Equilibrium in Misspecified Markov Decision Processes [*]

Ignacio Esponda
(WUSTL)

Demian Pouzo
(UC Berkeley)

January 10, 2017

Abstract

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. Unlike the case for static environments, however, an equilibrium approach for the dynamic setting cannot be used to characterize those steady states where the agent perceives that learning is incomplete. Two key features of our approach is that it distinguishes between the agent's simplified model and the true primitives and that the agent's belief is determined endogenously in equilibrium.

[*] We thank Vladimir Asriyan, Hector Chade, Xiaohong Chen, Emilio Espino, Drew Fudenberg, Bruce Hansen, Philippe Jehiel, Bart Lipman, Jack Porter, Philippe Rigollet, Tom Sargent, Philipp Strack, Iván Werning, and several seminar participants for helpful comments. Esponda: Olin Business School, Washington University in St. Louis, 1 Brookings Drive, Campus Box 1133, St. Louis, MO 63130, iesponda@wustl.edu; Pouzo: Department of Economics, UC Berkeley, 530-1 Evans Hall #3880, Berkeley, CA 94720, dpouzo@econ.berkeley.edu.

Contents

1 Introduction

Early interest on 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 (2012, 2016), Eyster and Piccione (2013), Spiegler (2013, 2016a, 2016b), and Heidhues et al. (2016). 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.[1]

The previous literature focuses 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 (other than a belief), are ubiquitous in economics. The main goal of this paper is to provide a tractable framework to study dynamic settings where the agent learns with 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 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 following Bellman equation,

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

where s is the current state, x is the agent's choice variable, $\Gamma(s)$ is the set of feasible actions, 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:

[1] We take the misspecified model as a primitive and assume that agents learn and behave optimally given their model. In contrast, Hansen and Sargent (2008) study optimal behavior of agents who have a preference for robustness because they are aware of the possibility of model misspecification.

a potentially large state space (i.e., the curse of dimensionality) and uncertainty about the transition probability function. For example, equation (1) may represent a dynamic savings problem where the agent decides every period what fraction x of her income 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 equation (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 equation (1). One of the challenges is to determine which transition function should be used to replace Q in equation (1) as a function of the agent's simplified view of her environment.

Before we describe our approach, consider first the way that this problem is often handled in the literature. The modeler simplifies the problem by eliminating some state variables from the state space and then solves the problem as if it were true that the true transition does not depend on the eliminated variables. The typical justification is that the original problem is too complex to solve for the modeler, and so it is reasonable that the agent would perform a similar simplification. This approach has two disadvantages. The first is that we should distinguish between the possibly simplified model of the agent and the true primitives, since it is the combination of the agent's policy function and the true primitives that determines outcomes. The second is that it is not at all obvious how to simplify Q. We will show through examples that, in some cases, there is no a priori reasonable candidate while, in other cases, the obvious candidate to replace Q does not seem appropriate.

We propose an alternative approach that distinguishes between the simplified model of the agent and the true primitives. This distinction turns out to be crucial for determining how to simplify the true Q. 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.

The agent has a prior μ over Θ, and this belief 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. The 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:

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

The main question that we ask is whether we can characterize the asymptotic behavior of this agent by studying the problem without learning, as represented by equation (1), where the agent holds some limiting belief $\mu^* \in \Delta(\Theta)$ and the true transition Q is replaced by the agent's simplified version of the transition, $\bar{Q}_{\mu^*} = \int_\Theta Q_\theta \mu^*(d\theta)$. In addition to gaining tractability, there are two reasons why we focus on asymptotic behavior. The first is that, as reviewed below, there is a long tradition in statistics and economics that highlights the advantages of studying asymptotic or equilibrium behavior, and we can contrast our results to this literature. The second is that we want to make sure that mistakes are not driven by lack of opportunities to learn.

In the static settings studied by previous literature (both in games and decision settings), the answer to the question that we ask is positive under mild assumptions, meaning that we can characterize steady-state beliefs and behavior by studying a problem where beliefs are fixed. In particular, we do not need to tackle the problem of belief updating in order to characterize limiting behavior. This type of result is quiet remarkable, but it has been customary to expect such a result to hold anytime a new equilibrium concept is postulated.

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 we restrict attention to a class of steady states that satisfy a property that we call exhaustive learning. This property says that the agent perceives that she has nothing else to learn in steady state. This property is satisfied, for example, if we are interested in behavior that is robust to a small amount of exogenous experimentation.[2] Steady states that do not satisfy exhaustive learning, however, cannot generally be characterized by an equilibrium approach with fixed beliefs. In contrast, the modeler is forced to consider the more complicated problem with belief updating, as represented by equation (2). As we explain in Section 5, the difference in results between the static and 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.

We define a notion of equilibrium for the dynamic environment and show that this notion captures the set of steady states with exhaustive learning. We call this notion a Berk-Nash equilibrium because, in the special case where the environment is static, it collapses to the single-agent version of Berk-Nash equilibrium, a concept introduced by Esponda and Pouzo (2016) to characterize steady state behavior in static environments with misspecified agents. A strategy in a Markov decision process (MDP) is a mapping from (non-belief) states to actions. For a given strategy and true transition probability function, the stochastic process for states and actions in an MDP is a Markov chain and has a corresponding stationary distribution that can be interpreted as the steady-state distribution over outcomes. A strategy and corresponding stationary distribution is a Berk-Nash equilibrium if there exists a belief μ^* over the parameter space such that: (i) the strategy is optimal for an MDP with transition probability function $\bar{Q}_{\mu^*}$, and (ii) μ^* puts probability one on the set of parameter values that yield transition probability functions that are "closest" to the true transition probability function. The notion of "closest" is given by a weighted version of the Kullback-Leibler divergence that depends on the equilibrium stationary distribution.

Our asymptotic characterization of beliefs and actions contributes to the literature that studies asymptotic beliefs and/or behavior under Bayesian learning. Table 1 categorizes some of the more relevant papers in connection to our work. The table on the left includes papers where the agent learns from data that is exogenous in the sense that she does not affect the stochastic properties of the data. This topic has mostly been tackled by statisticians for both correctly-specified and misspecified

[2] The property is weaker, however, and allows for beliefs to be incorrect due to lack of experimentation, which is a hallmark of the bandit (e.g., Rothschild (1974b), McLennan (1984), Easley and Kiefer (1988)) and self-confirming equilibrium (e.g., Battigalli (1987), Fudenberg and Levine (1993), Dekel et al. (2004), Fershtman and Pakes (2012)) literatures.

	Correctly Specified	Misspecified
i.i.d.	Schwartz [65] Freedman [63] Diaconis-Freedman [86]	Berk [65] Bunke-Milhaud [98]
non-i.i.d.	Ghosal-Van der Vaart [07]	Shalizi [09] Vayanos-Rabin [10]^

Exogenous Data

	Correctly Specified	Misspecified
Static	Rothschild [74]^ Gittins [79]^ McLennan [84]^ Easley-Kiefer [88] Aghion et al [91]	Nyarko [91]^ Esponda [08]^ Esponda-Pouzo [16] Heidhues et al [16]^
Dynamic	Freixas [81]^ Koulovatianos et al [09]^ **This paper**	Fudenberg et al [16]^ **This paper**

Endogenous Data

Table 1: Literature on Bayesian Learning

models and for both i.i.d. and non-i.i.d. data. The table on the right includes papers where the agent learns from data that is endogenous in the sense that it is driven by the agent's actions, a topic that has been studied by economists mostly in static settings. By static we mean that the problem reduces to a static optimization problem if stripped of the learning dynamics.[3]

Table 1 also differentiates between two complementary approaches to studying asymptotic beliefs and/or behavior. The first approach is to focus on specific settings and provide a complete characterization of asymptotic actions and beliefs, including convergence results; these papers are marked with a superscript ^ in Table 1. Some papers pursue this approach in dynamic and correctly specified stochastic growth models (e.g., Freixas (1981), Koulovatianos et al. (2009)). In static misspecified settings, Nyarko (1991), Esponda (2008), and Heidhues et al. (2016) study passive learning problems where there is no experimentation motive. Fudenberg et al. (2016) is the only paper that provides a complete characterization in a dynamic decision problem with active learning.[4],[5] The second approach, which we follow in this paper and we followed earlier for the static case (Esponda and Pouzo, 2016) is to study general

[3] Formally, we say a problem is static if, for a fixed strategy and belief over the transition probability function, outcomes (states and actions) are independent across time.

[4] Under active learning, different actions convey different amount of information and a non-myopic agent takes the exploitation vs. experimentation tradeoff into account. There can be passive or active learning in both static and dynamic settings.

[5] The environment in Fudenberg et al. (2016) is dynamic because the agent controls the drift of a Brownian motion, even though the only relevant state variable for optimality ends up being the agent's belief.

settings and focus on characterizing the set of steady states.[6]

The paper is also related to the literature which 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. Jehiel (1995) considers the class of repeated alternating-move games and assumes that players only forecast a limited number of time periods into the future; see Jehiel (1998) for a learning foundation. 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.[7]

The framework and equilibrium notion are presented in Sections 2 and 3. In Section 4, we work through several examples. We provide a foundation for equilibrium in Section 5 and study equilibrium refinements in Section 6.

2 Markov Decision Processes

We begin by describing the environment faced by the agent.

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

- $\mathbb{S}$ is a nonempty and finite set of states
- $\mathbb{X}$ is a nonempty and finite set of actions
- $\Gamma : \mathbb{S} \to 2^{\mathbb{X}}$ is a non-empty constraint correspondence
- $q_0 \in \Delta(\mathbb{S})$ is a probability distribution on the initial state
- $Q : Gr(\Gamma) \to \Delta(\mathbb{S})$ is a transition probability function[8]
- $\pi : Gr(\Gamma) \times \mathbb{S} \to \mathbb{R}$ is a per-period payoff function

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

[7] 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.

[8] For a correspondence $\Gamma : \mathbb{S} \to 2^{\mathbb{X}}$, its graph is defined by $Gr(\Gamma) \equiv \{(s, x) \in \mathbb{S} \times \mathbb{X} : x \in \Gamma(s)\}$.

- $\delta \in [0,1)$ 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 $t = 0, 1, 2, ...$, the agent observes state $s_t \in \mathbb{S}$ and chooses a feasible action $x_t \in \Gamma(s_t) \subset \mathbb{X}$. 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.

The agent facing an MDP chooses a policy rule that specifies at each point in time a (possibly random) action as a function of the history of states and actions observed up to that point. 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

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

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

Definition 2. A **strategy** σ is a distribution over actions given states, $\sigma : \mathbb{S} \to \Delta(\mathbb{X})$, that satisfies $\sigma(s) \in \Gamma(s)$ for all s.

Let Σ denote the space of all strategies and let $\sigma(x \mid s)$ denote the probability that the agent chooses x when the state is s.[9]

Definition 3. A strategy $\sigma \in \Sigma$ is **optimal** for an MDP(Q) if, for all $s \in \mathbb{S}$ and all $x \in \mathbb{X}$ such that $\sigma(x \mid s) > 0$,

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

Let $\Sigma(Q)$ be the set of all strategies that are optimal for an MDP(Q).

[9] A standard result is the existence of a *deterministic* optimal strategy. Nevertheless, allowing for randomization will be important in the case where the transition probability function is uncertain.

Lemma 1. *(i) There is a unique solution V_Q to the Bellman equation in (3), and it is continuous in Q for all $s \in \mathbb{S}$; (ii) The correspondence of optimal strategies $Q \mapsto \Sigma(Q)$ is non-empty, compact-valued, convex-valued, and upper hemicontinuous.*

Proof. The proof is standard and relegated to the Online Appendix. □

A strategy determines the transitions in the space of states and actions and, consequently, the set of stationary distributions over states and actions. For any strategy σ and transition probability function Q, define a **transition kernel** $M_{\sigma,Q} : Gr(\Gamma) \to \Delta\left(Gr(\Gamma)\right)$ by letting

$$M_{\sigma,Q}(s', x' \mid s, x) = \sigma(x' \mid s')Q(s' \mid s, x) \tag{4}$$

for all $(s, x), (s', x') \in Gr(\Gamma)$. The transition kernel $M_{\sigma,Q}$ is the transition probability function over $Gr(\Gamma)$ given strategy σ and transition probability function Q.

For any $m \in \Delta(Gr(\Gamma))$, let $M_{\sigma,Q}[m] \in \Delta(Gr(\Gamma))$ denote the probability measure

$$\sum_{(s,x)\in Gr(\Gamma)} M_{\sigma,Q}(\cdot, \cdot \mid s, x)m(s, x).$$

Definition 4. A distribution $m \in \Delta(Gr(\Gamma))$ is a **stationary (or invariant) distribution** given (σ, Q) if $m = M_{\sigma,Q}[m]$.

A stationary distribution represents the steady-state distribution over outcomes (i.e, states and actions) when the agent follows a given strategy. Let $I_Q(\sigma) \equiv \{m \in \Delta(Gr(\Gamma)) \mid m = M_{\sigma,Q}[m]\}$ denote the set of stationary distributions given (σ, Q).

Lemma 2. *The correspondence of stationary distributions $\sigma \mapsto I_Q(\sigma)$ is non-empty, compact-valued, convex-valued, and upper hemicontinuous.*

Proof. See the Appendix. □

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 5. A **Subjective Markov Decision Process** (SMDP) is an MDP, $\langle \mathbb{S}, \mathbb{X}, \Gamma, q_0, Q, \pi, \delta \rangle$, and a nonempty family of transition probability functions, $\mathcal{Q}_\Theta = \{Q_\theta : \theta \in \Theta\}$, where each transition probability function $Q_\theta : Gr(\Gamma) \to \Delta(\mathbb{S})$ is indexed by a parameter $\theta \in \Theta$.

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 6. A **Regular Subjective Markov Decision Process** (regular-SMDP) is an SMDP that satisfies the following conditions

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

The first two conditions in Definition 6 place parametric and continuity assumptions on the subjective models.[10] The last condition plays two roles. First, it rules out 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 is upper hemicontinuous. Esponda and Pouzo (2016) provide a simple (non-dynamic) example where this assumption does not hold and equilibrium fails to exist.

[10] 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.

3.2 Equilibrium

The goal of this section is to define the notion of Berk-Nash equilibrium of an SMDP. The next definition is used to place constraints on the belief $\mu \in \Delta(\Theta)$ that the agent may hold if m is the stationary distribution over outcomes.

Definition 7. The **weighted Kullback-Leibler divergence** (wKLD) is a mapping $K_Q\colon \Delta(Gr(\Gamma)) \times \Theta \to \bar{\mathbb{R}}_+$ such that for any $m \in \Delta(Gr(\Gamma))$ and $\theta \in \Theta$,

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

The **set of closest parameter values given** $m \in \Delta(Gr(\Gamma))$ is the set

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

The set $\Theta_Q(m)$ contains the parameter values constitute the best fit with the true transition probability function Q when outcomes are drawn from the distribution m.

Lemma 3. *(i) For every $m \in \Delta(Gr(\Gamma))$ 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 non-empty, compact valued, and upper hemicontinuous.*

Proof. See the Appendix. □

We now define equilibrium.

Definition 8. A strategy and probability distribution $(\sigma, m) \in \Sigma \times \Delta(Gr(\Gamma))$ is a **Berk-Nash equilibrium** of the SMDP$(Q, \mathcal{Q}_\Theta)$ if there exists a belief $\mu \in \Delta(\Theta)$ such that

(i) σ is an optimal strategy for the MDP$(\bar{Q}_\mu)$, where $\bar{Q}_\mu = \int_\Theta Q_\theta \mu(d\theta)$,

(ii) $\mu \in \Delta(\Theta_Q(m))$, and

(iii) $m \in I_Q(\sigma)$.

Condition (i) in the definition of Berk-Nash equilibrium requires σ to be an optimal strategy in the MDP where the transition probability function is $\int_\Theta Q_\theta \mu(d\theta)$. Condition (ii) requires that the agent only puts positive probability on the set of closest parameter values given m, $\Theta_Q(m)$. Finally, condition (iii) requires m to be a stationary distribution given (σ, Q).

Remark 1. In Section 5, we interpret the set of equilibria as the set of steady states of a learning environment where the agent is uncertain about Q. The main advantage of the equilibrium approach is that it allows us to replace a difficult learning problem with a simpler MDP with a fixed transition probability function. The cost of this approach is that it can only be used to characterize asymptotic behavior, as opposed to the actual dynamics starting from the initial distribution over states, $q_0 \in \Delta(\mathbb{S})$. This explains why q_0 does not enter the definition of equilibrium, and why a mapping between q_0 and the set of corresponding equilibria cannot be provided in general.

Remark 2. In the special case of a static environment, Definition 8 reduces to Esponda and Pouzo's (2016) definition of Berk-Nash equilibrium for a single agent. In the dynamic environment, outcomes follow a Markov process and we must keep track not only of strategies but also of the corresponding stationary distribution over outcomes.

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

Theorem 1. *For any regular SMDP, there exists a Berk-Nash equilibrium.*

Proof. See the Appendix. □

The standard approach to proving existence begins by defining a "best response correspondence" in the space of strategies. This approach does not work here because the possible non-uniqueness of beliefs implies that the correspondence may not be convex valued. The trick we employ is to define equilibrium via a correspondence on the space of strategies, stationary distributions, and beliefs, and then use Lemmas 1, 2 and 3 to show that this correspondence satisfies the assumptions of a generalized version of Kakutani's fixed point theorem.[11]

[11] Esponda and Pouzo (2016) rely on perturbations to show existence of equilibrium in a static setting. In contrast, our approach does not require the use of perturbations.

3.3 Correctly specified and identified SMDPs

An SMDP is correctly specified if the set of subjective models contains the true model.

Definition 9. An SMDP$(Q, \mathcal{Q}_\Theta)$ is **correctly specified** if $Q \in \mathcal{Q}_\Theta$; otherwise, it is misspecified.

In decision problems, data is endogenous and so, following Esponda and Pouzo (2016), it is natural to consider two notions of identification: weak and strong identification. These definitions distinguish between outcomes on and off the equilibrium path. In a dynamic environment, the right object to describe what happens on and off the equilibrium path is not the strategy but rather the stationary distribution over outcomes m.

Definition 10. An SMDP is **weakly identified given m** $\in \Delta(Gr(\Gamma))$ if $\theta, \theta' \in \Theta_Q(m)$ implies that $Q_\theta(\cdot \mid s, x) = Q_{\theta'}(\cdot \mid s, x)$ for all $(s, x) \in Gr(\Gamma)$ such that $m(s, x) > 0$; if the condition is satisfied for all $(s, x) \in Gr(\Gamma)$, we say that the **SMDP is strongly identified given m**. An SMDP is weakly (strongly) identified if it is weakly (strongly) identified for all $m \in \Delta(Gr(\Gamma))$.

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. It is a condition that turns out to be important for proving the existence of equilibria that are robust to experimentation (see Section 6) and is always satisfied in correctly specified SMDPs.[12] Strong identification strengthens the condition by requiring that beliefs are unique also off the equilibrium path.

Proposition 1. *Consider a correctly specified and strongly identified SMDP with corresponding MDP(Q). A strategy and probability distribution $(\sigma, m) \in \Sigma \times \Delta(Gr(\Gamma))$ is a Berk-Nash equilibrium of the SMDP if and only if σ is optimal given MDP(Q) and m is a stationary distribution given σ.*

[12] The following is an example where weak identification fails. Suppose an unbiased coin is tossed every period, but the agent believes that the coin comes up heads with probability 1/4 or 3/4, but not 1/2. Then both 1/4 and 3/4 minimize the Kullback-Leibler divergence, but they imply different distributions over outcomes. Relatedly, Berk (1966) shows that beliefs do not converge.

Proof. Only if: Suppose (σ, m) is a Berk-Nash equilibrium. Then there exists μ such that σ is optimal given MDP($\bar{Q}_{\mu}$), $\mu \in \Delta(\Theta(m))$, and $m \in I_Q(\sigma)$. Because the SMDP is correctly specified, there exists θ^* such that $Q_{\theta^*} = Q$ and, therefore, by Lemma 3(i), $\theta^* \in \Delta(\Theta(m))$. Then, by strong identification, any $\hat{\theta} \in \Theta(m)$ satisfies $Q_{\hat{\theta}} = Q_{\theta^*} = Q$, implying that σ is also optimal given MDP(Q). *If*: Let $m \in I_Q(\sigma)$, where σ is optimal given MDP(Q). Because the SMDP is correctly specified, there exists θ^* such that $Q_{\theta^*} = Q$ and, therefore, by Lemma 3(i), $\theta^* \in \Delta(\Theta(m))$. Thus, σ is also optimal given Q_{θ^*}, implying that (σ, m) is a Berk-Nash equilibrium. □

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 strongly identified, then Berk-Nash equilibrium corresponds to the solution of the MDP under correct beliefs about the transition probability function. If one drops the assumption that the SMDP is strongly identified, then the "if" part of the proposition continues to hold but the "only if" condition does not hold. In other words, there may be Berk-Nash equilibria of correctly-specified SMDPs in which the agent has incorrect beliefs off the equilibrium path. This feature of equilibrium is analogous to the main ideas of the bandit and self-confirming equilibrium literatures.

4 Examples

We use three classic examples to illustrate how easy it is to use our framework to expand the scope of the classical dynamic programming approach.

4.1 Monopolist with unknown dynamic demand

The problem of a monopolist facing an unknown, *static* demand function was first studied by Rothschild (1974b) and Nyarko (1991) in correctly and misspecified settings, respectively. In the following example, the monopolist faces a dynamic demand function but incorrectly believes that demand is static.

MDP: In each period t, a monopolist chooses price $x_t \in \mathbb{X} = \{L, H\}$, where $0 < L < H$. It then sells $s_{t+1} \in \mathbb{S} = \{0, 1\}$ units at zero cost and obtains profit $\pi(x_t, s_{t+1}) = x_t s_{t+1}$. The probability that $s_{t+1} = 1$ is $q_{sx} \equiv Q(1 \mid s_t = s, x_t = x)$, where $0 < q_{sx} < 1$ for all $(s, x) \in Gr(\Gamma) = \mathbb{S} \times \mathbb{X}$.[13] The monopolist wants to

[13] The set of feasible actions is independent of the state, i.e., $\Gamma(s) = \mathbb{X}$ for all $s \in \mathbb{S}$.

maximize expected discounted profits, with discount factor $\delta \in [0,1)$.

Demand is dynamic in the sense that a sale yesterday increases the probability of a sale today: $q_{1x} > q_{0x}$ for all $x \in \mathbb{X}$. Moreover, a higher price reduces the probability of a sale: $q_{sL} > q_{sH}$ for all $s \in \mathbb{S}$. Finally, for concreteness, we assume that

$$\frac{q_{1L}}{q_{1H}} < \frac{H}{L} < \frac{q_{0L}}{q_{0H}}. \tag{5}$$

Expression (5) implies that current-period profits are maximized by choosing price L if there was no sale last period and price H otherwise (i.e., $Lq_{0L} > Hq_{0H}$ and $Hq_{1H} > Lq_{1L}$). Thus, the optimal strategy of a myopic monopolist (i.e., $\delta = 0$) who knows the primitives is $\sigma(H \mid 0) = 0$ and $\sigma(H \mid 1) = 1$. If, however, the monopolist is sufficiently patient, it is optimal to always choose price L.[14]

SMDP. The monopolist does not know Q and believes, incorrectly, that demand is not dynamic. Formally, $\mathcal{Q}_\Theta = \{Q_\theta : \theta \in \Theta\}$, where $\Theta = [0,1]^2$ and, for all $\theta = (\theta_L, \theta_H) \in \Theta$, $Q_\theta(1 \mid s, L) = \theta_L$ and $Q_\theta(1 \mid s, H) = \theta_H$ for all $s \in \mathbb{S}$. In particular, θ_x is the probability that a sale occurs given price $x \in \{L, H\}$, and the agent believes that it does not depend on s. Note that this SMDP is regular. For simplicity, we restrict attention to equilibria in which the monopolist does not condition on last period's state, and denote a strategy by σ_H, the probability that price H is chosen.

Equilibrium. *Optimality.* Because the monopolist believes that demand is static, the optimal strategy is to choose the price that maximizes current period's profit. Let

$$\Delta(\theta) \equiv H\theta_H - L\theta_L$$

denote the perceived expected payoff difference of choosing H vs. L under the belief that the parameter value is $\theta = (\theta_L, \theta_H)$ with probability 1. If $\Delta(\theta) > 0$, $\sigma_H = 1$ is the unique optimal strategy; if $\Delta(\theta) < 0$, $\sigma_H = 0$ is the unique optimal strategy; and if $\Delta(\theta) = 0$, any $\sigma_H \in [0,1]$ is optimal.

Beliefs. For any $m \in \Delta(\mathbb{S} \times \mathbb{X})$, the wKLD simplifies to

$$K_Q(m,\theta) = \sum_{x \in \{L,H\}} m_{\mathbb{X}}(x) \left\{ \bar{s}_x(m) \ln \theta_x + (1 - \bar{s}_x(m)) \ln(1 - \theta_x) \right\} + Const,$$

[14] Formally, there exists $C_\delta \in [q_{1L}/q_{1H}, q_{0L}/q_{0H}]$, where $C_0 = q_{1L}/q_{1H}$ and $\delta \mapsto C_\delta$ is increasing, such that, if $H/L < C_\delta$, the optimal strategy is $\sigma(H \mid 0) = \sigma(H \mid 1) = 0$.

where $\bar{s}_x(m) = m_{\mathbb{S}|\mathbb{X}}(0 \mid x)q_{0x} + m_{\mathbb{S}|\mathbb{X}}(0 \mid x)q_{1x}$ is the probability of a sale given x.

If $\sigma_L > 0$ and $\sigma_H > 0$, $\theta_Q(m) \equiv (\bar{s}_L(m), \bar{s}_H(m))$ is the unique parameter value that minimizes the wKLD function. If, however, one of the prices is chosen with zero probability, there are no restrictions on beliefs for the corresponding parameter, i.e., the set of minimizers is $\Theta_Q(m) = \{(\theta_L, \theta_H) \in \Theta : \theta_H = \bar{s}_H(m)\}$ if $\sigma_L = 0$ and $\Theta_Q(m) = \{(\theta_L, \theta_H) \in \Theta : \theta_L = \bar{s}_L(m)\}$ if $\sigma_H = 0$.

Stationary distribution. Fix a strategy σ_H and denote a corresponding stationary distribution by $m(\cdot; \sigma_H) \in \Delta(\mathbb{S} \times \mathbb{X})$. Since the strategy does not depend on the state, $m_{\mathbb{S}|\mathbb{X}}(\cdot \mid x; \sigma_H)$ does not depend on x and, therefore, coincides with the marginal stationary distribution over $\mathbb{S}$, denoted by $m_{\mathbb{S}}(\cdot; \sigma_H) \in \Delta(\mathbb{S})$. This distribution is unique and given by the solution to

$$m_{\mathbb{S}}(1; \sigma_H) = (1 - m_{\mathbb{S}}(1; \sigma_H))((1 - \sigma_H)q_{0L} + \sigma_H q_{0H}) + m_{\mathbb{S}}(1; \sigma_H)((1 - \sigma_H)q_{1L} + \sigma_H q_{1H}).$$

Equilibrium. We restrict attention to equilibria that are robust to experimentation (i.e., perfect equilibria; see Section 6) by focusing on the belief $\theta(\sigma_H) = (\theta_L(\sigma_H), \theta_H(\sigma_H)) \equiv \theta_Q(m(\cdot; \sigma_H))$ for a given strategy $\sigma_H \in [0, 1]$.[15] Next, let $\Delta(\theta(\sigma_H))$ be the *perceived* expected payoff difference for a given strategy σ_H. Note that $\sigma_H \mapsto \Delta(\theta(\sigma_H))$ is decreasing[16], which means that a higher probability of choosing price H leads to more pessimistic beliefs about the benefit of choosing H vs. L. Therefore, there exists a unique (perfect) equilibrium strategy. Figure 1 depicts an example where the equilibrium is in mixed strategies.[17] Since $\Delta(\theta(0)) > 0$, an agent who always chooses a low price must believe in equilibrium that setting a high price would instead be optimal. Similarly, $\Delta(\theta(1)) < 0$ implies that an agent who always chooses a high price must believe in equilibrium that settings a low price would instead be optimal. Therefore, in equilibrium, the agent chooses a strictly mixed strategy $\sigma_H^* \in (0, 1)$ such that $\Delta(\theta(\sigma_H^*)) = 0$.[18]

[15] Both $\sigma_H = 0$ and $\sigma_H = 1$ are Berk-Nash equilibria supported by beliefs $\theta_H(0) = 0$ and $\theta_L(1) = 0$, respectively. These outcomes, however, are not robust to experimentation, and are eliminated by requiring $\theta_H(0) = \lim_{\sigma_H \to 0} \bar{s}_H(m(\cdot; \sigma_H)) = \bar{s}_H(m(\cdot; 0))$, and similarly for $\theta_L(1)$.

[16] The reason is that $\frac{d}{d\sigma_H}\Delta(\theta(\sigma_H)) = \frac{d}{d\sigma_H} m_{\mathbb{S}}(1; \sigma_H)\,(H(q_{1H} - q_{0H}) + L(q_{1L} - q_{0L})) > 0$, since $\frac{d}{d\sigma_H} m_{\mathbb{S}}(1; \sigma_H) < 0$ and $q_{1x} > q_{0x}$ for all $x \in \{L, H\}$.

[17] See Esponda and Pouzo (2016) for the importance of mixed strategies in misspecified settings.

[18] More generally, the unique equilibrium is $\sigma_H = 0$ if $\Delta(\theta(0)) < 0$ (i.e., $\frac{H}{L} \leq D_1 \equiv \frac{q_{0L}}{(1-q_{1L})q_{0H} + q_{1H}q_{0L}}$), $\sigma_H = 1$ if $\Delta(\theta(1)) > 0$ (i.e., $\frac{H}{L} \geq D_2 \equiv (1 - q_{1H})\frac{q_{0L}}{q_{0H}} + q_{1L}$), and $\sigma_H^* \in (0, 1)$ the solution to $\Delta(\theta(\sigma_H^*)) = 0$ if $D_1 < \frac{H}{L} < D_2$, where $\frac{q_{1L}}{q_{1H}} < D_1 < D_2 < \frac{q_{0L}}{q_{1H}}$.

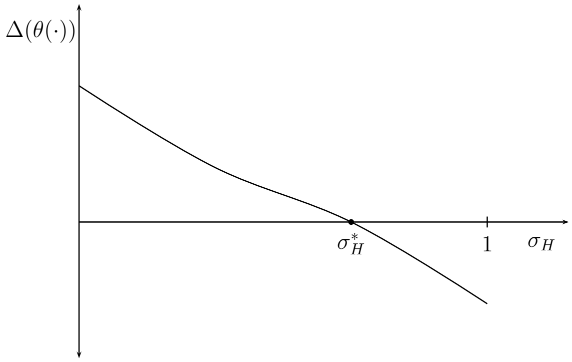

Figure 1: Equilibrium of the monopoly example

The misspecified monopolist may end up choosing higher prices than optimal, since she fails to realize that high prices today cost her in the future. But, a bit more surprisingly, she also may end up choosing lower prices for some primitives.[19] The reason is that her failure to realize that H does relatively better in state $s = 1$ makes H unattractive to her.

4.2 Search with uncertainty about future job offers

Search-theoretic models have been central to understanding labor markets since McCall (1970). Most of the literature assumes that the worker knows all the primitives. Exceptions include Rothschild (1974a) and Burdett and Vishwanath (1988), wherein the worker does not know the wage distribution but has a correctly-specified model. In contrast, we study a worker or entrepreneur who *knows* the distribution of wages or returns for new projects but does not know the probability that she would be able to find a new job or fund a new project. The worker or entrepreneur, however, does not realize that she is fired or her project fails with higher probability in times in which it is actually harder to find a new job or fund a new project. We show that the worker or entrepreneur becomes pessimistic about the chances of finding new prospects and sub-optimally accepts prospects with low returns in equilibrium.

MDP. At the beginning of each period t, a worker (or entrepreneur) faces a wage offer (or a project with returns) $w_t \in \mathbb{S} = [0, 1]$ and decides whether to reject or accept it, $x_t \in \mathbb{X} = \{0, 1\}$.[20] Her payoff in period t is $\pi(w_t, x_t) = w_t x_t$; i.e, she earns w_t if

[19] This happens if $C_\delta < H/L < D_1$; see footnotes 14 and 18.

[20] The set of feasible actions is independent of the state, i.e., $\Gamma(w) = \mathbb{X}$ for all $w \in \mathbb{S}$.

she accepts and zero otherwise. After making her decision, an economic fundamental $z_t \in \mathbb{Z}$ is drawn from an i.i.d. distribution G.[21] If the worker is employed, she is fired (or the project fails) with probability $\gamma(z_t)$. If the worker is unemployed (either because she was employed and then fired or because she did not accept employment at the beginning of the period), then with probability $\lambda(z_t)$ she draws a new wage $w_{t+1} \in [0,1]$ according to some absolutely continuous distribution F with density f; wages are independent and identically distributed across time. With probability $1-\lambda(z_t)$, the unemployed worker receives no wage offer, and we denote the corresponding state by $w_{t+1}=0$ without loss of generality. The worker will have to decide whether to accept or reject w_{t+1} at the beginning of next period. If the worker accepted employment at wage w_t at the beginning of time t and was not fired, then she starts next period with wage offer $w_{t+1}=w_t$ and will again have to decide whether to quit or remain in her job at that offer.[22] The agent wants to maximize discounted expected utility with discount factor $\delta \in [0,1)$. Suppose that $\overline{\gamma} \equiv E[\gamma(Z)] > 0$ and $\overline{\lambda} \equiv E[\lambda(Z)] > 0$.

We assume that $Cov(\gamma(Z), \lambda(Z)) < 0$; for example, the worker is more likely to get fired and less likely to receive an offer when economic fundamentals are strong, and the opposite holds when fundamentals are weak.

SMDP. The worker knows all the primitives except $\lambda(\cdot)$, which determines the probability of receiving an offer. The worker has a misspecified model of the world and believes $\lambda(\cdot)$ does not depend on the economic fundamental, i.e., $\lambda(z)=\theta$ for all $z \in \mathbb{Z}$, where $\theta \in [0,1]$ is the unknown parameter.[23] The transition probability function $Q_\theta(w' \mid w, x)$ is as follows: If $x=1$, then $w'=w$ with probability $1-\theta$, w' is a draw from F with probability $\theta\overline{\gamma}$, and $w'=0$ with probability $(1-\theta)\overline{\gamma}$; If $x=0$, then w' is a draw from F with probability θ and $w'=0$ with probability $1-\theta$.

Equilibrium. *Optimality.* Suppose that the worker believes that the true parameter is θ with probability 1. The value of receiving wage offer $w \in \mathbb{S}$ is

$$\begin{aligned} V(w) = \max\{&w + \delta\left((1-\overline{\gamma})V(w) + (1-\theta)\overline{\gamma}V(0) + \theta\overline{\gamma}E[V(W')]\right), \\ &0 + \delta\left(\theta E[V(W')] + (1-\theta)V(0)\right)\}. \end{aligned}$$

[21] To simplify the notation, we assume the fundamental is unobserved, although the results are identical if it is observed, since it is i.i.d. and it is realized after the worker makes her decision.

[22] Formally, $Q(w' \mid w, x)$ is as follows: If $x=1$, then $w'=w$ with probability $1-\overline{\gamma}$, w' is a draw from F with probability $E[\gamma(Z)\lambda(Z)]$, and $w'=0$ with probability $E[\gamma(Z)(1-\lambda(Z))]$; If $x=0$, then w' is a draw from F with probability $\overline{\lambda}$ and $w'=0$ with probability $1-\overline{\lambda}$.

[23] The results are identical if the agent is also uncertain of $\gamma(\cdot)$; given the current misspecification, the agent only cares about the expectation of γ and will have correct beliefs about it.

By standard arguments, her optimal strategy is a stationary reservation wage strategy $w(\theta)$ that solves the following equation:

$$w(\theta)(1-\delta+\delta\overline{\gamma}) = \delta\theta(1-\overline{\gamma})\int_{w>w(\theta)} (w-w(\theta))\,F(dw). \tag{6}$$

The worker accepts wages above the reservation wage and rejects wages below it. Also, $\theta \mapsto w(\theta)$ is increasing: The higher is the probability of receiving a wage offer, then the more she is willing to wait for a better offer in the future. Figure 2 depicts an example.

Beliefs. For any $m \in \Delta(\mathbb{S}\times\mathbb{X})$, the wKLD simplifies to

$$\begin{aligned} K_Q(m,\theta) &= \int_{\mathbb{S}\times\mathbb{X}} E_{Q(\cdot|\tilde{w},x)}\Big[\ln\frac{Q(W' \mid \tilde{w},x)}{Q_\theta(W' \mid \tilde{w},x)}\Big] m(d\tilde{w},dx) \\ &= \Big\{E[\gamma\lambda]\ln\frac{E[\gamma\lambda]}{\overline{\gamma}\theta} + E[\gamma(1-\lambda)]\ln\frac{E[\gamma(1-\lambda)]}{\overline{\gamma}(1-\theta)}\Big\} m_{\mathbb{X}}(1) \\ &+ \Big\{\overline{\lambda}\ln\frac{\overline{\lambda}}{\theta} + (1-\overline{\lambda})\ln\frac{1-\overline{\lambda}}{1-\theta}\Big\} m_{\mathbb{X}}(0), \end{aligned}$$

where the density of W' cancels out because the workers knows it and where $m_{\mathbb{X}}$ is the marginal distribution over $\mathbb{X}$. In the Online Appendix, we show that the unique parameter that minimizes $K_Q(m,\cdot)$ is

$$\theta_Q(m) \equiv \frac{m_{\mathbb{X}}(0)}{m_{\mathbb{X}}(0)+m_{\mathbb{X}}(1)\overline{\gamma}}\overline{\lambda} + \Big(1-\frac{m_{\mathbb{X}}(0)}{m_{\mathbb{X}}(0)+m_{\mathbb{X}}(1)\overline{\gamma}}\Big)\Big(\overline{\lambda}+\frac{Cov(\gamma,\lambda)}{\overline{\gamma}}\Big). \tag{7}$$

To see the intuition behind equation (7), note that the agent only observes the realization of λ, i.e., whether she receives a wage offer, when she is unemployed. Unemployment can be voluntary or involuntary. In the first case, the agent rejects the offer and, since this decision happens before the fundamental is realized, it is independent of getting or not an offer. Thus, with conditional on unemployment being voluntary, the agent will observe an unbiased average probability of getting an offer, $\overline{\lambda}$ (see the first term in the RHS of (7)). In the second case, the agent accepts the offer but is then fired. Since $Cov(\gamma,\lambda) < 0$, she is less likely to get an offer in periods in which she is fired and, because she does not account for this correlation, she will have a more pessimistic view about the probability of receiving a wage offer relative to the average probability $\overline{\lambda}$ (the second term in the RHS of (7) captures this bias).

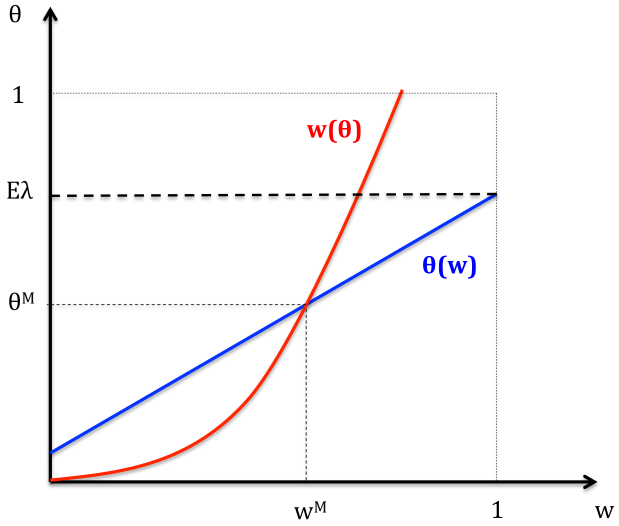

Figure 2: Equilibrium of the search model

Stationary distribution. Fix a reservation wage strategy w and denote the marginal over $\mathbb{X}$ of the corresponding stationary distribution by $m_{\mathbb{X}}(\cdot; w) \in \Delta(\mathbb{X})$. In the Online Appendix, we characterize $m_{\mathbb{X}}(\cdot; w)$ and show that $w \mapsto m_{\mathbb{X}}(0; w)$ is increasing. Intuitively, the more selective the worker, the higher the chance of being unemployed.

Equilibrium. Let $\theta(\omega) \equiv \theta_Q(m(\cdot; w))$ denote the equilibrium belief for an agent following reservation wage strategy w. The weight on $\overline{\lambda}$ in equation (7) represents the probability of voluntary unemployment conditional on unemployment. This weight is increasing in ω because $w \mapsto m_{\mathbb{X}}(0; w)$ is increasing. Therefore, $w \mapsto \theta(w)$ is increasing. In the extreme case in which $w = 1$, the worker rejects all offers, unemployment is always voluntary, and the bias disappears, $\theta(1) = \overline{\lambda}$. An example of the schedule $\theta(\cdot)$ is depicted in Figure 2. The set of Berk-Nash equilibria is given by the intersection of $w(\cdot)$ and $\theta(\cdot)$. In the example depicted in Figure 2, there is a unique equilibrium strategy $w^M = w(\theta^M)$, where $\theta^M < \overline{\lambda}$.

We conclude by comparing Berk-Nash equilibria to the optimal strategy of a worker who knows the primitives, w^*. By standard arguments, w^* is the unique solution to

$$w^*(1 - \delta + \delta\overline{\gamma}) = \delta(\overline{\lambda} - E[\gamma\lambda]) \int_{w > w^*} (w - w^*) \, F(dw). \tag{8}$$

The only difference between equations (6) and (8) appears in the term multiplying the RHS, which captures the cost of accepting a wage offer. In the misspecified case, this term is $\delta\theta(1-\overline{\gamma})$; in the correct case, it is $\delta(\overline{\lambda}-E[\gamma\lambda]) = \delta\overline{\lambda}(1-\overline{\gamma})-\delta Cov(\gamma,\lambda)$. The misspecification affects the optimal threshold in two ways. First, the misspecified agent estimates the mean of λ incorrectly, i.e., $\theta < \overline{\lambda}$; therefore, she (incorrectly) believes that, in expectation, offers arrive with lower probability. Second, she does not realize that, because $Cov(\gamma,\lambda) < 0$, she is less likely to receive an offer when fired. Both effects go in the same direction and make the option to reject and wait for the possibility of drawing a new wage offer next period less attractive for the misspecified worker. Formally, $\theta\delta(1-\overline{\gamma}) < \delta\overline{\lambda}(1-\overline{\gamma}) - \delta Cov(\gamma,\lambda)$ and so $w^M < w^*$.

4.3 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, but assume that these shocks are (positively) correlated. We show that agents who fail to account for the correlation of shocks underinvest in equilibrium.

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 income from the previous period and z_t is a current utility shock, and chooses how much income to save, $x_t \in \Gamma(y_t, z_t) = [0, y_t] \subseteq \mathbb{X} = \mathbb{R}_+$, consuming the rest. Current period utility is $\pi(y_t, z_t, x_t) = z_t \ln(y_t - x_t)$. Income next period, y_{t+1}, is given by

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

where $\varepsilon_t = \gamma^* z_t + \xi_t$ is an unobserved productivity shock, $\xi_t \sim N(0,1)$, and $0 < \delta\beta^* < 1$, where $\delta \in [0,1)$ is the discount factor. We assume that $\gamma^* > 0$, so that the utility and productivity shocks are positively correlated. Let $0 < L < H$ and let $q \in (0,1)$ be the probability that the shock is H.[24]

SMDP. The agent believes that

$$\ln y_{t+1} = \alpha + \beta \ln x_t + \varepsilon_t, \tag{10}$$

[24] 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.

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 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

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

and it is straightforward to verify that the optimal strategy is to invest a fraction of income that depends on the utility shock and the unknown parameter β, i.e., $x = A_z(\beta) \cdot y$, where $A_L(\beta) = \frac{\delta\beta((1-q)L+qH)}{(1-\delta\beta(1-q))H+\delta\beta(1-q)}$ and $A_H(\beta) = \frac{\delta\beta((1-q)L+qH)}{\delta\beta qH+(1-\delta\beta q)L} < A_L(\beta)$. 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 stationary distribution. Let $A = (A_L, A_H)$, with $A_H < A_L$, represent a strategy, where A_z is the proportion of income invested given utility shock z. Because the agent believes that ε_t is independent of the utility shock and normally distributed, minimizing the wKLD function is equivalent to performing an OLS regression of equation (10). Thus, for a strategy represented by $A = (A_L, A_H)$, the parameter value $\hat{\beta}(A)$ that minimizes wKLD is

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

where Cov and Var are taken with respect to the (true) stationary distribution of (Y, Z). Since $A_H < A_L$, then $Cov(Z, \ln A_Z) < 0$. 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 income 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 function $\hat{\beta}(A_L(\beta), A_H(\beta))$.[25] The function is continuous in β 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.[26] 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 under-confident about productivity).

5 Equilibrium foundation

In this section, we provide a learning foundation for the notion of Berk-Nash equilibrium of SMDPs. We fix an SMDP and assume that the agent is Bayesian and starts with a prior $\mu_0 \in \Delta(\Theta)$ over her set of models of the world. She observes past actions and states and uses this information to update her beliefs about Θ in every period.

Definition 11. For any $(s, x, s') \in Gr(\Gamma) \times \mathbb{S}$, let $B(s, x, s', \cdot) : D_{s,x,s'} \to \Delta(\Theta)$ denote the Bayesian operator: For all $A \subseteq \Theta$ Borel

$$B(s, x, s', \mu)(A) = \frac{\int_A Q_\theta(s' \mid s, x)\mu(d\theta)}{\int_\Theta Q_\theta(s' \mid s, x)\mu(d\theta)}. \tag{11}$$

for any $\mu \in D_{s,x,s'}$, where $D_{s,x,s'} = \{p \in \Delta(\Theta) \colon \int_\Theta Q_\theta(s' \mid s, x)p(d\theta) > 0\}$.

Definition 12. A **Bayesian Subjective Markov Decision Process** (Bayesian-SMDP) is an SMDP$(Q, \mathcal{Q}_\Theta)$ together with a prior $\mu_0 \in \Delta(\Theta)$ and the Bayesian operator B (see Definition 11). It is said to be **regular** if the corresponding SMDP is regular.

[25] Our existence theorem is not directly applicable because we have assumed, for convenience, nonfinite state and action spaces.

[26] 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 (i.e., it is not perfect; see Section 6).

By the Principle of Optimality, the agent's problem in a Bayesian-SMDP can be cast recursively as

$$W(s,\mu) = \max_{x \in \Gamma(s)} \int_{\mathbb{S}} \{\pi(s,x,s') + \delta W(s',\mu')\} \bar{Q}_\mu(ds'|s,x), \tag{12}$$

where $\bar{Q}_\mu = \int_\Theta Q_\theta \mu(d\theta)$, $\mu' = B(s,x,s',\mu)$ is 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 (12). 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 13. A **policy function** is a function $f : \Delta(\Theta) \to \Sigma$ mapping beliefs into strategies (recall that a strategy is a mapping $\sigma : \mathbb{S} \to \Delta(\mathbb{X})$). For any belief $\mu \in \Delta(\Theta)$, state $s \in \mathbb{S}$, and action $x \in \mathbb{X}$, let $f(x \mid s,\mu)$ denote the probability that the agent chooses x when selecting policy function f. A policy function f is **optimal** for the Bayesian-SMDP if, for all $s \in \mathbb{S}$, $\mu \in \Delta(\Theta)$, and $x \in \mathbb{X}$ such that $f(x \mid s,\mu) > 0$,

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

For each $\mu \in \Delta(\Theta)$, let $\bar{\Sigma}(\mu) \subseteq \Sigma$ denote the set of all strategies that are induced by a policy that is *optimal*, i.e.,

$$\bar{\Sigma}(\mu) = \big\{\sigma \in \Sigma : \exists \text{ optimal } f \text{ such that } \sigma(\cdot \mid s) = f(\cdot \mid s,\mu) \text{for all } s \in \mathbb{S}\big\}.$$

Lemma 4. *(i) There is a unique solution W to the Bellman equation in (12), and it is continuous in μ for all $s \in \mathbb{S}$; (ii) The correspondence of optimal strategies $\mu \mapsto \bar{\Sigma}(\mu)$ is non-empty, compact-valued, convex-valued, and upper hemicontinuous.*

Proof. The proof is standard and relegated to the Online Appendix. □

Let $h^\infty = (s_0, x_0, ..., s_t, x_t, ...)$ represent the infinite history or outcome path of the dynamic optimization problem and let $\mathbb{H}^\infty \equiv (Gr(\Gamma))^\infty$ represent the space of infinite histories. For every t, let $\mu_t : \mathbb{H}^\infty \to \Delta(\Theta)$ denote the agent's Bayesian beliefs, defined recursively by $\mu_t = B(s_{t-1}, x_{t-1}, s_t, \mu_{t-1})$ whenever $\mu_{t-1} \in D_{s_{t-1},x_{t-1},s_t}$ (see Definition

11), and arbitrary otherwise. We assume that the agent follows some policy function f. In each period t, there is a state s_t and a belief μ_t, and the agent chooses a (possibly mixed) action $f(\cdot \mid s_t, \mu_t) \in \Delta(\mathbb{X})$. 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 beliefs to μ_{t+1} using Bayes' rule. The primitives of the Bayesian-SMDP (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 $\mathbb{H}^\infty$ that is defined in a standard way; let $\boldsymbol{P}^f$ denote this probability distribution over $\mathbb{H}^\infty$.

We now define strategies and outcomes as random variables. For a fixed policy function f and for every t, let $\sigma_t : \mathbb{H}^\infty \to \Sigma$ denote the strategy of the agent, defined by setting

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

Finally, for every t, let $m_t : \mathbb{H}^\infty \to \Delta(Gr(\Gamma))$ be such that, for all t, h^∞, and $(s,x) \in Gr(\Gamma)$,

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

is the frequency of times that the outcome (s,x) occurs up to time t.

One reasonable criteria to claim that the agent has reached a steady-state is that her strategy and the time average of outcomes converge.

Definition 14. A strategy and probability distribution $(\sigma, m) \in \Sigma \times \Delta(Gr(\Gamma))$ is **stable** for a Bayesian-SMDP with prior μ_0 and policy function f if there is a set $\mathcal{H} \subseteq \mathbb{H}$ with $\mathbf{P}^f(\mathcal{H}) > 0$ such that, for all $h^\infty \in \mathcal{H}$, as $t \to \infty$,

$$\sigma_t(h^\infty) \to \sigma \quad \text{and} \quad m_t(h^\infty) \to m. \tag{13}$$

If, in addition, there exists a belief μ^* and a subsequence $(\mu_{t(j)})_j$ such that,

$$\mu_{t(j)}(h^\infty) \overset{w}{\to} \mu^* \tag{14}$$

and, for all $(s,x) \in Gr(\Gamma)$, $\mu^* = B(s, x, s', \mu^*)$ for all $s' \in \mathbb{S}$ such that $\bar{Q}_{\mu^*}(s' \mid s, x) > 0$, then (σ, m) is called **stable with exhaustive learning**.

Condition (13) requires that strategies and the time frequency of outcomes stabilize. By compactness, there exists a subsequence of beliefs that converges. The additional requirement of exhaustive learning says that the limit point of one of the subsequences, μ^*, is perceived to be a fixed point of the Bayesian operator, implying that no matter what state and strategy the agent contemplates, she does not *expect* her belief to change. Thus, the agent believes that all learning possibilities are exhausted under μ^*. The condition, however, does not imply that the agent has correct beliefs in steady state.

The next result establishes that, if the time average of outcomes stabilize to m, then beliefs become increasingly concentrated on $\Theta_Q(m)$.

Lemma 5. *Consider a regular Bayesian-SMDP with true transition probability function Q, full-support prior $\mu_0 \in \Delta(\Theta)$, and policy function f. 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 $\mathcal{H}$.

Proof. See the Appendix. □

The proof of Lemma 5 clarifies the origin of the wKLD function in the definition of Berk-Nash equilibrium. The proof adapts the proof of Lemma 2 by Esponda and Pouzo (2016) to dynamic environments. Lemma 5 extends results from the statistics of misspecified learning (Berk (1966), Bunke and Milhaud (1998), Shalizi (2009)) by considering a setting where agents learn from data that is endogenously generated by their own actions in a Markovian setting.

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

Theorem 2. *There exists $\bar{\delta} \in [0,1]$ such that:*

(i) for all $\delta \le \bar{\delta}$, if (σ, m) is stable for a regular Bayesian-SMDP with full-support prior μ_0 and policy function f that is optimal, then (σ, m) is a Berk-Nash equilibrium of the SMDP.

(ii) for all $\delta > \bar{\delta}$, if (σ, m) is stable with exhaustive learning for a regular Bayesian-SMDP with full-support prior μ_0 and policy function f that is optimal, then (σ, m) is a Berk-Nash equilibrium of the SMDP.

Proof. See the Appendix. □

Theorem 2 provides a learning justification for Berk-Nash equilibrium. 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 5 and the fact that behavior converges to σ, it follows that σ must solve the dynamic optimization problem for beliefs converging to $\mu^* \in \Theta_Q(m)$. In addition, by convergence of σ_t to σ and continuity of the transition kernel $\sigma \mapsto M_{\sigma,Q}$, an application of the martingale convergence theorem implies that m_t is asymptotically equal to $M_{\sigma,Q}[m_t]$. This fact, linearity of the operator $M_{\sigma,Q}[\cdot]$, and convergence of m_t to m then imply that m is an invariant distribution given σ.

The proof concludes by showing that σ not only solves the optimization problem for beliefs converging to μ^* but also solves the MDP, where the belief is forever fixed at μ^*. This is true, of course, if the agent is sufficiently impatient, which explains why part (i) of Theorem 2 holds. For sufficiently patient agents, the result relies on the assumption that the steady state satisfies exhaustive learning. We now illustrate and discuss the role of this assumption.

EXAMPLE. At the initial period, a risk-neutral agent has four investment choices: A, B, S, and O. Action A pays $1 - \theta^*$, action B pays θ^*, and action S pays a safe payoff of $2/3$ in the initial period, where $\theta^* \in \{0, 1\}$. For any of these three choices, the decision problem ends there and the agent makes a payoff of zero in all future periods. Action O gives the agent a payoff of $-1/3$ in the initial period and the option to make an investment next period, where there are two possible states, s_A and s_B. State s_A is realized if $\theta^* = 1$ and state s_B is realized if $\theta^* = 0$. In each of these states, the agent can choose to make a risky investment or a safe investment. The safe investment gives a payoff of $2/3$ in both states, and a subsequent payoff of zero in all future periods. The risky investment gives the agent a payoff that is thrice the payoff she would have gotten from choice A, that is, $3(1 - \theta^*)$, if the state is s_A, and it gives the agent thrice the payoff she would have gotten from choice B, that is, $3\theta^*$, if the state is s_B; the payoff is zero is all future periods.

Suppose that the agent knows all the primitives except the value of θ^*. Let $\Theta = \{0, 1\}$; in particular, the SMDP is correctly specified.

This problem is simple enough that we can directly characterize a steady-state and then check if it is a Berk-Nash equilibrium. Suppose the (Bayesian) agent who starts with a prior $\mu = \Pr(\theta = 1) \in (0, 1)$ and updates her belief. The value of action

O is

$$-\frac{1}{3} + \delta\left(\mu W(s_A, 1) + (1-\mu) W(s_B, 0)\right) = -\frac{1}{3} + \delta\frac{2}{3} < \frac{2}{3}, \tag{15}$$

where we have used the fact that $W(s_A, 1) = W(s_B, 0) = 2/3$. In other words, the agent realizes that if the state s_A is realized, then she will update her belief to $\mu' = 1$, which implies that the safe investment is optimal in state s_A; a similar argument holds for state s_B. She then finds it optimal to choose action A if $\mu \leq 1/3$, B if $\mu \geq 2/3$, and S if $\mu \in [1/3, 2/3]$. In particular, choosing S is a steady state outcome for any prior in $[1/3, 2/3]$.

We now show that the safe action, S, is *not* a Berk-Nash equilibrium if the agent is sufficiently patient. Let $\mu \in [0, 1]$ denote the agent's equilibrium belief about the probability that $\theta^* = 1$. For action S to be preferred to A and B, it must be the case that $\mu \in [1/3, 2/3]$. But, for a fixed μ, the perceived benefit from action O is

$$\begin{aligned} -\frac{1}{3} + \delta\left(\mu W(s_A, \mu) + (1-\mu) W(s_B, \mu)\right) &= -\frac{1}{3} + \delta\left(\mu \max\{\frac{2}{3}, 3(1-\mu)\} + (1-\mu)\max\{\frac{2}{3}, 3\mu\}\right) \\ &\geq -\frac{1}{3} + \delta 6\mu(1-\mu), \end{aligned} \tag{16}$$

which is strictly higher than 2/3, the payoff from action S, for all $\mu \in [1/3, 2/3]$ provided that $\delta > \bar{\delta} = 3/4$. Thus, for a sufficiently patient agent, there is no belief that makes action S optimal and, therefore, S is not chosen in any Berk-Nash equilibrium. The belief supporting S, however, does not satisfy exhaustive learning, since the agent believes that any other action would completely reveal all uncertainty. □

The previous example illustrates an important tension that arises when an equilibrium concept–where strategies are optimal given a *fixed* equilibrium belief–is intended to represent the steady state of a dynamic environment where beliefs are being updated. This tension, however, has not been recognized in the past, where equilibrium concepts have been shown to successfully capture steady-state behavior. The reason is that the tension illustrated by the previous example does not arise in static environments (where the only link between periods is the updating of a belief).

We will now explain why the tension described above does not arise in static environments, why it does arise in the type of dynamic environments that we study in this paper, and how the property of exhaustive learning is used in the proof of

Theorem 2 to deliver the intended result. We call an action a steady-state action if it is in the support of a stable strategy and we call it a non steady-state action otherwise. A key step is to show that, if a steady-state action is better than a non steady-state action when beliefs are updated, it will also be better when beliefs are fixed.

Consider first an inherently static environment (Esponda and Pouzo (2016)). Suppose that x and not y is a steady-state action, implying that x yields a higher payoff than y :

$$E_{Q_\mu(\cdot|x)}\left[\pi(x,S')+\delta V(B(x,S',\mu))\right]\geq E_{Q_\mu(\cdot|y)}\left[\pi(y,S')+\delta V(B(y,S',\mu))\right]. \tag{17}$$

In particular, the value function, V, only depends on the agent's belief. If we assume weak identification, then $B(x,s',\mu)=\mu$ for all s' that occur with positive probability according to μ, and so the LHS of (17) becomes $E_{Q_\mu(\cdot|x)}\left[\pi(x,S')+\delta V(\mu))\right]$. Next, we add and subtract $\delta V(\mu)$ from the RHS of (17) to obtain

$$E_{Q_\mu(\cdot|y)}\left[\pi(y,S')+\delta V(\mu)\right]+\delta E_{Q_\mu(\cdot|y)}\left[V(B(y,S',\mu))-V(\mu)\right]. \tag{18}$$

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 next period with updated belief $B(y,S',\mu)$, which depends on the action y and the random realization of S', and starting next period with the current belief μ. By the Martingale property of Bayesian updating and the convexity of the value function, it follows that the value of experimentation is nonnegative.[27] It then follows that (17) implies $E_{Q_\mu(\cdot|x)}\left[\pi(x,S')\right]\geq E_{Q_\mu(\cdot|y)}\left[\pi(y,S')\right]$. Thus, an action that is dynamically optimal is also optimal when the belief is fixed.

The previous argument does not carry over to an inherently dynamic environment. Suppose that x and not y is a steady-state action, implying that x yields a higher payoff than y :

$$E_{Q_\mu(\cdot|s,x)}\left[\pi(s,x,S')+\delta W(S',B(s,x,S',\mu))\right]\geq E_{Q_\mu(\cdot|s,y)}\left[\pi(s,y,S')+\delta W(S',B(s,y,S',\mu))\right]. \tag{19}$$

The value function, W, now also depends on a non-belief state, S'. As before, weak identification implies that the LHS of (19) is equivalent to $E_{Q_\mu(\cdot|s,x)}\left[\pi(s,x,S')+\delta W(S',\mu)\right]$.

[27] Formally, $E_{Q_\mu(\cdot|y)}\left[V(B(y,S',\mu))-V(\mu)\right]\geq V[E_{Q_\mu(\cdot|y)}B(y,S',\mu)]-V(\mu)=0$.

Next, we add and subtract $\delta E_{Q_\mu(\cdot|s,y)}\left[W(S',\mu)\right]$ from the RHS of (19) to obtain

$$E_{Q_\mu(\cdot|s,y)}\left[\pi(s,y,S')+\delta W(S',\mu)\right]+\delta E_{Q_\mu(\cdot|s,y)}\left[W(S',B(s,y,S',\mu))-W(S',\mu)\right]. \quad (20)$$

The second term in (20) is the difference in net present value between starting next period with non-belief state S' and updated belief $B(y,S',\mu)$ and starting next period with non-belief state S' and belief μ. This expression no longer represents what is traditionally understood as the value of experimentation because one also has to take into account that the non-belief state is changing. In fact, as the previous example illustrates, this second term may actually be negative (see equations (15) and (16)). The role of exhaustive learning is to guarantee that this second term is equal to zero. When this term is zero, (19) implies

$$E_{Q_\mu(\cdot|s,x)}\left[\pi(s,x,S')+\delta W(S',\mu)\right]\geq E_{Q_\mu(\cdot|s,y)}\left[\pi(s,y,S')+\delta W(S',\mu)\right],$$

and, therefore, an action that is dynamically optimal in the dynamic environment is also dynamically optimal when the belief is fixed.

We conclude with additional remarks about Theorem 2.

Remark 3. *Discount factor*: In the proof of Theorem 2, we provide an exact value for $\bar{\delta}$ as a function of primitives. This bound, however, may not be sharp. As illustrated by the above example, to compute a sharp bound we would have to solve the dynamic optimization problem with learning, which is precisely what we are trying to avoid by focusing on Berk-Nash equilibrium.

Convergence: Theorem 2 does not imply that behavior will necessarily stabilize 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. So one cannot hope to have general statements regarding convergence of outcomes—this is also true, for example, in the related context of learning to play Nash equilibrium in games.[28] Thus, the theorem leaves open the question of convergence in specific settings, a question that requires other tools (e.g., stochastic approximation) and is best tackled by explicitly studying the dynamics of specific classes of environments (see the references in the introduction).

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

Mixed strategies: Theorem 2 also raises the question of how a mixed strategy could ever become stable, given that, in general it is unlikely that agents will hold beliefs that make them exactly indifferent at any point in time. Fudenberg and Kreps (1993) asked the same question in the context of learning to play mixed strategy *Nash* equilibria, and answered it by adding small payoff perturbations a la Harsanyi (1973): 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 followed this approach in the paper that introduced Berk-Nash equilibrium in static contexts (Esponda and Pouzo, 2016). The same idea applies here, but we omit payoff perturbations to reduce the notational burden.[29]

6 Equilibrium refinements

Theorem 2 implies that, for sufficiently patient players, we should be interested in the following refinement of Berk-Nash equilibrium.

Definition 15. A strategy and probability distribution $(\sigma, m) \in \Sigma \times \Delta(Gr(\Gamma))$ is a **Berk-Nash equilibrium with exhaustive learning** of the SMDP if it is a Berk-Nash equilibrium that is supported by a belief $\mu^* \in \Delta(\Theta)$ such that, for all $(s, x) \in Gr(\Gamma)$,

$$\mu^* = B(s, x, s', \mu^*)$$

for all $s' \in \mathbb{S}$ such that $\bar{Q}_{\mu^*}(s' \mid s, x) > 0$.

In an equilibrium with exhaustive learning, there is a supporting belief that is perceived to be a fixed point of the Bayesian operator, implying that no matter what state and strategy the agent contemplates, she does not *expect* her belief to change. The requirement of exhaustive learning does not imply robustness to experimentation. For example, in the monopoly problem studied in Section 4.1, choosing low price with probability 1 is an equilibrium with exhausted learning which is supported by the belief that, with probability 1, $\theta_L^* = 0$. We rule out equilibria that are not robust to experimentation by introducing a further refinement.

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

Definition 16. An ε-perturbed SMDP is an SMDP wherein strategies are restricted to belong to

$$\Sigma^{\varepsilon} = \{\sigma \in \Sigma : \sigma(x \mid s) \geq \varepsilon \text{ for all } (s, x) \in Gr(\Gamma)\}.$$

Definition 17. A strategy and probability distribution $(\sigma, m) \in \Sigma \times \Delta(Gr(\Gamma))$ is a **perfect Berk-Nash equilibrium** of an SMDP if there exists a sequence $(\sigma^{\varepsilon}, m^{\varepsilon})_{\varepsilon>0}$ of Berk-Nash equilibria with exhaustive learning of the ε-perturbed SMDP that converges to (σ, m) as $\varepsilon \to 0$.[30]

Selten (1975) introduced the idea of perfection in extensive-form games. By itself, however, perfection does not guarantee that all $(s, x) \in Gr(\Gamma)$ are reached in an MDP. The next property guarantees that all states can be reached when the agent chooses all strategies with positive probability.

Definition 18. An MDP(Q) satisfies **full communication** if, for all $s_0, s' \in \mathbb{S}$, there exist finite sequences $(s_1, ..., s_n)$ and $(x_0, x_1, ..., x_n)$ such that $(s_i, x_i) \in Gr(\Gamma)$ for all $i = 0, 1, ..., n$ and

$$Q(s' \mid s_n, x_n)Q(s_n \mid s_{n-1}, x_{n-1})...Q(s_1 \mid s_0, x_0) > 0.$$

An SMDP satisfies full communication if the corresponding MDP satisfies it.

Full communication is standard in the theory of MDPs and holds in all of the examples in Section 4. It guarantees that there is a single recurrent class of states for all ε-perturbed environments. In cases where it does not hold and there is more than one recurrent class of states, one can still apply the following results by focusing on one of the recurrent classes and ignoring the rest as long as the agent correctly believes that she cannot go from one recurrent class to the other.

Full communication guarantees that there are no off-equilibrium outcomes in a perturbed SMDP. It does not, however, rule out the desire for experimentation on the equilibrium path. We rule out the latter by requiring weak identification.

[30] Formally, in order to have a sequence, we take $\varepsilon > 0$ to belong to the rational numbers; hereinafter we leave this implicit to ease the notational burden.

Proposition 2. *Suppose that an SMDP is weakly identified, ε-perturbed, and satisfies full communication.*

(i) If the SMDP is regular and if (σ, m) is stable for the Bayesian-SMDP, it is also stable with exhaustive learning.

(ii) If (σ, m) is a Berk-Nash equilibrium, it is also a Berk-Nash equilibrium with exhaustive learning.

Proof. See the Appendix. □

Proposition 2 provides conditions such that a steady state satisfies exhaustive learning and a Berk-Nash equilibrium can be supported by a belief that satisfies the exhaustive learning condition. Under these conditions, we can find equilibria that are robust to experimentation, i.e., perfect equilibria, by considering perturbed environments and taking the perturbations to zero (see the examples in Section 4).

The next proposition shows that perfect Berk-Nash is a refinement of Berk-Nash with exhaustive learning. As illustrated by the monopoly example in Section 4.1, it is a strict refinement.

Proposition 3. *Any perfect Berk-Nash equilibrium of a regular SMDP is a Berk-Nash equilibrium with exhaustive learning.*

Proof. See the Appendix. □

We conclude by showing existence of perfect Berk-Nash equilibrium (hence, of Berk-Nash equilibrium with exhaustive learning, by Proposition 3).

Theorem 3. *For any regular SMDP that is weakly identified and satisfies full communication, there exists a perfect Berk-Nash equilibrium.*

Proof. See the Appendix. □

7 Conclusion

We provide a framework for modeling the behavior of an agent who holds a simplified view of a recursive dynamic optimization problem. The agent faces a Markov decision process and has a prior over a set of possible transition probability functions. This set captures the agent's simplified view of her environment; in particular, the agent has a misspecified model if the set does not include the true transition function. We focus on asymptotic behavior of an agent who updates her beliefs using Bayes' rule. In particular, we define an equilibrium notion, Berk-Nash equilibrium, in order to capture the agent's steady state behavior. Two key features of our approach is that it distinguishes between the agent's simplified model and the true primitives and that the agent's belief is determined endogenously in equilibrium. Moreover, the framework can be used to tackle applications that remained previously inaccessible

We show that a Berk-Nash equilibrium does indeed capture steady state behavior provided that the agent is sufficiently impatient. If the agent is patient, however, our equilibrium concept only captures those steady states that satisfy a property that we call exhaustive learning. This property says that the agent perceives that she has nothing else to learn in steady state. This property is satisfied, for example, if we are interested in behavior that is robust to a small amount of exogenous experimentation.

Steady states that do not satisfy exhausted learning, however, cannot generally be characterized by an equilibrium approach with fixed beliefs. For such cases, the modeler is forced to consider the more complicated problem where the agent's belief is part of the state variable. This is a feature of the dynamic environment that is not present in the static case, and it informs us of the limitations of using an equilibrium approach to study behavior in dynamic environments.

References

Aliprantis, C.D. and K.C. Border, *Infinite dimensional analysis: a hitchhiker's guide*, Springer Verlag, 2006.

Arrow, K. and J. Green, "Notes on Expectations Equilibria in Bayesian Settings," *Institute for Mathematical Studies in the Social Sciences Working Paper No. 33*, 1973.

Battigalli, P., *Comportamento razionale ed equilibrio nei giochi e nelle situazioni sociali*, Universita Bocconi, Milano, 1987.

Berk, R.H., "Limiting behavior of posterior distributions when the model is incorrect," *The Annals of Mathematical Statistics*, 1966, *37* (1), 51–58.

Brock, W. A. and L. J. Mirman, "Optimal economic growth and uncertainty: the discounted case," *Journal of Economic Theory*, 1972, *4* (3), 479–513.

Bunke, O. and X. Milhaud, "Asymptotic behavior of Bayes estimates under possibly incorrect models," *The Annals of Statistics*, 1998, *26* (2), 617–644.

Burdett, K. and T. Vishwanath, "Declining reservation wages and learning," *The Review of Economic Studies*, 1988, *55* (4), 655–665.

Dekel, E., D. Fudenberg, and D.K. Levine, "Learning to play Bayesian games," *Games and Economic Behavior*, 2004, *46* (2), 282–303.

Diaconis, P. and D. Freedman, "On the consistency of Bayes estimates," *The Annals of Statistics*, 1986, pp. 1–26.

Doraszelski, U. and J. F. Escobar, "A theory of regular Markov perfect equilibria in dynamic stochastic games: Genericity, stability, and purification," *Theoretical Economics*, 2010, *5* (3), 369–402.

Easley, D. and N.M. Kiefer, "Controlling a stochastic process with unknown parameters," *Econometrica*, 1988, pp. 1045–1064.

Esponda, I., "Behavioral equilibrium in economies with adverse selection," *The American Economic Review*, 2008, *98* (4), 1269–1291.

_ **and D. Pouzo**, "Conditional retrospective voting in large elections," *forthcoming in American Economic Journal: Microeconomics*, 2012.

Esponda, Ignacio and Demian Pouzo, "Berk–Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models," *Econometrica*, 2016, *84* (3), 1093–1130.

Evans, G. W. and S. Honkapohja, *Learning and Expectations in Macroeconomics*, Princeton University Press, 2001.

Eyster, E. and M. Piccione, "An approach to asset-pricing under incomplete and diverse perceptions," *Econometrica*, 2013, *81* (4), 1483–1506.

_ **and M. Rabin**, "Cursed equilibrium," *Econometrica*, 2005, *73* (5), 1623–1672.

Fershtman, C. and A. Pakes, "Dynamic games with asymmetric information: A framework for empirical work," *The Quarterly Journal of Economics*, 2012, p. qjs025.

Freedman, D.A., "On the asymptotic behavior of Bayes' estimates in the discrete case," *The Annals of Mathematical Statistics*, 1963, *34* (4), 1386–1403.

Freixas, X., "Optimal growth with experimentation," *Journal of Economic Theory*, 1981, *24* (2), 296–309.

Fudenberg, D. and D. Kreps, "Learning Mixed Equilibria," *Games and Economic Behavior*, 1993, *5*, 320–367.

_ **and D.K. Levine**, "Self-confirming equilibrium," *Econometrica*, 1993, pp. 523–545.

_ **and** _ , *The theory of learning in games*, Vol. 2, The MIT press, 1998.

_ **, G. Romanyuk, and P. Strack**, "Active Learning with Misspecified Beliefs," *Working Paper*, 2016.

Hall, R. E., "Macroeconomic fluctuations and the allocation of time," Technical Report 1 1997.

Hansen, L.P. and T.J. Sargent, *Robustness*, Princeton Univ Pr, 2008.

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

Heidhues, P., B. Koszegi, and P. Strack, "Unrealistic Expectations and Misguided Learning," *Working Paper*, 2016.

Hofbauer, J. and W.H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, 2002, *70* (6), 2265–2294.

Jehiel, P., "Limited horizon forecast in repeated alternate games," *Journal of Economic Theory*, 1995, *67* (2), 497–519.

_ , "Learning to play limited forecast equilibria," *Games and Economic Behavior*, 1998, *22* (2), 274–298.

_ , “Analogy-based expectation equilibrium,” *Journal of Economic theory*, 2005, *123* (2), 81–104.

_ **and D. Samet**, “Valuation equilibrium,” *Theoretical Economics*, 2007, *2* (2), 163–185.

_ **and F. Koessler**, “Revisiting games of incomplete information with analogy-based expectations,” *Games and Economic Behavior*, 2008, *62* (2), 533–557.

Kagel, J.H. and D. Levin, “The winner’s curse and public information in common value auctions,” *The American Economic Review*, 1986, pp. 894–920.

Kirman, A. P., “Learning by firms about demand conditions,” in R. H. Day and T. Groves, eds., *Adaptive economic models*, Academic Press 1975, pp. 137–156.

Koulovatianos, C., L. J. Mirman, and M. Santugini, “Optimal growth and uncertainty: learning,” *Journal of Economic Theory*, 2009, *144* (1), 280–295.

McCall, J. J., “Economics of information and job search,” *The Quarterly Journal of Economics*, 1970, pp. 113–126.

McLennan, A., “Price dispersion and incomplete learning in the long run,” *Journal of Economic Dynamics and Control*, 1984, *7* (3), 331–347.

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

Piccione, M. and A. Rubinstein, “Modeling the economic interaction of agents with diverse abilities to recognize equilibrium patterns,” *Journal of the European economic association*, 2003, *1* (1), 212–223.

Rothschild, M., “Searching for the lowest price when the distribution of prices is unknown,” *Journal of Political Economy*, 1974, *82* (4), 689–711.

_ , “A two-armed bandit theory of market pricing,” *Journal of Economic Theory*, 1974, *9* (2), 185–202.

Sargent, T. J., *The Conquest of American Inflation*, Princeton University Press, 1999.

Selten, R., "Reexamination of the perfectness concept for equilibrium points in extensive games," *International journal of game theory*, 1975, *4* (1), 25–55.

Shalizi, C. R., "Dynamics of Bayesian updating with dependent data and misspecified models," *Electronic Journal of Statistics*, 2009, *3*, 1039–1074.

Sobel, J., "Non-linear prices and price-taking behavior," *Journal of Economic Behavior & Organization*, 1984, *5* (3), 387–396.

Spiegler, R., "Placebo reforms," *The American Economic Review*, 2013, *103* (4), 1490–1506.

_ , "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics*, 2016a, *131*, 1243–â1290.

_ , "On the "Limited Feedback" Foundation of Boundedly Rational Expectations," *Working Paper*, 2016b.

Appendix

Proof of Lemma 2. *$I_Q(\sigma)$ is nonempty*: $M_{\sigma,Q}$ is a linear (hence continuous) self-map on a convex and compact subset of an Euclidean space (the set of probability distributions over the finite set $Gr(\Gamma)$); hence, Brower's theorem implies existence of a fixed point.

$I_Q(\sigma)$ is convex valued: For all $\alpha \in [0,1]$ and $m_1, m_2 \in \Delta(Gr(\Gamma))$, $\alpha M_{\sigma,Q}[m_1]+(1-\alpha)M_{\sigma,Q}[m_2] = M_{\sigma,Q}[\alpha m_1 + (1-\alpha)m_2]$. Thus, if $m_1 = M_{\sigma,Q}[m_1]$ and $m_2 = M_{\sigma,Q}[m_2]$, then $\alpha m_1 + (1-\alpha)m_2 = M_{\sigma,Q}[\alpha m_1 + (1-\alpha)m_2]$.

$I_Q(\sigma)$ is upper hemicontinuous and compact valued: Fix any sequence $(\sigma_n, m_n)_n$ in $\Sigma \times \Delta(Gr(\Gamma))$ such that $\lim_{n\to\infty}(\sigma_n, m_n) = (\sigma, m)$ and such that $m_n \in I_Q(\sigma_n)$ for all n. Since $M_{\sigma_n,Q}[m_n] = m_n$, $||m - M_{\sigma,Q}[m]|| \leq ||m - m_n|| + ||M_{\sigma_n,Q}[m_n - m]|| + ||M_{\sigma_n,Q}[m] - M_{\sigma,Q}[m]||$. The first term in the RHS vanishes by the hypothesis. The second term satisfies $||M_{\sigma_n,Q}[m_n - m]|| \leq ||M_{\sigma_n,Q}|| \times ||m_n - m||$ and also vanishes.[31] For the third term, note that $\sigma \mapsto M_{\sigma,Q}[m]$ is a linear mapping and $\sup_\sigma ||M_{\sigma,Q}[m]|| \leq \max_{s'} |\sum_{(s,x)\in Gr(\Gamma)} Q(s' \mid s, x)m(s,x)| < \infty$. Thus $||M_{\sigma_n,Q}[m] - M_{\sigma,Q}[m]|| \leq K \times ||\sigma_n -$

[31] For a matrix A, $||A||$ is understood as the operator norm.

$\sigma||$ for some $K < \infty$, and so it also vanishes. Therefore, $m = M_{\sigma,Q}[m]$; thus, $I_Q(\cdot)$ has a closed graph and so $I_Q(\sigma)$ is a closed set. Compactness of $I_Q(\sigma)$ follows from compactness of $\Delta(Gr(\Gamma))$. Therefore, $I_Q(\cdot)$ is upper hemicontinuous (see Aliprantis and Border (2006), Theorem 17.11). □

The proof of Lemma 3 relies on the following claim. The proofs of Claims A, B, and C in this appendix appear in the Online Appendix .

Claim A. **(i)** *For any regular SMDP, there exists* $\theta^* \in \Theta$ *and* $K < \infty$ *such that, for all* $m \in \Delta(Gr(\Gamma))$, $K_Q(m, \theta^*) \leq K$. ***(ii)*** *Fix any* $\theta \in \Theta$ *and a sequence* $(m_n)_n$ *in* $\Delta(Gr(\Gamma))$ *such that* $Q_\theta(s' \mid s, x) > 0$ *for all* $(s', s, x) \in \mathbb{S} \times Gr(\Gamma)$ *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 semicontinuous: 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* $\liminf_{n\to\infty} K_Q(m_n, \theta_n) \geq K_Q(m, \theta)$.

Proof of Lemma 3. (i) By Jensen's inequality and strict concavity of $\ln(\cdot)$, $K_Q(m, \theta) \geq -\sum_{(s,x)\in Gr(\Gamma)} \ln(E_{Q(\cdot|s,x)}[\frac{Q_\theta(S'|s,x)}{Q(S'|s,x)}])m(s, x) = 0$, 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) $\Theta_Q(m)$ *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.

$\Theta_Q(\cdot)$ *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 $\Theta(\cdot)$ has a closed graph and, by compactness of Θ, it is uhc). Suppose, in order 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 that* $Q(s' \mid s, x) > 0$. We will show that there is an element of the sequence, $\hat{\theta}_J$, that "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 $\left|K_Q(m, \hat{\theta}_J) - K_Q(m, \hat{\theta})\right| \leq \varepsilon/2$. Moreover, Claim A(ii) applied to $\theta = \hat{\theta}_J$ implies that there exists $N_{\varepsilon,J}$ such that, for all $n \geq N_{\varepsilon,J}$, $\left|K_Q(m_n, \hat{\theta}_J) - K_Q(m, \hat{\theta}_J)\right| \leq \varepsilon/2$. Thus, for all $n \geq N_{\varepsilon,J}$, $\left|K_Q(m_n, \hat{\theta}_J) - K_Q(m, \hat{\theta})\right| \leq \left|K_Q(m_n, \hat{\theta}_J) - K_Q(m, \hat{\theta}_J)\right| + \left|K_Q(m, \hat{\theta}_J) - K_Q(m, \hat{\theta})\right| \leq \varepsilon$ and, therefore,

$$K_Q(m_n, \hat{\theta}_J) \leq K_Q(m, \hat{\theta}) + \varepsilon \leq K_Q(m, \theta) - 2\varepsilon. \tag{21}$$

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 (21), 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). □

Proof of Theorem 1. Let $\mathbb{W} = \Sigma \times \Delta(Gr(\Gamma)) \times \Delta(\Theta)$ and endow it with the product topology (given by the Euclidean one for $\Sigma \times \Delta(Gr(\Gamma))$ and the weak topology for $\Delta(\Theta)$). Clearly, $\mathbb{W} \neq \{\emptyset\}$. Since Θ is compact, $\Delta(\Theta)$ is compact under the weak topology; Σ and $\Delta(Gr(\Gamma))$ are also compact. Thus by Tychonoff's theorem (see Aliprantis and Border (2006)), $\mathbb{W}$ is compact under the product topology. $\mathbb{W}$ is also convex. Finally, $\mathbb{W} \subseteq \mathbb{M}^2 \times rca(\Theta)$ where $\mathbb{M}$ is the space of $|\mathbb{S}| \times |\mathbb{X}|$ real-valued matrices and $rca(\Theta)$ is the space of regular Borel signed measures endowed with the weak topology. The space $\mathbb{M}^2 \times rca(\Theta)$ is locally convex with a family of seminorms $\{(\sigma, m, \mu) \mapsto p_f(\sigma, m, \mu) = ||(\sigma, m)|| + |\int_\Omega f(x)\mu(dx)| \colon \ f \in \mathbb{C}(\Omega)\}$ ($\mathbb{C}(\Omega)$ is the space of real-valued continuous and bounded functions and $||.||$ is understood as the spectral norm). Also, we observe that $(\sigma, m, \mu) = 0$ iff $p_f(\sigma, m, \mu) = 0$ for all $f \in \mathbb{C}(\Omega)$, thus $\mathbb{M}^2 \times rca(\Theta)$ is also Hausdorff.

Let $\mathcal{T} : \mathbb{W} \to 2^{\mathbb{W}}$ be such that $\mathcal{T}(\sigma, m, \mu) = \Sigma(\bar{Q}_\mu) \times I_Q(\sigma) \times \Delta(\Theta_Q(m))$. Note that if (σ^*, m^*, μ^*) is a fixed point of $\mathcal{T}$, then m^* is a Berk-Nash equilibrium. By Lemma 1, $\Sigma(\cdot)$ is nonempty, convex valued, compact valued, and upper hemicontinuous. Thus, for every $\mu \in \Delta(\Theta)$, $\Sigma(\bar{Q}_\mu)$ is nonempty, convex valued, and compact valued. Also, because Q_θ is continuous in θ (by regularity assumption), then $\bar{Q}_\mu$ is continuous (under the weak topology) in μ. Since $Q \mapsto \Sigma(Q)$ is upper hemicontinuous, then $\Sigma(\bar{Q}_\mu)$ is also upper hemicontinuous as a function of μ. By Lemma 2, $I_Q(\cdot)$ is nonempty, convex valued, compact valued and upper hemicontinuous. By Lemma 3 and the regularity condition, the correspondence $\Theta_Q(\cdot)$ is nonempty, compact valued, and upper hemicontinuous; hence, the correspondence $\Delta(\Theta_Q(\cdot))$ is nonempty, upper hemicontinuous (see Aliprantis and Border (2006), Theorem 17.13), compact valued (see Aliprantis and Border (2006), Theorem 15.11) and, trivially, convex valued. Thus, the correspondence $\mathcal{T}$ is nonempty, convex valued, compact valued (by Tychonoff's Theorem), and upper hemicontinuous (see Aliprantis and Border (2006), Theorem 17.28) under the product topology; hence, it has a closed graph (see Aliprantis and Border (2006), Theorem 17.11). Since $\mathbb{W}$ is a nonempty compact convex subset of a locally Hausdorff space, then there exists a fixed point of $\mathcal{T}$ by the Kakutani-Fan-Glicksberg theorem (see Aliprantis and Border (2006), Corollary 17.55). □

For the proof of Lemma 5, we rely on the following definitions and Claim. Let $K^*(m) = \inf_{\theta\in\Theta} K_Q(m,\theta)$ 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 $(s',s,x) \in \mathbb{S}\times Gr(\Gamma)$ such that $Q(s' \mid s,x) > 0$. Existence of such a set $\hat{\Theta}$ follows from the regularity assumption.

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

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

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

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

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

Proof of Lemma 5. 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 $\mathcal{H}$. Let $K^*(m) \equiv K_Q(m,\Theta_Q(m))$. 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 condition 3 of Definition 6, i.e., regularity). We now show that $\mu_0(\hat{\Theta}_\eta(m)) > 0$. By Lemma 3, $\Theta_Q(m)$ 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 non-empty 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 3 and an application of the Theorem of the Maximum). Compactness of $\{\Theta\colon d_m(\theta)\geq\epsilon\}$ and lower semi-continuity of $\theta\mapsto K_Q(m,\theta)$ (see Claim A(iii)) imply that $\inf_{\Theta:\, d_m(\theta)\geq\epsilon} K_Q(m,\theta) = \min_{\Theta:\, d_m(\theta)\geq\epsilon} K_Q(m,\theta) > K^*(m)$. Let $\alpha \equiv (\min_{\Theta:\, d_m(\theta)\geq\epsilon} K_Q(m,\theta) - K^*(m))/3 > 0$. Also, let $\eta > 0$ be chosen such that $K_Q(m,\theta) \leq K^*(m) + 0.25\alpha$ 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 B hold; note that $\mathbf{P}^f\left(\mathcal{H} \setminus \mathcal{H}_1\right) = 0$. Henceforth, fix $h^\infty \in \mathcal{H}_1$; 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 $C > 0$,

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

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

$$\limsup_{t\to\infty}\left\{\exp\left\{t\left(K^*(m)+0.5\alpha\right)\right\}A_t(\epsilon)\right\} = 0 \tag{22}$$

and

$$\liminf_{t\to\infty}\left\{\exp\left\{t\left(K^*(m)+0.5\alpha\right)\right\}B_t(\eta)\right\} = \infty. \tag{23}$$

Regarding equation (22), we first show that

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

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

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

Therefore,

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

Regarding equation (23), by Fatou's lemma and some algebra it suffices to show that

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

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

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

By Claim B(i),

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

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

Proof of Theorem 2. For any $s \in \mathbb{S}$ and $\mu \in \Delta(\Theta)$, let

$$\begin{aligned}
x(s,\mu) &\equiv \arg\max_{x\in\Gamma(s)} E_{\bar{Q}_\mu(\cdot|s,x)}\left[\pi(s,x,S')\right]\\
\tilde{\delta}(s,\mu) &\equiv \min_{x\in\Gamma(s)\backslash x(s,\mu)}\Big\{\max_{x\in\Gamma(s)} E_{\bar{Q}_\mu(\cdot|s,x)}\left[\pi(s,x,S')\right] - E_{\bar{Q}_\mu(\cdot|s,x)}\left[\pi(s,x,S')\right]\Big\}\\
\hat{\delta} &\equiv \max\big\{\min_{s,\mu}\tilde{\delta}(s,\mu),0\big\}\\
\bar{\delta} &\equiv \max\big\{\delta\geq 0 \mid \hat{\delta} - 2\frac{\delta}{1-\delta}M \geq 0\big\} = \frac{\hat{\delta}/M}{2+\hat{\delta}/M},
\end{aligned}$$

where $M \equiv \max_{(s,x)\in Gr(\Gamma),s\in\mathbb{S}'}|\pi(s,x,s')|$.

By Lemma 5, for all open sets $U \supseteq \Theta_Q(m)$, $\lim_{t\to\infty}\mu_t(U) = 1$ a.s.-$\mathbf{P}^f$ in $\mathcal{H}$. Also Let $g_\tau(h^\infty)(s,x) = \mathbf{1}_{(s,x)}(s_\tau,x_\tau) - M_{\sigma_\tau}(s,x \mid s_{\tau-1},x_{\tau-1})$ for any τ and $(s,x) \in Gr(\Gamma)$ and $h^\infty \in \mathbb{H}$. The sequence $(g_\tau)_\tau$ is a Martingale difference and by analogous arguments to those in the proof of Claim B: $\lim_{t\to\infty}||t^{-1}\sum_{\tau=0}^{t} g_\tau(h^\infty)|| = 0$ a.s.-$\mathbf{P}^f$. Let $\mathcal{H}^*$ be the set in $\mathcal{H}$ such that for all $h^\infty \in \mathcal{H}^*$ the following holds: for all open sets $U \supseteq \Theta_Q(m)$, $\lim_{t\to\infty}\mu_t(U) = 1$ and $\lim_{t\to\infty}||t^{-1}\sum_{\tau=0}^{t} g_\tau(h^\infty)|| = 0$. Note that $\mathbf{P}^f(\mathcal{H}\setminus\mathcal{H}^*) = 0$. Henceforth, fix an $h^\infty \in \mathcal{H}^*$, which we omit from the notation.

We first establish that $m \in I_Q(\sigma)$. Note that

$$\|m - M_{\sigma,Q}[m]\| \leq \|m - m_t\| + \|m_t - M_{\sigma,Q}[m]\|$$

where $(s,x) \mapsto M_{\sigma,Q}[p](s,x) \equiv \sum_{\tilde{s},\tilde{x} \in Gr(\Gamma)} M_\sigma(s,x|\tilde{s},\tilde{x})p(\tilde{s},\tilde{x})$ for any $p \in \Delta(Gr(\Gamma))$. By stability, the first term in the RHS vanishes, so it suffices to show that $\lim_{t\to\infty} ||m_t - M_{\sigma,Q}[m]|| = 0$. The fact that $\lim_{t\to\infty} ||t^{-1}\sum_{\tau=0}^{t} g_\tau|| = 0$ and the triangle inequality imply

$$\begin{aligned}
\lim_{t\to\infty}\big\|m_t - M_{\sigma,Q}[m]\big\| \leq & \lim_{t\to\infty}\big\|m_t - t^{-1}\sum_{\tau=1}^{t} M_{\sigma_\tau,Q}(\cdot,\cdot \mid s_{\tau-1},x_{\tau-1})\big\| \\
& + \lim_{t\to\infty}\big\|t^{-1}\sum_{\tau=1}^{t} M_{\sigma_\tau,Q}(\cdot,\cdot \mid s_{\tau-1},x_{\tau-1}) - M_{\sigma,Q}[m]\big\| \\
= & \lim_{t\to\infty}\big\|t^{-1}\sum_{\tau=1}^{t} g_\tau\big\| + \lim_{t\to\infty}\big\|t^{-1}\sum_{\tau=1}^{t} M_{\sigma_\tau,Q}(\cdot,\cdot \mid s_{\tau-1},x_{\tau-1}) - M_{\sigma,Q}[m]\big\| \\
\leq & \lim_{t\to\infty}\big\|t^{-1}\sum_{\tau=1}^{t} M_{\sigma_\tau,Q}(\cdot,\cdot \mid s_{\tau-1},x_{\tau-1}) - M_{\sigma,Q}\big[t^{-1}\sum_{\tau=1}^{t}\mathbf{1}_{(\cdot,\cdot)}(s_{\tau-1},x_{\tau-1})\big]\big\| \\
& + \lim_{t\to\infty}\big\|M_{\sigma,Q}\big[t^{-1}\sum_{\tau=1}^{t}\mathbf{1}_{(\cdot,\cdot)}(s_{\tau-1},x_{\tau-1})\big] - M_{\sigma,Q}[m]\big\|. \qquad (24)
\end{aligned}$$

Moreover, by definition of $M_{\sigma,Q}$ (see equation (4)), for all $(s,x) \in Gr(\Gamma)$,

$$t^{-1}\sum_{\tau=1}^{t} M_{\sigma_\tau,Q}(s,x \mid s_{\tau-1},x_{\tau-1}) = \sum_{\tilde{s},\tilde{x}\in Gr(\Gamma)} Q(s|\tilde{s},\tilde{x})t^{-1}\sum_{\tau=1}^{t}\sigma_\tau(x|s)\mathbf{1}_{(\tilde{s},\tilde{x})}(s_{\tau-1},x_{\tau-1}) \tag{25}$$

$$M_{\sigma,Q}\big[t^{-1}\sum_{\tau=1}^{t}\mathbf{1}_{(\cdot,\cdot)}(s_{\tau-1},x_{\tau-1})\big] = \sum_{\tilde{s},\tilde{x}\in Gr(\Gamma)} Q(s \mid \tilde{s},\tilde{x})t^{-1}\sum_{\tau=1}^{t}\sigma(x \mid s)\mathbf{1}_{(\tilde{s},\tilde{x})}(s_{\tau-1},x_{\tau-1}). \tag{26}$$

Equations (25) and (26) and stability ($\sigma_t \to \sigma$) imply that the first term in the RHS of 24 vanishes. The second term in the RHS also vanishes under stability due to continuity of the operator $M_\sigma[.]$ and the fact that $t^{-1}\sum_{\tau=1}^{t}\mathbf{1}_{(\cdot,\cdot)}(s_{\tau-1},x_{\tau-1}) = \frac{t-1}{t}m_{t-1}(\cdot,\cdot)$. Thus, $\|m - M_{\sigma,Q}[m]\| = 0$, and so $m \in I_Q(\sigma)$.

Therefore, for proving cases (i) and (ii), we need to establish that, for each case, there exists $\mu \in \Delta(\Theta_Q(m))$ such that σ is an optimal strategy for the MDP$(\bar{Q}_\mu)$.

(i) Consider any $\delta \in [0, \bar{\delta}]$. Since $\Delta(\Theta)$ is compact under the weak topology, there exists a subsequence of $(\mu_t)_t$ — which we still denote as $(\mu_t)_t$ — such that $\mu_t \overset{w}{\to} \mu_\infty$ and $\mu_\infty \in \Delta(\Theta_Q(m))$. Since $\sigma_t \in \bar{\Sigma}(\mu_t)$ for all t and $\bar{\Sigma}$ is uhc (see Lemma 4), stability $(\sigma_t \to \sigma)$ implies $\sigma \in \bar{\Sigma}(\mu_\infty)$. We conclude by showing that σ is an optimal strategy for the MDP$(\bar{Q}_{\mu_\infty})$. If $\delta = \bar{\delta} = 0$, this assertion is trivial. If $\bar{\delta} \geq \delta > 0$, it suffices to show that

$$\begin{aligned} x(s, \mu_\infty) &= \arg\max_{x \in \Gamma(s)} \int_{\mathbb{S}} \{\pi(s, x, s') + \delta W(s', B(s, x, s', \mu_\infty))\} \, \bar{Q}_{\mu_\infty}(ds'|s, x) \\ &= \arg\max_{x \in \Gamma(s)} \int_{\mathbb{S}} \{\pi(s, x, s') + \delta W(s', \mu_\infty)\} \, \bar{Q}_{\mu_\infty}(ds'|s, x). \end{aligned} \tag{27}$$

We conclude by establishing (27). Note that, since $\bar{\delta} > 0$, it follows that $\hat{\delta} > 0$, which in turn implies that $x(s, \mu_\infty)$ is a singleton. The first equality in (27) holds because, by definition of $\bar{\delta}$,

$$E_{\bar{Q}_{\mu_\infty}(\cdot|s, x(s,\mu_\infty))}\left[\pi(s, x(s, \mu_\infty), S')\right] - E_{\bar{Q}_{\mu_\infty}(\cdot|s,x)}\left[\pi(s, x, S')\right] \geq \hat{\delta} \geq 2\frac{\delta M}{1-\delta} > 0$$

for all $x \in \Gamma(s) \backslash \{x(s, \mu_\infty)\}$, and, by definition of M,

$$2\frac{\delta M}{1-\delta} \geq \delta \int_{\mathbb{S}} \Big\{ W(s', B(s, x, s', \mu_\infty)) \bar{Q}_{\mu_\infty}(ds'|s, x) - W(s', B(s, x(s, \mu_\infty), s', \mu_\infty)) \bar{Q}_{\mu_\infty}(ds'|s, x(s, \mu_\infty)) \Big\}.$$

The second equality in (27) holds by similar arguments.

(ii) By stability with exhaustive learning, there exists a subsequence $(\mu_{t(j)})_j$ such that $\mu_{t(j)} \overset{w}{\to} \mu^*$. This fact and the fact that for all open $U \supseteq \Theta_Q(m)$, $\lim_{t\to\infty} \mu_{t(j)}(U) = 1$, imply that $\mu^* \in \Delta(\Theta_Q(m))$. Since $\sigma_{t(j)} \in \bar{\Sigma}(\mu_{t(j)})$ for all j and $\bar{\Sigma}$ is uhc (see Lemma 4), stability $(\sigma_t \to \sigma)$ implies $\sigma \in \bar{\Sigma}(\mu^*)$. Moreover, by condition of stability with exhaustive learning (i.e., $\mu^* = B(s, x, s', \mu^*)$ for all $(s, x) \in Gr(\Gamma)$ and $s' \in supp(\bar{Q}_{\mu^*}(\cdot|s, x))$), $W(s, \mu^*) = \max_{x \in \Gamma(s)} \int_{\mathbb{S}} \{\pi(s, x, s') + \delta W(s', \mu^*)\} \, \bar{Q}_{\mu^*}(ds'|s, x)$ for all $s \in \mathbb{S}$. Then, by uniqueness of the value function, σ is an optimal strategy for the MDP$(\bar{Q}_{\mu_*})$. □

The proof of Proposition 2 relies on the following claim.

Claim C. If $(\sigma, m) \in \Sigma \times \Delta(\mathbb{S} \times \mathbb{X})$ is such that $\sigma \in \Sigma^\varepsilon$ and $m \in I_Q(m)$ with Q satisfying the full communication condition in Definition 18, then $m(s, x) > 0$ for all

$(s,x) \in Gr(\Gamma)$.

Proof of Proposition 2. (i) We show that, if (σ, m) is stable for a Bayesian-SMDP that is ε-perturbed, weakly identified and satisfies full communication (and has a prior μ_0 and policy function f), then (σ, m) is stable with exhaustive learning. That is, we must find a subsequence $(\mu_{t(j)})_j$ such that $\mu_{t(j)}$ converges weakly to μ^* and $\mu^* = B(s, x, s', \mu^*)$ for any $(s,x) \in Gr(\Gamma)$ and $s' \in supp(\bar{Q}_{\mu^*}(\cdot \mid s,x))$. By compactness of $\Delta(\Theta)$, there always exists a convergent subsequence with limit point $\mu^* \in \Delta(\Theta)$. By Lemma 5, $\mu^* \in \Delta(\Theta_Q(m))$. By assumption, $\sigma \in \Sigma^\varepsilon$ and, by the arguments given in the proof of Theorem 2, $m \in I_Q(\sigma)$. Since the SMDP satisfies full-communication, by Claim C, $supp(m) = Gr(\Gamma)$. This result, the fact that $\mu^* \in \Delta(\Theta_Q(m))$, and weak identification imply strong identification, i.e., for any θ_1 and θ_2 in the support of μ^*, $Q_{\theta_1}(\cdot \mid s,x) = Q_{\theta_2}(\cdot \mid s,x)$ for all $(s,x) \in Gr(\Gamma)$. Hence, it follows that, for all $A \subseteq \Theta$ Borel and for all $(s,x) \in Gr(\Gamma)$ and $s' \in \mathbb{S}$ such that $\bar{Q}_{\mu^*}(s' \mid s,x) > 0$ (i.e., $\int_\Theta Q_\theta(s' \mid s,x)\mu^*(d\theta) > 0$),

$$B(s,x,s',\mu^*)(A) = \frac{\int_A Q_\theta(s' \mid s,x)\mu^*(d\theta)}{\int_\Theta Q_\theta(s' \mid s,x)\mu^*(d\theta)} = \mu^*(A).$$

Thus, μ^* satisfies the desired condition.

(ii) We prove that if (σ, m) is a Berk-Nash equilibrium, then it is also a Berk-Nash equilibrium with exhaustive learning. Let μ be the supporting equilibrium belief. By Claim C and weak identification, it follows that there is strong identification, and so for any θ_1 and θ_2 in the support of μ, $Q_{\theta_1}(\cdot \mid s,x) = Q_{\theta_2}(\cdot \mid s,x)$ for all $(s,x) \in Gr(\Gamma)$. It follows that, for all $A \subseteq \Theta$ Borel and for all $(s,x) \in Gr(\Gamma)$ and $s' \in \mathbb{S}$ such that $\bar{Q}_\mu(s' \mid s,x) > 0$ (i.e., $\int_\Theta Q_\theta(s' \mid s,x)\mu(d\theta) > 0$),

$$B(s,x,s',\mu)(A) = \frac{\int_A Q_\theta(s' \mid s,x)\mu(d\theta)}{\int_\Theta Q_\theta(s' \mid s,x)\mu(d\theta)} = \mu(A).$$

Thus, (σ, m) is a Berk-Nash equilibrium with exhaustive learning. □

Proof of Proposition 3. Suppose (σ, m) is a perfect Berk-Nash equilibrium and let $(\sigma^\varepsilon, m^\varepsilon, \mu^\varepsilon)_\varepsilon$ be the associated sequence of equilibria with exhausted learning such that $\lim_{\varepsilon \to 0}(\sigma^\varepsilon, m^\varepsilon) = (\sigma, m)$. By possibly going to a sub-sequence, let $\mu = \lim_{\varepsilon \to 0} \mu^\varepsilon$ (under the weak topology). By the upper hemicontinuity of the equilibrium correspondence $\mathcal{T}(\sigma, m, \mu) = \Sigma(\bar{Q}_\mu) \times I_Q(\sigma) \times \Delta(\Theta_Q(m))$ (see the proof of Theorem 1),

(σ, m) is a Berk-Nash equilibrium with supporting belief μ. We conclude by showing that (σ, m) is a Berk-Nash equilibrium with exhaustive learning.

For all $(s, x) \in Gr(\Gamma)$ and $s' \in supp\left(\bar{Q}_{\mu}(\cdot|s, x)\right)$, and for all $f : \Theta \to \mathbb{R}$ bounded and continuous, $\left|\int f(\theta)\mu(d\theta) - \int f(\theta)B(s, x, s', \mu)(d\theta)\right| \leq \left|\int f(\theta)\mu(d\theta) - \int f(\theta)\mu^{\varepsilon}(d\theta)\right| + \left|\int f(\theta)\mu^{\varepsilon}(d\theta) - \int f(\theta)B(s, x, s', \mu)(d\theta)\right|$. The first term in the RHS vanishes as $\varepsilon \to 0$ by definition of weak convergence. For the second term, note that, for sufficiently small ε, $s' \in supp\left(\bar{Q}_{\mu^{\varepsilon}}(\cdot|s, x)\right)$, and so, since $\mu^{\varepsilon} = B(s, x, s', \mu^{\varepsilon})$ for any $(s, x) \in Gr(\Gamma)$ and $s' \in supp\left(\bar{Q}_{\mu^{\varepsilon}}(\cdot|s, x)\right)$, we can replace $\int f(\theta)\mu^{\varepsilon}(d\theta)$ with $\int f(\theta)B(s, x, s', \mu^{\varepsilon})(d\theta)$. Thus, the second term vanishes by continuity of the Bayesian operator. Therefore, by a standard argument[32], $\mu(A) = B(s, x, s', \mu)(A)$ for all $A \subseteq \Theta$ Borel and all $(s, x) \in Gr(\Gamma)$ and $s' \in supp\left(\bar{Q}_{\mu}(\cdot|s, x)\right)$, which implies that (σ, m) is a Berk-Nash equilibrium with exhaustive learning.□

Proof of Theorem 3. Existence of a Berk-Nash equilibrium of an ε-perturbed environment, $(\sigma^{\varepsilon}, m^{\varepsilon})$, follows for all $\varepsilon \in (0, \bar{\varepsilon}]$, where $\bar{\varepsilon} = 1/(|\mathbb{X}| + 1)$, from the same arguments used to establish existence for the case $\varepsilon = 0$ (see Theorem 1). Weak identification, full communication and Proposition 2(ii) imply that there exists a sequence $(\sigma^{\varepsilon}, m^{\varepsilon})_{\varepsilon > 0}$ of Berk-Nash equilibrium with exhaustive learning. By compactness of $\Sigma \times \Delta(Gr(\Gamma))$, there is a convergent subsequence, which is a perfect Berk-Nash equilibrium by definition. □

[32] Suppose μ_1, μ_2 in $\Delta(\Theta)$ are such that $\left|\int f(\theta)\mu_1(d\theta) - \int f(\theta)\mu_2(d\theta)\right| = 0$ for any f bounded and continuous. Then, for any $F \subseteq \Theta$ closed, $\mu_1(F) - \mu_2(F) \leq E_{\mu_1}[f_F(\theta)] - \mu_2(F) = E_{\mu_2}[f_F(\theta)] - \mu_2(F)$, where f_F is any continuous and bounded and $f_F \geq 1_{\{F\}}$; we call the class of such functions C_F. Thus, $\mu_1(F) - \mu_2(F) \leq \inf_{f \in C_F} E_{\mu_2}[f(\theta)] - \mu_2(F) = 0$, where the equality follows from an application of the monotone convergence theorem. An analogous trick yields the reverse inequality and, therefore, $\mu_1(F) = \mu_2(F)$ for any $F \subseteq \Theta$ closed. Borel measures over Θ are inner regular (also known as tight; see Aliprantis and Border (2006), Ch. 12, Theorem 12.7). Thus, for any Borel set $A \subseteq \Theta$ and any $\epsilon > 0$, there exists a $F \subseteq A$ compact such that $\mu_i(A \setminus F) < \epsilon$ for all $i = 1, 2$. Therefore $\mu_1(A) - \mu_2(A) \leq \mu_1(A) - \mu_2(F) \leq \mu_1(F) - \mu_2(F) + \epsilon$. By our previous result, it follows that $\mu_1(A) - \mu_2(A) \leq \epsilon$. A similar trick yields the reverse inequality and, since ϵ is arbitrary, this implies that $\mu_1(A) = \mu_2(A)$ for all $A \subseteq \Theta$ Borel.