribution over these experienced costs depends on the stationary distribution over all shocks. Comparing (13) and (14), we also observe that the misspecified agent chooses actions lower than optimal if $z_j \leq E_q[z]/2$ and higher than optimal if $z_j \geq E_q[z]/2$. Intuitively,

the agent overestimates the marginal cost of low actions, and these low actions are taken when the shock is low. Similarly, the agent underestimates the marginal cost of high actions, and these actions are taken when the shock is high.

5. Equilibrium foundation

Following the tradition of providing learning foundations for equilibrium concepts, in this section we study the problem of an agent who faces a regular SMDP, starts with a prior $\mu_0 \in \Delta(\Theta)$ over the set of models of the world Θ, and updates the prior in each period as a result of observing the current state, her action, and the new state. Our main objective is to understand under which conditions the agent's steady-state behavior can be represented by a Berk–Nash equilibrium.

5.1 *Bayesian learning in SMDPs*

Consider an agent who faces a regular SMDP and has a prior $\mu_0 \in \Delta(\Theta)$, which is assumed to have full support. The prior is updated in each period using Bayes' rule, where $\mu' = B(s, x, s', \mu)$ is the posterior for any prior μ, current state s, action x, and realized future state s', and for any $(s, x, s') \in \mathbb{S} \times \mathbb{X} \times \mathbb{S}$, the Bayesian operator $B(s, x, s', \cdot) : D_{s,x,s'} \to \Delta(\Theta)$ is defined as follows: For all $A \subseteq \Theta$ Borel, $B(s, x, s', \mu)(A) = \int_A Q_\theta(s' \mid s, x)\mu(d\theta) / \int_\Theta Q_\theta(s' \mid s, x)\mu(d\theta)$ for any $\mu \in D_{s,x,s'}$, where $D_{s,x,s'} = \{p \in \Delta(\Theta) : \int_\Theta Q_\theta(s' \mid s, x)\, p(d\theta) > 0\}$.

By the principle of optimality, the agent's problem can be cast recursively as

$$W(s, \mu) = \max_{x \in \mathbb{X}} \int_{\mathbb{S}} \big\{\pi\big(s, x, s'\big) + \delta W\big(s', \mu'\big)\big\} \bar{Q}_\mu\big(ds' | s, x\big), \tag{15}$$

where $\bar{Q}_\mu = \int_\Theta Q_\theta \mu(d\theta)$, $\mu' = B(s, x, s', \mu)$ is the next period's belief, updated using Bayes' rule, and $W : \mathbb{S} \times \Delta(\Theta) \to \mathbb{R}$ is the (unique) solution to the Bellman equation (15). Compared to the case where the agent knows the transition probability function, the agent's belief about Θ is now part of the state space.

Definition 9. A *policy function* is a function $f : \mathbb{S} \times \Delta(\Theta) \to \Delta(\mathbb{X})$, where $f(x \mid s, \mu)$ denotes the probability that the agent chooses x if she is in state s and her belief is μ. A policy function f is *optimal* if, for all $s \in \mathbb{S}$, $\mu \in \Delta(\Theta)$, and $x \in \mathbb{X}$ such that $f(x \mid s, \mu) > 0$,

$$x \in \arg\max_{\hat{x} \in \mathbb{X}} \int_{\mathbb{S}} \big\{\pi\big(s, \hat{x}, s'\big) + \delta W\big(s', B\big(s, \hat{x}, s', \mu\big)\big)\big\} \bar{Q}_\mu\big(ds' | s, \hat{x}\big).$$

Let $h = (s_0, x_0, \ldots, s_t, x_t, \ldots)$ represent an infinite history of state–action pairs and let $\mathbb{H} \equiv (\mathbb{S} \times \mathbb{X})^\infty$ represent the space of infinite histories. For every t, let $\mu_t : \mathbb{H} \to \Delta(\Theta)$ denote the agent's belief at time t, defined recursively by $\mu_t(h) = B(s_{t-1}, x_{t-1}, s_t, \mu_{t-1}(h))$ whenever B is the Bayesian operator and arbitrary otherwise. Henceforth, we drop the history h from the notation.

In each period t, there is a state s_t and a belief μ_t, and the agent chooses a (possibly mixed) action $f(\cdot \mid s_t, \mu_t) \in \Delta(\mathbb{X})$.[8] After an action x_t is realized, the state s_{t+1} is drawn from the true transition probability. The agent observes the realized action and the new state and updates her belief to μ_{t+1} using Bayes' rule. The primitives of the problem (including the initial distribution over states, q_0, and the prior, $\mu_0 \in \Delta(\Theta)$) and a policy function f induce a probability distribution over $\mathbb{H}$ that is defined in a standard way; let $\boldsymbol{P}^f$ denote this probability distribution over $\mathbb{H}$.

We now define outcomes as random variables. For every t, we define the frequency of state–action pairs at time t to be a function $m_t : \mathbb{H} \to \Delta(\mathbb{S} \times \mathbb{X})$ such that for all h and $(s, x) \in \mathbb{S} \times \mathbb{X}$,

$$m_t(h)(s, x) = \frac{1}{t} \sum_{\tau=0}^{t} \mathbf{1}_{(s,x)}(s_\tau, x_\tau)$$

is the frequency of times that the outcome (s, x) occurs up to time t. One reasonable criterion to claim that the agent has reached a steady state is that the time average of outcomes converges.

The next result establishes that if the frequency of state–action pairs converges to m, then beliefs become increasingly concentrated on $\Theta_Q(m)$.

LEMMA 2. *Let Q denote the true transition probability function and let f denote the policy function. Suppose that $(m_t)_t$ converges to m for all histories in a set $\mathcal{H} \subseteq \mathbb{H}$ such that $\mathbf{P}^f(\mathcal{H}) > 0$. Then, for all open sets $U \supseteq \Theta_Q(m)$, $\lim_{t\to\infty} \mu_t(U) = 1$ $\mathbf{P}^f$-a.s. in $\mathcal{H}$.*

The proof adapts the proof of Lemma 2 by Esponda and Pouzo (2016) to dynamic environments, and the reader is referred to that paper for an intuitive explanation of the result.[9]

The following result provides a learning foundation for the notion of Berk–Nash equilibrium of an SMDP.

THEOREM 2. *Let f be an optimal policy function. Suppose that $(m_t)_t$ converges to m with $\mathbf{P}^f$-positive probability and that the SMDP is weakly identified given m. Suppose also that one of the following conditions holds:*

(i) The SMDP is subjectively static.

(ii) The SMDP is identified given m.

Then m is a Berk–Nash equilibrium of the SMDP.

Theorem 2 provides a learning justification for Berk–Nash equilibrium. The main idea behind the proof is as follows. For each state–action pair (s, x) in the support of m,

[8] In particular, it would be straightforward to introduce payoff perturbations to our environment so that the agent's behavior at time t would be given by a nondegenerate distribution over actions.

[9] The seminal result that provides asymptotic characterization of Bayesian beliefs when the data generating process is exogenous (i.e., absent any actions) is due to Berk (1966); see also Bunke and Milhaud (1998) and Shalizi (2009) for extensions.

there exists a subsequence of state–action pairs and beliefs such that (s, x) is played along the entire subsequence. Moreover, we can find a sub-subsequence where the belief converges; let $\mu_{s,x} \in \Delta(\Theta)$ denote this limiting belief under which (s, x) realizes. Since $(m_t)_t$ converges to m, we can apply Lemma 2 to conclude that $\mu_{s,x} \in \Delta(\Theta_Q(m))$. Thus, by optimality of f and the upper hemicontinuity of the correspondence of optimal actions, it follows that for any state s and any action x in the support of $m(\cdot \mid s)$, x is optimal in the dynamic optimization problem with current belief $\mu_{s,x}$, i.e.,

$$x \in \arg\max_{\hat{x}\in\mathbb{X}} \int_{\mathbb{S}} \{\pi(s, \hat{x}, s') + \delta W(s', \mu')\} \bar{Q}_{\mu_{s,x}}(ds'|s, \hat{x}). \tag{16}$$

Consider first the case where the SMDP is subjectively static. In this case, the value function W depends only on the agent's belief and, by slightly abusing notation, (16) implies that

$$\begin{aligned} &E_{\bar{Q}_{\mu_{s,x}}(\cdot|x)}\big[\pi(x, S') + \delta W(B(x, S', \mu_{s,x}))\big] \\ &\quad \geq E_{\bar{Q}_{\mu_{s,x}}(\cdot|y)}\big[\pi(y, S') + \delta W(B(y, S', \mu_{s,x}))\big] \end{aligned} \tag{17}$$

for any other action y. By weak identification, $B(x, s', \mu_{s,x}) = \mu_{s,x}$ for all s' that occur with positive probability according to $\mu_{s,x}$, and so the left-hand side of (17) becomes $E_{\bar{Q}_{\mu_{s,x}}(\cdot|x)}[\pi(x, S') + \delta W(\mu_{s,x}))]$. Next, we add and subtract $\delta W(\mu_{s,x})$ from the right-hand side of (17) to obtain

$$E_{\bar{Q}_{\mu_{s,x}}(\cdot|y)}\big[\pi(y, S') + \delta W(\mu_{s,x})\big] + \delta E_{\bar{Q}_{\mu_{s,x}}(\cdot|y)}\big[W(B(y, S', \mu_{s,x})) - W(\mu_{s,x})\big]. \tag{18}$$

The second term in (18) is what is known in the literature as the value of experimentation: It is the difference in net present value between starting the next period with updated belief $B(y, S', \mu_{s,x})$, which depends on the action y and the random realization of S', and starting the period with the current belief $\mu_{s,x}$. By the Martingale property of Bayesian updating and the convexity of the value function, it follows that the value of experimentation is nonnegative; formally, $E_{\bar{Q}_{\mu_{s,x}}(\cdot|y)}[W(B(y, S', \mu_{s,x})) - W(\mu_{s,x})] \geq W(E_{\bar{Q}_{\mu_{s,x}}(\cdot|y)}[B(y, S', \mu_{s,x})]) - W(\mu_{s,x}) = 0$. It then follows that $E_{\bar{Q}_{\mu_{s,x}}(\cdot|x)}[\pi(x, S')] \geq E_{\bar{Q}_{\mu_{s,x}}(\cdot|y)}[\pi(y, S')]$. Thus, for any (s, x) in the support of m, there exists a belief $\mu_{s,x}$ such that x is optimal when the belief is fixed at $\mu_{s,x}$. Finally, weak identification implies that all the beliefs in $\{\mu_{s,x} : m(s, x) > 0\}$ yield the same probability distribution over the next period's state conditional on an action in the support of $m_{\mathbb{X}}$. Therefore, we can replace all these beliefs with a single belief that belongs to $\Delta(\Theta_Q(m))$, so that conditions (i) and (ii) in the definition of Berk–Nash equilibrium (Definition 7) are satisfied for the special case of subjectively static SMDPs.

More generally, we can prove the same result by assuming identification. If the SMDP is identified, we can essentially think of $\Delta(\Theta_Q(m))$ as being a degenerate belief on a specific parameter value, which in turn implies two properties. First, $\mu_{s,x}$ does not depend on s, x; denote it by μ. Second, since the belief μ is degenerate, it forever remains fixed, and so (16) implies that x is optimal given s in the MDP($\bar{Q}_\mu$), where the

belief is fixed at μ. Thus, once again, conditions (i) and (ii) in the definition of Berk–Nash equilibrium are satisfied.

Finally, the reason why condition (iii) in the definition of Berk–Nash equilibrium holds can be described as follows. If the agent were using strategy m_t to make decisions, then the probability distribution over states next period would be given by $Q[m_t](\cdot) \equiv \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} Q(\cdot \mid s,x)m_t(s,x)$. Since m_t converges to m and the operator $Q[\cdot]$ is continuous, the asymptotic evolution of the state is given by the probability distribution $Q[m](\cdot)$. Since m_t converges, then it must converge to a stationary distribution of the Markov process over states defined by this operator.

In the remainder of this section, we investigate the extent to which we can extend the previous arguments to cases where identification fails or the SMDP is not subjectively static. We begin by noting that the definition of steady state used in Section 5.1 (the convergence of time averages) is different from the definition used elsewhere. In previous work (e.g., Fudenberg and Kreps 1993, Esponda and Pouzo 2016), it is common to define a steady state as a situation where the agent's intended behavior converges. In Theorem 2, all we need is that the time average converges, but because of the dynamic nature of the environment, we need the convergence of the frequency of state–action pairs, not just of the actions. In particular, this type of convergence does not guarantee that the agent's intended behavior converges, but only that its frequency does. We now show that if we strengthen the notion of steady state to require that both intended behavior and time averages converge, then a steady state corresponds to a Berk–Nash equilibrium provided that all states are visited with positive probability.[10]

We define a strategy $\sigma : \mathbb{S} \to \Delta(\mathbb{X})$ to be a mapping between states and probability distribution over actions. Let Σ denote the set of all strategies. For a fixed policy function f and for every t, let $\sigma_t : \mathbb{H} \to \Sigma$ denote the (time-t intended) strategy of the agent, defined by setting

$$\sigma_t(h) = f\big(\cdot \mid \cdot, \mu_t(h)\big) \in \Sigma.$$

THEOREM 3. *Let f be an optimal policy function. Suppose that $(\sigma_t)_t$ converges and $(m_t)_t$ converges to m with $\mathbf{P}^f$-positive probability. Suppose also that the SMDP is weakly identified given m and that $m(s) > 0$ for all $s \in \mathbb{S}$. Then m is a Berk–Nash equilibrium of the SMDP.*

The main idea behind the proof is as follows. We can always find a subsequence of posteriors that converges to some μ^* and, by Lemma 2 and the fact that the agent's intended strategy $(\sigma_t)_t$ converges to some σ, it follows that σ must solve the dynamic optimization problem for beliefs converging to $\mu^* \in \Delta(\Theta_Q(m))$. A key difference with the proof of Theorem 3 is that we can use the fact that the agent's intended behavior converges to conclude that the same belief μ^* justifies all of the agent's limiting actions. Next, it is not difficult to show that the limiting behavior of the agent in state s must correspond to the conditional distribution of the limiting time average, i.e., $\sigma(\cdot \mid s) = m(\cdot \mid s)$. Since all states are visited with positive probability according to m, it follows

[10]We are unable to show whether this result is also true when intended behavior does not converge.

that there exists a belief μ^* such that for every (s, x) with $m(s, x) > 0$, x is optimal in the dynamic optimization problem with current belief μ^*. The final step is to show that this type of optimality implies optimality in the dynamic optimization problem where the belief is *fixed* at μ^*.

For this final step, we rely on the assumption that all states are visited with positive probability; the argument is as follows. For each $\tilde{s}$, let $x_{\tilde{s}}$ denote an action that is played in the limit when the state is $\tilde{s}$, i.e., $m(\tilde{s}, x_{\tilde{s}}) > 0$. Consider the strategy where the agent plays $x_{\tilde{s}}$ in each state $\tilde{s}$. By weak identification, the belief never changes and the value of following this strategy does not depend on the specific belief in $\Delta(\Theta_Q(m))$, since, by weak identification, all parameter values in $\Theta_Q(m)$ give rise to the same distribution over the next period's states. By the previous optimality argument, we know that action x_s is optimal in state s given belief μ^*. This means that x_s maximizes the sum of today's payoff and the continuation value, where the continuation value is the value of playing $x_{\tilde{s}}$ in each state $\tilde{s}$ in the future. Consider an alternative action y. This alternative action yields some payoff today and then a continuation value where it is possible that the agent's belief changes. This possibly new belief, call it μ', must still have support in $\Theta_Q(m)$, since the original belief μ^* has support in $\Theta_Q(m)$. Consider the continuation value of this action y with a new belief μ' and a new state. The agent can still, from that moment on, follow the strategy of playing $x_{\tilde{s}}$ in each state $\tilde{s}$ in the future. Thus, the continuation value from playing y is at least the same or higher as the continuation value from x_s. Therefore, the fact that x_s is optimal when the nonnegative value of information from playing a different action y is taken into account implies that x_s must also be optimal when the belief is fixed at μ^* and there is no further value from learning.

The argument in the proof of Theorem 3 relies on the assumption that all states are visited with positive probability. This assumption allows us to construct a strategy (to play $x_{\tilde{s}}$ in each state $\tilde{s}$) that provides a lower bound to the payoff that the agent could obtain from choosing an action that could potentially lead to an updated belief. We conclude with an example that illustrates that this assumption is important. In particular, the following example shows a case where only one state is reached in steady state, and even though the agent's behavior and the time average converge, this steady state is not a Berk–Nash equilibrium.

EXAMPLE. There are five states: s^I, s_0, s_1, s_k, and s^{opt}. In states s_0 and s_1, the agent gets utility 0 and 1, respectively, and then returns to the initial state s^I. In state s_k, the agent gets utility k and then returns to the initial state s^I. In the initial state s^I, the agent has four possible actions: A, B, S, and O. Irrespective of her action, she gets utility $2/3$ in state s^I. If she chooses A, she goes to state s_0 with probability θ and to s_1 with probability $1 - \theta$, while if she chooses B, she goes to s_0 with probability $1 - \theta$ and to s_1 with probability θ. If she chooses S, she remains in state s^I. In other words, A and B are risky alternatives that yield utility 0 or 1 tomorrow, and S is a safe action that yields $2/3$ tomorrow. Moreover, the agent eventually returns to s^I. Formally, the payoffs are $\pi(s^I, x) = 2/3$ and $\pi(s_j, x) = j$ for all x, and the transitions are $Q_\theta(s_0 \mid s^I, A) = Q_\theta(s_1 \mid s^I, B) = \theta$, $Q_\theta(s_1 \mid s^I, A) = Q_\theta(s_0 \mid s^I, B) = Q_\theta(s_0 \mid s^I, O) = 1 - \theta$, and $Q_\theta(s^I \mid s_j, x) = 1$ for all $j \in \{0, 1, k\}$ and all x.

The agent can also take action O in state s^I, which potentially generates the *option* to make a risky but more profitable investment that yields k in the future. Taking action O in state s^I leads the agent to state s^{opt} with probability θ and to state s_0 with probability $1-\theta$. In state s^{opt}, the agent gets a utility cost (loses) $1/3$ irrespective of her action. If she chooses to make a risky investment (R, which we can associate with actions A, B, and O so as to have the same set of actions for all states), with probability $1-\theta$ she goes to state s_k and, therefore, gets utility k, and with probability θ, she goes to state s_0 and, therefore, gets utility 0. If she chooses the safe option (S), then she goes to state s^I next period. In any case, she always ends up returning to state s^I. Formally, the payoffs are $\pi(s^{\text{opt}}, x) = -1/3$ for all x, and the transitions are $Q_\theta(s^{\text{opt}} \mid s^I, O) = Q_\theta(s_0 \mid s^{\text{opt}}, R) = \theta$, $Q_\theta(s_k \mid s^{\text{opt}}, R) = 1-\theta$, and $Q_\theta(s^I \mid s^{\text{opt}}, S) = 1$. Figure 2 depicts all the states, actions, and transitions for this example.

Suppose that the agent knows all the primitives except the value of θ. Moreover, suppose that the true value of θ is either 0 or 1, and that the SMDP is correctly specified, i.e., $\Theta = \{0, 1\}$, thus, highlighting that the new issue present in dynamic environments is not due to misspecification. We also assume that the agent is patient, but not too patient, $\delta \in (0, \sqrt{1/3})$, and that the return from the risky investment in state s^{opt} is high enough relative to the rate of impatience, $k > 2 + 4/\delta$.

This problem is simple enough that we can directly characterize the steady state and then check if it is a Berk–Nash equilibrium. Consider a (Bayesian) agent who starts with

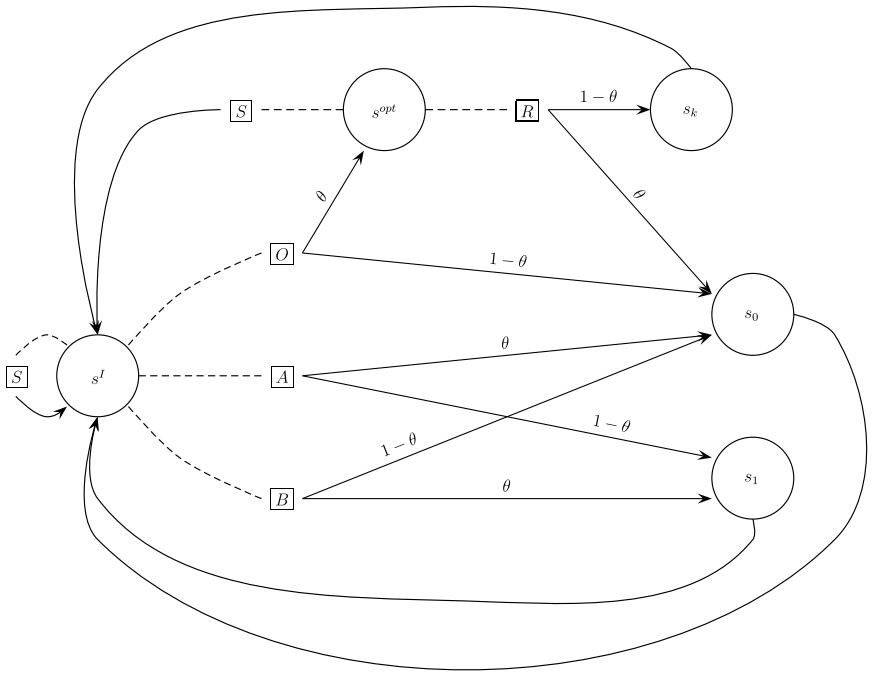

FIGURE 2. Example: Steady state is not a Berk–Nash equilibrium. States are depicted with circles and actions are depicted with squares. For each state, dashed lines indicate the actions that can be taken in the state. Arrows indicate transition probabilities given each state–action pair.

a prior $\mu = \Pr(\theta = 1) \in (0, 1)$ in state s^I. If she chooses action O, then, beginning next period (recall that all actions yield the same current payoff of $2/3$ in state s^I), she will get

$$\mu W\big(s^{\text{opt}}, 1\big) + (1 - \mu)W(s_0, 0). \tag{19}$$

Crucially, if action O takes her to state s^{opt}, then she learns that $\theta = 1$, so that $\mu' = 1$. In this case, it is optimal to take the safe action and return to s^I next period, since taking the risky action would lead to a zero payoff with probability 1 and a delay of one period in getting back to s^I. Therefore, $W(s^{\text{opt}}, 1) = -1/3 + \delta W(s^I, 1)$. Also, if she ends up in state s_0, she gets 0 and then goes on to state s^I, i.e., $W(s_0, 0) = 0 + \delta W(s^I, 0)$. Moreover, if the agent is in state s^I and has certainty about the state, i.e., $\mu' = 0$ or 1, then it is optimal for her to choose either action A or B, respectively, and her payoff alternates between $2/3$ and 1 forever, i.e., $W(s^I, 1) = W(s^I, 0) =: W^* = (2/3 + \delta)/(1 - \delta^2)$. Therefore, expression (19) becomes

$$-(1/3)\mu + \delta W^*. \tag{20}$$

Consider instead the case where the agent chooses action A in state s^I. Then next period she gets

$$(1 - \mu)W(s_1, 0) + \mu W(s_0, 1), \tag{21}$$

where $W(s_1, 0) = 1 + \delta W(s^I, 0) = 1 + \delta W^*$ and $W(s_0, 1) = 0 + \delta W(s^I, 1) = \delta W^*$. Thus, expression (21) becomes

$$(1 - \mu) + \delta W^*. \tag{22}$$

Similarly, if the agent chooses action B, then next period she will get

$$\mu + \delta W^*. \tag{23}$$

Finally, choosing action S in state s^I keeps the agent in state s^I and results in no information about θ being revealed. If S is optimal at s^I, then it is optimal to choose it in every period, in which case the agent earns a payoff of $2/3$ in each period and her discounted payoff beginning next period is

$$\frac{2/3}{1 - \delta}. \tag{24}$$

Comparing (20) and (22), it follows that action A is better than action O for any belief μ, implying that the agent never picks O in state s^I. Intuitively, the agent realizes that if she picks O and ends up in state s^{opt}, then she will infer that the risky alternative will deliver a zero payoff for sure and so there is no point in picking O to begin with. Also, by comparing (22), (23), and (24), it follows that if the agent starts in state s^I with a prior μ that satisfies

$$\frac{1/3}{1 - \delta^2} \le \mu \le \frac{2/3 - \delta^2}{1 - \delta^2},$$

then it is optimal for her to choose S and stay at s^I forever. (Such a set of priors is nonempty because $\delta \in (0, \sqrt{1/3})$). Therefore, repeatedly choosing S and staying at s^I is a

steady-state outcome. Note, however, that Theorem 2 does not apply to this steady state because (i) the SMDP is not subjectively static, and (ii) identification does not hold, because the agent learns nothing about θ by playing S at s^I. Theorem 3 also does not apply here, because in this steady-state outcome, only state s^I is visited. In fact, we now show that this steady-state outcome cannot arise in a Berk–Nash equilibrium, suggesting a limitation of equilibrium analysis in dynamic settings.

To analyze Berk–Nash equilibria, let μ denote the agent's equilibrium belief and consider the agent's choice in state s^I. Let us first find the set of μs such that action S is preferred to both A and B, ignoring action O. If the agent takes action S, then, beginning next period (recall that all actions yield the same current payoff), she goes back to s^I and obtains

$$W\big(s^I,\mu\big).$$

Action A, alternatively, yields

$$\mu W(s_0,\mu)+(1-\mu)W(s_1,\mu), \tag{25}$$

where, importantly, the agent does *not* update her equilibrium belief upon moving to state s_0 or s_1, as the definition of equilibrium requires optimization with respect to a single, fixed equilibrium belief. As before, we have $W(s_0,\mu)=0+\delta W(s^I,\mu)$ and $W(s_1,\mu)=1+\delta W(s^I,\mu)$. Therefore, expression (25) becomes

$$(1-\mu)+\delta W\big(s^I,\mu\big). \tag{26}$$

Similarly, action B yields

$$\mu+\delta W\big(s^I,\mu\big). \tag{27}$$

Finally, note that if S is optimal, then the agent stays always in s^I and earns $2/3$ in every period; therefore, $W(s^I,\mu)=(2/3)/(1-\delta)$. It then follows from (25), (26), and (27) that S can be optimal only if $1/3\le\mu\le 2/3$. We show, however, that under any such μ, the agent prefers action O to action S. Therefore, S cannot arise as a Berk–Nash equilibrium outcome. To establish this claim, let us assume that S is optimal. A deviation to action O yields

$$\mu W\big(s^{\mathrm{opt}},\mu\big)+(1-\mu)W(s_0,\mu), \tag{28}$$

where $W(s^{\mathrm{opt}},\mu)=-1/3+\delta(\mu W(s_0,\mu)+(1-\mu)W(s_k,\mu))$. Note that we use the fact that, in deviating to O, the agent would pick the risky alternative in state s^{opt}; otherwise, it could never be optimal to choose O. By also using the fact that $W(s_j,\mu)=j+\delta W(s^I,\mu)$ for $j\in\{0,k\}$, expression (28) becomes

$$-(1/3)\mu+\delta\mu(1-\mu)k+\big(\mu\delta+(1-\mu)\big)\delta W\big(s^I,\mu\big). \tag{29}$$

Using the fact that $W(s^I,\mu)=(2/3)/(1-\delta)$ if S is optimal, we can compare $W(s^I,\mu)$ with (29) and use algebra to conclude that it is strictly lower (hence, the agent prefers to

deviate from S to O) for all values of μ between 1/3 and 2/3 given the assumption that $k > 2 + 4/\delta$.[11] ◊

5.2 *Discussion*

We conclude with additional remarks about the above results.

Guidance for using the equilibrium concept Theorems 2 and 3 suggest that the equilibrium approach is valid in SMDPs that are not subjectively static provided that either identification holds or all states are visited with positive probability (the latter is the case, for example, if every state can be reached from any other state, irrespective of the agent's actions). Alternatively, if either of these conditions fails, the modeler can add small perturbations that either guarantee that identification holds (as we did, for example, in Section 4.1) or small perturbations that guarantee that all states can be reached with positive probability. Of course, there are environments where these perturbations are not justifiable, such as in bandit problems, where the only way to learn about the consequence of an action is to take that action. To the extent to which those environments are not subjectively static, then our results suggest that the equilibrium approach is of limited use in those cases.

Convergence Theorems 2 and 3 do not imply that behavior necessarily stabilizes in an SMDP. In fact, it is well known from the theory of Markov chains that even if no decisions affect the relevant transitions, outcomes need not stabilize without further assumptions; this is also true, for example, in the related context of learning to play Nash equilibrium in games.[12] Thus, the question of convergence remains open at this level of generality. Recently there has been progress tackling convergence, but all in the context of static environments where the only relevant state variable is the agent's belief (Fudenberg et al. 2017, Heidhues et al. 2018, 2021, Esponda et al. 2019, Frick et al. 2020a, and Fudenberg et al. 2020).

Mixed strategies Theorem 3 also suggests that we can interpret a mixed strategy as the limit of the frequency of actions. In particular, even if the agent's action may not settle down, the frequency of actions may; see Esponda et al. (2019) for a formalization of this idea. Alternatively, we can interpret a mixed strategy following the approach of Fudenberg and Kreps (1993), who show that adding small payoff perturbations à la Harsanyi (1973) can provide a *learning* foundation for mixed-strategy Nash equilibria: Agents do not actually mix; instead, every period their payoffs are subject to small perturbations, and what we call the mixed strategy is simply the probability distribution generated by playing *pure* strategies and integrating over the payoff perturbations. We also followed this approach in the paper that introduced Berk–Nash equilibrium in static contexts

[11] The term $W(s^I, \mu)$ is less than expression (29) whenever $2/3 + \mu((2/3)\delta + 1/3) - \delta\mu(1-\mu)k < 0$. For $1/3 \leq \mu \leq 2/3$, the left-hand side of this last expression is largest when $\mu = 2/3$, and replacing this value in the expression, we obtain $k > 2 + 4/\delta$.

[12] For example, in the game-theory literature, general global convergence results have been obtained only in special classes of games, e.g., zero-sum, potential, and supermodular games (Hofbauer and Sandholm 2002).

(Esponda and Pouzo 2016). The same idea applies here at the expense of additional notational burden.[13]

Appendix

A.1 *Proving Lemma 1*

The proof of Lemma 1 relies on the following claim.

Claim A. *(i) For any regular SMDP, there exists $\theta^* \in \Theta$ and $K < \infty$ such that, for all $m \in \Delta(\mathbb{S} \times \mathbb{X})$, $K_Q(m, \theta^*) \leq K$. (ii) Fix any $\theta \in \Theta$ and a sequence $(m_n)_n$ in $\Delta(\mathbb{S} \times \mathbb{X})$ such that $Q_\theta(s' \mid s, x) > 0$ for all $(s', s, x) \in \mathbb{S} \times \mathbb{S} \times \mathbb{X}$ such that $Q(s' \mid s, x) > 0$ and $\lim_{n\to\infty} m_n = m$. Then $\lim_{n\to\infty} K_Q(m_n, \theta) = K_Q(m, \theta)$. (iii) K_Q is (jointly) lower semi-continuous: Fix any $(m_n)_n$ and $(\theta_n)_n$ such that $\lim_{n\to\infty} m_n = m$ and $\lim_{n\to\infty} \theta_n = \theta$. Then $\liminf_{n\to\infty} K_Q(m_n, \theta_n) \geq K_Q(m, \theta)$. (iv) For all $m \in \Delta(\mathbb{S} \times \mathbb{X})$, $\theta \mapsto K_Q(m, \theta)$ is continuous at every $\theta \in \Theta$ such that $K_Q(m, \theta) < \infty$.*

Proof. The proof is very similar to the proof of Claim A in Esponda and Pouzo (2016), so we present only a sketch. Part (i) follows from the third condition in the definition of regular SMDP. Part (ii) follows standard continuity arguments. For part (iii), observe that $K_Q(m_n, \theta_n) = \sum_{s,x} E_{Q(\cdot|s,x)}[\log \frac{Q(S'|s,x)}{Q_{\theta_n}(S'|s,x)}] m_n(s, x)$. It follows that $\sum_{s,x} E_{Q(\cdot|s,x)}[\log Q(S'|s, x)] m_n(s, x) \to \sum_{s,x} E_{Q(\cdot|s,x)}[\log Q(S'|s, x)] m(s, x)$, so it remains to study $\liminf_{n\to\infty} -\sum_{s,x} E_{Q(\cdot|s,x)}[\log Q_{\theta_n}(S'|s, x)] m_n(s, x)$. Suppose the liminf is finite (if not, the result holds trivially). As $\theta \mapsto Q_\theta$ is continuous, then if $m(s, x) > 0$, it follows that $E_{Q(\cdot|s,x)}[\log Q_{\theta_n}(S'|s, x)] m_n(s, x) \to E_{Q(\cdot|s,x)}[\log Q_\theta(S'|s, x)] m(s, x)$. If $m(s, x) = 0$, it follows that $E_{Q(\cdot|s,x)}[\log Q_{\theta_n}(S'|s, x)] m_n(s, x) \to 0 \geq -E_{Q(\cdot|s,x)}[\log Q_\theta(S'|s, x)] m(s, x)$ (by convention $0 \log 0 = 0$). Thus, the desired result holds.

Part (iv). Since $\sum_{s,x} E_{Q(\cdot|s,x)}[\log \frac{Q(S'|s,x)}{Q_\theta(S'|s,x)}] m(s, x) < \infty$, continuity follows from continuity of $\theta \mapsto \log \frac{Q(s'|s,x)}{Q_\theta(s'|s,x)} Q(s'|s, x) m(s, x)$ and the fact that $\mathbb{S} \times \mathbb{X}$ is finite. □

Proof of Lemma 1. (i) By Jensen's inequality and strict concavity of $\ln(\cdot)$, $K_Q(m, \theta) \geq -\sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} \ln(E_{Q(\cdot|s,x)}[\frac{Q_\theta(S'|s,x)}{Q(S'|s,x)}]) m(s, x) = 0$, with equality if and only if $Q_\theta(\cdot \mid s, x) = Q_\theta(\cdot \mid s, x)$ for all (s, x) such that $m(s, x) > 0$.

(ii) The term $\Theta_Q(m)$ *is nonempty.* By Claim A(i), there exists $K < \infty$ such that the minimizers are in the constraint set $\{\theta \in \Theta : K_Q(m, \theta) \leq K\}$. Because $K_Q(m, \cdot)$ is continuous over a compact set, a minimum exists.

The term $\Theta_Q(\cdot)$ *is uhc* and compact-valued. Fix any $(m_n)_n$ and $(\theta_n)_n$ such that $\lim_{n\to\infty} m_n = m$, $\lim_{n\to\infty} \theta_n = \theta$, and $\theta_n \in \Theta_Q(m_n)$ for all n. We establish that $\theta \in \Theta_Q(m)$ (so that $\Theta_Q(\cdot)$ has a closed graph and, by compactness of Θ, it is uhc). Suppose, to obtain a contradiction, that $\theta \notin \Theta_Q(m)$. Then, by Claim A(i), there exists $\hat{\theta} \in \Theta$ and $\varepsilon > 0$ such that $K_Q(m, \hat{\theta}) \leq K_Q(m, \theta) - 3\varepsilon$ and $K_Q(m, \hat{\theta}) < \infty$. By regularity, there exists $(\hat{\theta}_j)_j$ with $\lim_{j\to\infty} \hat{\theta}_j = \hat{\theta}$ and, for all j, $Q_{\hat{\theta}_j}(s' \mid s, x) > 0$ for all $(s', s, x) \in \mathbb{S}^2 \times \mathbb{X}$ *such*

[13] Doraszelski and Escobar (2010) incorporate payoff perturbations in a dynamic environment.

that $Q(s' \mid s, x) > 0$. We show that there is an integer J such that $\hat{\theta}_J$ "does better" than θ_n given m_n, which is a contradiction. Because $K_Q(m, \hat{\theta}) < \infty$, continuity of $K_Q(m, \cdot)$ implies that there exists J large enough such that $|K_Q(m, \hat{\theta}_J) - K_Q(m, \hat{\theta})| \leq \varepsilon/2$. Moreover, Claim A(ii) applied to $\theta = \hat{\theta}_J$ implies that there exists $N_{\varepsilon,J}$ such that, for all $n \geq N_{\varepsilon,J}$, $|K_Q(m_n, \hat{\theta}_J) - K_Q(m, \hat{\theta}_J)| \leq \varepsilon/2$. Thus, for all $n \geq N_{\varepsilon,J}$, $|K_Q(m_n, \hat{\theta}_J) - K_Q(m, \hat{\theta})| \leq |K_Q(m_n, \hat{\theta}_J) - K_Q(m, \hat{\theta}_J)| + |K_Q(m, \hat{\theta}_J) - K_Q(m, \hat{\theta})| \leq \varepsilon$ and, therefore,

$$K_Q(m_n, \hat{\theta}_J) \leq K_Q(m, \hat{\theta}) + \varepsilon \leq K_Q(m, \theta) - 2\varepsilon. \tag{30}$$

Suppose $K_Q(m, \theta) < \infty$. By Claim A(iii), there exists $n_\varepsilon \geq N_{\varepsilon,J}$ such that $K_Q(m_{n_\varepsilon}, \theta_{n_\varepsilon}) \geq K_Q(m, \theta) - \varepsilon$. This result, together with (30), implies that $K_Q(m_{n_\varepsilon}, \hat{\theta}_J) \leq K_Q(m_{n_\varepsilon}, \theta_{n_\varepsilon}) - \varepsilon$. But this contradicts $\theta_{n_\varepsilon} \in \Theta_Q(m_{n_\varepsilon})$. Finally, if $K_Q(m, \theta) = \infty$, Claim A(iii) implies that there exists $n_\varepsilon \geq N_{\varepsilon,J}$ such that $K_Q(m_{n_\varepsilon}, \theta_{n_\varepsilon}) \geq 2K$, where K is the bound defined in Claim A(i). But this also contradicts $\theta_{n_\varepsilon} \in \Theta_Q(m_{n_\varepsilon})$. Thus, $\Theta_Q(\cdot)$ has a closed graph and so $\Theta_Q(m)$ is a closed set. Compactness of $\Theta_Q(m)$ follows from compactness of Θ. Therefore, $\Theta_Q(\cdot)$ is upper hemicontinuous (see Aliprantis and Border 2006, Theorem 17.11). □

A.2 *Proof of Theorem 1*

Let $\mathbb{W} = \Delta(\mathbb{S} \times \mathbb{X}) \times \Delta(\Theta)$ and endow it with the product topology (given by the Euclidean topology for $\Delta(\mathbb{S} \times \mathbb{X})$ and the weak topology for $\Delta(\Theta)$). Clearly, $\mathbb{W} \neq \{\varnothing\}$. Since Θ is compact, $\Delta(\Theta)$ is compact under the weak topology; Σ and $\Delta(\mathbb{S} \times \mathbb{X})$ are also compact. Thus, $\mathbb{W}$ is compact under the product topology and is also convex. Finally, $\mathbb{W} \subseteq \mathbb{M} \times \text{rca}(\Theta)$, where $\mathbb{M}$ is the space of $|\mathbb{S}| \times |\mathbb{X}|$ real-valued matrices and $\text{rca}(\Theta)$ is the space of regular Borel signed measures endowed with the weak topology. The space $\mathbb{M} \times \text{rca}(\Theta)$ is locally convex with a family of seminorms $\{(m, \mu) \mapsto p_f(m, \mu) = \|m\| + |\int_\Omega f(x)\mu(dx)| \colon f \in \mathbb{C}(\Omega)\}$ ($\mathbb{C}(\Omega)$ is the space of real-valued continuous and bounded functions, and $\|.\|$ is understood as the spectral norm). Also, we observe that $(m, \mu) = 0$ if and only if $p_f(m, \mu) = 0$ for all $f \in \mathbb{C}(\Omega)$; thus, $\mathbb{M} \times \text{rca}(\Theta)$ is also Hausdorff.

Let $\mathcal{T} : \mathbb{W} \to 2^{\mathbb{W}}$ be such that $\mathcal{T}(m, \mu) = \mathcal{M}(m, \mu) \times \Delta(\Theta_Q(m))$, where

$$(m, \mu) \mapsto \mathcal{M}(m, \mu) \equiv \left\{m' \in \Delta(\mathbb{S} \times \mathbb{X}) \colon m' \in \mathcal{O}(\mu) \text{ and } m'_{\mathbb{S}} = Q[m]\right\},$$

where for any $\mu \in \Delta(\Theta)$, $\mathcal{O}(\mu)$ is the set of all $m' \in \Delta(\mathbb{S} \times \mathbb{X})$ that satisfy optimality (i.e., for all $(s, x) \in \mathbb{S} \times \mathbb{X}$ such that $m(s, x) > 0$, x is optimal given s in the MDP($\bar{Q}_\mu$), where $\bar{Q}_\mu = \int_\Theta Q_\theta \mu(d\theta)$) and $m \mapsto Q[m](\cdot) = \sum_{(s,x) \in \mathbb{S} \times \mathbb{X}} Q(\cdot \mid s, x) m(s, x) \in \Delta(\mathbb{S})$.

Hence, to show the existence of an equilibrium, it is sufficient to show that $\mathcal{T}$ has a fixed point. Since $\mathbb{W}$ is a nonempty, compact, convex subset of a locally Hausdorff space, there exists a fixed point of $\mathcal{T}$ by the Kakutani–Fan–Glicksberg theorem (see Aliprantis and Border 2006, Corollary 17.55) if $\mathcal{T}$ is nonempty, convex-valued, compact-valued, and upper hemicontinuous under the product topology (and, hence, it has a closed graph (see Aliprantis and Border 2006, Theorem 17.11)).

Nonempty We show that, for every $(m,\mu)\in\mathbb{W}$, $\mathcal{M}(m,\mu)$ and $\Theta_Q(m)$ are nonempty and, thus, so is $\mathcal{T}(m,\mu)$. Nonemptiness of $\Theta_Q(m)$ follows from Lemma 1. For nonemptiness of $\mathcal{M}(m,\mu)$, note that, for each s, the argmax of the MDP($\bar{Q}_\mu$) is nonempty; in particular, there exists $m'_{\mathbb{X}|\mathbb{S}}$ such that, for each s, any action in the support of $m'_{\mathbb{X}|\mathbb{S}}(\cdot\mid s)$ is optimal. Then $m'=m'_{\mathbb{X}|\mathbb{S}}Q[m]\in\Delta(\mathbb{S}\times\mathbb{X})$ is an element of $\mathcal{M}(m,\mu)$.

Convex-valued It suffices to show that for every $(m,\mu)\in\mathbb{W}$, both $\Delta(\Theta_Q(m))$ and $\mathcal{M}(m,\mu)$ are convex. Convexity of the former is obvious. To show convexity of $\mathcal{M}(m,\mu)$, take any m_1 and m_2 in $\mathcal{M}(m,\mu)$. For any $\lambda\in[0,1]$, it is clear that $\lambda m_{\mathbb{S},1}+(1-\lambda)m_{\mathbb{S},2}=Q[m]$. Also, any (s,x) in the support of $\lambda m_1+(1-\lambda)m_2$ has to be in the support of either m_1 or m_2, and, thus, x is optimal given s in the MDP($\bar{Q}_\mu$). Therefore, $\lambda m_1+(1-\lambda)m_2\in\mathcal{M}(m,\mu)$.

Compact-valued For every $(m,\mu)\in\mathbb{W}$, $\Delta(\Theta_Q(m))$ is compact (under the weak topology) because $\Theta_Q(m)$ is compact (see Aliprantis and Border 2006, Theorem 15.11). The set $\Delta(\mathbb{S}\times\mathbb{X})$ is compact, so to show compactness of $\mathcal{M}(m,\mu)$, it suffices to show that it is closed. Take any convergent (to some m') sequence $(m'_n)_n$ in $\mathcal{M}(m,\mu)$. It is clear that $m'=Q[m]$. Taking any (s,x) in the support of m', it follows that for sufficiently large n, (s,x) are in the support of m'_n and so x is optimal given s in the MDP($\bar{Q}_\mu$). Thus, $\mathcal{T}$ is compact-valued under the product topology.

Upper hemicontinuity By Aliprantis and Border (2006, Theorem 17.28), to show upper hemicontinuity of $\mathcal{T}$ under the product topology, it suffices to show that both $m\mapsto\Delta(\Theta_Q(m))$ and $\mathcal{M}$ are uhc. The correspondence $\Theta_Q(\cdot)$ is upper hemicontinuous; hence, the correspondence $\Delta(\Theta_Q(\cdot))$ is too (see Aliprantis and Border 2006, Theorem 17.13). To show upper hemicontinuity of $\mathcal{M}$, take a sequence $(m'_n,m_n,\mu_n)_n$ in Graph($\mathcal{M}$) that converges to (m',m,μ). It is clear that $m'_{\mathbb{S}}=Q[m]$, so we need to show only that $\mathcal{O}$ is uch.

Claim B. *The action $\mathcal{O}$ is uhc.*

Proof. Take any sequence $(m'_n,\mu_n)_n$ in Graph($\mathcal{O}$) that converges to (m',μ). Take any (s,x) in the support of m'. Then, for sufficiently large n, (s,x) are in the support of m'_n and, therefore, x is optimal given s in the MDP($\bar{Q}_{\mu_n}$). By standard arguments, $(s,Q)\mapsto M(s,Q)\equiv\arg\max_{\hat{x}\in\mathbb{X}}\int_{\mathbb{S}}\{\pi(s,\hat{x},s')+\delta V(s')\}Q(ds'|s,\hat{x})$ is uhc (since $\mathbb{S}\times\mathbb{X}$ are finite, Q belongs to the space of real-valued matrices with its natural topology). Since $\theta\mapsto Q_\theta$ is bounded and continuous, $\mu\mapsto\bar{Q}_\mu$ is continuous under the weak topology. Thus, $(s,\mu)\mapsto M(s,\bar{Q}_\mu)$ is uhc. Since $x\in M(s,\bar{Q}_{\mu_n})$ for all n, it follows that $x\in M(s,\bar{Q}_\mu)$; therefore, x is optimal given s in the MDP($\bar{Q}_\mu$), as desired. □

A.3 *Proof of Lemma 2*

For the proof of Lemma 2, we rely on the following definitions and the claim below. Define $K^*_Q(m)\equiv\inf_{\theta\in\Theta}K_Q(m,\theta)$ and let $\hat{\Theta}\subseteq\Theta$ be a dense set such that, for all $\theta\in\hat{\Theta}$, $Q_\theta(s'\mid s,x)>0$ for all $(s,x,s')\in\mathbb{S}\times\mathbb{X}\times\mathbb{S}$ such that $Q(s'\mid s,x)>0$. Existence of such a set $\hat{\Theta}$ follows from the regularity assumption.

Claim C. *Suppose* $\lim_{t\to\infty}\|m_t - m\| = 0$ *a.s.-*$\mathbf{P}^f$. *Then (i) for all* $\theta \in \hat{\Theta}$,

$$\lim_{t\to\infty} t^{-1}\sum_{\tau=1}^{t}\log\frac{Q(s_\tau|s_{\tau-1},x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1})} = \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|s,x)}\left[\log\frac{Q(S'|s,x)}{Q_\theta(S'|s,x)}\right]m(s,x)$$

a.s.-$\mathbf{P}^f$*; (ii) for* $\mathbf{P}^f$*-almost all* $h \in \mathbb{H}$, *and for any* $\epsilon > 0$ *and* $\alpha = (\inf_{\Theta:\, d_m(\theta)\geq\epsilon} K_Q(m,\theta) - K_Q^*(m))/3$, *there exists* T *such that, for all* $t \geq T$,

$$t^{-1}\sum_{\tau=1}^{t}\log\frac{Q(s_\tau|s_{\tau-1},x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1})} \geq K_Q^*(m) + \frac{3}{2}\alpha$$

for all $\theta \in \{\Theta:\, d_m(\theta) \geq \epsilon\}$, *where* $d_m(\theta) = \inf_{\tilde{\theta}\in\Theta_Q(m)}\|\theta - \tilde{\theta}\|$.

Proof. (The proof is similar to the proof of Claim B in Esponda and Pouzo 2016.) We first show that for $\mathbf{P}^f$-almost all histories and any $\epsilon > 0$, there exists a M_ϵ such that

$$|t^{-1}\sum_{\tau=1}^{t}\log Q(s_\tau|s_{\tau-1},x_{\tau-1}) = \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|s,x)}\big[\log Q\big(S'|s,x\big)\big]m(s,x)| < \epsilon$$

for all $t \geq M_\epsilon$. To do this, for any $\tau \in \{1,2,\ldots\}$, let $l_\tau \equiv \log Q(s_\tau|s_{\tau-1},x_{\tau-1}) - E_{Q(\cdot|s_{\tau-1},x_{\tau-1})}[\log Q(S'|s_{\tau-1},x_{\tau-1})]$. Observe that for all $z \in \mathbb{S}^2\times\mathbb{X}$, $E_{\mathbf{P}^f(\cdot|h^t)}[l_{t+1}] = 0$ a.s.-$\mathbf{P}^f$, where $\mathbf{P}^f(\cdot|h^t)$ denotes the conditional probability induced by $\mathbf{P}^f$ given the partial history h^t. Moreover, $\sup_t E_{\mathbf{P}^f}[l_t^2] \leq \sup_t \sum_{\tau=1}^{t}\tau^{-2}E[\sum_{s'\in\mathbb{S}}(\log Q(s'|S,X))^2 Q(s'\mid S,X)] < \infty$ because $x \mapsto (\log x)^2 x$ is bounded and $\sum_\tau \tau^{-2} < \infty$. Thus, an application of the the Martingale Convergence Theorem and Kronecker's lemma imply that

$$\lim_{t\to\infty} t^{-1}\sum_{\tau=1}^{t}\big(\log Q(s_\tau|s_{\tau-1},x_{\tau-1}) - E_{Q(\cdot|s_{\tau-1},x_{\tau-1})}\big[\log Q\big(S'|s_{\tau-1},x_{\tau-1}\big)\big]\big) = 0$$

a.s.-$\mathbf{P}^f$. Therefore, to establish the desired result, it suffices to show that

$$\begin{aligned}\lim_{t\to\infty} t^{-1}&\sum_{\tau=1}^{t} E_{Q(\cdot|s_{\tau-1},x_{\tau-1})}\big[\log Q\big(S'|s_{\tau-1},x_{\tau-1}\big)\big]\\ &- \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|s,x)}\big[\log Q\big(S'|s,x\big)\big]m(s,x) = 0\end{aligned} \tag{31}$$

a.s.-$\mathbf{P}^f$. Observe that

$$\begin{aligned}&t^{-1}\sum_{\tau=1}^{t} E_{Q(\cdot|s_{\tau-1},x_{\tau-1})}\big[\log Q\big(S'|s_{\tau-1},x_{\tau-1}\big)\big]\\ &= \sum_{s,x\in\mathbb{S}\times\mathbb{X}} t^{-1}\sum_{\tau=1}^{t} 1_{(s,x)}(s_{\tau-1},x_{\tau-1})E_{Q(\cdot|s,x)}\big[\log Q\big(S'|s,x\big)\big]\\ &= \sum_{s,x\in\mathbb{S}\times\mathbb{X}} m_t(s,x)E_{Q(\cdot|s,x)}\big[\log Q\big(S'|s,x\big)\big].\end{aligned}$$

Equation (31) follows because $\lim_{t\to\infty}\|m_t - m\| = 0$ a.s.-$\mathbf{P}^f$ and $E_{Q(\cdot|s,x)}[\log Q(S'|s,x)] = \sum_{s'\in\mathbb{S}} \log Q(s'|s,x)Q(s'|s,x)$ is bounded for all $(s,x)\in\mathbb{S}\times\mathbb{X}$. So, to establish parts (i) and (ii), it remains to control only the expression

$$-\lim_{t\to\infty} t^{-1}\sum_{\tau=1}^{t}\big(\log Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1}) - E_{Q(\cdot|s_{\tau-1},x_{\tau-1})}\big[\log Q_\theta\big(S'|s_{\tau-1},x_{\tau-1}\big)\big]\big).$$

Part (i). Pointwise over $\hat{\Theta}$,

$$\lim_{t\to\infty} t^{-1}\sum_{\tau=1}^{t}\big(\log Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1}) - E_{Q(\cdot|s_{\tau-1},x_{\tau-1})}\big[\log Q_\theta\big(S'|s_{\tau-1},x_{\tau-1}\big)\big]\big) = 0$$

a.s.-$\mathbf{P}^f$ by essentially the same arguments used in the first part of the proof.

Part (ii). For any $\xi > 0$, let $\Theta_\xi \subseteq \Theta$ such that $\theta\in\Theta_\xi$ if and only if $Q_\theta(s'|s,x)\geq\xi$ for all (s',s,x) such that $P_m(s',s,x) > 0$. Also, observe that

$$\lim_{t\to\infty} t^{-1}\sum_{\tau=1}^{t}\log Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1}) = \sum_{s',s,x\in\mathbb{S}^2\times\mathbb{X}} \mathrm{freq}_t\big(s',s,x\big)\log Q_\theta\big(s'|s,x\big),$$

where $z\mapsto \mathrm{freq}_t(z) \equiv t^{-1}\sum_{\tau=1}^{t} 1_z(s_\tau,s_{\tau-1},x_{\tau-1})$. Let $(s',s,x)\mapsto P_m(s',s,x)\equiv Q(s'|s,x)m(s,x)$. By essentially the same argument used in the first part of the proof, it follows that for any $\zeta>0$ and $\mathbf{P}^f$-almost any h, there exists a T_ζ such that $\max_{z\in\mathbb{S}^2\times\mathbb{X}}|\mathrm{freq}_t(z) - P_m(z)| < \zeta$ for all $t\geq T_\zeta$.

Hence, for any $\theta\in\{\Theta\setminus\Theta_\xi\}\cap\{\Theta\colon d_m(\theta)\geq\epsilon\}$,

$$\begin{aligned}
&\sum_{(s',s,x)\in\mathbb{S}^2\times\mathbb{X}} \mathrm{freq}_t\big(s',s,x\big)\log Q_\theta\big(s'|s,x\big)\\
&\quad\leq \sum_{(s',s,x)\colon P_m(s',s,x)>0} \big(P_m\big(s',s,x\big)-\zeta\big)\log Q_\theta\big(s'|s,x\big)\\
&\quad\leq \sum_{s,x\in\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|s,x)}\big[\log Q_\theta\big(s'|s,x\big)\big]m(s,x)\\
&\qquad -\zeta\sum_{(s',s,x)\colon P_m(s',s,x)>0}\log Q_\theta\big(s'|s,x\big)
\end{aligned}$$

for all $t\geq T_\zeta$. Therefore,

$$t^{-1}\sum_{\tau=1}^{t}\log\frac{Q(s_\tau|s_{\tau-1},x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1})} \geq K_Q(m,\theta) + \zeta\sum_{(s',s,x)\colon P_m(s',s,x)>0}\log Q_\theta\big(s'|s,x\big)$$

for any $t\geq\max\{T_\zeta,M_\alpha\}$. By definition of $\{\Theta\colon d_m(\theta)\geq\epsilon\}$, it follows that

$$t^{-1}\sum_{\tau=1}^{t}\log\frac{Q(s_\tau|s_{\tau-1},x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1})} \geq K_Q^*(m) + 2\alpha + \zeta\sum_{(s',s,x)\colon P_m(s',s,x)>0}\log Q_\theta\big(s'|s,x\big)$$

for any $t \geq T_\zeta$. Since $\theta \in \{\Theta \setminus \Theta_\xi\} \cap \{\Theta\colon d_m(\theta) \geq \epsilon\}$, let $z_\theta = (s'_\theta, s_\theta, x_\theta)$ be such that $Q_\theta(s'_\theta|s_\theta, x_\theta) < \xi$ and $P_m(z_\theta) > 0$, and note that $\zeta \sum_{(s',s,x)\colon P_m(s',s,x)>0} \log Q_\theta(s'|s,x) \leq \zeta \log \xi p_L$, where $p_L \equiv \min\{P_m(z)\colon P_n(z) > 0\}$. This implies that there exists a ζ^* such that $\zeta^* \log \xi p_L \leq -0.5\alpha$ and so

$$t^{-1}\sum_{\tau=1}^{t} \log \frac{Q(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})} \geq K_Q^*(m) + \frac{3}{2}\alpha$$

for any $t \geq \max\{T_{\zeta^*}, M_\alpha\}$.

For any $\theta \in \Theta_\xi \cap \{\Theta\colon d_m(\theta) \geq \epsilon\}$, it follows that $\sum_{(s',s,x)\in\mathbb{S}^2\times\mathbb{X}} \text{freq}_t(s', s, x) \log Q_\theta(s'|s, x) \leq \ln \xi$,

$$t^{-1}\sum_{\tau=1}^{t} \log \frac{Q(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})} \geq -\ln \xi + \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|s,x)}\big[\log Q\big(S'|s, x\big)\big]m(s, x) - 1$$

for any $t \geq M_1$. Since $\sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|s,x)}[\log Q(S'|s, x)]m(s, x)$ is finite, we can choose ξ such that the right-hand side is greater than or equal to $K_Q^*(m) + \frac{3}{2}\alpha$.

We thus showed that for $\mathbf{P}^f$-almost all $h \in \mathbb{H}$ and for any $\epsilon > 0$, there exists T such that, for all $t \geq T$,

$$t^{-1}\sum_{\tau=1}^{t} \log \frac{Q(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})} \geq K_Q^*(m) + \frac{3}{2}\alpha$$

for all $\theta \in \{\Theta\colon d_m(\theta) \geq \epsilon\}$, as desired. □

Proof of Lemma 2. It suffices to show that $\lim_{t\to\infty} \int_\Theta d_m(\theta)\mu_t(d\theta) = 0$ a.s.-$\mathbf{P}^f$ over $\mathcal{H}$. For any $\eta > 0$, let $\Theta_\eta(m) = \{\theta \in \Theta : d_m(\theta) < \eta\}$ and $\hat{\Theta}_\eta(m) = \hat{\Theta} \cap \Theta_\eta(m)$ (the set $\hat{\Theta}$ is defined in the third condition of Definition 5, i.e., regularity). We now show that $\mu_0(\hat{\Theta}_\eta(m)) > 0$. By Lemma 1, $\Theta_Q(m)$ is nonempty. By denseness of $\hat{\Theta}$, $\hat{\Theta}_\eta(m)$ is nonempty. Nonemptiness and continuity of $\theta \mapsto Q_\theta$, imply that there exists a nonempty open set $U \subseteq \hat{\Theta}_\eta(m)$. By full support, $\mu_0(\hat{\Theta}_\eta(m)) > 0$. Also, observe that for any $\epsilon > 0$, $\{\Theta\colon d_m(\theta) \geq \epsilon\}$ is compact. This follows from compactness of Θ and continuity of $\theta \mapsto d_m(\theta)$ (which follows by Lemma 1 and an application of the theorem of the maximum). Compactness of $\{\Theta\colon d_m(\theta) \geq \epsilon\}$ and lower semicontinuity of $\theta \mapsto K_Q(m, \theta)$ (see Claim A(iii)) imply that $\inf_{\Theta\colon d_m(\theta)\geq\epsilon} K_Q(m, \theta) = \min_{\Theta\colon d_m(\theta)\geq\epsilon} K_Q(m, \theta) > K_Q^*(m)$. Let $\alpha \equiv (\min_{\Theta\colon d_m(\theta)\geq\epsilon} K_Q(m, \theta) - K_Q^*(m))/3 > 0$. Also, let $\eta > 0$ be chosen such that $K_Q(m, \theta) \leq K_Q^*(m) + 0.25\alpha$ for all $\theta \in \Theta_\eta(m)$ (such η always exists by continuity of $\theta \mapsto K_Q(m, \theta)$).

Let $\mathcal{H}_1$ be the subset of $\mathcal{H}$ for which the statements in Claim C hold; note that $\mathbf{P}^f(\mathcal{H} \setminus \mathcal{H}_1) = 0$. Henceforth, fix $h \in \mathcal{H}_1$; we omit h from the notation to ease the notational burden. By simple algebra and the fact that d_m is bounded in Θ, it follows that, for all

$\epsilon > 0$ and some finite $C > 0$,

$$\int_\Theta d_m(\theta)\mu_t(d\theta) = \frac{\int_\Theta d_m(\theta)Q_\theta(s_t \mid s_{t-1}, x_{t-1})\mu_{t-1}(d\theta)}{\int_\Theta Q_\theta(s_t \mid s_{t-1}, x_{t-1})\mu_{t-1}(d\theta)} = \frac{\int_\Theta d_m(\theta)Z_t(\theta)\mu_0(d\theta)}{\int_\Theta Z_t(\theta)\mu_0(d\theta)}$$
$$\leq \epsilon + C\frac{\int_{\{\Theta:\, d_m(\theta)\geq\epsilon\}} Z_t(\theta)\mu_0(d\theta)}{\int_{\hat\Theta_\eta(m)} Z_t(\theta)\mu_0(d\theta)} \equiv \epsilon + C\frac{A_t(\epsilon)}{B_t(\eta)},$$

where

$$Z_t(\theta) \equiv \prod_{\tau=1}^t \frac{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q(s_\tau|s_{\tau-1}, x_{\tau-1})} = \exp\left\{-\sum_{\tau=1}^t \log\left(\frac{Q(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})}\right)\right\}.$$

Hence, it suffices to show that

$$\limsup_{t\to\infty}\left\{\exp\left\{t\left(K^*(m) + 0.5\alpha\right)\right\}A_t(\epsilon)\right\} = 0 \tag{32}$$

and

$$\liminf_{t\to\infty}\left\{\exp\left\{t\left(K^*_Q(m) + 0.5\alpha\right)\right\}B_t(\eta)\right\} = \infty. \tag{33}$$

Regarding (32), we first show that

$$\lim_{t\to\infty} \sup_{\{\Theta:\, d_m(\theta)\geq\epsilon\}} \left\{\left(K^*_Q(m) + 0.5\alpha\right) - t^{-1}\sum_{\tau=1}^t \log\frac{Q(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})}\right\} \leq \text{const} < 0.$$

To show this, note that, by Claim C(ii) there exists a T, such that for all $t \geq T$, $t^{-1}\sum_{\tau=1}^t \log(Q(s_\tau|s_{\tau-1}, x_{\tau-1})/Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})) \geq K^*_Q(m) + \frac{3}{2}\alpha$ for all $\theta \in \{\Theta:\, d_m(\theta) \geq \epsilon\}$. Thus,

$$\lim_{t\to\infty} \sup_{\{\Theta:\, d_m(\theta)\geq\epsilon\}} \left\{K^*_Q(m) + \frac{\alpha}{2} - t^{-1}\sum_{\tau=1}^t \log\frac{Q(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})}\right\} \leq -\alpha.$$

Therefore,

$$\begin{aligned}&\limsup_{t\to\infty}\left\{\exp\left\{t\left(K^*_Q(m) + 0.5\alpha\right)\right\}A_t(\epsilon)\right\}\\ &\leq \limsup_{t\to\infty} \sup_{\{\Theta:\, d_m(\theta)\geq\epsilon\}} \exp\left\{t\left(\left(K^*_Q(m) + 0.5\alpha\right) - t^{-1}\sum_{\tau=1}^t \log\frac{Q(s_\tau|s_{\tau-1}, x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1}, x_{\tau-1})}\right)\right\}\\ &= 0.\end{aligned}$$

Regarding (33), by Fatou's lemma and some algebra, it suffices to show that

$$\liminf_{t\to\infty} \exp\left\{t\left(K^*_Q(m) + 0.5\alpha\right)\right\}Z_t(\theta) = \infty > 0$$

(pointwise on $\theta \in \hat{\Theta}_\eta(m)$) or, equivalently,

$$\liminf_{t\to\infty}\left(K_Q^*(m) + 0.5\alpha - t^{-1}\sum_{\tau=1}^{t}\log\frac{Q(s_\tau|s_{\tau-1},x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1})}\right) > 0.$$

By Claim C(i),

$$\liminf_{t\to\infty}\left(K_Q^*(m) + 0.5\alpha - t^{-1}\sum_{\tau=1}^{t}\log\frac{Q(s_\tau|s_{\tau-1},x_{\tau-1})}{Q_\theta(s_\tau|s_{\tau-1},x_{\tau-1})}\right) = K_Q^*(m) + 0.5\alpha - K_Q(m,\theta)$$

(pointwise on $\theta \in \hat{\Theta}_\eta(m)$). By our choice of η, the right-hand side is greater than 0.25α and our desired result follows. □

A.4 *Proof of Theorem 2*

Let $\mathcal{H}$ be the set of histories such that $(m_t)_t$ converges to m. By hypothesis, $\mathbf{P}^f(\mathcal{H}) > 0$. By Lemma 2, there exists a set $\mathcal{H}'$ with $\mathbf{P}^f(\mathcal{H}') = \mathbf{P}^f(\mathcal{H}) > 0$ such that every history in $\mathcal{H}'$ satisfies the result stated in Lemma 2. Throughout, we fix a history $h \in \mathcal{H}'$. Henceforth, we omit the history from the notation.

Also, let $(\mu, s) \mapsto M(s,\mu) \equiv \arg\max_{x\in\mathbb{X}} \int_{\mathbb{S}}\{\pi(s,x,s') + \delta W(s', B(s,x,s',\mu))\}\bar{Q}_\mu(ds'|s, x)$, which by standard arguments is uhc.

We first establish conditions (i) and (ii) in the definition of Berk–Nash equilibrium (Definition 7). Let (s,x) be such that $m(s,x) > 0$. Since $(m_t)_t$ converges to m, (s,x) occurs infinitely often along the history, so we can find a subsequence along which (s,x) occurs along the entire subsequence: $(s_{t(j)}, x_{t(j)}) = (s,x)$ for all j. By compactness of $\Delta(\Theta)$, we can take a further subsequence such that $\mu_{s,x} = \lim_{k\to\infty}\mu_{t(j(k))}$ exists. By our choice of history (see the beginning of the proof) and Lemma 2, $\mu_{s,x} \in \Delta(\Theta_Q(m))$. Also, since $x \in M(s, \mu_{t(j(k))})$ for all k and $\lim_{k\to\infty}\mu_{t(j(k))} = \mu_{s,x}$, upper hemicontinuity of $M(s,\cdot)$ implies that $x \in M(s,\mu_{s,x})$. Thus, we have shown that, for any (s,x) such that $m(s,x) > 0$, there exists $\mu_{s,x} \in \Delta(\Theta_Q(m))$ such that

$$x \in \arg\max_{\hat{x}\in\mathbb{X}} \int_{\mathbb{S}}\big\{\pi\big(s,\hat{x},s'\big) + \delta W\big(s', B\big(s,\hat{x},s',\mu_{s,x}\big)\big)\big\}\bar{Q}_{\mu_{s,x}}\big(ds'|s,\hat{x}\big). \tag{34}$$

We now consider each case in Theorem 2 separately. Consider first the case where identification holds. Identification implies that there exists Q_m^* such that, for all $\mu \in \Delta(\Theta_Q(m))$, $\bar{Q}_\mu = Q_m^*$. Note also that the posterior given $\mu \in \Delta(\Theta_Q(m))$ must also be in $\Delta(\Theta_Q(m))$, and so expression (34) implies that x is optimal in the MDP(Q_m^*). Thus, picking any $\mu \in \Delta(\Theta_Q(m))$, we have shown that, for all (s,x) in the support of $m(s,x)$, condition (i) is satisfied. Because $\mu \in \Delta(\Theta_Q(m))$, condition (ii) is also satisfied.

Consider next the case where the SMDP is subjectively static. In this case, the payoff function, the value function, the Bayesian operator, and the subjective transition probability function do not depend on s, and so, in a slight abuse of notation, we drop s from

subsequent expressions. For any $x' \in \mathbb{X}$,

$$\begin{aligned}
&\int_{\mathbb{S}} \{\pi(x,s') + \delta W(B(x,s',\mu_{s,x}))\}\bar{Q}_{\mu_{s,x}}(ds'|x) \\
&\quad = \int_{\mathbb{S}} \pi(x,s')\bar{Q}_{\mu_{s,x}}(ds'|x) + \delta W(\mu_{s,x}) \\
&\quad \geq \int_{\mathbb{S}} \{\pi(x',s') + \delta W(B(x',s',\mu_{s,x}))\}\bar{Q}_{\mu_{s,x}}(ds'|x') \\
&\quad \geq \int_{\mathbb{S}} \pi(x',s')\bar{Q}_{\mu_{s,x}}(ds'|x') + \delta W(\mu_{s,x}),
\end{aligned}$$

where the first line follows from weak identification (since (s,x) is in the support of m, weak identification implies $B(x,s',\mu_{s,x}) = \mu_{s,x}$ for all s' in the support of $\bar{Q}_{\mu_{s,x}}(ds'|x)$), the second line follows from (34), and the third line follows from the convexity of the value function $\mu \mapsto W(\mu)$ (which we prove at the end of this proof) and the Martingale property of Bayesian updating (which imply, using Jensen's inequality, $\int_{\mathbb{S}} W(B(x',s',\mu_{s,x})\bar{Q}_{\mu_{s,x}}(ds'|x') \geq W(\int_{\mathbb{S}} B(x',s',\mu_{s,x})\bar{Q}_{\mu_{s,x}}(ds'|x')) = W(\mu_{s,x})$.) Therefore,

$$x \in \arg\max_{\hat{x}\in\mathbb{X}} \int_{\mathbb{S}} \pi(\hat{x},s')\bar{Q}_{\mu_{s,x}}(ds'|\hat{x}). \tag{35}$$

Thus, for the subjectively static SMDP, we have shown that, for any (s,x) in the support of m, there exists a belief $\mu_{s,x} \in \Delta(\Theta_Q(m))$ such that (35) is satisfied (which, for this special case, means that x is optimal given s in the MDP($\bar{Q}_{\mu_{s,x}}$)).

It remains to establish that we can pick $\mu_{s,x}$ to be the same for all (s,x) in the support of m. We use the assumption of weak identification to establish this claim. Let (s^*,x^*) be any other element in the support of m. By repeating the argument above, there exists $\mu_{s^*,x^*} \in \Delta(\Theta_Q(m))$ such that

$$x^* \in \arg\max_{\hat{x}\in\mathbb{X}} \int_{\mathbb{S}} \pi(\hat{x},s')\bar{Q}_{\mu_{s^*,x^*}}(ds'|\hat{x}). \tag{36}$$

By weak identification and the fact that both $\mu_{s,x}$ and μ_{s^*,x^*} belong to $\Delta(\Theta_Q(m))$, then $\bar{Q}_{\mu_{s^*,x^*}}(\cdot|\tilde{s},\tilde{x}) = \bar{Q}_{\mu_{s,x}}(\cdot|\tilde{s},\tilde{x})$ for all $(\tilde{s},\tilde{x})$ in the support of m. Therefore, for any $x' \in \mathbb{X}$,

$$\begin{aligned}
\int_{\mathbb{S}} \pi(x^*,s')\bar{Q}_{\mu_{s,x}}(ds'|x^*) &= \int_{\mathbb{S}} \pi(x^*,s')\bar{Q}_{\mu_{s^*,x^*}}(ds'|x^*) \\
&\geq \int_{\mathbb{S}} \pi(x,s')\bar{Q}_{\mu_{s^*,x^*}}(ds'|x) \\
&= \int_{\mathbb{S}} \pi(x,s')\bar{Q}_{\mu_{s,x}}(ds'|x) \\
&\geq \int_{\mathbb{S}} \pi(x',s')\bar{Q}_{\mu_{s,x}}(ds'|x'),
\end{aligned}$$

where the two equalities follow from the implication of weak identification mentioned above and the two inequalities follow from (36) and (35), respectively. Thus, we can use the same belief $\mu_{s,x}$ to support any state–action pair (s^*,x^*) in the support of m.

We conclude by showing condition (iii) in the definition of Berk–Nash equilibrium. Let $m \mapsto Q[m](s') \equiv \sum_{(s,x)\in\mathbb{S}\times\mathbb{X}} Q(s' \mid s,x)m(s,x)$ for any $s' \in \mathbb{S}$. We want to show that $m_{\mathbb{S}} = Q[m]$. By the triangle inequality,

$$\big\|m_{\mathbb{S}} - Q[m]\big\| \le \Big\|m_{\mathbb{S}}(\cdot) - \sum_{x\in\mathbb{X}} m_{t+1}(\cdot,x)\Big\| + \Big\|\sum_{x\in\mathbb{X}} m_{t+1}(\cdot,x) - Q[m_t]\Big\| + \big\|Q[m_t] - Q[m]\big\|.$$

As $(m_t)_t$ converges to m, the first and the third terms in the right-hand side vanish. We now show that the second term also vanishes and, thus, conclude the verification of condition (iii). Observe that for any $s' \in \mathbb{S}$,

$$\begin{aligned}
&\sum_{x\in\mathbb{X}} m_{t+1}\big(s',x\big) - Q[m_t]\big(s'\big)\\
&\quad= (t+1)^{-1}\sum_{\tau=1}^{t+1} \mathbf{1}_{s'}(s_\tau) - t^{-1}\sum_{\tau=1}^{t} Q\big(s' \mid s_\tau, x_\tau\big)\\
&\quad= t^{-1}\sum_{\tau=1}^{t}\big\{\mathbf{1}_{s'}(s_{\tau+1}) - Q\big(s' \mid s_t,x_t\big)\big\} + \frac{\mathbf{1}_{s'}(s_1) + t^{-1}\sum_{\tau}^{t}\mathbf{1}_{s'}(s_{\tau+1})}{t+1}.
\end{aligned}$$

The second summand of the right-hand side vanishes as $t \to \infty$. Regarding the first one, observe that for any $t \in \mathbb{N}$, $E_{\mathbf{P}^f}[\mathbf{1}_s(s_{t+1}) \mid h^t] = Q(s' \mid s_t, x_t)$, where $E_{\mathbf{P}^f}[\cdot \mid h^t]$ is the conditional expectation under $\mathbf{P}^f$ given history h^t. Let $\zeta_t \equiv \sum_{\tau=1}^{t} \tau^{-1}\{\mathbf{1}_{s'}(s_{\tau+1}) - Q(s' \mid s_t, x_t)\}$ and note that $\sup_t E_{\mathbf{P}^f}[\zeta_t^2] \le 2\sup_t \sum_{\tau=1}^{t}\tau^{-2} < \infty$. Thus, by the Martingale convergence theorem, the process $(\zeta_t)_{t=1}^{\infty}$ converges $\mathbf{P}^f$-a.s. to ζ. Kronecker's lemma implies that $\lim_{t\to\infty} t^{-1}\sum_{\tau=1}^{t}\{\mathbf{1}_{s'}(s_{\tau+1}) - Q(s' \mid s_t,x_t)\} = 0$ $\mathbf{P}^f$-a.s. Without loss of generality, we assume the history h satisfies this limit and, thus, $\lim_{t\to\infty} \| \sum_{x\in\mathbb{X}} m_{t+1}(\cdot,x) - Q[m_t]\| = 0$.

Proof that $\mu \mapsto W(\mu)$ is convex: The value function is unique, so it suffices to show that the Bellman operator maps convex functions into themselves. To do this, let μ_1 and μ_2 be in $\Delta(\Theta)$; for any $\lambda \in (0,1)$, let $\mu_\lambda \equiv \lambda\mu_1 + (1-\lambda)\mu_2$ and $\mu \mapsto G(\mu)$ be convex. Define

$$B[G](\mu_\lambda) \equiv \max_{x\in\mathbb{X}} \int \big\{\pi\big(x,s'\big) + \delta G\big(B\big(x,s',\mu_\lambda\big)\big)\big\}\bar{Q}_{\mu_\lambda}\big(ds' \mid x\big).$$

Note that

$$\begin{aligned}
&\big(x,s'\big) \mapsto B\big(x,s',\mu_\lambda\big)\\
&\qquad= \lambda\frac{\int Q_\theta\big(s' \mid x\big)\mu_1(d\theta)}{\int Q_\theta\big(s' \mid x\big)\mu_\lambda(d\theta)}B\big(x,s',\mu_1\big) + (1-\lambda)\frac{\int Q_\theta\big(s' \mid x\big)\mu_2(d\theta)}{\int Q_\theta\big(s' \mid x\big)\mu_\lambda(d\theta)}B\big(x,s',\mu_2\big).
\end{aligned}$$

By convexity of G,

$$\int G\big(B(x,s',\mu_\lambda)\big)\bar{Q}_{\mu_\lambda}(ds' \mid x)$$
$$\leq \lambda \int G\big(B(x,s',\mu_1)\big)\bar{Q}_{\mu_1}(ds' \mid x) + (1-\lambda)\int G\big(B(x,s',\mu_2)\big)\bar{Q}_{\mu_2}(ds' \mid x).$$

Therefore,

$$\begin{aligned} B[G](\mu_\lambda) &\leq \max_{x\in\mathbb{X}} \lambda \int \left\{\pi(x,s') + \delta \int G\big(B(x,s',\mu_1)\big)\right\}\bar{Q}_{\mu_1}(ds' \mid x) \\ &\quad + (1-\lambda)\int \left\{\pi(x,s') + \delta \int G\big(B(x,s',\mu_2)\big)\right\}\bar{Q}_{\mu_2}(ds' \mid x) \\ &\leq \lambda B[G](\mu_1) + (1-\lambda)B[G](\mu_2) \end{aligned}$$

as desired.

A.5 *Proof of Theorem 3*

Consider the set $\mathcal{H}'$ introduced at the beginning of the proof of Theorem 2, and recall that $\mathbf{P}^f(\mathcal{H}') > 0$. Observe that for any history and any $t \in \{0,1,\ldots\}$, $\mathbf{P}^f(s',x' \mid h^t) = \sigma_t(h)(x'|s')Q(s'|s_t,x_t)$. Thus, by the MCT, there exists a set $\mathcal{M}$ of histories such that, for each $h \in \mathcal{M}$,

$$\lim_t \left\| m_t(h) - t^{-1}\sum_{\tau=1}^{t} \sigma_t(h)(\cdot|\cdot)Q(\cdot|s_t,x_t) \right\| = 0$$

and $\mathbf{P}^f(\mathcal{M}) = 1$. Throughout, we fix a history $h \in \mathcal{H}' \cap \mathcal{M}$ and note that $\mathbf{P}^f(\mathcal{H}' \cap \mathcal{M}) > 0$. Henceforth, we omit the history from the notation. Also, define $M(s,\mu)$ as in the proof of Theorem 2.

We already proved condition (iii) of the definition of Berk–Nash equilibrium when we proved Theorem 2, so here we prove conditions (i) and (ii).

We first show $\sigma(\cdot|\cdot) = m(\cdot|\cdot)$. To do this, observe that $t^{-1}\sum_{\tau=1}^{t} Q(\cdot|s_t,x_t) = \sum_{s,x} Q(\cdot|s,x)m_t(s,x)$ and so

$$\lim_{t\to\infty} \left\| t^{-1}\sum_{\tau=1}^{t} \sigma_t(h)(\cdot|\cdot)Q(\cdot|s_t,x_t) - \sigma(\cdot|\cdot)\sum_{s,x} Q(\cdot|s,x)m(s,x) \right\| = 0.$$

By our choice of history, this implies that $m(s',x') = \sigma(x'|s')\sum_{s,x} Q(s'|s,x)m(s,x)$ for any $(s',x') \in \Delta(\mathbb{S}\times\mathbb{X})$. By condition (iii), it follows that $m(s',x') = \sigma(x'|s')m(s')$, which implies that $m(\cdot|\cdot) = \sigma(\cdot|\cdot)$, as desired.

Next, note that by compactness of $\Delta(\Theta)$, we can find a subsequence of beliefs $(\mu_{t(k)})_k$ that converges to some μ^*. By our choice of history (see the beginning of the proof) and Lemma 2, $\mu^* \in \Delta(\Theta_Q(m))$. Next consider any (s,x) such that $m(s,x) > 0$, which readily implies that $\sigma(x \mid s) > 0$. By convergence of $\sigma_{t(k)}$ to σ, $\sigma_{t(k)}(x \mid s) = f(x \mid s,\mu_{t(k)}) > 0$ for all sufficiently large k. By optimality of f, it follows that $x \in M(s,\mu_{t(k)})$

for all sufficiently large k. By the upper hemicontinuity of M and convergence of $\mu_{t(k)}$ to μ^*, it follows that $x \in M(s, \mu^*)$. Thus, it follows that there exists $\mu^* \in \Delta(\Theta_Q(m))$ such that, for any (s, x) in the support of m,

$$x \in \arg\max_{\hat{x}\in\mathbb{X}} \int_{\mathbb{S}} \{\pi(s, \hat{x}, s') + \delta W(s', B(s, \hat{x}, s', \mu^*))\} \bar{Q}_{\mu^*}(ds'|s, \hat{x}).$$

We conclude by establishing that x is optimal given s in the MDP where the belief is fixed at μ^*. That is,

$$x \in \arg\max_{\hat{x}\in\mathbb{X}} \int_{\mathbb{S}} \{\pi(s, \hat{x}, s') + \delta V_{\mu^*}(s')\} \bar{Q}_{\mu^*}(ds'|s, \hat{x}),$$

where $s \mapsto V_{\mu^*}(s) = \max_{\hat{x}\in\mathbb{X}} \int \{\pi(s, \hat{x}, s') + \delta V_{\mu^*}(s')\} \bar{Q}_{\mu^*}(ds'|s, \hat{x})$.

Since $m(s) > 0$ for all s, it follows that for any s and for any x such that $m(x \mid s) = \sigma(x \mid s) > 0$,

$$\begin{aligned} W(s, \mu^*) &= \int_{\mathbb{S}} \{\pi(s, x, s') + \delta W(s', B(s, x, s', \mu^*))\} \bar{Q}_{\mu^*}(ds'|s, x) \\ &= \int_{\mathbb{S}} \{\pi(s, x, s') + \delta W(s', \mu^*)\} \bar{Q}_{\mu^*}(ds'|s, x), \end{aligned}$$

where the second line follows from $\mu^* \in \Delta(\Theta_Q(m))$ and weak identification. Therefore, by the uniqueness of the value function, $s \mapsto W(s, \mu^*) = V_{\mu^*}(s)$.

Hence, it suffices to show that for any $\hat{x} \in \mathbb{X}$,

$$\int_{\mathbb{S}} \{\pi(s, x, s') + \delta V_{\mu^*}(s')\} \bar{Q}_{\mu^*}(ds'|s, x) \geq \int_{\mathbb{S}} \{\pi(s, \hat{x}, s') + \delta V_{\mu^*}(s')\} \bar{Q}_{\mu^*}(ds'|s, \hat{x}).$$

For this, let $s \mapsto x(s)$ be such that $\sigma(x(s)|s) > 0$ for all $s \in \mathbb{S}$. Observe that

$$\begin{aligned} &\int_{\mathbb{S}} \{\pi(s, x(s), s') + \delta V_{\mu^*}(s')\} \bar{Q}_{\mu^*}(ds'|s, x(s)) \\ &\quad \geq \int_{\mathbb{S}} \{\pi(s, \hat{x}, s') + \delta W(s', B(s, \hat{x}, s', \mu^*))\} \bar{Q}_{\mu^*}(ds'|s, \hat{x}). \end{aligned}$$

By weak identification and the fact that $(s', x(s')) \in \text{supp}(m)$, it follows that

$$\begin{aligned} &W(s', B(s, \hat{x}, s', \mu^*)) \\ &\quad \geq \int_{\mathbb{S}} \{\pi(s', x(s'), s'') + \delta W(s'', B(s, \hat{x}, s', \mu^*))\} \bar{Q}_{B(s,\hat{x},s',\mu^*)}(ds''|s', x(s')) \\ &\quad = \int_{\mathbb{S}} \{\pi(s', x(s'), s'') + \delta W(s'', B(s, \hat{x}, s', \mu^*))\} \bar{Q}_{\mu^*}(ds''|s', x(s')), \end{aligned}$$

where the second line follows because $B(s, \hat{x}, s', \mu^*) \in \Delta(\Theta_Q(m))$ and under weak identification this implies that $s \mapsto \bar{Q}_{B(s,\hat{x},s',\mu^*)}(\cdot|s, x(s)) = \bar{Q}_{\mu^*}(\cdot|s, x(s))$ for any $(s, x(s)) \in \text{supp}(m)$. By applying this inequality over and over to $W(\cdot, B(s, \hat{x}, s', \mu^*))$, it follows that

$$W(s', B(s, \hat{x}, s', \mu^*)) \geq \sum_{j=0}^{\infty} \delta^j \bar{\mathbf{Q}}^j_{\mu^*} \left[\int \pi(\cdot, x(\cdot), s'') \bar{Q}_{\mu^*}(ds''|\cdot, x(\cdot)) \right](s'),$$

where $g \mapsto \bar{\mathbf{Q}}_{\mu^*}[g](s) \equiv \int g(s')\bar{Q}_{\mu^*}(ds'|s, x(s))$ for any $s \in \mathbb{S}$. By uniqueness of the value function, the right-hand side equals $V_{\mu^*}(s')$ and, thus,

$$W\big(s', B\big(s, \hat{x}, s', \mu^*\big)\big) \geq V_{\mu^*}\big(s'\big)$$

for any $s' \in \mathbb{S}$, thereby implying the desired result.

References

Aliprantis, Charalambos D. and Kim C. Border (2006), *Infinite Dimensional Analysis: A Hitchhiker's Guide*, third edition. Springer, Berlin. [743, 744]

Arrow, Kenneth and Jerry Green (1973), *Notes on Expectations Equilibria in Bayesian Settings*. Working paper, Institute for Mathematical Studies in the Social Sciences Working Paper No. 33. [717]

Battigalli, Pierpaolo (1987), *Comportamento razionale ed equilibrio nei giochi e nelle situazioni sociali*. Universita Bocconi, Milano. [720]

Bencivenga, Valerie R. (1992), "An econometric study of hours and output variation with preference shocks." *International Economic Review*, 33, 449–471. [729]

Berk, Robert (1966), "Limiting behavior of posterior distributions when the model is incorrect." *Annals of Mathematical Statistics*, 37, 51–58. [725, 734]

Blume, Lawrence E. and David Easley (1998), "Rational expectations and rational learning." Organizations with incomplete information: Essays in economic analysis: A tribute to Roy Radner 61–109, Cambridge University Press, New York. [720]

Bohren, J. Aislinn and Daniel N. Hauser (2017), *Social Learning With Model Misspecification: A Framework and a Robustness Result*. Working paper, PIER Working Paper No. 17-007. [720]

Brock, William A. and Leonard J. Mirman (1972), "Optimal economic growth and uncertainty: The discounted case." *Journal of Economic Theory*, 4, 479–513. [728]

Bunke, Olaf and Xavier Milhaud (1998), "Asymptotic behavior of Bayes estimates under possibly incorrect models." *Annals of Statistics*, 26, 617–644. [734]

Dekel, Eddie, Drew Fudenberg, and David K. Levine (2004), "Learning to play Bayesian games." *Games and Economic Behavior*, 46, 282–303. [720]

Diaconis, Persi and David Freedman (1986), "On the consistency of Bayes estimates." *Annals of Statistics*, 14, 1–26. [722]

Doraszelski, Ulrich and Juan F. Escobar (2010), "A theory of regular Markov perfect equilibria in dynamic stochastic games: Genericity, stability, and purification." *Theoretical Economics*, 5, 369–402. [742]

Easley, David and Nicholas M. Kiefer (1988), "Controlling a stochastic process with unknown parameters." *Econometrica*, 56, 1045–1064. [720]

Eliaz, Kfir and Ran Spiegler (2020), "A model of competing narratives." *American Economic Review*, 110, 3786–3816. [717]

Esponda, Ignacio (2008), "Behavioral equilibrium in economies with adverse selection." *American Economic Review*, 98, 1269–1291. [717]

Esponda, Ignacio and Demian Pouzo (2016), "Berk–Nash equilibrium: A framework for modeling agents with misspecified models." *Econometrica*, 84, 1093–1130. [717, 719, 722, 724, 734, 736, 742, 745]

Esponda, Ignacio and Demian Pouzo (2017), "Conditional retrospective voting in large elections." *American Economic Journal: Microeconomics*, 9, 54–75. [717]

Esponda, Ignacio and Demian Pouzo (2019), "Retrospective voting and party polarization." *International Economic Review*, 60, 157–186. [717]

Esponda, Ignacio, Demian Pouzo, and Yuichi Yamamoto (2019), "Asymptotic behavior of Bayesian learners with misspecified models." Unpublished paper, arXiv:1904.08551. [741]

Evans, George W. and Seppo Honkapohja (2001), *Learning and Expectations in Macroeconomics*. Princeton University Press. [720]

Eyster, Erik and Michele Piccione (2013), "An approach to asset-pricing under incomplete and diverse perceptions." *Econometrica*, 81, 1483–1506. [717]

Eyster, Erik and Matthew Rabin (2005), "Cursed equilibrium." *Econometrica*, 73, 1623–1672. [717]

Fershtman, Chaim and Ariel Pakes (2012), "Dynamic games with asymmetric information: A framework for empirical work." *Quarterly Journal of Economics*, 127, 1611–1661. [720]

Freedman, David (1963), "On the asymptotic behavior of Bayes' estimates in the discrete case." *Annals of Mathematical Statistics*, 34, 1386–1403. [722]

Freixas, Xavier (1981), "Optimal growth with experimentation." *Journal of Economic Theory*, 24, 296–309. [728]

Frick, Mira, Ryota Iijima, and Yuhta Ishii (2020a), *Stability and Robustness in Misspecified Learning Models*. Working Paper, Cowles Foundation Discussion Paper No. 2235. [741]

Frick, Mira, Ryota Iijima, and Yuhta Ishii (2020b), "Misinterpreting others and the fragility of social learning." *Econometrica*, 88, 2281–2328. [720]

Fudenberg, Drew and David M. Kreps (1993), "Learning mixed equilibria." *Games and Economic Behavior*, 5, 320–367. [719, 736, 741]

Fudenberg, Drew and David M. Kreps (1995), "Learning in extensive-form games I. Self-confirming equilibria." *Games and Economic Behavior*, 8, 20–55. [719]

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack (2020), "Limits points of endogenous misspecified learning." Unpublished paper, SSRN 3553363. [741]

Fudenberg, Drew and David K. Levine (1993), "Self-confirming equilibrium." *Econometrica*, 61, 523–545. [720]

Fudenberg, Drew and David K. Levine (1998), *The Theory of Learning in Games*. MIT Press, Cambridge, Massachusetts. [720]

Fudenberg, Drew, Gleb Romanyuk, and Philipp Strack (2017), "Active learning with a misspecified prior." *Theoretical Economics*, 12, 1155–1189. [717, 741]

Hall, Robert E. (1997), *Macroeconomic Fluctuations and the Allocation of Time*. Technical Report 1. [728]

Harsanyi, John C. (1973), "Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points." *International Journal of Game Theory*, 2, 1–23. [741]

He, Kevin (2018), "Mislearning from censored data: The gambler's fallacy in optimal-stopping problems." Unpublished paper, arXiv:1803.08170. [720]

Heidhues, Paul, Botond Koszegi, and Philipp Strack (2021), "Convergence in models of misspecified learning." *Theoretical Economics*, 16, 73–99. [717, 741]

Heidhues, Paul, Botond Kőszegi, and Philipp Strack (2018), "Unrealistic expectations and misguided learning." *Econometrica*, 86, 1159–1214. [717, 741]

Hofbauer, Josef and William H. Sandholm (2002), "On the global convergence of stochastic fictitious play." *Econometrica*, 70, 2265–2294. [741]

Jéhiel, Philippe (1995), "Limited horizon forecast in repeated alternate games." *Journal of Economic Theory*, 67, 497–519. [720]

Jéhiel, Philippe (1998), "Learning to play limited forecast equilibria." *Games and Economic Behavior*, 22, 274–298. [720]

Jehiel, Philippe (2005), "Analogy-based expectation equilibrium." *Journal of Economic Theory*, 123, 81–104. [717, 721]

Jehiel, Philippe and Frédéric Koessler (2008), "Revisiting games of incomplete information with analogy-based expectations." *Games and Economic Behavior*, 62, 533–557. [717, 721]

Jehiel, Philippe and Dov Samet (2007), "Valuation equilibrium." *Theoretical Economics*, 2, 163–185. [720]

Kagel, John H. and Dan Levin (1986), "The winner's curse and public information in common value auctions." *American Economic Review*, 76, 894–920. [717]

Kirman, Alan P. (1975), "Learning by firms about demand conditions." In *Adaptive Economic Models* (R. H. Day and T. Groves, eds.), 137–156, Academic Press. [717]

Koulovatianos, Christos, Loenard J. Mirman, and Marc Santugini (2009), "Optimal growth and uncertainty: Learning." *Journal of Economic Theory*, 144, 280–295. [720, 728]

McLennan, Andrew (1984), "Price dispersion and incomplete learning in the long run." *Journal of Economic Dynamics and Control*, 7, 331–347. [720]

Molavi, Pooya (2019), "Macroeconomics with learning and misspecification: A general theory and applications." Unpublished paper, MIT. [720]

Nyarko, Yaw (1991), "Learning in mis-specified models and the possibility of cycles." *Journal of Economic Theory*, 55, 416–427. [717]

Ortoleva, Pietro and Erik Snowberg (2015), "Overconfidence in political behavior." *American Economic Review*, 105, 504–535. [720]

Piccione, Michele and Ariel Rubinstein (2003), "Modeling the economic interaction of agents with diverse abilities to recognize equilibrium patterns." *Journal of the European economic association*, 1, 212–223. [717]

Rabin, Matthew and Dimitri Vayanos (2010), "The Gambler's and hot-hand fallacies: Theory and applications." *Review of Economic Studies*, 77, 730–778. [720]

Rothschild, Michael (1974), "A two-armed bandit theory of market pricing." *Journal of Economic Theory*, 9, 185–202. [720]

Sargent, Thomas J. (1999), *The Conquest of American Inflation*. Princeton University Press, Princeton, New Jersey. [717, 720]

Shalizi, Cosma Rohilla (2009), "Dynamics of Bayesian updating with dependent data and misspecified models." *Electronic Journal of Statistics*, 3, 1039–1074. [734]

Sobel, Joel (1984), "Non-linear prices and price-taking behavior." *Journal of Economic Behavior & Organization*, 5, 387–396. [717]

Spiegler, Ran (2013), "Placebo reforms." *American Economic Review*, 103, 1490–1506. [717]

Spiegler, Ran (2016), "Bayesian networks and boundedly rational expectations." *Quarterly Journal of Economics*, 131, 1243–1290. [717, 721]

Spiegler, Ran (2017), "Data monkeys: A procedural model of extrapolation from partial statistics." *Review of Economic Studies*, 84, 1818–1841. [717, 721]

Co-editor Ran Spiegler handled this manuscript.

Manuscript received 29 June, 2019; final version accepted 5 August, 2020; available online 12 August, 2020.