Learning to Optimize*

George W. Evans
University of Oregon
University of St. Andrews

Bruce McGough
University of Oregon

This Version: April 18, 2018

Abstract

We consider decision-making by boundedly-rational agents in dynamic stochastic environments. The behavioral primitive is anchored to the shadow price of the state vector. Our agent forecasts the value of an additional unit of the state tomorrow using estimated models of shadow prices and transition dynamics, and uses this forecast to choose her control today. The control decision, together with the agent's forecast of tomorrow's shadow price, are then used to update the perceived shadow price of today's states. By following this boundedly-optimal procedure the agent's decision rule converges over time to the optimal policy. Specifically, within standard linear-quadratic environments, we obtain general conditions for asymptotically optimal decision-making: agents learn to optimize. Our results carry over to closely related procedures based on value-function learning and Euler-equation learning. We provide examples of shadow-price learning, and show how it can be embedded in a market equilibrium setting.
JEL Classifications: E52; E31; D83; D84

Key Words: Learning, Optimization, Bellman Systems

1 Introduction

A central paradigm of modern macroeconomics is the need for micro-foundations. Macroeconomists construct their models by aggregating the behavior of individual agents who are assumed "rational" in two important ways: they form forecasts optimally; and, given these forecasts, they make choices by maximizing their objective. Together with simple market

*This paper has benefited from the many comments from participants at numerous seminars and conferences. We would specifically like to thank, without in any way implicating, Dean Corbae, Richard Dennis, Eric Leeper and Randy Wright, for a number of helpful suggestions. SES-1025011 is gratefully acknowledged.

structures, and sometimes institutional frictions, it is this notion of rationality that identifies a micro-founded model. While assuming rationality is at the heart of much economic theory, the implicit sophistication required of agents in the benchmark "rational expectations" equilibrium,[1] both as forecasters and as decision theorists, is substantial: they must be able to form expectations conditional on the true distributions of the endogenous variables in the economy; and they must be able to make choices – i.e. solve infinite horizon programming problems – given these expectations.

The criticism that the ability to make optimal forecasts requires an unrealistic level of sophistication has been leveled repeatedly; and, in response to this criticism, a literature on bounded rationality and adaptive learning has developed. Boundedly rational agents are not assumed to know the true distributions of the endogenous variables; instead, they have forecasting models that they use to form expectations. These agents update their forecasting models as new data become available, and through this updating process the dynamics of the associated economy can be explored. In particular, the asymptotic behavior of the economy can be analyzed, and if the economy converges in some natural sense to a rational expectations equilibrium, then we may conclude that agents in the economy are able to learn to forecast optimally.

In this way, the learning literature has provided a response to the criticism that rational expectations is unrealistic. Early work on least-squares (and, more generally, adaptive) learning in macroeconomics includes Bray (1982), Bray and Savin (1986) and Marcet and Sargent (1989); for a systematic treatment, see Evans and Honkapohja (2001). Convergence to rational expectations is not automatic and "expectational stability" conditions can be computed to determine local stability. Recent applications have emphasized the possibility of novel learning dynamics that may also arise in some models.

Increasingly, the adaptive learning approach has been applied to dynamic stochastic general equilibrium (DSGE) models by incorporating learning into a system of expectational difference equations obtained from linearizing conditions that capture optimizing behavior and market equilibrium. We will discuss this procedure later, but for now we emphasize that because the representative agents in these models typically live forever, they are being assumed to be optimal decision makers, solving difficult stochastic dynamic optimization problems, despite having bounded rationality as forecasters. We find this discontinuity in sophistication unsatisfactory as a model of individual agent decision-making. The difficulty that subjects have in making optimal decisions, given their forecasts, has lead experimental researchers to distinguish between "learning to forecast" and "learning to optimize" experiments.[2] For example, in recent experimental work, Bao, Duffy, and Hommes (2013) find that in a cobweb setting making optimal decisions is as difficult as making optimal forecasts.

To address this discontinuity we define the notion of *bounded optimality*. We imagine

[1] Seminal papers of the rational expectations approach include, e.g., Muth (1961), Lucas (1972) and Sargent (1973).

[2] This issue is discussed in Marimon and Sunder (1993), Marimon and Sunder (1994) and Hommes (2011). The distinction was also noted in Sargent (1993).

our agents facing a sequence of decision problems in an uncertain environment: not only is there uncertainty in that the environment is inherently stochastic, but also our agents do not fully understand the conditional distributions of the variables requiring forecasts. One option when modeling agent decisions in this type of environment is to assume that agents are Bayesian and that, given their priors, they are able to fully solve their dynamic programming problems. However, we feel this level of sophistication is extreme, and instead, we prefer to model our agents as relying on decidedly simpler behavior. Informally, we assume that each day our agents act as if they face a two-period optimization problem: they think of the first period as "today" and the second period as "the future," and use one-period-ahead forecasts of shadow prices to measure the trade-off between choices today and the impact of these choices on the future. We call our implementation of bounded optimality *shadow price learning* (SP-learning).

Our notion of bounded optimality is inexorably linked to bounded rationality: agents in our economy are not assumed to fully understand the conditional distributions of the economy's variables, or, in the context of an individual's optimization problem, the conditional distributions of the state variables. Instead, consistent with the adaptive learning literature, we provide our agents with forecasting models, which they re-estimate as new data become available. Our agents use these estimated models to make one-period forecasts, and then use these one-period forecasts to make decisions.

We find our learning mechanism appealing for a number of reasons: it requires only simple econometric modeling and thus is consistent with the learning literature; it assumes agents make only one-period-ahead forecasts instead of establishing priors over the distributions of all future endogenous variables; and it imposes only that agents make decisions based on these one-period-ahead forecasts, rather than requiring agents to solve a dynamic programming problem with parameter uncertainty. Finally, SP-learning postulates that, fundamentally, agents make decisions by facing suitable prices for their trade-offs. This is a hallmark of economics. The central question that we address is whether SP-learning can converge asymptotically to fully optimal decision making. This is the analog of the original question, posed in the adaptive learning literature, of whether least-squares learning can converge asymptotically to rational expectations. Our main result is that convergence to fully optimal decision-making can indeed be demonstrated in the context of the standard linear-quadratic setting for dynamic decision-making.

Although we focus on SP-learning, we also consider two alternative implementations of bounded optimality: value-function learning and Euler-equation learning. Under value-function learning agents estimate and update (a model) of the value function, and make decisions based on the implied shadow prices given by the derivative of the estimated value function. With Euler-equation learning agents bypass the value-function entirely and instead make decisions based on an estimated model of their own policy rule. We establish that our central convergence results extend to these alternative implementations.

Our paper is organized as follows. In Section 2 we provide an overview of alternative

approaches and introduce our technique. In Section 3 we investigate the agent's problem in a standard linear-quadratic framework. We show, under quite general conditions, that the policy rule employed by our boundedly optimal agent converges to the optimal policy rule: following our simple behavioral primitives, our agent learns to optimize. This is our central theoretical result, given as Theorem 4 in Section 3. Section 4 provides a general comparison of SP-learning with alternative implementations, including value-function learning and Euler-equation learning. Theorem 5 establishes the corresponding convergence result for value-function learning and Theorem 6 for Euler-equation learning. We note that while there are many applications of Euler-equation learning in the literature, our Theorem 6 is the first to establish its asymptotic optimality at the agent level in a general setting. Sections 5 and 6 illustrate the technique in two separate modeling environments: an LQ Robinson Crusoe economy and a model of investment under uncertainty. The Robinson Crusoe economy can be used to illustrate the difference between Euler-equation and SP learning. The investment application is used to show how SP learning can be applied in non-LQ as well as LQ settings, and how SP learning embeds naturally into market equilibrium, and hence general equilibrium, settings. Section 7 concludes.

2 Background and Motivation

Before turning to a systematic presentation of our results we first, in this Section, review the most closely related approaches available in the literature, and we then introduce and motivate our general methodology and discuss how it relates to the existing literature.

2.1 Agent-level learning and decision-making

We are, of course, not the first to address the issues outlined in the Introduction. A variety of agent-level learning and decision-making mechanisms, differing both in imposed sophistication and conditioning information, have been advanced. Here we briefly summarize these contributions, beginning with those that make the smallest departure from the benchmark rational expectations hypothesis.

Cogley and Sargent (2008) consider Bayesian decision making in a permanent-income model with risk aversion. In their set-up, income follows a two-state Markov process with unknown transition probabilities, which implies that standard dynamic programming techniques are not immediately applicable. A traditional bounded rationality approach is to embrace Kreps's' "anticipated utility" model, in which agents determine their program given their current estimates of the unknown parameters. Instead, Cogley and Sargent (2008) treat their agents as Bayesian econometricians, who use recursively updated sufficient statistics as part of an expanded state space to specify their programming problem's time-invariant transition law. In this way agents are able to compute the fully optimal decision rule. The authors find that the fully optimal solution in their set-up is only a marginal improvement

on the boundedly optimal procedure of Kreps. This is particularly interesting because to obtain their fully optimal solution Cogley and Sargent (2008) need to assume a finite planning horizon as well as a two-state Markov process for income, and even then, computation of the optimal decision rule requires a great deal of technical expertise.

The approach taken by Adam and Marcet (2011), like Cogley and Sargent (2008), requires that agents solve a dynamic programming problem given their beliefs. These beliefs take the form of a fully specified distribution over all potential future paths of those variables taken as external to the agents. This is somewhat more general than Cogley and Sargent (2008) in that the distribution may or may not involve parameters that need to be estimated. Adam and Marcet (2011) analyze a basic asset pricing model with heterogeneous agents, incomplete markets, linear utility and limit constraints on stock holding. Within this model, they define an "internally rational" expectations equilibrium (IREE) as characterized by a sequence of pricing functions mapping the fundamental shocks to prices, such that markets clear, given agents' beliefs and corresponding optimal behavior.

In the Adam and Marcet (2011) approach, agents may be viewed quite naturally as Bayesians, i.e., they may have forecasting models in mind with distributions over the models' parameters. In this sense agents are adaptive learners in a manner consistent with forming forecasts optimally against the implied conditional distributions obtained from a "well-defined system of subjective probability beliefs." An REE is an IREE in which agents' "internal" beliefs are consistent with the external "truth," that is, with the objective equilibrium distribution of prices. Since they require that, in equilibrium, the pricing function is a map from shocks to prices, it follows that agents must hold the belief that prices are functions only of the shocks – in this way, REE beliefs reflect a singularity: the joint distribution of prices and shocks is degenerate, placing weight only on the graph of the price function. Their particular set-up has one other notable feature, that the optimal decisions of each agent require only one-step ahead forecasts of prices and dividend. This would not generally hold for risk averse agents, as can be seen from the set-up of Cogley and Sargent (2008), in which a great deal of sophistication is required to solve for the optimal plans.

Using the "anticipated utility" framework, Preston (2005) develops an infinite-horizon (or long-horizon) approach, in which agents use past data to estimate a forecasting model; then, treating these estimated parameters as fixed, agents make time t decisions that are fully optimal. This decision-making is optimal in the sense that it incorporates the (perceived) lifetime budget constraint (LBC) and the transversality condition (TVC). However, in this approach agents ignore the knowledge that their estimated forecasting model will change over time. Applications of the approach include, for example, Eusepi and Preston (2010) and Evans, Honkapohja, and Mitra (2009). Long-horizon forecasts were also emphasized in Bullard and Duffy (1998).

A commonly used approach known as Euler equation learning, developed e.g. in Evans and Honkapohja (2006), takes the Euler equation of a representative agent as the behavioral primitive and assumes that agents make decisions based on the boundedly rational forecasts

required by the Euler equation.[3] As in the other approaches, agents use estimated forecast models, which they update over time, to form their expectations. In contrast to infinite-horizon learning, agents are behaving in a simple fashion, forecasting only one period in advance. Thus they focus on decisions on this margin and ignore their LBC and TVC. Despite these omissions, when Euler equation learning is stable the LBC and TVC will typically be satisfied.[4] Euler-equation learning is usually done in a linear framework. An application that retains the nonlinear features is Howitt and Özak (2014).[5]

Euler equation learning can be viewed as an agent-level justification for "reduced-form learning," which is widely used, especially in applied work.[6] Under the reduced-form implementation, one starts with the system of expectational difference equations obtained by linearizing and reducing the equilibrium equations implied by RE, and then replaces RE with subjective one-step ahead forecasts based on a suitable linear forecasting model updated over time using adaptive learning. This approach leads to a particularly simple stability analysis,[7] but often fails to make clear the explicit connection to agent-level decision making.

The above procedures all involve forecasting, and thus require an estimate of the transition dynamics of the economy. This estimation step can be avoided using an approach called Q-learning, developed originally by Watkins (1989) and Watkins and Dayan (1992). Under Q-learning, which is most often used in finite-state environments, an agent estimates the "quality values" associated with each state/action pair. One advantage of Q-learning is that it eliminates the need to form forecasts by updating quality measures ex-post. To pursue the details some notation will be helpful. Let $x \in X$ represent a state and $a \in A$ represent an action. The usual Bellman system has the form $V(x) = \max_{a \in A} (r(x,a) + \beta \sum P_{xy}(a) V(y))$, where r captures the instantaneous return and $P_{xy}(a)$ is the probability of moving from state x to state y given action a. The quality function $Q : X \times A \to \mathbb{R}$ is defined as

$$Q(x,a) = r(x,a) + \beta \sum P_{xy}(a) V(y).$$

Under Q-learning, given $Q_{t-1}(x,a)$, the estimate of the quality function at (the beginning of) time t, and given the state x at t, the agent chooses the action a with the highest quality, i.e. $a = \max_{a' \in A} Q_{t-1}(x,a')$. At the beginning of time $t+1$, the estimate of Q is updated recursively as follows:

$$Q_t(x,a) = Q_{t-1}(x,a) + \frac{1}{t} I \left(a = \max_{a' \in A} Q_{t-1}(x,a') \right) \left(r(x,a) + \beta \max_{b \in A} Q_{t-1}(y,b) \right),$$

[3] See also Honkapohja, Mitra, and Evans (2013)

[4] A finite-horizon extension of Euler-equation learning is developed in Branch, Evans, and McGough (2013).

[5] Howitt and Özak (2014) study boundedly optimal decision making in a non-linear consumption/savings model. Within a finite-state model, agents are assumed to use decision rules that are linear in wealth and updated so as to minimize the squared ex-post Euler equation error, i.e. the squared difference between marginal utility yesterday and discounted marginal utility today, accounting for growth. They find numerically that agents quickly learn to use rules that result in small welfare losses relative to the optimal decision rule.

[6] An early example is Bullard and Mitra (2002)

[7] See Chapter 10 of Evans and Honkapohja (2001).

where I is the indicator function and y is the state that is realized in $t+1$. Notice that Q_t does not require knowledge of the state's transition function. Provided the state and action spaces are finite, Watkins and Dayan (1992) show $Q_t \to Q$ almost surely under a key assumption, which requires in particular each state/action pair is visited infinitely many times. We note that this assumption is not easily generalized to the continuous state and action spaces that are standard in the macroeconomic literature.

A related approach to boundedly rational decision making uses classifier systems. An early well-known economic application is Marimon, McGrattan, and Sargent (1990). They introduce classifier system learning into the model of money and matching due to Kiyotaki and Wright (1989). They consider two types of classifier systems. In the first, there is a complete enumeration of all possible decision rules. This is possible in the Kiyotaki-Wright set-up because of the simplicity of that model. The second type of classifier system instead uses rules that do not necessarily distinguish each state, and which uses genetic algorithms to periodically prune rules and generate new ones. Using simulations Marimon et al. show that learning converges to a stationary Nash equilibrium in the Kiyotaki-Wright model, and that, when there are multiple equilibria, learning selects the fundamental low-cost solution.

Lettau and Uhlig (1999) incorporate rules of thumb into dynamic programming using classifier systems. In their "general dynamic decision problem" they consider agents maximizing expected discounted utility, where agents make decisions using rules of thumb (a mapping from a subset of states into the action space, giving a specified action for specified states within this subset). Each rule of thumb has an associated strength. Learning takes place via updating of strengths. At time t the classifier with highest strength among all applicable classifiers is selected and the corresponding action is undertaken. After the return is realized *and* the state in $t+1$ is (randomly) generated, the strength of the classifier used in t is updated (using a gain sequence) by the return plus β times the strength of the strongest applicable classifier in $t+1$. Lettau and Uhlig give a consumption decision example, with two rules of thumb, the optimal decision rule based on dynamic programming and another non-optimal rule of thumb, applicable only in high-income states, in which agents consume all their income. They showed that convergence to this suboptimal rule of thumb is possible.[8][9]

Our SP-learning framework shares various characteristics of the alternative implementations of agent-level learning discussed above. Like Q-learning and the related approaches based on classifier systems, SP-learning builds off of the intuition of the Bellman equation. (In fact, what we will call value-function learning explicitly establishes the connection.) As

[8] Lettau and Uhlig discuss the relationship of their decision rule to Q-learning in their footnote 11, p. 165: they state that (i) Q-learning also introduces action mechanisms that ensure enough exploration so that all (x, a) combinations are triggered infinitely often, and (ii) in Q-learning the value $Q(x, a)$ is assigned and updated for *every* state-action pair (x, a). This corresponds to classifiers that are only applicable in a single state. In general, classifiers are allowed to cover more general sets of state-action pairs.

[9] A recent related approach is sparse dynamic programing in which agents may choose to use summary variables rather than the complete state vector. See Gabaix (2014).

in infinite-horizon learning, we employ the anticipated utility approach rather than the more sophisticated Bayesian perspective. Like Euler-equation learning, it is sufficient for agents to look only one step ahead. While each of the alternative approaches has advantages, we find SP-learning persuasive in many applications due to its simplicity, generality and economic intuition.

2.2 Shadow-price learning

Returning to the current paper, our objective is to develop a general approach for boundedly rational decision-making in a dynamic stochastic environment. While particular examples would include the optimal consumption-savings problems summarized above, the technique is generally applicable and can be embedded in standard general equilibrium macro models. To illustrate our technique, consider a standard dynamic programming problem

$$\begin{aligned} V^*(x_0) &= \max E_0 \sum_{t\geq 0} \beta^t r(x_t, u_t) && (1) \\ \text{subject to } x_{t+1} &= g(x_t, u_t, \varepsilon_{t+1}) && (2) \end{aligned}$$

and $\bar{x}_0$ given. Here $u_t \in \Gamma(x_t) \subseteq \mathbb{R}^m$ is the vector of controls (with $\Gamma(x_t)$ compact), $x_t \in \mathbb{R}^n$ is the vector of (endogenous and exogenous) states variables, and ε_{t+1} is white noise. Our approach is based on the standard first-order conditions derived from the Lagrangian[10]

$$\mathcal{L} = E_0 \sum\nolimits_{t\geq 0} \beta^t \left(r(x_t, u_t) + \lambda_t^{*\prime} \left(g(x_{t-1}, u_{t-1}, \varepsilon_t) - x_t \right) \right), \text{ namely}$$

$$\begin{aligned} \lambda_t^* &= r_x(x_t, u_t)' + \beta E_t g_x(x_t, u_t, \varepsilon_{t+1})' \lambda_{t+1}^* && (3) \\ 0 &= r_u(x_t, u_t)' + \beta E_t g_u(x_t, u_t, \varepsilon_{t+1})' \lambda_{t+1}^*. && (4) \end{aligned}$$

Under the SP-learning approach we replace λ_t^* with λ_t, representing the perceived shadow price of the state, and we treat equations (3)-(4) as the basis of a *behavioral* decision rule.

To implement SP-learning (3)-(4) need to be supplemented with forecasting equations for the required expectations. In line with the adaptive learning literature, assume that the transition equation (2) is unknown, and must be estimated, and that agents do so by approximating the transition equation using a linear specification of the form[11]

$$x_{t+1} = Ax_t + Bu_t + C\varepsilon_{t+1},$$

and thus the agents approximate $g_x(x_t, u_t, \varepsilon_{t+1})$ by A and $g_u(x_t, u_t, \varepsilon_{t+1})$ by B. The coefficient matrices A, B are estimated and updated over time using recursive least squares (RLS). We

[10] For $t = 0$ the last term in the sum is replaced by $\lambda_0^{*\prime}(\bar{x}_0 - x_0)$, where $\bar{x}_0$ is the initial state vector.

[11] Here we have expanded the state vector x_t to include a constant.

also assume that agents believe the perceived shadow price λ_t is (or can be approximated by) a linear function of state, up to white noise, i.e.

$$\lambda_t = Hx_t + \mu_t,$$

where the matrix H also is estimated. Finally, we assume that agents know their preference function $r(x_t, u_t)$. Then, given the state x_t and estimates for A, B, H the decision procedure is obtained by solving the system

$$\begin{aligned} r_u(x_t, u_t)' &= -\beta B' \hat{E}_t \lambda_{t+1} \\ \hat{E}_t \lambda_{t+1} &= H\left(Ax_t + Bu_t\right) \end{aligned} \tag{5}$$

for u_t and the forecasted shadow price, $\hat{E}_t\lambda_{t+1}$. Here $\hat{E}_t$ denotes the conditional expectation of the agent based on his forecasting model. These values can then be used with (3) to obtain an updated estimate of the current shadow price

$$\lambda_t = r_x(x_t, u_t)' + \beta A' \hat{E}_t \lambda_{t+1}. \tag{6}$$

Finally, the data (x_t, u_t, λ_t) can be used to recursively update the parameter estimates (A, B, H) over time. Taken together this procedure defines a natural implementations of the SP-learning approach.

As we will see, under more specific assumptions, this implementation of boundedly optimal decision making leads to asymptotically optimal decisions. In this sense shadow-price learning is reasonable from an agent perspective. Our approach has a number of strengths. Particularly attractive, we think, is the pivotal role played by shadow prices. In economics prices are central because agents use them to assess trade-offs. Here the perceived shadow price of next period's state vector, together with the estimated transition dynamics, measures the intertemporal trade-offs and thereby determines the agent's choice of control vector today. The other feature that we find compelling is the simplicity of the required behavior: agents make decisions as if they face a two-period problem. In this way we eliminate the discontinuity between the sophistication of agents as forecasters and agents as decision-makers. In addition, SP-learning incorporates the RLS updating of parameters that is the hallmark of the adaptive learning approach. Finally, this version of bounded optimality is applicable to the general stochastic regulator problem, and can be embedded in market equilibrium models.

While we view SP-learning as a very natural implementation of bounded optimality, there are some closely related variations that also yield asymptotic optimality. In Section 6 of his seminal paper on asset pricing, discussing stability analysis, Lucas (1978) briefly outlines how agents might update over time their subjective value function. In Section 4.2 we show how to specify a real-time procedure for updating an agent's value function. Our Theorem 5 implies that this procedure converges asymptotically to the true value function of an optimizing agent. From Section 4.2 it can seen that another variation, Euler-equation learning, is in some cases equivalent to SP-learning. Indeed, Theorem 6 establishes the first

formal general convergence result for Euler-equation learning by establishing its connections to SP-learning.

Having found that SP-learning is reasonable from an agent's perspective, in that he can expect to eventually behave optimally, we embed shadow price learning into a simple economy consistent with our quadratic regulator environment. We consider a Robinson Crusoe economy, with quadratic preferences and linear technology: see Hansen and Sargent (2014) for many examples of these types of economies, including one of the examples we give. By including production lags we provide a simple example of a multivariate model in which SP-learning and Euler-equation learning differ. We use the Crusoe economy to walk carefully through the boundedly optimal behavior displayed by our agent, thus providing examples of, and intuition for the behavioral assumptions made in Section 3.1.

While our formal results are proved for the Linear-Quadratic framework, as we have stressed, the techniques and intuition can be applied in a general setting and to equilibrium market settings. To illustrate theses point we conclude with an application to a model of investment under uncertainty, first at the agent level, using both LQ and non-LQ environments, and then in a market equilibrium setting.

3 Learning to optimize

We begin be specifying the programming problem of interest. We focus on the behavior of a decision maker with a quadratic objective function and who faces a linear transition equation; the linear-quadratic (LQ) set-up allows us to exploit certainty equivalence and to conduct parametric analysis.[12] The specification of our LQ problem, which is standard, is taken from Hansen and Sargent (2014); see also Stokey and Lucas Jr. (1989), and Bertsekas (1987).

3.1 Linear quadratic dynamic programing

The "sequence problem" is to determine a sequence of controls u_t that solves

$$\begin{aligned} V^*(x_0) = \max \quad & -E_0 \sum \beta^t \left(x_t' R x_t + u_t' Q u_t + 2 x_t' W u_t \right) \\ s.t. \quad & x_{t+1} = A x_t + B u_t + C \varepsilon_{t+1}. \end{aligned} \tag{7}$$

Here Q is symmetric positive definite and $R - W'Q^{-1}W$ is symmetric positive semi-definite, which ensure that the period objective is concave. These conditions, as well as further restrictions on R, Q, W, A and B, will be discussed in detail below: see LQ.1-LQ.3. The initial condition x_0 is taken as given. As with the general dynamic programming problem (1) - (2), we assume $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$, with the matrices conformable. To allow for a

[12] The LQ set-up can also be used to approximate more general nonlinear environments.

constant in the objective and in the transition, we assume that $x_{1t} = 1$. It follows that first row of A is $(1, 0, \ldots, 0)$ and that the first row of B is a $1 \times m$ vector of zeros and the first row of C is a $1 \times k$ vector of zeros. We further assume that $\varepsilon_t \in \mathbb{R}^k$ is a zero-mean i.i.d. process with $E\varepsilon_t\varepsilon_t' = \sigma_\varepsilon^2 I$ and compact support. The assumptions on ε_t are convenient but can be relaxed considerably: for example, Theorem 2 only requires that ε_t be a martingale difference sequence with (finite) time-invariant second moments; and, Theorem 4 holds if the assumption of compact support is replaced by the existence of finite absolute moments.

The sequence problem is commonly analyzed by considering the associated Bellman functional equation. The Principle of Optimality states that the solution to the sequence problem V^* satisfies

$$V^*(x) = \max_u -\left(x'Rx + u'Qu + 2x'Wu\right) + \beta E\left(V^*(Ax + Bu + C\varepsilon)|x, u\right). \tag{8}$$

The Bellman system (8) may be analyzed using the Riccati equation

$$P = R + \beta A'PA - \left(\beta A'PB + W\right)\left(Q + \beta B'PB\right)^{-1}\left(\beta B'PA + W'\right). \tag{9}$$

Under certain conditions that will be discussed in detail below, this non-linear matrix equation has a unique, symmetric, positive semi-definite solution P^*.[13] The matrix P^*, interpreted as a quadratic form, identifies the solution to the Bellman system (8), and allows for the computation of the feedback matrix F^* that provides the sequence of controls solving the programming problem (7). Specifically, Theorem 2 states that

$$V^*(x) = -x'P^*x - \frac{\beta}{1-\beta}\text{tr}\left(\sigma_\varepsilon^2 P^* CC'\right) \tag{10}$$

$$F^* = \left(Q + \beta B'P^*B\right)^{-1}\left(\beta B'P^*A + W'\right), \tag{11}$$

where tr denotes the trace of a matrix, and where $u_t = -F^*x_t$ solves (7). Note that the optimal policy matrix F^* depends on the matrix P^*, but not on σ_ε^2 or C. This is an illustration of certainty equivalence in the LQ-framework: the optimal policy rule is the same in the deterministic and stochastic settings.

3.1.1 Perceptions and realizations: the T-map

Solving Bellman systems in general, and Riccati equations in particular, is often approached recursively: given an approximation V_n to the solution V^*, a new approximation, V_{n+1}, may be obtained using the right-hand-side of (8):

$$V_{n+1}(x) = \max_u -\left(x'Rx + u'Qu + 2x'Wu\right) + \beta E\left(V_n(Ax + Bu + C\varepsilon)|x, u\right).$$

[13] Solving the Riccati equation is not possible analytically; however, a variety of numerical methods are available.

This approach has particular appeal to us because it has the flavor a learning algorithm: given a *perceived* value function V_n we may compute the corresponding *induced* value function V_{n+1}. In this Section we work out the initial implications of this viewpoint.

We start with the deterministic case in which $C = 0$, thus shutting down the stochastic shocks. We imagine the decision-making behavior of a boundedly rational agent, who perceives that the value V of the state tomorrow (which here we denote $\tilde{x}$) is represented as a quadratic form: $V(\tilde{x}) = -\tilde{x}'P\tilde{x}$. To ensure that the agent's objective is concave, we assume that P is symmetric positive semi-definite. For convenience we will refer to P as the perceived value function. The agent chooses u to solve

$$V^P(x) = \max_u -(x'Rx + u'Qu + 2x'Wu) - \beta(Ax + Bu)'P(Ax + Bu), \tag{12}$$

where V^P is the value function induced by perceptions P. We note that the induced value function V^P is determined by combining the perceived value of the future state with the realized return given by the decisions based on those perceptions.

The following Lemma characterizes the agent's control decision and the induced value function.[14]

Lemma 1 *Consider the deterministic problem (12). If P is symmetric positive semi-definite then*

1. *The unique optimal control decision for perceptions P is given by $u = -F(P)x$, where*

$$F(P) = (Q + \beta B'PB)^{-1}(\beta B'PA + W').$$

2. *The induced value function for perceptions P is given by $V^P(x) = -x'T(P)x$, where*

$$T(P) = R + \beta A'PA - (\beta A'PB + W)(Q + \beta B'PB)^{-1}(\beta B'PA + W'). \tag{13}$$

We note that the right-hand-side of $T(P)$ is given by the right-hand-side of the Riccati equation. We conclude that the fixed point P^* of the T-map identifies the solution to our agent's optimal control problem. For general perceptions P the boundedly optimal control decision is given by $u = -F(P)x$.

Remark 1 *Since the first row of B is zero, it follows that $B'P$ does not depend on P_{11}. Hence the $(1,1)$ entry of perceptions P does not affect the control decision.*

Here and in the sequel it will sometimes be convenient to allow for more general perceptions P. To this end we define $\mathcal{U}$ to be the open set of all $n \times n$ matrices P for which

[14] See Appendix A for proofs of Lemmas and Theorems.

$\det(Q + \beta B'PB) \neq 0$. Since Q is positive definite, $\mathcal{U}$ is not empty. It follows that T is well-defined on $\mathcal{U}$.

The same contemplation may be considered in the stochastic case. Again, consider a boundedly rational agent who perceives that the value V of the state tomorrow is represented as a quadratic form: $V(\tilde{x}) = -\tilde{x}'P\tilde{x}$ for some symmetric positive semi-definite P. The agent now chooses u to solve

$$\begin{aligned} V_\varepsilon^P(x) &= \max_u [-(x'Rx + u'Qu + 2x'Wu) \\ &\quad -\beta E\left((Ax + Bu + C\varepsilon)'P(Ax + Bu + C\varepsilon)|x, u\right)]. \end{aligned} \tag{14}$$

In the stochastic case the result corresponding to Lemma 1 is the following.

Lemma 2 *Consider the stochastic problem (14). If P is symmetric positive semi-definite then*

1. *The optimal control decision for perceptions P is given by $u = -F_\varepsilon(P)x$, where*
$$F_\varepsilon(P) = (Q + \beta B'PB)^{-1}(\beta B'PA + W').$$
2. *The induced value function for perceptions P is given by $V_\varepsilon^P(x) = -x'T^\varepsilon(P)x$, where $T^\varepsilon(P) = T(P) - \beta\Delta(P)$ and $\Delta(P) = -tr\left(\sigma_\varepsilon^2 PCC'\right) \oplus 0_{n-1\times n-1}$.*

Furthermore, if $\tilde{P} \in \mathcal{U}$ and $T(\tilde{P}) = \tilde{P}$ then $T^\varepsilon(\tilde{P}_\varepsilon) = \tilde{P}_\varepsilon$, where

$$\tilde{P}_\varepsilon = \tilde{P} - \frac{\beta}{1-\beta}\Delta(\tilde{P}). \tag{15}$$

Here $\oplus$ denotes the direct sum of two matrices, i.e. for matrices M_1 and M_2 we define $M_1 \oplus M_2$ as the block-diagonal matrix

$$M_1 \oplus M_2 = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}.$$

Fully optimal decision-making is determined by the fixed point P_ε^* of the map T^ε. This fixed point is related to the solution P^* to the Riccati equation, and hence to the solution to the deterministic problem, by the following equation:

$$P_\varepsilon^* = P^* - \frac{\beta}{1-\beta}\Delta(P^*).$$

In this way, the solution to the non-stochastic problem yields the solution to the stochastic problem. We note that the map T^ε will be particularly useful when analyzing value-function learning in Section 4.2.1.

It is well-known that LQ problems exhibit certainty equivalence, i.e. the optimal control decision is independent of C. Certainty equivalence carries over to boundedly optimal decision making, but the manifestation is distinct:

- Under fully optimal decision making, $F_\varepsilon(P_\varepsilon^*) = F(P^*)$.
- Under boundedly optimal decision making, and given perceptions P, we have $F_\varepsilon(P) = F(P)$. In the sequel we will therefore use $F(P)$ for $F_\varepsilon(P)$ whenever convenient.

Note that in both cases the control decision given the state x is the same for the stochastic and deterministic problems.

3.1.2 The LQ assumptions

We now consider conditions sufficient to guarantee the sequence problem (7) has a unique solution, which, via the principle of optimality, guarantees that the Riccati equation has a unique positive semi-definite solution. We introduce the needed concepts informally first, and then turn to their precise definitions.

- **Concavity.** To apply the needed theory of dynamic programming, the instantaneous objective must be bounded above, which, due to its quadratic nature, requires concavity. Intuitively, the agent should not be able to attain infinite value.
- **Stabilizability.** To be representable as a quadratic form, V^* must not diverge to negative infinity. This condition is guaranteed by the ability to choose a bounded control sequence u_t so that the corresponding trajectory of the state is also bounded. Intuitively, the agent should be able to stabilize the state.
- **Detectability**. Stabilizability implies that avoiding unbounded paths is feasible, but does not imply that the agent will want to stabilize the state. A further condition, detectability, is needed: explosive paths should be "detected" by the objective. Specifically, if the instantaneous objective gets large (in magnitude) whenever the state does then a stabilized state trajectory is desirable. Intuitively, the agent should want to stabilize the state.

To make these notions precise, it is helpful to consider the non-stochastic problem, which we transform to eliminate the state-control interaction in the objective and discounting: see Hansen and Sargent (2014), Chapter 3 for the many details. To this end, first notice

$$\begin{aligned} & x'Rx + u'Qu + 2x'Wu \qquad (16)\\ = \;& x'Rx - x'WQ^{-1}W'x + x'WQ^{-1}QQ^{-1}W'x + u'Qu + u'QQ^{-1}W'x + x'WQ^{-1}Qu \\ = \;& x'\left(R - WQ^{-1}W'\right)x + (u + Q^{-1}W'x)'Q(u + Q^{-1}W'x) \\ = \;& x'\hat{R}x + (u + Q^{-1}W'x)'Q(u + Q^{-1}W'x), \end{aligned}$$

where $\hat{R} = R - WQ^{-1}W'$. Next, let

$$\hat{x}_t = \beta^{\frac{t}{2}} x_t \text{ and } \hat{u}_t = \beta^{\frac{t}{2}}(u_t + Q^{-1}W'x_t).$$

Then

$$\beta^t \left(x_t' R x_t + u_t' Q u_t + 2 x_t' W u_t \right) = \hat{x}_t' \hat{R} \hat{x}_t + \hat{u}_t' Q \hat{u}_t. \tag{17}$$

Finally, we compute

$$\begin{aligned} \hat{x}_{t+1} &= \beta^{\frac{t+1}{2}} x_{t+1} = \beta^{\frac{1}{2}} \left(A \beta^{\frac{t}{2}} x_t + B \beta^{\frac{t}{2}} u_t \right) \\ &= \beta^{\frac{1}{2}} \left(A \beta^{\frac{t}{2}} x_t + B (\hat{u}_t - Q^{-1} W' \beta^{\frac{t}{2}} x_t) \right) \\ &= \beta^{\frac{1}{2}} \left(A - B Q^{-1} W' \right) \hat{x}_t + \beta^{\frac{1}{2}} B \hat{u}_t = \hat{A} \hat{x}_t + \hat{B} \hat{u}_t, \end{aligned}$$

where the last equality defines notation. It follows that the non-stochastic version of the LQ problem (7) is equivalent to the transformed problem

$$\begin{aligned} \max \quad & -\sum \left(\hat{x}_t' \hat{R} \hat{x}_t + \hat{u}_t' Q \hat{u}_t \right) \\ s.t. \quad & \hat{x}_{t+1} = \hat{A} \hat{x}_t + \hat{B} \hat{u}_t, \end{aligned} \tag{18}$$

where

$$\begin{aligned} \hat{R} &= R - W Q^{-1} W' \\ \hat{A} &= \beta^{\frac{1}{2}} \left(A - B Q^{-1} W' \right) \\ \hat{B} &= \beta^{\frac{1}{2}} B. \end{aligned}$$

The T-map of the transformed problem is computed as before, and will be of considerable importance:

$$-\hat{x}' \hat{T}(P) \hat{x} = \max_u - \left(\hat{x}' \hat{R} \hat{x} + \hat{u}' Q \hat{u} \right) - (\hat{A} \hat{x} + \hat{B} \hat{u})' P (\hat{A} \hat{x} + \hat{B} \hat{u}), \tag{19}$$

where P is a symmetric positive semi-definite matrix representing the agent's perceived value function. Letting

$$\hat{F}(P) = \left(Q + \hat{B}' P \hat{B} \right)^{-1} \hat{B}' P \hat{A},$$

as shown in Lemma 1, the solution to the right-hand-side of (19) is given by $\hat{u} = -\hat{F}(P) \hat{x}$.

It will be useful to identify the state dynamics that would obtain if the perceptions P were to be held constant over time. To this end, let $\Omega(P) = \hat{A} - \hat{B} \hat{F}(P)$. It follows that the state dynamics $\hat{x}_t$ for transformed problem, and the state dynamics x_t for the original problem would be provided by the following equations, respectively:

$$\hat{x}_t = \Omega(P) \hat{x}_{t-1} \text{ and } x_t = \beta^{-1/2} \Omega(P) x_{t-1}. \tag{20}$$

These equations will be useful when we later study the decisions and evolution of the state x_t as perceptions P are updated over time. As shown in Lemma 3, the matrix $\Omega(P)$ also provides a very useful alternative representation of $T(P)$.

Lemma 3 *Let $P \in \mathcal{U}$. Then*

1. $T(P) = \hat{T}(P)$.
2. $\hat{T}(P) = \hat{R} + \hat{F}(P')'Q\hat{F}(P) + \Omega(P')'P\Omega(P)$.

Note that Lemma 3 holds for matrices that are not necessarily symmetric, positive semi-definite. However, we also note that the T-map preserves both symmetry and positive definiteness.

Because of item 1 of this Lemma, we will drop the hat on the T-map, even when explicitly considering the transformed problem. Also, in the sequel, while hatted matrices will correspond to the transformed problem, to reduce clutter, and because they refer to vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, whenever convenient we drop the hats from the states and controls.

We now turn to the formal conditions defining stabilizability and detectability. The latter is stated in terms of the rank decomposition of $\hat{R}$. Specifically, below in LQ.1 we assume that $\hat{R}$ is symmetric positive semidefinite. Thus, by the rank decomposition, $\hat{R}$ can be factored as $\hat{R} = \hat{D}\hat{D}'$, where $\text{rank}(\hat{R}) = r$ and $\hat{D}$ is $n \times r$.[15] With this notation, we say that:

- A matrix is stable if its eigenvalues have modulus less than one.
- The matrix pair $(\hat{A}, \hat{B})$ is *stabilizable* if there exists a matrix K such that $\hat{A} + \hat{B}K$ is stable.
- The matrix pair $(\hat{A}, \hat{D})$ is *detectable* provided that whenever y is a (nonzero) eigenvector of $\hat{A}$ associated with the eigenvalue μ and $\hat{D}'y = 0$ it follows that $|\mu| < 1$. Intuitively, $\hat{D}'$ acts as a factor of the objective function's quadratic form $\hat{R}$: if $\hat{D}'y = 0$ then y is not detected by the objective function; in this case, the associated eigenvalue must be contracting.

With these definitions in hand, we may formally state the assumptions we make concerning the matrices identifying the LQ problem.

LQ.1: The matrix $\hat{R}$ is symmetric positive semi-definite and the matrix Q is symmetric positive definite.

LQ.2: The system $(\hat{A}, \hat{B})$ is stabilizable.

LQ.3: The system $(\hat{A}, \hat{D})$ is detectable.

[15] Any positive semi-definite matrix X may be factored as $X = U\Lambda U'$, where U is a unitary matrix. The rank decomposition $X = DD'$ obtains by writing $\Lambda = \Lambda_1 \oplus 0$, with Λ_1 invertible, and letting $D = (U_{11}', U_{21}')'\sqrt{\Lambda_1}$.

This list provides the formal assumptions corresponding to the concepts of concavity, stabilizability, and detectability discussed informally above. By (17) LQ.1 imparts the appropriate concavity assumptions on the objective, and LQ.2 says that it is possible to find a set of controls driving the state to zero in the transformed problem. Finally, by LQ.3, $(\hat{A}, \hat{D})$ is detectable and the control path must be chosen to counter dynamics in the explosive eigenspaces of $\hat{A}$.[16] To illustrate, suppose z is an eigenvector of A with associated eigenvalue μ, suppose that $|\mu| > 1$, and finally assume that $x_0 = z$. If the control path is not chosen to mitigate the explosive dynamics in the eigenspace associated to μ then the state vector will diverge in norm. Furthermore, because $(\hat{A}, \hat{D})$ is detectable, we know that $\hat{D}'z \neq 0$. Taken together, these observations imply that an explosive state is suboptimal:

$$-x_t'\hat{R}x_t = -\mu^{2t}z'\hat{D}\hat{D}'z = -\left(|\mu|^t|\hat{D}'z|\right)^2 \to -\infty.$$

Hansen and Sargent (see Appendix A of Ch. 3 in Hansen and Sargent (2014)) put it more concisely (and eloquently): If $(\hat{A}, \hat{B})$ is stabilizable then it is feasible to stabilize the state vector; if $(\hat{A}, \hat{D})$ is detectable then it is desirable to stabilize the state vector.

The detectability of $(\hat{A}, \hat{D})$ plays another, less-obvious role in our analysis: it is needed for the stability at P^* of the following (soon-to-be-very-important!) matrix-valued differential equation:

$$\dot{P} = T(P) - P. \tag{21}$$

Here we view P as a function of a notional time variable τ and $\dot{P}$ denotes $dP/d\tau$. Under LQ.1-LQ.3 the stability of (21) at P^* is proved formally established using Theorem 1 below, but some intuition is available. For arbitrary state vector x, we may apply the envelope theorem to the maximization problem

$$-x'T(P)x = \max_u -\left(x'\hat{R}x + u'Qu + 2x'Wu\right) - \beta(Ax + Bu)'P(Ax + Bu)$$

to get

$$\begin{aligned} x'dTx &= \beta(Ax + Bu)'dP(Ax + Bu) = x'\Omega(P)'dP\Omega(P)x, \text{ or} & (22)\\ dT &= \Omega(P)'dP\Omega(P). & (23) \end{aligned}$$

Here the controls u in the middle expression of (22) are chosen optimally, the second equality of (22) follows from (20), and the equality (23) holds because x is arbitrary.

As is discussed in more detail in the next paragraph, the stability of the matrix system (21) turns on the Jacobian of its vectorization, which may be determined by applying the

[16] The rank decomposition of a matrix may not be unique (it is if the matrix is symmetric, positive definite). If $R = DD' = SS'$ comprises two distinct rank decompositions of a symmetric, positive semi-definite matrix R, and if (A, D) is detectable then (A, S) is also detectable. Indeed, if y is an eigenvalue of A and $S'y = 0$ then $Ry = 0$, so that $DD'y = 0$. Since D is $n \times r$ and of full rank, it follows that it acts injectively on the range of D'; therefore, $DD'y = 0$ implies $D'y = 0$, which, by the detectability of (A, D), means the eigenvalue associated to y must be contracting.

"vec" operator to (23).[17] Since

$$vec\left(\Omega(P)'dP\Omega(P)\right) = \left(\Omega(P)'\otimes\Omega(P)'\right)vec\left(dP\right),$$

and since the set of eigenvalues of $\Omega(P)'\otimes\Omega(P)'$ is the set of all products of the eigenvalues of $\Omega(P)$, it can be seen that P^* is a stable rest point of (21) whenever the eigenvalues of $\Omega(P^*)$ are smaller than one in modulus, that is, whenever $\Omega(P^*)$ is a stable matrix. This is precisely where detectability plays its central role. As discussed above, by LQ.3, an agent facing the transformed problem desires to "stabilize the state," that is, send $\hat{x}_t \to 0$. Also, by (20), the state dynamics in the transformed problem are given by $\hat{x}_{t+1} = \Omega(P^*)x_t$. It follows that $\Omega(P^*)$ must be a stable matrix.

Formal analysis of stability requires additional machinery. Before stating the principal result, we first secure the notation needed to compute derivatives when we have matrix-valued functions and matrix-valued differential equations. If $f:\mathbb{R}^p\to\mathbb{R}^q$ then $D(f)$ is the matrix of first partials, and for $x\in\mathbb{R}^p$, the notation $D(f)(x)$ emphasizes that the partials are evaluated at the vector x. Notice that D is an operator that acts on vector-valued functions – we do not apply D to matrix-valued functions. The analysis of matrix-valued differential equations is conducted by working through the vec operator. If $f:\mathbb{R}^{p\times p}\to\mathbb{R}^{q\times q}$ then we define $f_v:\mathbb{R}^{p^2}\to\mathbb{R}^{q^2}$ by $f_v = vec\circ f\circ vec^{-1}$, where the dimensions of the domain and range of the vec operators employed are understood to be determined by f. Thus suppose $f:\mathbb{R}^{p\times p}\to\mathbb{R}^{p\times p}$, assume $f(x^*)=0$, and consider the matrix-valued differential equation $\dot{x}=f(x)$, where $\dot{x}$ denotes the derivative with respect to time. Let $y=vec(x)$, and note that $\dot{y}=vec(\dot{x})$. Then

$$\begin{aligned}\dot{x}=f(x) \quad &\Longrightarrow \quad vec(\dot{x})=vec(f(x))\\ &\Longrightarrow \quad \dot{y}=\left(vec\circ f\circ vec^{-1}\right)(vec(x)) \implies \dot{y}=f_v(y).\end{aligned}$$

Hence if $y^*=vec(x^*)$ then Lyapunov stability of x^* may be assessed by determining the eigenvalues of $D(f_v)(y^*)$. As is well known, Lyapunov stability holds if these eigenvalues have negative real parts. A sufficient condition for this is that $D(f_v)(y^*)+I_{p^2}$ is a stable matrix.

The T-map and its fixed point P^* are central to our analysis. The relevant properties are summarized by the following theorem.

[17] The vec operator is the standard isomorphism coupling $\mathbb{R}^{n\times m}$ with $\mathbb{R}^{nm}$. Intuitively, the vectorization of a matrix Z is obtained by simply stacking its columns. More formally, let $Z\in\mathbb{R}^{n\times m}$. For each $1\le k\le nm$, use the division algorithm to uniquely write $k=jn+i$, for $0\le j\le m$ and $0\le i<n$. Then

$$vec(Z)_k=\begin{cases} Z_{i1} & \text{if} \quad j=0\\ Z_{nj} & \text{if} \quad i=0\\ Z_{i,j+1} & \text{else}\end{cases}.$$

Theorem 1 *Assume LQ.1 – LQ.3. There exists an $n \times n$, symmetric, positive semi-definite matrix P^* such that for any $n \times n$, symmetric, positive semi-definite matrix P_0, we have $T^m(P_0) \to P^*$ as $m \to \infty$. Further,*

1. $T(P^*) = P^*$.
2. $D(T_v)(vec(P^*))$ *is stable.*
3. P^* *is the unique fixed point of T among the class of $n \times n$, symmetric, positive semi-definite matrices.*

Corollary 1 $D(T_v^\varepsilon)(vec(P_\varepsilon^*))$ *is stable.*

The theorem and corollary are proven in the appendix.

3.1.3 The LQ solution

Theorem 1 concerns the T-map and its fixed point P^*. The connection between the T-map and the LQ-problem (7) is given by the Bellman equation. In fact, Theorem 1 can be used to prove:

Theorem 2 *Under assumptions LQ.1 – LQ.3, the Riccati equation (9) has a unique symmetric positive semi-definite solution, P^*, and iteration of the Riccati equation yields convergence to P^* if initialized at any positive semi-definite matrix P_0. The value function for the optimization problem (7) is given by*

$$V^*(x) = -x'P^*x - \frac{\beta}{1-\beta} tr\left(\sigma_\varepsilon^2 P^* CC'\right)$$

and $u = -F(P^*)x$ is the optimal policy rule.

Theorem 2 is, of course, not original to us. That LQ.1 – LQ.3 are sufficient to guarantee existence and uniqueness of a solution to (7) appears to be well known: see Bertsekas (1987) pp. 79-80 and Bertsekas and Shreve (1978), Chapters 7 and 9. We include the statement and key elements of the proof of Theorem 2 for completeness, and because these elements, together with Theorem 1 and Corollary 1, are foundational for our main results. Hansen and Sargent (2014), Ch. 3, discuss the stability results under the assumptions LQ.1 – LQ.3.[18]

Proving Theorem 2 involves applying the general theory of dynamic programming to the sequence problem and showing that analyzing the Bellman system corresponds to analyzing

[18] Other useful references are Anderson and Moore (1979), Stokey and Lucas Jr. (1989), Lancaster and Rodman (1995) and Kwakernaak and Sivan (1972). Alternative versions of Theorem 2 often use the somehwat stronger assumptions of controllability and observability.

the T-map. The challenge concerns the optimality of a linear policy rule: this optimality must be demonstrated by considering non-linear policy rules, which, in effect, eliminates the technical advantage of having a quadratic objective. The stochastic case is further complicated by issues of measurability: even if the perceived value function (which, when corresponding to a possibly non-linear policy rule, cannot be assumed quadratic) is Borel measurable, the induced value function may not be. Additional technical machinery involving the theories of universal measurability and lower-semi-analytic functions is required to navigate these nuances. In the Appendix we work through the deterministic case in detail. The stochastic case is then addressed, providing a road-map to the literature.

3.2 Bounded optimality: shadow-price learning

For an agent to solve the programming problem (7) as described above, he must understand the quadratic nature of his value function as captured by the matrix P^*, he must know the relationship of this matrix to the Riccati equation, he must be aware that iteration on the Riccati equation provides convergence to P^*, and finally, he must know how to deduce the optimal control path given P^*. Furthermore, this behavior is predicated upon the assumption that he knows the conditional means of the state variables, that is, he knows A and B.

We modify the primitives identifying agent behavior, first by imposing bounded rationality and then by assuming bounded optimality. It is natural to assume that while our agent knows the impact of his control decisions on the state, i.e. he knows B, the agent is not assumed to know the parameters of the state-contingent transition dynamics, i.e. he must estimate A.[19] Our agent is also not assumed to know how to solve his programming problem: he does not know Theorem 2. Instead, he uses a simple forecasting model to estimate the value of a unit of state tomorrow, and then he uses this forecast, together with his estimate of the transition equation, to determine his control today. Based on his control choice and his forecast of the value of a unit of state tomorrow, he revises the value of a unit of state today. This provides him new data to update his state-value forecasting model.

We develop our analysis of the agent's boundedly rational behavior in two stages. In the first stage, which we call "stylized learning," we avoid the technicalities introduced by the stochastic nature of data realization and forecast-model estimation; instead, we simply assume that the agent's beliefs evolve according to a system of differential equations that have a natural and intuitive appeal. In the second stage, we will then formally connect these equations to the asymptotic dynamics of the agent's beliefs under the assumption that he is recursively estimating and updating his forecasting models, in real time, and behaving accordingly.

[19] There may be circumstances in which the agent does not know B, and hence will need to estimate it as well. We discuss this briefly later.

3.2.1 Stylized shadow-price learning

To facilitate intuition for our learning mechanism, we reconsider the above problem using a Lagrange multiplier formulation. The Lagrangian is given by

$$\mathcal{L} = E_0 \sum_{t \geq 0} \beta^t \left(-x_t' R x_t - u_t' Q u_t - 2 x_t' W u_t + \lambda_t^{*\prime} (A x_{t-1} + B u_{t-1} + C \varepsilon_t - x_t) \right),$$

where again for $t = 0$ the last term in the sum is replaced by $\lambda_0^{*\prime}(\bar{x}_0 - x_0)$. As usual, λ_t^* may be interpreted as the shadow price of the state vector x_t along the optimal path. The first-order conditions provide

$$\begin{aligned} \mathcal{L}_{x_t} &= 0 \Rightarrow \lambda_t^{*\prime} = -2 x_t' R - 2 u_t' W' + \beta E_t \lambda_{t+1}^{*\prime} A \\ \mathcal{L}_{u_t} &= 0 \Rightarrow 0 = -2 u_t' Q - 2 x_t' W + \beta E_t \lambda_{t+1}^{*\prime} B. \end{aligned}$$

Transposing and combining with the transition equation yields the following dynamic system:

$$\lambda_t^* = -2 R x_t - 2 W u_t + \beta A' E_t \lambda_{t+1}^* \tag{24}$$

$$0 = -2 W' x_t - 2 Q u_t + \beta B' E_t \lambda_{t+1}^* \tag{25}$$

$$x_{t+1} = A x_t + B u_t + C \varepsilon_{t+1}. \tag{26}$$

This system, together with transversality, identifies the unique solution to (7). It also provides intuitive behavioral restrictions on which we base our notion of bounded optimality.

We now marry the assumption from the learning literature that agents make boundedly rational forecasts with a list of behavioral assumptions characterizing the decisions agents make given these forecasts; and, we do so in a manner that we feel imparts a level of sophistication consistent with bounded rationality. Much of the learning literature centers on equilibrium dynamics implied by one-step-ahead boundedly rational forecasts; we adopt and expand on this notion by developing assumptions consistent with the following intuition: agents make one-step-ahead forecasts and agents know how to solve a two-period optimization problem based on their forecasts. Formalizing this intuition, we make the following assumptions:

1. Agents know their individual instantaneous return function, that is, they know Q, R, and W;
2. Agents know the form of the transition law and estimate the coefficient matrix A;
3. Conditional on their perceived value of an additional unit of x tomorrow, agents know how to choose their control today;
4. Conditional on their perceived value of an additional unit of x tomorrow, agents know how to compute the value of an additional unit of x today.

Assumption one seems quite natural: if the agent is to make informed decisions about a certain vector of quantities u, he should at least be able to understand the direct impact of these decisions. Assumption two is standard in the learning literature: our agent needs to forecast the state vector, but is uncertain about its evolution; therefore, he specifies and estimates a forecasting model, which we take as having the same functional form as the linear transition equation, and forms forecasts accordingly. Denote by $\tilde{A}$ the agent's perception of A. As will be discussed below, under stylized learning, these perceptions are assumed to evolve over time according to a differential equation, whereas under real-time learning, the agent's perceptions are taken as estimates which he updates as new data become available.

Assumptions three and four require more explanation. Let λ_t be the agent's perceived shadow price of x_t *along the realized path of* x *and* u. One should not think of λ as identical to λ^*; indeed λ^* is the vector of shadow prices of x *along the optimal path of* x *and* u and the agent is not (necessarily) interested in this value. Let $\hat{E}_t\lambda_{t+1}$ be the agent's time t forecast of the time $t+1$ value of an additional increment of the state x. Assumption three says that given $\hat{E}_t\lambda_{t+1}$, the agent knows how to choose u_t, that is, he knows how to solve the associated two-period problem. And how is this choice made? The agent simply contemplates an incremental decrease du_i in u_i and equates marginal loss with marginal benefit. If r is the "rate function" $r(x,u) = -(x'Rx + u'Qu + 2x'Wu)$ then the marginal loss is $r_{u_i}du_i$. To compute the marginal gain, he must estimate the effect of du_i on the whole state vector tomorrow. This effect is determined by $B_i du_i$, where B_i is the i^{th}-column of B. To weigh this effect against the loss obtained in time t, he must then compute its inner product with the expected price vector, and discount. Thus

$$r_{u_i}du_i = \beta\hat{E}_t\left(\lambda_{it+1}\right)' B_i du_i.$$

Stacking, and imposing our linear-quadratic set-up, gives the bounded rationality equivalent to (25):

$$0 = -2W'x_t - 2Qu_t + \beta B'\hat{E}_t\lambda_{t+1}. \tag{27}$$

Equation (27) operationalizes assumption three.

To update their shadow-price forecasting model, the agent needs to determine the perceived shadow price λ_t. Assumption four says that given $\hat{E}_t\lambda_{t+1}$, the agent knows how to compute λ_t. And how is this price computed? The agent simply contemplates an incremental increase in x_{it} and evaluates the benefit. An additional unit of x_{it} affects the contemporaneous return and the conditional distribution of tomorrow's state; and the shadow price λ_t must encode both of these effects. Specifically, if r is the rate function then the benefit of dx_{it} is given by

$$\left(r_{x_i} + \beta\hat{E}_t\left(\lambda_{t+1}\right)'\tilde{A}_i\right)dx_{it} = \lambda_{it}dx_{it}$$

where the equality provides our definition of λ_{it}. Stacking, and imposing our linear quadratic set-up yields the bounded rationality equivalent to (24):

$$\lambda_t = -2Rx_t - 2Wu_t + \beta\tilde{A}'\hat{E}_t\lambda_{t+1}. \tag{28}$$

Equation (28) operationalizes assumption four.

Assumption three, as captured by (27), lies at the heart of bounded optimality: it provides that the agent makes one-step-ahead forecasts of shadow prices and makes decisions today based on those forecasted prices, just as he would if solving a two-period problem. Assumption four, as captured by (28), provides the mechanism by which the agent computes his revised evaluation of a unit of state at time t: the agent uses the forecast of prices at time $t+1$ and his control decision at time t to reassess the value of time t state; in this way, our boundedly optimal agent keeps track of his forecasting performance. Below, the agent uses λ_t to update his shadow-price forecasting model. We call boundedly optimal behavior, as captured by assumptions one through four, *shadow price learning*.

We now specify the shadow-price forecasting model, that is, the way our agent forms $\hat{E}_t\lambda_{t+1}$. Along the optimal path it is not difficult to show that $\lambda_t^* = -2P^*x_t$, and so it is natural to impose a forecasting model of this functional form. Therefore, we assume that at time t the agent believes that

$$\lambda_t = Hx_t + \mu_t \tag{29}$$

for some $n \times n$ matrix H (which we assume is near $-2P^*$) and some error term μ_t. Equation (29) has the feel of what is known in the learning literature as a perceived law of motion (PLM): the agent perceives that his shadow price exhibits a linear dependence on the state as captured by the matrix H. When engaged in real-time decision making, as considered in Section 3.2.2, our agent will revisit his belief H as new data become available. Under our stylized learning mechanism, H is taken to evolve according to a differential equation as discussed below.

We can now be precise about the agent's behavior. Given beliefs $\tilde{A}$ and H, expectations are formed using (29):

$$\hat{E}_t\lambda_{t+1} = H(\tilde{A}x_t + Bu_t). \tag{30}$$

Equations (27) and (30) jointly determine the agent's time t forecast $\hat{E}_t\lambda_{t+1}$ and time t control decision u_t.[20] Finally, (28) is used to determine the agent's perceived shadow price of the state λ_t^*.

It follows that the evolution of u_t and λ_t satisfy

$$\begin{aligned} u_t &= (2Q - \beta B'HB)^{-1}(\beta B'H\tilde{A} - 2W')x_t \\ &\equiv F^{SP}(H, \tilde{A}, B)x_t, \text{ and} \end{aligned} \tag{31}$$

$$\lambda_t = \left(-2R - 2WF^{SP}(H, \tilde{A}, B) + \beta\tilde{A}'H\left(\tilde{A} + BF^{SP}(H, \tilde{A}, B)\right)\right)x_t \tag{32}$$

$$\equiv T^{SP}(H, \tilde{A}, B)x_t, \tag{33}$$

[20] The control decision is determined by the first-order condition (27), which governs optimality provided suitable second-order conditions hold. These are satisfied if the perceptions matrix H is negative semi-definite, which we assume unless otherwise stated.

where the second lines of each equation define notation.[21] Here it is convenient to keep the explicit dependence of F^{SP} and T^{SP} on B as well as on the beliefs $(H, \tilde{A})$. The decision rule (31) is of course closely related to the decision rule given in Part 1 of Lemma 1, specifically:

$$F^{SP}(H, A, B) = -F(-H/2).$$

Note that it is not necessary to assume, nor do we assume, that our agent computes the map T^{SP}.

We now turn to stylized learning, which dictates how our agent's beliefs, as summarized by the collection $(H, \tilde{A})$, evolve over time. In order to draw comparisons and promote intuition for our learning model, it is useful first to succinctly summarize the corresponding notion from the macroeconomics adaptive learning literature. There, boundedly rational forecasters are typically assumed to form expectations using a forecasting model, or PLM, that represents the believed dependence of relevant variables (i.e. variables that require forecasting) on regressors; the actions these agents take then generate an implied dependence, or actual law of motion (ALM), and the map taking perceptions, say as summarized by a vector Θ, to the implied dynamics, is denoted with a $\tilde{T}$. This map captures the model's "expectational feedback" in that it measures how agents' perceptions of the relationship between the relevant variables and the regressors feeds back to the realized relationship.

A fixed point of $\tilde{T}$, say Θ^*, corresponds to a rational expectations equilibrium (REE): agents' perceptions then coincide with the true data-generating process, so that expectations are being formed optimally. Conversely, the discrepancy between perceptions and reality, as measured by $\tilde{T}(\Theta) - \Theta$, captures the direction and magnitude of the misspecification in agents' beliefs. Under a stylized learning mechanism, agents are assumed to modify their beliefs in response to this discrepancy according to the expectational stability (E-stability) equation $d\Theta/d\tau = \tilde{T}(\Theta) - \Theta$. Notice that Θ^* represents a fixed point of this differential equation. If this fixed point is Lyapunov stable then the corresponding REE is said to be E-stable, and stability indicates that an economy populated with stylized learners would eventually be in the REE.[22]

Returning to our environment in which a single agent makes boundedly optimal decisions, we observe that shadow-price learning is quite similar to the model of boundedly rational forecasting just discussed. Equation (29) has already been interpreted as a PLM, and equation (33) acts as an ALM and reflects the model's feedback: the agent's beliefs, choices and evaluations result in an actual linear dependence of his perceived shadow price on the state vector as captured by $T^{SP}(H, \tilde{A}, B)$. Notice that the actual dependence of x_{t+1} on x_t and

[21] Since Q is positive definite and P^* is positive semi-definite, it follows that $Q + \beta B' P^* B$ is invertible; thus $2Q - \beta B' H B$ is invertible for H near $-2P^*$.

[22] Besides having an intuitive appeal, stylized learning is closely connected to real-time learning through the E-stability Principle, which states that REE that are stable under the E-stability differential equation are (locally) stable under recursive least-squares or closely related adaptive learning rules. See Evans and Honkapohja (2001). Formally establishing that the E-stability Principle holds for a given model requires the theory of stochastic recursive algorithms, as discussed and employed in Section 3.2.2 below.

u_t is independent of the agent's perceptions and actions: there is no feedback along these beliefs' dimensions. For this reason, we assume for now that $\tilde{A} = A$, and, abusing notation, we suppress the corresponding dependency in the T-map: $T^{SP}(H) \equiv T^{SP}(H, A, B)$. We will return to this point when we consider real-time learning in Section 3.2.2.

Letting $H^* = -2P^*$, it follows that $T^{SP}(H^*) = H^*$. With beliefs H^*, our agent correctly anticipates the dependence of his perceived shadow price on the state vector; also, since

$$F(H^*, A, B) = -\left(Q + \beta B' P^* B\right)^{-1} \left(\beta B' P^* A + W'\right),$$

it follows from equation (11) that with beliefs H^*, our agent makes control choices optimally: a fixed point of the map T^{SP} corresponds to optimal beliefs and associated behaviors. Conversely, the discrepancy between the perceived and realized dependence of λ_t on x_t, as measured by $T^{SP}(H) - H$, captures the direction and magnitude of the misspecification in our agent's beliefs. Whereas in the literature on adaptive learning this discrepancy arises because the agent does not fully understand the conditional distributions of the economy's aggregate variables, here the discrepancy reflects our agent's limited sophistication: he does not fully understand his dynamic programming problem, and instead bases his decisions on his best measure of the trade-offs he faces, as reflected by his belief matrix H.

In Theorem 3 we embrace the concept of stylized learning presented above and assume our agent updates his beliefs according to the matrix-valued differential equation $dH/d\tau = T^{SP}(H) - H$. This system can be viewed as the bounded optimality counterpart of the E-stability equations used to study whether expectations updated by least-squares learning converge to RE. Note that H^* is a fixed point of this dynamic system. The following theorem together with Theorem 4, which demonstrates stability under real-time learning, constitute the core results of the paper.

Theorem 3 *If LQ.1 – LQ.3 are satisfied then H^* is a Lyapunov stable fixed point of $dH/d\tau = T^{SP}(H) - H$. That is, H^* is stable under stylized learning.*

The proof of Theorem 3, which is given in Appendix A, simply involves connecting the maps T and T^{SP}, and then using Theorem 1 to assess stability. While Theorem 3 provides a stylized learning environment, the main result of our paper, captured by Theorem 4, considers real-time learning. In essence Theorem 4 shows that the stability result of Theorem 3 carries over to the real-time learning environment.

3.2.2 Real-time shadow-price learning

To establish that stability under stylized learning carries over to stability under real-time learning, we now assume that our agent uses available data to estimate his forecasting model, and then uses this estimated model to form forecasts and make decisions, thereby generating

new data. The forecasting model may be written

$$\begin{aligned} x_{t+1} &= A_t x_t + B u_t + \text{error} \\ \lambda_t &= H_t x_t + \text{error}, \end{aligned}$$

where the coefficient matrices are time t estimates obtained using recursive least-squares (RLS).[23] Because we assume the agent knows B, to obtain the estimates A_t he regresses $x_t - Bu_{t-1}$ on x_{t-1}, using data $\{x_t, x_{t-1}, u_{t-1}, \ldots, x_0, u_0\}$. To estimate the shadow-price forecasting model at time t, and thus obtain the estimate H_t, we assume our agent regresses λ_{t-1} on x_{t-1} using data $\{x_{t-1}, \ldots, x_0, \lambda_{t-1}, \ldots, \lambda_0\}$.[24]

For the real-time learning results we require that the regressors x_t be non-explosive and not exhibit asymptotic perfect multicollinearity. The assumptions LQ.1 - LQ.3 do not address perfect multicollinearity and imply only that the state variables either do not diverge or that they diverge less rapidly than $\beta^{-1/2}$. We therefore now impose the additional assumptions:

LQ.RTL The eigenvalues of $A+BF(H^*, A, B)$ not corresponding to the constant term have modulus less than one, and the associated asymptotic second-moment matrix for the process $x_t = (A + BF(H^*, A, B))\, x_{t-1} + C\varepsilon_t$ is non-singular.

We note that under LQ.RTL fully optimal behavior leads to stationary state dynamics and precludes asymptotic perfect multicollinearity; however, LQ.RTL does not require that the dimension of ε_t be smaller than n. For an example, see Section 5. More explicit conditions that guarantee satisfaction of LQ.RTL are given in Appendix A. We note that asymptotic stationarity of regressors, as implied by LQ.RTL, is a standard assumption in the learning literature. Econometric analysis of explosive regressors involves considerable complexity, which would require a non-trivial extension of the treatment here.

We may describe the evolution of the estimate of A_t and H_t over time using RLS. The following dynamic system, written in recursive causal ordering, captures the evolution of

[23] An alternative to RLS is "constant gain" learning (CGL), which discounts older data. Under CGL asymptotic results along the lines of the following Theorem provide for weak convergence to a distribution centered on optimal behavior. See Ch. 7 of Evans and Honkapohja (2001) for some general results, and for applications and results concerning transition dynamics, see Williams (2014) and Cho, Williams, and Sargent (2002).

[24] As is standard in the learning literature, when analyzing real-time learning, the agent is assumed not to use current data on λ_t to form current estimates of H as this avoids technical difficulties with the recursive formulation of the estimators. See Marcet and Sargent (1989) for discussion and details.

agent behavior under bounded optimality:

$$\begin{aligned} x_t &= Ax_{t-1} + Bu_{t-1} + C\varepsilon_t \\ \mathcal{R}_t &= \mathcal{R}_{t-1} + \gamma_t \left(x_t x_t' - \mathcal{R}_{t-1}\right) \\ H_t' &= H_{t-1}' + \gamma_t \mathcal{R}_{t-1}^{-1} x_{t-1} \left(\lambda_{t-1} - H_{t-1} x_{t-1}\right)' \\ A_t' &= A_{t-1}' + \gamma_t \mathcal{R}_{t-1}^{-1} x_{t-1} \left(x_t - Bu_{t-1} - A_{t-1} x_{t-1}\right)' \\ u_t &= F^{SP}(H_t, A_t, B) x_t \\ \lambda_t &= T^{SP}(H_t, A_t, B) x_t \\ \gamma_t &= \kappa (t+N)^{-\vartheta}. \end{aligned} \tag{34}$$

Here γ_t is a standard specification of a decreasing "gain" sequence that measures the response of estimates to forecast errors. We assume that $0 < \vartheta, \kappa \leq 1$ and N is a non-negative integer. It is standard under LS learning to set $\vartheta = 1$.

Theorem 4 (Asymptotic Optimality of SP-learning) *If LQ.1 – LQ.3 and LQ.RTL are satisfied then, locally,*

$$\begin{aligned} (H_t, A_t) &\xrightarrow{a.s.} (H^*, A) \\ F^{SP}(H_t, A_t, B) &\xrightarrow{a.s.} -F^* \end{aligned}$$

when the recursive algorithm is augmented with a suitable projection facility.

See Appendix A for the proof, including a more careful statement of the Theorem, a construction of the relevant neighborhood, and a discussion of the "projection facility," which essentially prevents the estimates from wandering too far away from the fixed point.[25] A detailed discussion of real-time learning in general and projection facilities in particular is provided by Marcet and Sargent (1989), Evans and Honkapohja (1998) and Evans and Honkapohja (2001). We conclude that under quite general conditions, our simple notion of boundedly optimal behavior is asymptotically optimal, that is, shadow-price learners learn to optimize.

While Theorem 4 is a strong result, as it provides for almost sure convergence, one might wonder whether convergence is global or whether the projection facility is necessary. The theory of stochastic recursive algorithms does provide results in this direction. However, in our set-up dispensing with the projection facility is not generally possible. An unusual sequence of random shocks may push perceptions H and $\tilde{A}$ into regions that impart explosive dynamics to the state x_t. Numerical investigation suggests that our learning algorithm is remarkably

[25] Theorem 4 does not require the initial perception H_0 to be negative semi-definite, which would ensure that the agent's second-order condition holds. However, for γ_1 sufficiently small, if H_0 is negative semi-definite then H_t will be negative semi-definite for all $t \geq 1$.

robust. Not surprisingly we observe that stability without a projection facility is governed in large part by the maximum modulus of the eigenvalues of both the derivative of the T-map, $DT_H(H^*, A, B)$, and the matrix governing the state dynamics, $A + BF(H^*, A, B)$: the further these eigenvalues are below one the less likely a projection facility will be activated.

To illustrate these points, we examine several examples. First, consider the simple univariate case, with $A = 0, B = 1, R = 2, Q = 1$, and $W = 1$. A straightforward computation shows that $T(H) = 4(-1 + (2 - \beta H)^{-1})$. With $\beta = .95$, we find that $H^* \approx -3.2$, and that there is another, unstable fixed point $H^u \approx 1.31$. Figure 1 plots the T-map and the 45-degree line, and illustrates that, under the iteration dynamic $H_n = T(H_{n-1})$, the fixed point H^* is asymptotically obtained provided that $H_0 < H^u$. We note also that for the agent corresponding to this specification of the decision problem, more x_0 is welfare reducing, thus it is reasonable to assume that even boundedly optimal agents would hold initial beliefs satisfying $H_0 < 0 < H^u$. Finally, since behavior under the iteration dynamic is closely related to behavior under real-time learning, this result suggests that the dynamic (34) is likely to converge without a projection facility if agents hold reasonable initial beliefs, i.e. $H_0 < 0 < H^u$, and we find this to be borne out under repeated simulation.

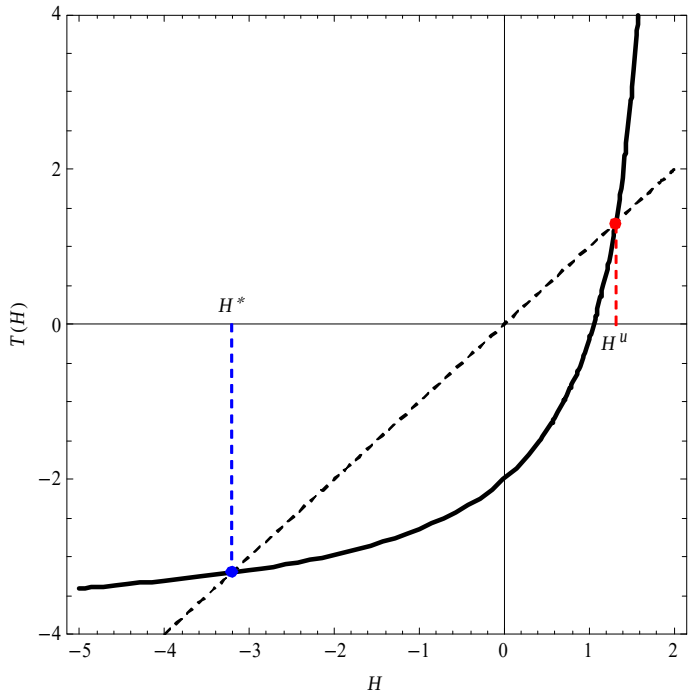

Figure 1: Univariate T-map

Next, we consider a two-dimensional example. A simple notation is helpful: if M is a square matrix then $\rho(M)$ is the spectral radius, that is, the maximum modulus of the

eigenvalues. In this example, the primitive matrices are chosen so as to promote instability:

$$\rho\left(DT_H(H^*, A, B)\right) = .9916 \text{ and } \rho\left(A + BF(H^*, A, B)\right) = .9966.$$

The first equality implies that the T-map does not correct errors quickly and the second equality means that along the optimal path the state dynamics are very nearly explosive. When conditions like these hold, even small shocks can place beliefs within a region resulting in explosive state dynamics, which, when coupled with an unresponsive T-map, leads to unstable learning paths. We find, using simulations, that even if the dynamic (34) is initialized at the optimum, 80% of the learning paths diverge within 300 periods.[26]

Finally, we consider another two-dimensional example. In this case, the primitive matrices are chosen to be less extreme than the previous example, but still have reasonably high spectral radii:

$$\rho\left(DT_H(H^*, A, B)\right) = .8387 \text{ and } \rho\left(A + BF(H^*, A, B)\right) = .8429.$$

Under this specification we find that the dynamic (34) is very stable. To examine this, we drop the projection facility and we adopt a constant gain formulation (setting $\kappa = 0.01$ and $\vartheta = 0$ so that $\gamma_t = 0.01$) so that estimates continue to fluctuate asymptotically, increasing the likelihood of instability. We ran 500 simulations for 2500 periods, each using randomly drawn initial conditions, and found that all 500 paths converged toward and stayed near the optimum.

Returning now to our general discussion, the principal and striking result of the adaptive learning literature is that boundedly rational agents, who update their forecasting models in natural ways, can learn to forecast optimally. Theorem 4 is complementary to this principal result, and equally striking: boundedly optimal decisions can converge asymptotically to fully optimal decisions. By estimating shadow prices, our agent has converted an infinite-horizon problem into a two-period optimization problem, which, given his beliefs, is comparatively straightforward to solve. The level of sophistication needed for boundedly optimal decision-making appears to be quite natural: our agent understands simple dynamic trade-offs, can solve simultaneous linear equations and can run simple regressions. Remarkably, with this level of sophistication, the agent can learn over his lifetime how to optimize based on a single realization of his decisions and the resulting states.

4 Extensions

We next consider several extensions of our approach. The first adapts our analysis to take explicit account of exogenous states. This will be particularly convenient for our later ap-

[26] Here we have employed a constant gain learning algorithm. Also, by diverge we mean that the paths escape a pre-defined neighborhood of the optimum.

plications. We then show how to modify our approach to cover value-function learning and Euler-equation learning.

4.1 Exogenous states

Some state variables are exogenous: their conditional distributions are unaffected by the control choices of the agent. To make boundedly optimal decisions, the agent must forecast future values of exogenous states, but it is not necessary that he track the corresponding shadow prices: there is no trade-off between the agent's choice and expected realizations of an exogenous state. In our work above, to make clear the connection between shadow-price learning and the Riccati equation, we have ignored the distinction between exogenous and endogenous states; however, in our examination of Euler equation learning in Section 4.2.2, and for the applications in Sections 5 and 6, it is helpful to leverage the simplicity afforded by conditioning only on endogenous states. In this subsection we show that the analogue to Theorem 3 holds when the agent restricts his shadow-price PLM to include only endogenous states as regressors.

Assume that the first n_1 entries of the of the state vector x are exogenous and that $n_2 = n - n_1$. For the remainder of this subsection, we use the notation

$$X = \left(\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right)$$

to emphasize the block decomposition of the $n \times n$ matrix X, with X_{11} being an $n_1 \times n_1$ matrix, and the remaining matrices conformable. Also, for $n \times m$ matrix X, we use

$$X = \left(\begin{array}{c} X_1 \\ \hline X_2 \end{array}\right), \text{ and } X' = \left(\begin{array}{c|c} X_1' & X_2' \end{array}\right) \tag{35}$$

with X_1 an $n_1 \times m$ matrix and X_2 conformable. With this notation, we may write the matrices capturing the transition dynamics as

$$A = \left(\begin{array}{c|c} A_{11} & 0 \\ \hline A_{21} & A_{22} \end{array}\right) \text{ and } B = \left(\begin{array}{c} 0 \\ \hline B_2 \end{array}\right),$$

where the zeros represent conformable matrices with zeros in all entries, and thus capture the exogeneity of x_{1t}.[27] Since it is always possible to reorder the components of the states, we will assume this ordering whenever exogenous states are present.

Endogenous-state SP learners are assumed to know the structure of A and B, i.e. to know which states are exogenous; and we assume they use this knowledge when making

[27] We assume that the exogenous states are either stationary or unit root: the eigenvalues of A_{11} lie within the closed unit disk.

their control decisions, and updating perceived shadow prices, based on the perceived trade-offs between today and tomorrow. These trade-offs are reflected in the following marginal conditions, which we take as behavioral primitives for these agents:

$$\begin{aligned} 0 &= -2W'\check{x}_t - 2Q\check{u}_t + \beta B_2'\hat{E}_t\check{\lambda}_{2t+1} \\ \check{\lambda}_{2t} &= -2R\check{x}_t - 2W\check{u}_t + \beta\tilde{A}_{22}'\hat{E}_t\check{\lambda}_{2t+1}, \end{aligned}$$

where $\tilde{A}$ is the perceived transition matrices, which are assumed to satisfy $\tilde{A}_{12} = 0$ and $B_1 = 0$. Here the notation $\check{u}_t, \check{x}_t, \check{\lambda}_{2t}$ is used to indicate the control decisions, state realizations and perceived shadow prices of the endogenous SP learner. Note that these primitives are precisely the equations that would be obtained from (27)-(28) under the exogeneity restrictions assumed for the transition dynamics.

For simplicity we focus on stylized learning in which the agent knows the transition dynamics A and B. To make his control decisions, he forecasts the future values of the endogenous states' shadow prices, which we denote by $\check{\lambda}_{2t}$, using the PLM

$$\check{\lambda}_{2t} = \check{H}\check{x}_t + \text{noise}. \tag{36}$$

Notice that the price of an additional unit of endogenous state may depend on the realization of an exogenous state; hence, we condition $\check{\lambda}_{2t}$ on the entire state vector $\check{x}_t$.

We now compare the behavior of an endogenous state SP learner with that of a "full-state" SP learner of Section 3.2, who uses the PLM $\lambda_t = Hx_t +$ noise. Since under the structural assumptions on A and B

$$B'\hat{E}_t\lambda_{t+1} = B_2'\hat{E}_t\lambda_{2t+1} \text{ and } (A'\hat{E}_t\lambda_{t+1})_2 = A_{22}'\hat{E}_t\lambda_{2t+1},$$

where the notation $(\star)_2$ identifies the last n_2 rows of the matrix $(\star)$, it follows that $u_t = \check{u}_t, x_t = \check{x}_t$ and $\lambda_{2t} = \check{\lambda}_{2t}$ whenever $H_2 = \check{H}$. Thus Theorem 3 applies: by only forecasting endogenous states our agent learns to optimize. Furthermore, it is straightforward to generalize this framework and result to allow for real-time estimation of the transition dynamics and the shadow-prices PLM.

In fact it is clear that this result is more general. Continuing to assume that all agents know the transition dynamics A and B, two agents forecasting distinct sets of shadow prices will make the same control decisions, provided that both sets include all endogenous states and the agents' forecasting models of the endogenous shadow prices coincide. In particular, it is natural to assume that agents do not forecast the shadow price of the constant term and when convenient we will make this assumption.

SP learners who forecast some or all exogenous states will additionally asymptotically obtain the shadow prices of these exogenous states, but these shadow prices are not needed or used for decision-making. Also, SP learners of any these types, whether they forecast all, some or none of the exogenous states, will learn to optimize even if they do not know which states are exogenous.

The results of this Section are useful for applications in which particular states are exogenous and in which it is natural to assume that agents understand and impose knowledge of this exogeneity. Because our stability results are unaffected by whether the agent forecasts exogenous shadow prices, in the sequel we will frequently assume that agents make use of this knowledge.

4.2 Bounded optimality: alternative implementations

Although shadow-price learning provides an appealing way to implement bounded rationality, our discussion in the Introduction suggests that there may be alternatives. In this section we show that our basic approach can easily be extended to encompass two closely related alternatives: value-function learning and Euler-equation learning. Shadow-price learning focuses attention on the marginal value of the state, as this is the information required to chose controls, and under SP-learning the agent estimates the shadow prices directly. In contrast, value-function learning infers the shadow prices from an estimate of the value function itself. A second alternative, when the endogenous states exhibit no lagged dependence, is for the agent to estimate shadow prices simply using marginal returns r_x, leading to Euler-equation learning. We consider these two alternatives in turn.

4.2.1 Value-function learning

Our implementation of value-function learning leverages the intuition developed in Section 3.1.1: given a perceived value function V, represented by the symmetric positive semidefinite matrix P, i.e. $V(x) = -x'Px$, the agent chooses control $u = -F(P)x$, which results in an induced value function V^P represented by $T^\varepsilon(P)$, i.e. $V^P(x) = -x'T^\varepsilon(P)x$. See Lemma 2 for the result and corresponding formulae.

As with shadow-price learning, here we assume that the agent updates his perceptions in a manner consistent with the use of regression analysis, and to model this, some additional notation is required. Let $z(x) = (x_i x_j)_{1\le i\le j\le n} \in \mathbb{R}^N$, where $N = n(n+1)/2$, be the vector of relevant regressors, i.e. the collection of all possible pairwise products of states. We remark that $x_1 = 1$, so that z includes a constant and all linear terms x_i. Let $\mathcal{S}(n) \subset \mathbb{R}^{n\times n}$ be the vector space of symmetric matrices, and let $\mathcal{M} : \mathbb{R}^N \to \mathcal{S}(n)$ be the vector space isomorphism that provides the following correspondence:

$$q \in \mathbb{R}^N \Longrightarrow q'z(x) = -x'\mathcal{M}(q)x.$$

with this notation, the perceived value function corresponding to perceptions P may be written as depending on analogous perceptions $q = \mathcal{M}^{-1}(P)$:

$$V(x) = -x'Px = q'z(x), \text{ where } \mathcal{M}(q) = P,$$

and because of this, we will speak of perceptions $q \in \mathbb{R}^N$.

To update his perceptions q, the agent regresses estimates $\hat{V}(x)$ of the value function on states $z(x)$. Now note that given perceptions q, the agent's control decision is given by $u = -F(\mathcal{M}(q))x$, which may then be used to obtain the estimate $\hat{V}(x)$:

$$\hat{V}(x) = -x'Rx - u'Qu - 2x'Wu + \beta\hat{E}q'z\,(Ax + Bu + C\varepsilon) = T^{VF}(q)'z(x),$$

where the second equality defines the map $T^{VF} : \mathbb{R}^N \to \mathbb{R}^N$. Now notice

$$\begin{aligned} T^{VF}(q)'z(x) &= -x'Rx - u'Qu - 2x'Wu - \beta\hat{E}(Ax + Bu + C\varepsilon)'\mathcal{M}(q)\,(Ax + Bu + C\varepsilon) \\ &= -x'T^{\varepsilon}(\mathcal{M}(q))x. \end{aligned}$$

It follows that $T^{VF} = \mathcal{M}^{-1} \circ T^{\varepsilon} \circ \mathcal{M}$, i.e. T^{VF} and T^{ε} are conjugate operators.

That T^{VF} is related to T^{ε} via conjugation has two immediate implications. First, if $q^* = \mathcal{M}^{-1}(P^*_{\varepsilon})$ then $T^{VF}(q^*) = q^*$ and q^* corresponds to fully optimal beliefs. Second, since conjugation preserves stability, by Corollary 1 we know that q^* is a Lyapunov stable fixed point of the differential equation $dq/d\tau = T^{VF}(q) - q$. We summarize these findings in the following theorem, which is the value-function learning analog of Theorem 3:

Theorem 5 *Assume LQ.1 – LQ.3 are satisfied. Then $q^* = \mathcal{M}^{-1}(P^*_{\varepsilon})$ is a Lyapunov stable fixed of the differential equation $dq/d\tau = T^{VF}(q) - q$, and $V^*(x) = q^{*'} \cdot z(x)$.*

We now briefly discuss the real-time algorithm that corresponds to value-function learning. Our agent is assumed to use available data to estimate the transition equation and his perceived value-function $V(x) = q'z(x)$, and then use these estimates to form forecasts and thereby generate new data. The transition equation is estimated just as in Section 3.2.2. To estimate the value-function at time t, and thus obtain an estimate of the beliefs coefficients q_t, we assume our agent regresses $\hat{V}_{t-1}$ on $z(x_{t-1})$ using data $\{x_{t-1}, \ldots, x_0, \hat{V}_{t-1}, \ldots, \hat{V}_0\}$. Given q_t, the agent's control choice is

$$u_t = F\,(H(q_t), A_t, B)\,x_t,$$

where F is given by (31), $H(q_t) = -2\mathcal{M}(q_t)$. Finally, the agent's estimated value at time t is given by

$$\hat{V}_t = T^{VF}\,(q_t, A_t, B)' \cdot z_t,$$

where we abuse notation somewhat by incorporating into the function T^{VF} the time-varying estimates of A. A causal recursive dynamic system analogous to (34) may then be used to state and prove a convergence theorem analogous to Theorem 4.

4.2.2 Euler-equation learning

Euler equation learning takes as primitive the agent's first-order conditions (FOC), written in terms of primal variables and derived using e.g. a variational argument, and assumes

decisions are taken to meet this condition subject to boundedly rational forecasts. In this section, we focus on decision-making environments that result in "one-step-ahead" Euler equations, that is, LQ-problems yielding FOCs with only one lead. Existence of one-step-ahead Euler equations is an important consideration, and it is closely related to the problem's state-control specification: a given dynamic programming problem may have several natural (and equivalent) state-control specifications, and a generic variational argument may yield a one-step-ahead Euler equation against one set of states and controls and not against another.

To address in generality the issues just laid out, in Appendix B we develop a *transformation* framework that will allow us to formally move between different state-control specifications; then we establish a general condition providing for the existence of a one-step-ahead Euler equation; next, we develop Euler equation learning assuming this condition holds, and show that LQ.1 - LQ.3 imply a stable learning environment. To facilitate the presentation here we focus on two special cases commonly met in practice.

The Bellman system associated with the standard dynamic programming problem (1)-(2) is given by

$$V(x) = \max_{u \in \Gamma(x)} r(x,u) + \beta E_\varepsilon V(g(x,u,\varepsilon)).$$

The first-order and envelope conditions are

$$\begin{aligned} 0 &= r_u(x_t,u_t)' + \beta E_t g_u(x_t,u_t,\varepsilon_{t+1})' V_x(x_{t+1})' && (37)\\ V_x(x_t)' &= r_x(x_t,u_t)' + \beta E_t g_x(x_t,u_t,\varepsilon_{t+1})' V_x(x_{t+1})'. && (38) \end{aligned}$$

Stepping (37) ahead one period and inserting into (38) yields

$$0 = r_u(x_t,u_t)' + \beta E_t \left(g_u(x_t,u_t,\varepsilon_{t+1})' \left(r_x(x_{t+1},u_{t+1})' + \beta g_x(x_{t+1},u_{t+1},\varepsilon_{t+2})' V_x(x_{t+2})' \right) \right).$$

Observe that if

$$E_t g_u(x_t,u_t,\varepsilon_{t+1})' g_x(x_{t+1},u_{t+1},\varepsilon_{t+2})' V_x(x_{t+2})' = 0 \tag{39}$$

then we obtain the usual Euler equation

$$0 = r_u(x_t,u_t)' + \beta E_t g_u(x_t,u_t,\varepsilon_{t+1})' r_x(x_{t+1},u_{t+1})'.$$

Equation (39) provides a natural condition for the existence of one-step-ahead Euler equations. Within the LQ framework, this condition becomes $E_t \lambda'_{t+2} AB du_t = 0$: the value two periods hence of a change in the control today is expected to be zero. The corresponding Euler equation is given by

$$Qu_t + W'x_t + \beta B' E_t (Rx_{t+1} + Wu_{t+1}) = 0. \tag{40}$$

Using the ordering of variables established in Section 4.1, we note that $A_{22} = 0$ provides a simple condition sufficient for the existence of one-step-ahead Euler equations within an LQ-framework.

If $A_{22} \neq 0$ it may still be possible to use variational arguments to derive a one-step ahead first-order condition that can be interpreted as an Euler equation. The key is whether one can choose a variation in controls at time t that leaves the state at time $t+2$ unaffected. Recalling the ordering convention of Section 4.1, consider the case in which $n_2 = m$, i.e. the number of controls equals the number of endogenous states, and $\det(B_2) \neq 0$. A variation du_t results in the change in endogenous states $dx_{2,t+1} = B_2 du_t$, which may be offset by the control choice

$$du_{t+1} = -B_2^{-1} A_{22} B_2 du_t,$$

so that $dx_{2,t+2} = 0$. The impact of this variation on the agent's objective, set equal to zero, yields the first-order condition

$$r_u(x_t, u_t)du_t + \beta E_t \left(r_x(x_{t+1}, u_{t+1})dx_{t+1} + r_u(x_{t+1}, u_{t+1})du_{t+1} \right) = 0,$$

where $dx_{1,t+1} = 0$. Combining with $dx_{2,t+1} = B_2 du_t$ and $du_{t+1} = -B_2^{-1} A_{22} B_2 du_t$, and using the quadratic form for $r(x, u)$ we get

$$Qu_t + W' x_t + \beta B' E_t \left(\breve{R} x_{t+1} + \breve{W} u_{t+1} \right) = 0, \tag{41}$$

where

$$\breve{R} = R - \begin{pmatrix} 0 \\ I_m \end{pmatrix} A_{22}' \left(B_2^{-1} \right)' W' \text{ and } \breve{W} = W - \begin{pmatrix} 0 \\ I_m \end{pmatrix} A_{22}' \left(B_2^{-1} \right)' Q. \tag{42}$$

Note that if $A_{22} = 0$ equation (40) is recovered.

We now turn to Euler equation learning based on (41). It is convenient to treat our two special cases simultaneously. Therefore, abusing notation somewhat, whenever $A_{22} = 0$, we set $\breve{R} = R$ and $\breve{W} = W$. Otherwise, we assume $n_2 = m$ and $\det(B_2) \neq 0$, and $\breve{R}$ and $\breve{W}$ are given by (42). We adopt stylized learning and for simplicity we assume the agent knows the transition matrix A as well as B. The agent is required to forecast his own future control decision, and we provide him a forecasting model that takes the same form as optimal behavior: $u_t = -F x_t$. The agent computes

$$\hat{E}_t x_{t+1} = A x_t + B u_t \text{ and } \hat{E}_t u_{t+1} = -F \hat{E}_t x_{t+1}, \tag{43}$$

which may then be used in conjunction with (41), with E_t replaced by $\hat{E}_t$, to determine his control decision. The following notation will be helpful: for "appropriate" $n \times n$ matrix X, set

$$\Phi(X) = (Q + \beta B' X B)^{-1} \text{ and } \Psi(X) = \beta B' X A + \breve{W}',$$

where X is appropriate provided that $\det(Q + \beta B' X B) \neq 0$. Then, combining (43) with (41) and simplifying yields

$$u_t = -T^{EL}(F) x_t, \text{ where } T^{EL}(F) = \Phi(\breve{R} - \breve{W} F)\Psi(\breve{R} - \breve{W} F). \tag{44}$$

Equation (44) may be interpreted as the actual law of motion given the agent's beliefs F. Finally, the agent updates his forecast of future behavior by regressing the control on

the state. This updating process results in a recursive algorithm analogous to (34), which identifies the agent's behavior over time.

Let $F^* = \Phi(P^*)\Psi(P^*)$, where P^* is the solution to the Riccati equation. By Theorem 2, we know that $u = -F^*x$ is the optimal feedback rule. In Appendix B we show $\Phi\left(\breve{R} - \breve{W}F^*\right) = \Phi(P^*)$ and $\Psi(\breve{R} - \breve{W}F^*) = \Psi(P^*)$, and thus it follows from (44) that $T^{EL}(F^*) = F^*$. It can be shown that F^* is stable under stylized learning. We have:

Theorem 6 *Assume LQ.1 – LQ.3 are satisfied, and that either $A_{22} = 0$ or else both $n_2 = m$ and $\det(B_2) \neq 0$. If agents behave as Euler equation learners with perceptions $u_t = -Fx_t$ and use (41) as their behavioral primitive then F^* is a Lyapunov stable fixed point of the differential equation $dF/d\tau = T^{EL}(F) - F$. That is, F^* is stable under stylized learning.*

We now briefly discuss the real-time algorithm that corresponds to Euler-equation learning. Our agent is assumed to use available data to estimate the transition equation and the coefficients F of his decision-rule $u = -Fx$, and then to use these estimates to form forecasts and make decisions, thereby generating new data. The transition equation is estimated just as in Section 3.2.2. To estimate the beliefs coefficients F, we assume our agent regresses u_{t-1} on x_{t-1} using data $\{x_{t-1}, \ldots, x_0, u_{t-1}, \ldots, u_0\}$. Given the time t estimate F_t, the agent's control choice satisfies

$$u_t = -T^{EL}\left(F_t, A_t, B\right) x_t,$$

where again we abuse notation somewhat by incorporating into the function T^{EL} the time-varying estimates of A. A causal recursive dynamic system along the lines of (34) may then be used to state and prove a convergence theorem for Euler-equation learning analogous to Theorem 4.

The above results on the Euler-equation learning procedure are based on first-order Euler-equation learning. Often it is possible in more general circumstances to derive an Euler equation that involves multiple leads. We illustrate this point in the context of the example given in Section 5.

4.3 Summary

Section 3, which presents our main results, provides theoretical justification for a class of boundedly rational and boundedly optimal decision rules, based on adaptive learning, in which an agent facing a dynamic stochastic optimization problem makes decisions at time t to meet his perceived optimality conditions, given his beliefs about the values of an extra unit of the state variables in the coming period and his perceived trade-off between controls and states between this period and the next. We fully develop the approach in the context of shadow-price learning in which our agent uses natural statistical procedure to update each period his estimates of the shadow prices of states and of the transition dynamics. Our results

show that in the standard Linear-Quadratic setting, an agent following our decision-making and updating rules will make choices that converge over time to fully optimal decision-making. In the current Section we have extended these convergence results to alternative variations based on value-function learning and Euler-equation learning. Taken together, our results are the bounded optimality counterpart of the now well-established literature on the convergence of least-squares learning to rational expectations. In the remaining sections we show how to apply our results to several standard economic examples.

5 Application 1: SP-learning in a Crusoe economy

We examine implementation of bounded optimality in two well-known models. In this Section we consider a single-agent LQ economy, the "Robinson Crusoe" model, in which the assumptions of Section 3 hold, so that convergence to optimal decision-making is guaranteed. In the next Section we will consider the Lucas-Sargent model of investment, which provides an equilibrium setting.

5.1 A Robinson Crusoe economy

A narrative approach may facilitate intuition. Thus, imagine Robinson Crusoe, a middle class Brit, finding himself marooned on a tropical island. An organized young man, he quickly takes stock of his surroundings. He finds that he faces the following problem:

$$\begin{aligned}
\max \quad & -\hat{E}\sum_{t\geq 0}\beta^t\left((c_t-b_t)^2+\phi l_t^2\right) && (45)\\
\text{s.t.} \quad & y_t = A_1 s_t + A_2 s_{t-1} + z_t && (46)\\
& s_{t+1} = y_t - c_t + \mu_{t+1} && (47)\\
& s_t = l_t && (48)\\
& b_t = b^* + \Delta(b_{t-1}-b^*) + \varepsilon_t \\
& z_t = \rho z_{t-1} + \eta_t, \\
& \text{with } s_{-1}, s_0, \text{ and } z_0, b_0 \text{ given.}
\end{aligned}$$

Here y_t is fruit and c_t is consumption of fruit. Equation (46) is Bob's production function – he can either plant the fruit or eat it, seeds and all – and the double lag captures the production differences between young and old fruit trees. All non-consumed seeds are planted. Some seasons, wind brings in additional seeds from nearby islands; other seasons, local voles eat some of the seeds: thus s_{t+1}, the number of young trees in time $t+1$, is given by equation (47), where the white noise term μ_{t+1} captures the variation due to wind and voles. Note that s_t is both the quantity of young trees in t and the quantity of old trees in $t+1$. Weeds are prevalent on the island: without weeding around all the young trees, the weeds rapidly spread everywhere and there is no production at all from any trees: see equation (48). This

is bad news for Bob as he's not fond of work: $\phi > 0$. Finally, z_t is a productivity shock (rabbits eat saplings and ancient seeds sprout) and b_t is stochastic bliss: see Ch. 5 of Hansen and Sargent (2014) for further discussion of this economy as well as many other examples of economies governed by quadratic objectives and linear transitions.[28]

Some comments on ϕ and the constraint (48) are warranted, as they play important roles in our analysis. Because $\phi > 0$ and $l_t = s_t$, it follows that an increased stock of productive trees reduces Bob's utility. In the language of LQ programming, these assumptions imply that a diverging state s_t is detected and must be avoided: specifically, $\phi > 0$ is necessary for the corresponding matrix pair to be detectable – see Assumption LQ.3 and Appendix C. In contrast, and somewhat improbably, Bob's disheveled American cousin Slob is not at all lazy: his ϕ is zero and his behavior, which we analyze in a companion paper, is quite different from British Bob's.

When he is first marooned, Bob does not know if there is a cyclic weather pattern; but he thinks that if last year was dry this year might be dry as well. Good with numbers, Bob decides to estimate this possible correlation using RLS. Bob also estimates the production function using RLS. Finally, Bob contemplates how much fruit to eat. He decides that his consumption choice should depend on the value of future fruit trees forgone. He concludes that the value of an additional tree tomorrow will depend (linearly) on how many trees there are, and makes a reasoned guess about this dependence. Given this guess, Bob estimates the value of an additional tree tomorrow, and chooses how much fruit to eat today.

Belly full, Bob pauses to reflect on his decisions. Bob realizes his consumption choice depended in part on his estimate about the value of additional trees tomorrow, and that perhaps he should revisit this estimate. He decides that the best way to do this is to contemplate the value of an additional tree today. Bob realizes that an additional young tree today requires weeding, but also provides additional young trees tomorrow (if he planted the young tree's fruit) and an old tree tomorrow, and that an additional old tree today provides young trees tomorrow (if he planted the old tree's fruit). Using his estimate of the value of additional trees tomorrow, Bob estimates the value of an additional young tree and an additional old tree today. He then uses these estimates to re-evaluate his guess about the dependence of tree-value on tree-stock. Exhausted by his efforts, Bob falls sound asleep. He should sleep well: Theorem 4 tells us that by following this simple procedure, Bob will learn to optimally exploit his island paradise.

This simple narrative describes the behavior of our boundedly optimal agent. It also points to a subtle behavioral assumption that is more easily examined by adding precision to the narrative. To avoid unnecessary complication, set $\Delta = 0, z_t = 0, \varepsilon_t = 0$. The

[28]The only novelty in our economy is the presence of a double lag in production. The double lag is a mechanism to expose the difference between Euler equation learning and SP-learning. Other mechanisms, such as the incorporation of habit persistence in the quadratic objective, yield similar results.

simplified problem becomes

$$\begin{aligned} \max \quad & -\hat{E}\sum_{t\geq 0}\beta^t\left((c_t-b^*)^2+\phi s_t^2\right) \\ \text{s.t.} \quad & s_{t+1}=A_1 s_t+A_2 s_{t-1}-c_t+\mu_{t+1} \end{aligned} \tag{49}$$

In notation of Section 3, the state vector is $x_t=(1,s_t,s_{t-1})'$ and the control is $u_t=c_t$. We assume that $\beta A_1+\beta^2 A_2>1$ and $\phi>0$ to guarantee that steady-state consumption is positive and below bliss: see Appendix C for a detailed analysis of the steady state and fully optimal solution to (49). In particular, it is shown there that LQ.1 - LQ.3 are satisfied. For the reasons given in Section 4.1 there is no need to forecast the shadow price of the intercept. Thus let λ_{1t} be the time t value of an additional new tree in time t and λ_{2t} the time t value of an additional old tree in time t. Bob guesses that λ_{it} depends on s_t and s_{t-1}:

$$\lambda_{it}=a_i+b_i s_t+d_i s_{t-1}, \text{ for } i=1,2. \tag{50}$$

He then forecasts λ_{it+1}:

$$\hat{E}_t\lambda_{it+1}=a_i+b_i\hat{E}_t s_{t+1}+d_i s_t, \text{ for } i=1,2. \tag{51}$$

Because he must choose consumption, and therefore savings, before output is realized, Bob estimates the production function and finds

$$\hat{E}_t s_{t+1}=A_{1t}s_t+A_{2t}s_{t-1}-c_t,$$

where A_t is obtained by regressing s_t on $(s_{t-1},s_{t-2})'$. He concludes that

$$\hat{E}_t\lambda_{it+1}=a_i+(b_i A_{1t}+d_i)s_t+b_i A_{2t}s_{t-1}-b_i c_t, \tag{52}$$

which, he notes, depends on his consumption choice today.

Now Bob contemplates his consumption decision. By increasing consumption by dc, Bob gains $-2(c_t-b^*)dc$ and loses $\beta\hat{E}_t\lambda_{1t+1}dc$. Bob equates marginal gain with marginal loss, and obtains

$$c_t=b^*-\frac{\beta}{2}\hat{E}_t\lambda_{1t+1}. \tag{53}$$

This equation together with equation (52) for $i=1$ is solved simultaneously by Bob to obtain numerical values for his consumption c_t and the forecasted shadow price $\hat{E}_t\lambda_{1t+1}$.[29]

Finally, Bob revisits his parameter estimates a_i, b_i, and d_i. He first thinks about the benefit of an additional new tree today: it would require weeding, but the fruits could be saved to produce an estimated A_{1t} new trees tomorrow, plus he gets an additional old tree tomorrow. He concludes

$$\lambda_{1t}=-2\phi s_t+\beta A_{1t}\hat{E}_t\lambda_{1t+1}+\beta\hat{E}_t\lambda_{2t+1}. \tag{54}$$

[29] See Appendix D for formal details linking this example to the set-up of Section 3.

Equations (53) and (54) reflect a novel feature in Bob's problem: a reduction in consumption in period t necessarily leads to additional old trees in period $t+2$, which means that Bob needs to track two shadow prices and cannot use one-step-ahead Euler equation learning.

To complete his contemplations, Bob finally considers the benefit of an additional old tree today: the fruits could be saved to produce an estimated A_{2t} new trees tomorrow. Thus

$$\lambda_{2t} = \beta A_{2t} \hat{E}_t \lambda_{1t+1}. \tag{55}$$

Because Bob has numerical values for $\hat{E}_t \lambda_{it+1}$, (54) and (55), together with the estimates A_{it}, generate numerical values for the perceived shadow prices. Bob will then use these data to update his estimates of the parameters a_i, b_i, and d_i. We have the following result.

Proposition 1 *Provided LQ.RTL holds, Robinson Crusoe learns to optimally consume fruit.*

A brief comment is required concerning the assumption LQ.RTL. As noted, LQ.1 - LQ.3 imply that the state dynamics explode no faster than $\beta^{-1/2}$. On the other hand, LQ.RTL requires that the state be asymptotically stationary. When $A_2 = 0$, and hence when A_2 is small and positive, it can be shown that the conditions $\beta A_1 + \beta^2 A_2 > 1$ and $\phi > 0$ imply LQ.RTL. However, there are cases with $A_2 > 0$ in which the optimal state dynamics are explosive. See Appendix D for more details.

This implementation of the narrative above highlights our view of Bob's behavior: he estimates forecasting models, makes decisions, and collects new data to update his models. Thus Bob understands simple trade-offs, can solve simultaneous linear equations and can run simple regressions. These skills are the minimal requirements for boundedly optimal decision-making in a dynamic, stochastic environment. Remarkably, they are also sufficient for asymptotic optimality.

One might ask whether Bob should be more sophisticated. For example Bob might search for a forecasting model that is consistent with the way shadow prices are subsequently revised: Bob could seek a fixed point of the T^{SP}-map. We view this alternative behavioral assumption as too strong, and somewhat unnatural, for two reasons. First, we doubt that in practice most boundedly rational agents explicitly understand the existence of a T^{SP}-map. Even if an agent did knew the form of the T^{SP}-map, would he recognize that a fixed point is what is wanted to ensure optimal behavior? Why would the agent think such a fixed point even exists? And if it did exist, how would the agent find it? Recognition that a fixed point is important, exists, and is computable is precisely the knowledge afforded those who study dynamic programming; our assumption is that our agent does not have this knowledge, even implicitly.

Our second reason for assuming Bob does not seek a fixed point to the T^{SP}-map relates to the above observation that obtaining such a fixed point is equivalent to full optimality given the "perceived transition equations." However, if the perceived coefficients A_{it} are far from the true coefficients, A_i, it is not clear that the behavior dictated by a fixed point to the

T^{SP}-map is superior to the behavior we assume. Given that computation is unambiguously costly, it makes more sense to us that Bob not iterate on the T^{SP}-map for fear that he might make choices based on magnified errors.

The central point of our paper is precisely that with limited sophistication, plausible and natural boundedly optimal decision-making rules converge to fully optimal decision-making.

5.2 Comparing learning mechanisms in a Crusoe economy

The simplified model (49) provides a nice laboratory to compare and contrast SP-learning with Euler equation learning. For simplicity we adopt stylized learning and assume that our agent knows the true values of A_i. Shadow price learning has been detailed in the previous section: the agent has PLM for shadow prices (50), and using this PLM, he forecasts future shadow prices: see (51). As just discussed, these forecasts are combined with (53) to yield his consumption decision

$$c_t = \phi_1(a, b, d) + \phi_2(a, b, d)s_t + \phi_3(a, b, d)s_{t-1}. \tag{56}$$

The control choice is then used to compute shadow price forecasts via (52). By Proposition 1 his decision-making is asymptotically optimal.

Turning to Euler-equation learning, if $A_2 = 0$ then there is a one-step ahead Euler equation that can be obtained using a suitable transform: see Appendix C. In this case Theorem 6 applies and Euler-equation learning provides an alternative asymptotically optimal implementation of bounded optimality. When $A_2 \neq 0$ it is not possible to obtain a one-step-ahead Euler equation; however, a simple variational argument leads to the following FOC:

$$c_t - \beta\phi\hat{E}_t s_{t+1} = \Psi + \beta A_1 \hat{E}_t c_{t+1} + \beta^2 A_2 \hat{E}_t c_{t+2}, \tag{57}$$

where $\Psi = b^*(1 - \beta A_1 - \beta^2 A_2)$. We refer to (57) as a *second*-order Euler equation.[30] We can then proceed analogously to Section 4.2.2 and implement Euler equation learning by taking (57) as a behavioral primitive: the agent is assumed to forecast his future consumption behavior and then choose consumption today based on these forecasts. The agent is assumed to form forecasts using a PLM that is functionally consistent with optimal behavior:

$$c_t = f_1 + f_2 s_t + f_3 s_{t-1}.$$

Using this forecasting model and the transition equation

$$s_{t+1} = A_1 s_t + A_2 s_{t-1} - c_t + \mu_{t+1},$$

the agent chooses c_t to satisfy (57). This behavior can then be used to identify the associated T^{EL} map. See the Appendix for a derivation of the T^{EL} map.

[30] The Euler equation can be derived by a direct variational argument. Alternatively it can be obtained from (53), (54) and (55).

Our interest here is to compare shadow price learning and Euler equation learning. Although we have not explored this point in the current paper, it is known from the literature on stochastic recursive algorithms that the speed of convergence of the real-time versions of our set-ups is governed by the maximum real parts of the eigenvalues of the T-map's derivative: this maximum needs to be less than one for stability and larger values lead to slower convergence. Here, if $A_2 = 0$ then the agent's problem has a one dimensional control and a two dimensional state, with one dimension corresponding to a constant: in this case the Euler equation is first-order and shadow price learning and Euler equation learning are equivalent. However, for $A_2 > 0$ the endogenous state's dimension is two and the equivalence may break down, as is evidenced by Figure 2, which plots the maximum real part of the eigenvalues for the respective T-maps' derivatives.

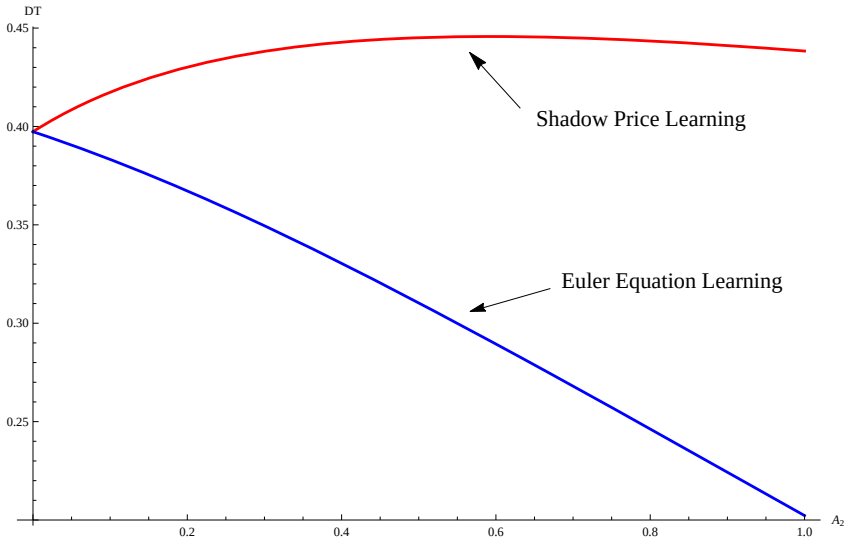

Figure 2: SP learning vs Euler learning: largest eigenvalue

The intuition for the inequivalence of shadow-price learning and Euler equation learning is straightforward: shadow price learning recognizes the two endogenous states and estimates the corresponding PLM. In contrast, under Euler equation learning agents need to understand and combine several structure relationships: they must understand the relationship between the two shadow prices;[31] and, they must combine this understanding with the first-order condition for the controls to eliminate the dependence on these shadow prices. In our simple example, Bob, as a shadow-price learner, would need to combine equations (54) and (55) with his decision rule to obtain the Euler equation (57). In this sense, shadow-price learning requires less structural information than Euler equation learning.

[31] We note that for the general LQ-problem, the relationship between shadow prices may be quite complicated.

6 Application 2: Investment under uncertainty

In this section we consider a version of the investment model of Lucas and Prescott (1971), as developed in Sargent (1987), Ch. XIV, and modified to include a time-to-build aspect. This provides a laboratory for examining SP-learning both from an agent's perspective and within an equilibrium context.

6.1 The model

There is a unit mass of firms indexed by $\omega \in \Omega$. All firms produce the same good. The quantity produced by firm ω is $y(\omega)$ and total economic output in time t is $y_t = \int_\Omega y_t(\omega)d\omega$. Firms sell their goods in a competitive market characterized by the exogenous (inverse) demand curve

$$p_t = \alpha_0 - \alpha_1 y_t + v_t, \tag{58}$$

where p_t is the price of goods and v_t is a stationary demand shock. To produce their goods $y_t(\omega)$, firms employ the same technology in which output is a function of installed capital $k_t(\omega)$. For simplicity we assume a Cobb-Douglass form:

$$y_t(\omega) = f\left(k_t(\omega)\right) = k_t(\omega)^\alpha$$

where $0 < \alpha \leq 1$.

6.1.1 The firm's problem

Firm ω's problem is to choose the level of investment $I_t(\omega)$ today that maximizes the subjectively-discounted value of current and future nominal profits.[32] More specifically, firm ω's problem is given by

$$\max_{I_t(\omega)} \quad \hat{E}(\omega)\sum_{t\geq 0}\beta^t\left(p_t y_t(\omega) - (J + q_t(\omega))I_t(\omega) - \frac{\gamma}{2}I_t(\omega)^2\right) \tag{59}$$

$$k_t(\omega) = (1-\delta)k_{t-1}(\omega) + \mu I_t(\omega) + (1-\mu)I_{t-1}(\omega) \tag{60}$$

$$v_{t+1} = \rho_v v_t + \varepsilon^v_{t+1} \text{ with } |\rho_v| < 1 \tag{61}$$

$$p_{t+1} = a_0 + a_1 p_t + a_2 v_t + \varepsilon^p_{t+1} \tag{62}$$

$$q_{t+1}(\omega) = q(\omega) + \varepsilon^q_{t+1}(\omega). \tag{63}$$

Here $\delta, \mu \in (0,1)$. Equation (60) captures a "time-to-build" feature. $J > 0$ is the exogenous market price of capital, assumed fixed for simplicity, and equation (61) provides our specification of the exogenous demand shock. The price process given by equation (62) is

[32] The subjective discount factor $0 < \beta < 1$ can be justified by assuming risk neutrality of the firm's owners.

treated as exogenous by the firm. The coefficients are assumed to impart stationarity. We have chosen a specification that is consistent with rational expectations equilibrium in the LQ case, i.e. when $\alpha = 1$.[33] Finally, $q_t(\omega)$ is an exogenous idiosyncratic transaction cost, where $\varepsilon^q_{t+1}(\omega)$ is mean zero, iid, and cross-sectionally independent.

It is convenient and natural to transform variables as follows. Let $z_t(\omega)$ be the quantity of installed capital in firm ω at the beginning of time t:

$$\begin{aligned} z_t(\omega) &= (1-\delta)k_{t-1}(\omega) + (1-\mu)I_{t-1}(\omega) \\ &= (1-\delta)z_{t-1}(\omega) + (1-\delta\mu)I_{t-1}(\omega). \end{aligned}$$

The firm's problem may then be written

$$\begin{aligned} \max_{I_t(\omega)} \quad & \hat{E}(\omega)\sum_{t\geq 0}\beta^t\left(p_t f(z_t(\omega)+\mu I_t(\omega)) - (J+q_t(\omega))I_t(\omega) - \frac{\gamma}{2}I_t(\omega)^2\right) \\ & z_{t+1}(\omega) = (1-\delta)z_t(\omega) + (1-\delta\mu)I_t(\omega) \\ & \& \text{ exogenous state dynamics} \end{aligned}$$

6.2 Temporary equilibrium

Firm ω's endogenous state in time t is $z_t(\omega)$, its control is $I_t(\omega)$, and the exogenous states, both individual and aggregate, are given by $(p_t, v_t, q_t(\omega))$. It follows that, regardless of whether the firm is modeled as rational or as an SP-learner, its investment decision takes the form $I_t(\omega) = I(p_t, v_t, q_t(\omega), z_t(\omega))$. These investment decisions are coordinated through price movements and goods-market clearing.

Coordination of investment decisions is made formal through the concept of a temporary equilibrium, which may be interpreted as a map taking time t exogenous and pre-determined endogenous variables to the remaining endogenous variables. For the model at hand,

$$\begin{aligned} \text{Exogenous and pre-determined endogenous variables} &: (v_t, \{z_t(\omega), q_t(\omega)\}) \\ \text{remaining endogenous variables} &: (p_t, \{I_t(\omega), z_{t+1}(\omega)\}), \end{aligned}$$

and the temporary equilibrium map is defined implicitly by

$$\begin{aligned} p_t &= \alpha_0 - \alpha_1 \int_\Omega f\left(z_t(\omega) + \mu I(p_t, v_t, q_t(\omega), z_t(\omega))\right) d\omega + v_t \\ z_{t+1}(\omega) &= (1-\delta)z_t(\omega) + (1-\delta\mu)I(p_t, v_t, q_t(\omega), z_t(\omega)). \end{aligned}$$

[33] For $\alpha < 1$ we choose parameters a, b so that the price process corresponds to the first-order approxmation of the REE. At this point it may seem surprising that more lags are not needed in the regression: this will be examined further below.

6.3 Rational Expectations

To obtain a benchmark, we first solve for the market equilibrium under rational expectations. To this end, we shut down the idiosyncratic shocks, $\varepsilon^q_{t+1}(\omega) = q(\omega) = 0$, and adopt a representative-agent perspective.

The firm's only time-t endogenous state variable is installed capital, z_t, with corresponding multiplier λ^z_t. Here for notational convenience we do not use an asterisk to distinguish between optimal and perceived shadow prices. Because firms take price as given, optimal decision making requires that the following conditions be met:

$$J + \gamma I_t = \mu p_t f'(z_t + \mu I_t) + \beta(1 - \delta\mu) E_t \lambda^z_{t+1} \tag{64}$$

$$\lambda^z_t = p_t f'(z_t + \mu I_t) + \beta(1 - \delta) E_t \lambda^z_{t+1}. \tag{65}$$

Equation (64) balances the cost of raising investment today, coming from purchasing price and installation, with the benefit of additional investment that arises through the corresponding increase in installed capital today and lagged investment tomorrow. Equation (65) is an envelope condition: the right-hand-side measures the benefit of marginal increases in the state.

Definition. An equilibrium is a collection $\{p_t, z_t, I_t, \lambda^z_t, v_t\}_{t\geq 0}$ satisfying

$$\begin{aligned}
p_t &= \alpha_0 - \alpha_1 f(z_t + \mu I_t) + v_t \\
z_t &= (1 - \delta) z_{t-1} + (1 - \delta\mu) I_{t-1} \\
J + \gamma I_t &= \mu p_t f'(z_t + \mu I_t) + \beta(1 - \delta\mu) E_t \lambda^z_{t+1} \\
\lambda^z_t &= p_t f'(z_t + \mu I_t) + \beta(1 - \delta) E_t \lambda^z_{t+1} \\
v_t &= \rho v_{t-1} + \varepsilon_t,
\end{aligned}$$

as well as additional boundary conditions, including z_{-1}, v_{-1} given.

Some observations are in order. The demand equation (58) provides a contemporaneous relationship between I_t, p_t, z_t, and v_t. Equations (64) and (65) can be combined to provide a contemporaneous relationship between λ_t, I_t, p_t and z_t. We may use these relationships to eliminate two variables, say λ_t and p_t, so that the linearized system involves only the jump variable I_t, the predetermined variable z_t and the exogenous variable v_t. In the determinate case, restriction to the stable saddle path implies that I_t is a function of z_t and u_t. It follows that the equilibrium (to the linearized model) is characterized by a VAR(1) in z_t and v_t. Alternatively, we may write the equilibrium as a VAR(1) in p_t and v_t, which may then be used to identify equilibrium consistent beliefs of the form (62).

When $\alpha = 1$, the firm's problem is LQ so that the economy is generically linear and the exact solution may be obtained. When $\alpha \in (0, 1)$, the economy is non-linear, and we approximate the REE to first-order. Note that this approximation becomes exact as $\alpha \to 1$.

Solving the linearized model proceeds in the usual fashion. Figure 3 provides impulse response functions giving the effect of a demand shock ε_t in proportional deviation form for two values of α.

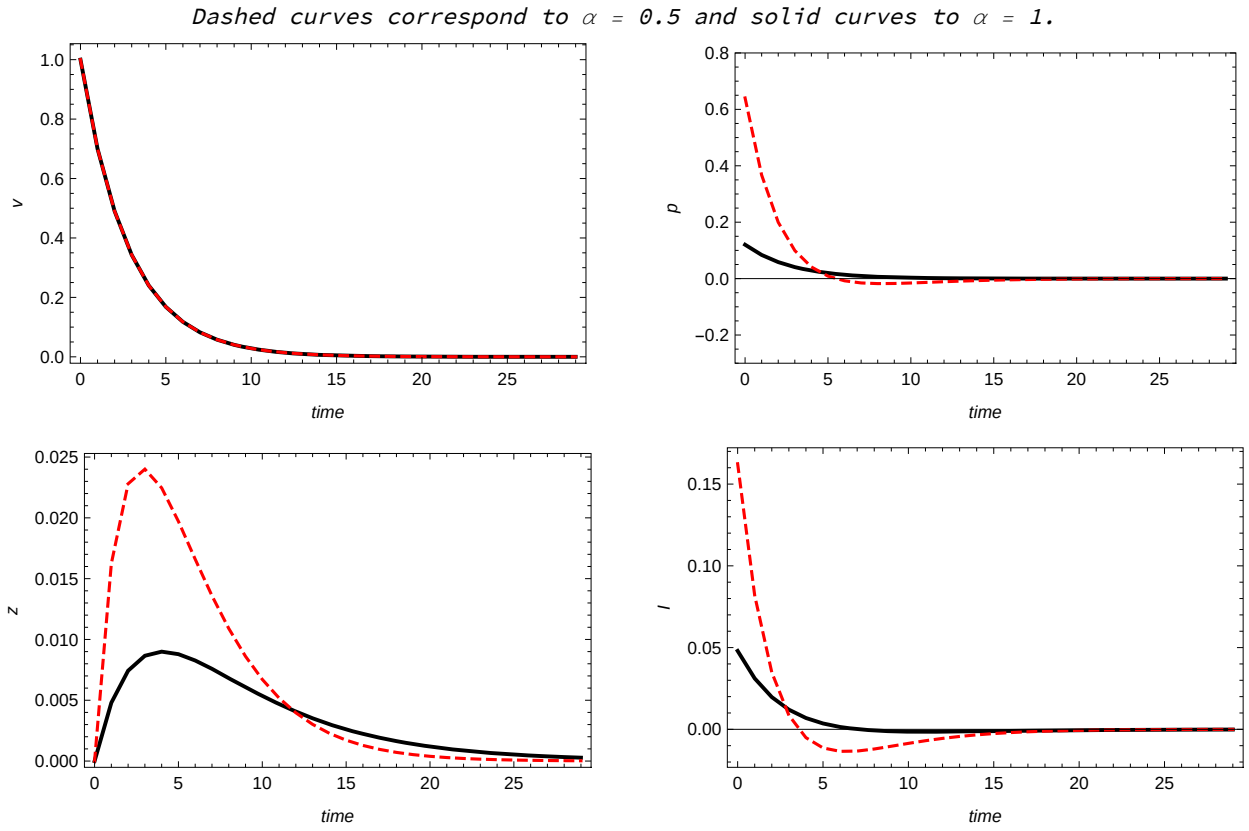

Figure 3: IRFs for demand shock in the REE

In the case $\alpha = 0.5$, diminishing marginal returns results in a smaller initial supply response, leading a larger initial increase in prices, which subsequently induces greater investment. Over time, as price returns to its steady state level, investment temporarily falls below its steady-state level.

6.4 SP-learning: the agent's problem

In this section, we analyze the decision problem of a given firm, taking the price process as exogenous. To avoid notational clutter, we will omit the ω index. In the next section, where we address general equilibrium consideration, a careful distinction between individual and aggregate variables will be needed. For this section, p_t and v_t are aggregates, taken as exogenous by the firm, and I_t and z_t are firm-specific endogenous variables. In this section we return to allowing for a random transaction cost q_t, which is also exogenous to the firm.

We reproduce the agent's problem for convenience:

$$\begin{aligned}
\max_{I_t} \quad & \hat{E}\sum_{t\geq 0}\beta^t\left(p_t f(z_t+\mu I_t)-(J+q_t)I_t-\frac{\gamma}{2}I_t^2\right) \qquad (66)\\
z_{t+1} &= (1-\delta)z_t+(1-\delta\mu)I_t\\
v_{t+1} &= \rho v_t+\varepsilon^v_{t+1}\\
q_{t+1} &= q+\varepsilon^q_{t+1}\\
p_{t+1} &= a_0+a_1p_t+a_2v_t+\varepsilon^p_{t+1}.
\end{aligned}$$

Using the notation from the theoretical development, the firm's states and controls are given as

$$\begin{aligned}
\text{Exogenous states:} \quad & x_{1t}=(1,v_t,q_t,p_t)\\
\text{Endogenous state:} \quad & x_{2t}=z_t\\
\text{Control:} \quad & u_t=I_t.
\end{aligned}$$

Our theoretical work established that we need only consider the shadow price of the endogenous state. Thus, Let λ^z_t be the perceived value in time t of an additional unit of pre-installed capital, z_t. It follows that his investment decision, I_t satisfies

$$J+q_t+\gamma I_t=\mu p_t f'(z_t+\mu I_t)+\beta(1-\delta\mu)E_t\lambda^z_{t+1}. \qquad (67)$$

Equation (67) may be interpreted as the firm's FOC: the LHS represents the cost of purchasing and installing an additional of capital, and the RHS measures the benefit in terms of current and expected future revenues.

The firm believes the value of an additional unit of pre-installed capital depends on the states: $\lambda^z_t=Hx_t$. To forecast λ^z_{t+1}, the firm must forecast x_{t+1} and for this we must establish the perceived transition equation. We may write

$$x_{t+1}=Ax_t+Bu_t+C\varepsilon_{t+1}$$

with

$$A=\begin{pmatrix} 1 & 0 & 0 & 0 & 0\\ 0 & \rho & 0 & 0 & 0\\ q & 0 & 0 & 0 & 0\\ a_0 & a_2 & 0 & a_1 & 0\\ 0 & 0 & 0 & 0 & 1-\delta \end{pmatrix} \text{ and } B=\begin{pmatrix} 0\\ 0\\ 0\\ 0\\ 1-\mu\delta \end{pmatrix}.$$

It follows that

$$\hat{E}_t\lambda^z_{t+1}=H\hat{E}_tx_{t+1}=H(Ax_t+BI_t), \qquad (68)$$

where H and some of the entries in A may be taken as beliefs, and thus require estimation. Equations (67) and (68) may be solved simultaneously to compute

$$\begin{aligned}
\hat{E}_t\lambda^z_{t+1} &= \lambda^{z,e}(x_t,H,A) \qquad (69)\\
I_t &= I(x_t,H,A). \qquad (70)
\end{aligned}$$

In particular, given beliefs (H, A) and the current state x_t, the firm's control choice I_t is determined.

The firm updates its perceived shadow price by contemplating the value of an additional unit of pre-installed capital, taking his investment decision as given. Formally,

$$\lambda_t^z = p_t f'(z_t + \mu I_t) + \beta(1-\delta)\hat{E}_t \lambda_{t+1}^z.$$

Using equations (69) and (70), we may write

$$\lambda_t^z = \lambda^z(x_t, H, A). \tag{71}$$

Thus given beliefs (H, A) and the current state x_t, the firm's shadow price is updated.

The behavior of a firm acting as a stylized SP learner may now be summarized: in time t the firm, holding beliefs summarized as (H, A), takes the good's price, p_t, capital cost, $J + q_t$, and demand shock, v_t, as well as its own pre-installed capital level, z_t, as given, and then chooses investment, I_t, according to (70) and updates its shadow price, λ_t^z, according to (71). In Section 6.4.1, assuming $\alpha = 1$ so that the investment problem is LQ in nature, we address whether this firm will learn to make optimal decisions. To model the firm as a real-time SP learner, its behavior must be augmented to include recursive estimation of H and A. This is taken up in Section 6.4.2.

6.4.1 SP-learning: the agent's problem in the LQ-case

We now let $\alpha = 1$ so that the firm's objective is quadratic. To write the objective as

$$p_t(z_t + \mu I_t) - (J + q_t)I_t - \frac{\gamma}{2} I_t^2 = -(x_t' R x_2 + u_t' Q u_t + 2 x_t' W u_t) \tag{72}$$

with R positive semi-definite, it is necessary to redefine the state slightly: Letting $x_t = (1, v_t, q_t, -p_t, z_t)$, $Q = \gamma/2$, and

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} \frac{J}{2} \\ 0 \\ \frac{1}{2} \\ \frac{\mu}{2} \\ 0 \end{pmatrix}$$

makes (72) hold. Also, with

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \rho & 0 & 0 & 0 \\ q & 0 & 0 & 0 & 0 \\ -a_0 & -a_2 & 0 & a_1 & 0 \\ 0 & 0 & 0 & 0 & 1-\delta \end{pmatrix}, \; B = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1-\mu\delta \end{pmatrix} \text{ and } C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

the transition dynamic is given by

$$x_{t+1} = Ax_t + Bu_t + C \begin{pmatrix} \varepsilon_{t+1}^v \\ \varepsilon_{t+1}^q \\ \varepsilon_{t+1}^p \end{pmatrix}.$$

The matrices (R, Q, W, A, B) characterize the firm's problem and may be used to conduct stability analysis.

To apply our main result, we transform the matrices and examine whether LQ.1 – LQ.3 hold. With $\hat{R} = R - WQ^{-1}W'$ we find

$$\hat{R} = \begin{pmatrix} -\frac{J^2}{2\gamma} & 0 & -\frac{J}{2\gamma} & -\frac{J\mu}{2\gamma} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{J}{2\gamma} & 0 & -\frac{1}{2\gamma} & -\frac{\mu}{2\gamma} & 0 \\ -\frac{J\mu}{2\gamma} & 0 & -\frac{\mu}{2\gamma} & -\frac{\mu^2}{2\gamma} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Translating $\hat{R}$ by $S \oplus 0$ with

$$S = \begin{pmatrix} \frac{J^2}{2\gamma} & 0 & \frac{J}{2\gamma} & \frac{J\mu}{2\gamma} \\ 0 & 0 & 0 & 0 \\ \frac{J}{2\gamma} & 0 & \frac{1}{2\gamma} & \frac{\mu}{2\gamma} \\ \frac{J\mu}{2\gamma} & 0 & \frac{\mu}{2\gamma} & \frac{\mu^2}{2\gamma} \end{pmatrix}$$

shows that LQ.1 holds.

Now setting $\hat{A} = \sqrt{\beta}\,(A - BQ^{-1}W')$ we have

$$\hat{A} = \begin{pmatrix} \sqrt{\beta} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\beta}\rho & 0 & 0 & 0 \\ \sqrt{\beta}q & 0 & 0 & 0 & 0 \\ -a_0\sqrt{\beta} & -a_2\sqrt{\beta} & 0 & a_1\sqrt{\beta} & 0 \\ -\frac{\sqrt{\beta}J(1-\delta\mu)}{\gamma} & 0 & -\frac{\sqrt{\beta}(1-\delta\mu)}{\gamma} & -\frac{\sqrt{\beta}\mu(1-\delta\mu)}{\gamma} & \sqrt{\beta}(1-\delta) \end{pmatrix}.$$

Since we have assumed p_t is stationary, we know that $|a_1| < 1$. It follows that the eigenvalues of $\hat{A}$ are stable which is sufficient to guarantee that LQ.2 and LQ.3 hold. We conclude that our main result is applicable and our firm's optimal investment strategy is stable under stylized learning.

6.4.2 SP-learning: the agent's problem in the general case

To address the general case with $\alpha \in (0, 1]$, as well as consider stability under real-time learning, we now set up and analyze the firm's recursive decision problem with beliefs updated over time as new data become available. Let $x_t = (1, v_t, q_t, p_t, z_t)$ and H_{t-1} be the estimate

obtained by regressing λ^z on x using the data $\{\lambda^z_{t-n}, x_{t-n}\}_{n\geq 1}$, and let R_{t-1} be the sample second-moment matrix of the regressors. Because the price process is exogenous in this exercise, we assume that the transition matrix A is known to agents.

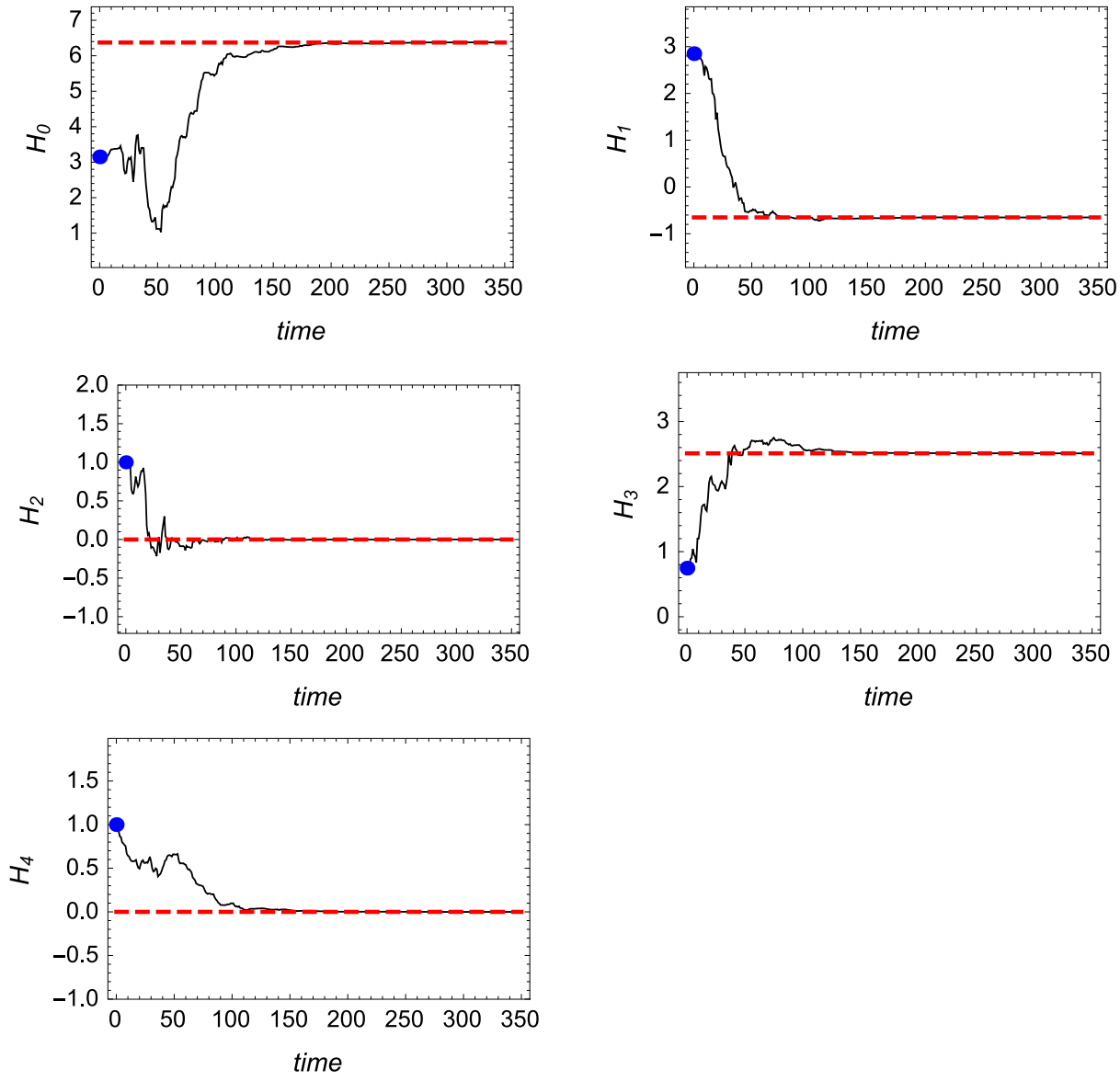

Fig. 4: Belief parameters H for shadow price with exogenous goods price. LQ case.

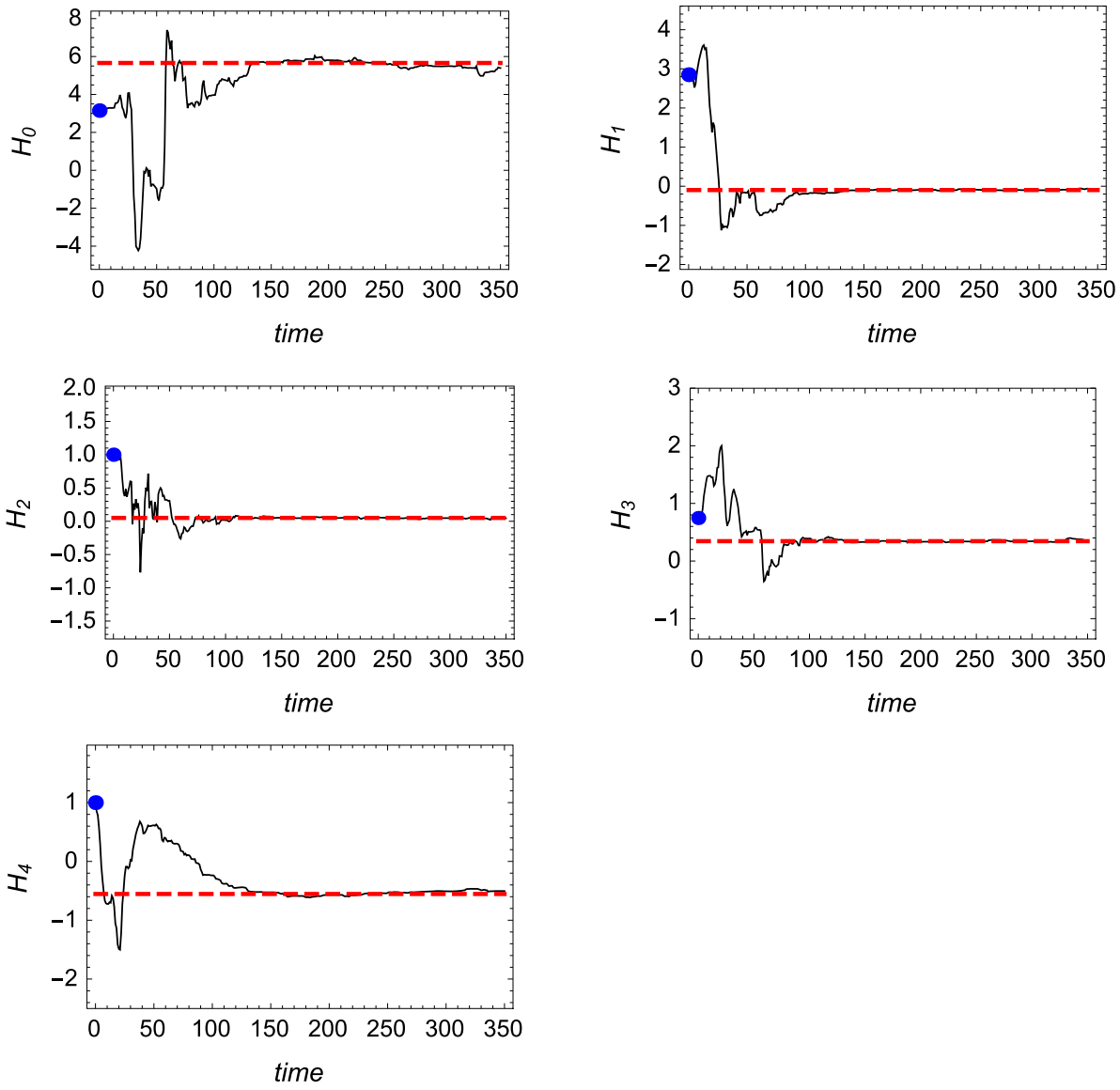

Figure 5: Beliefs parameters for shadow price with exogenous goods price process. Non-LQ case, with $\alpha = 0.3$.

At the beginning of period t the following data are known: (x_t, H_{t-1}, R_{t-1}). The causal

recursive system updating these data is given by

$$\begin{aligned}
z_{t+1} &= (1-\delta)z_t + (1-\mu\delta)I(x_t, H_{t-1}, A) \\
R_t &= R_{t-1} + \gamma_t(x_t x_t' - R_{t-1}) \\
H_t' &= H_{t-1}' + \gamma_t R_t^{-1} x_t \left(\lambda^z(x_t, H_{t-1}, A) - H_{t-1}x_t\right) \\
q_{t+1} &= \bar{q} + \varepsilon_{t+1}^q \\
p_{t+1} &= a_0 + a_1 p_t + a_2 v_t + \varepsilon_{t+1}^p \\
v_{t+1} &= \rho v_t + \varepsilon_{t+1}^v \\
x_{t+1} &= (1, v_{t+1}, q_{t+1}, p_{t+1}, z_{t+1})
\end{aligned}$$

We begin by considering real-time learning in the LQ-case, that is, with $\alpha = 1$. Initial beliefs are set quite far from optimal beliefs, state variables are at steady state, the price process is consistent with REE, and the gain is set at $\gamma_t = 0.1$. Figure 4 provides the time path of beliefs, with the red, dashed line indicating optimal beliefs.[34]

For the non-LQ case, Figure 5 shows the results for the case $\alpha = 0.3$ and constant gain $\gamma_t = 0.1$. The red dashed line show the optimal coefficients of the corresponding linearized model. The results strongly suggest that, to first order, agents learn over time how to optimize using SP learning.

6.4.3 SP-learning in market equilibrium

Now we endogenize the price process through market clearing. In this example we restrict attention to the LQ case $\alpha = 1$. For convenience we assume agents are homogeneous; thus, for notational simplicity we suppress the index ω, and we speak of the *representative agent*. To further simplify the system we set $q_t = 0$. The agent's state vector, then, is given by $x_t = (1, v_t, p_t, z_t)$; however, to avoid asymptotic multicollinearity, the agents uses only the subvector $\tilde{x}_t = (1, p_t, z_t)$ for forecasting his shadow price. Thus let H_t be the estimate obtained by regressing λ_{t-n} on $\tilde{x}_{t-n}$ for $n \geq 0$. Let a_t be the estimate obtained by regressing p_{t-n} on $(1, p_{t-n-1}, v_{t-n-1})$ for $n \geq 0$. Thus (H_{t-1}, a_{t-1}) summarize the agent's beliefs used for decision-making in time t.

Now define $\hat{x}_t = (1, v_t, z_t)$ to be the vector of variables which are observable to the representative agent when he makes conditional decisions. We write

$$\begin{aligned}
I_t &= I(x_t, H_{t-1}, A_{t-1}) = I(\hat{x}_t, p_t, H_{t-1}, a_{t-1}) \\
\lambda_t &= \lambda(x_t, H_{t-1}, A_{t-1}) = \lambda(\hat{x}_t, p_t, H_{t-1}, a_{t-1}),
\end{aligned}$$

[34] Because the exogenous price process is taken here to correspond to the REE in the representative-agent model, our agent's asymptotic decision-making corresponds to that of the representative agent, which makes perfectly collinear the state variable z_t with p_t, v_t, q_t in the LQ setting. For sufficiently long time series this could cause numerical problems. This issue does not arise in an equilibrium setting with heterogeneity.

where A_{t-1} is the transition matrix implied by the beliefs a_{t-1}. The advantage of this notation is that it emphasizes the conditional nature of I_t and λ_t: they depend on prices p_t which are determined in equilibrium.

The temporary equilibrium map $\mathcal{TE} : \mathbb{R}^3 \oplus \mathbb{R}^3 \oplus \mathbb{R}^2 \to \mathbb{R}$ is defined implicitly as the price that clears the goods market:

$$p = \alpha_0 - \alpha_1 f(z + I(\hat{x}, p, H, a)) + v \implies p = \mathcal{TE}(\hat{x}, H, a).$$

With this map in hand, we may now write down the causal recursive system identifying equilibrium behavior. Let $\mathcal{R}_t^H$ and $\mathcal{R}_t^a$ be the sample second moment matrices of $\tilde{x}_t$ and $(1, p_{t-1}, v_{t-1})$ respectively. At the beginning of time t, the following data are available:

$$H_{t-1}, a_{t-1}, v_{t-1}, p_{t-1}, z_{t-1}, I_{t-1}, \mathcal{R}_{t-1}^a, \mathcal{R}_{t-1}^H.$$

The recursive system is given by

$$\begin{aligned}
v_t &= \rho v_{t-1} + \varepsilon_t^v \\
z_t &= (1-\delta) z_{t-1} + (1 - \mu\delta) I_{t-1} \\
\hat{x}_t &= (1, v_t, z_t) \\
p_t &= \mathcal{TE}(\hat{x}_t, H_{t-1}, a_{t-1}) \\
I_t &= I(\hat{x}_t, p_t, H_{t-1}, a_{t-1}) \\
\lambda_t &= \lambda(\hat{x}_t, p_t, H_{t-1}, a_{t-1}) \\
\tilde{x}_t &= (1, p_t, z_t) \\
\mathcal{R}_t^H &= \mathcal{R}_{t-1}^H + \gamma_t \left(\tilde{x}_t \tilde{x}_t' - \mathcal{R}_{t-1}^H \right) \\
H_t' &= H_{t-1}' + \gamma_t (\mathcal{R}_t^H)^{-1} \tilde{x}_t (\lambda_t - H_{t-1} \tilde{x}_t) \\
\mathcal{R}_t^a &= \mathcal{R}_{t-1}^a + \gamma_t \left(\begin{pmatrix} 1 \\ p_{t-1} \\ v_{t-1} \end{pmatrix} \begin{pmatrix} 1 & p_{t-1} & v_{t-1} \end{pmatrix} - \mathcal{R}_{t-1}^a \right) \\
a_t' &= a_{t-1}' + \gamma_t (\mathcal{R}_t^a)^{-1} \begin{pmatrix} 1 \\ p_{t-1} \\ v_{t-1} \end{pmatrix} \left(p_t - a_{t-1} \begin{pmatrix} 1 \\ p_{t-1} \\ v_{t-1} \end{pmatrix} \right)
\end{aligned}$$

Figure 6 illustrates the results in the decreasing-gain case. Again the dashed red lines represent the REE values. The perceived price process parameters a, shown in the top panel, start from initial values that are significantly away from the REE values. For a period of time the coefficients continue to depart from REE values, before appearing to converge to REE. The bottom panel correspondingly shows apparent convergence of the shadow price parameters H to optimal decision-making.

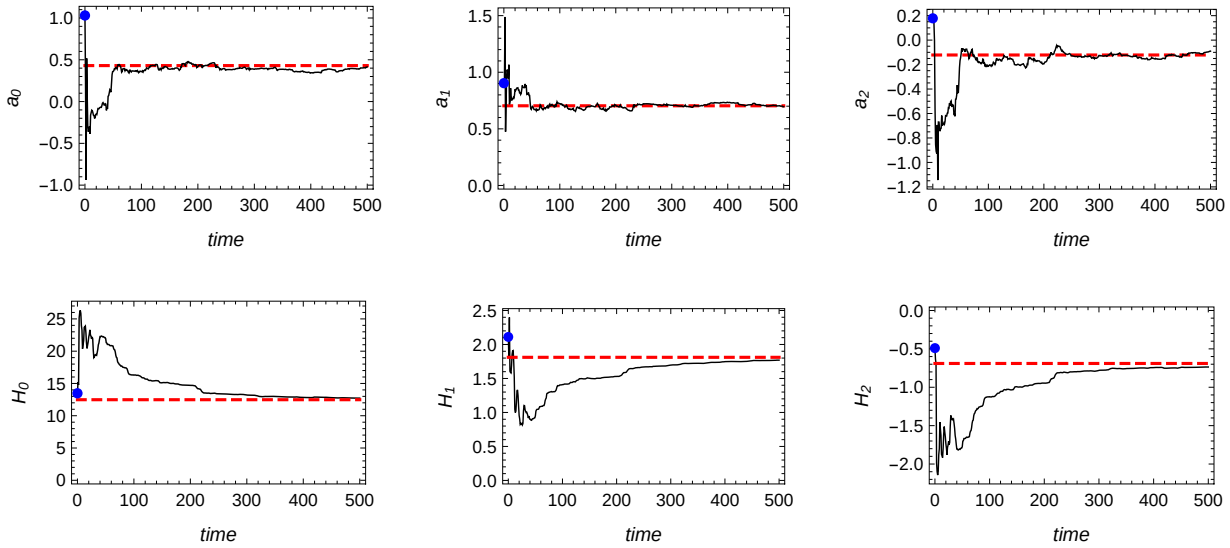

Figure 6: Beliefs parameters in market equilibrium. LQ case $\alpha = 1$. Top panel: market-price parameters. Bottom panel: shadow-price parameters.

This example shows how SP-learning can be embedded in an equilibrium setting, indicating its broad potential for inclusion in general equilibrium models. Applications to a range of dynamic macro general equilibrium models is investigated in our companion paper **?**.

7 Conclusion

The prominent role played by micro-foundations in modern macroeconomic theory has directed researchers to intensely scrutinize the assumption of rationality – an assumption on which these micro-foundations fundamentally rest; and, some researchers have criticized the implied level of sophistication demanded of agents in these micro-founded models as unrealistically high. Rationality on the part of agents consists of two central behavioral primitives: that agents are optimal forecasters; and that agents make optimal decisions given these forecasts. While the macroeconomics learning literature has defended the optimal forecasting ability of agents by showing that agents may learn the economy's rational expectations equilibrium, and thereby learn to forecast optimally, the way in which agents make decisions while learning to forecast has been given much less attention.

In this paper, we formalize the connection between boundedly rational forecasts and agents' choices by introducing the notion of bounded optimality. Our agents follow simple behavioral primitives: they use econometric models to forecast one-period ahead shadow prices; and they make control decisions today based on the trade-off implied by these forecasted prices. We call this learning mechanism shadow-price learning. We find our learning mechanism appealing for a number of reasons: it requires only simple econometric modeling

and thus is consistent with the learning literature; it assumes agents make only one-period-ahead forecasts instead of establishing priors over the distributions of all future relevant variables; and it imposes only that agents make decisions based on these one-period-ahead forecasts, rather than requiring agents to solve a dynamic programming problem with parameter uncertainty.

Investigation of SP-learning reveals that it is behaviorally consistent at the agent level: by following our simple behavioral assumptions, an individual facing a standard dynamic programming problem will learn to optimize. Our central stability results are shown to imply asymptotic optimality of alternative implementations of boundedly optimal decision-making, including value-function and Euler-equation learning. The application to investment under uncertainty shows how SP learning embeds naturally in non-LQ environments and market equilibrium settings.

It may appear tempting to take our results to suggest that agents should simply be modeled as fully optimizing. However, we think the implications of our results are far-ranging. Since agents have finite lives, learning dynamics will likely be significant in many settings. If there is occasional structural change, transitional dynamics will also be important; and if agents anticipate repeated structural change, say in the form of randomly switching structural transition dynamics, and if as a result they employ discounted least squares, as in Sargent (1999) and Williams (2014), then there will be perpetual learning that can include important escape dynamics. Finally, as we found in the investment model example of Section 6, agents with plausibly misspecified forecasting models will only approximate optimal decision making. Such misspecification can be along other dimensions as well, e.g. economizing on the components of the state vector. As in the adaptive learning literature, we anticipate that our convergence results will be the leading edge of a family of approaches to boundedly optimal decision making. We intend to explore this family of approaches in future work, as well as to study applications to range of DSGE models.

Appendix A: Proofs

Notation. Throughout the Appendices, we use without further comment the following notation: For $n \times n$ matrix P,

$$\Phi(P) = (Q + \beta B'PB)^{-1} \text{ and } \Psi(P) = \beta B'PA + W'.$$

Proof of Lemma 1: We reproduce the problem here for convenience:

$$V^P(x) = \max_u -(x'Rx + u'Qu + 2x'Wu) - \beta(Ax + Bu)'P(Ax + Bu).$$

The first-order condition, which is sufficient for optimality and uniqueness of solution because P is symmetric positive semi-definite, is given by

$$-2u'Q - 2x'W - 2\beta\left(x'A'PB + u'B'PB\right) = 0, \text{ or}$$

$$u = -\left(Q + \beta B'PB\right)^{-1}\left(\beta B'PA + W'\right)x = -\Phi(P)\Psi(P)x.$$

To compute the value function we insert this into the objective:

$$\begin{aligned}
V^P(x) &= -x'\left(R + \Psi'\Phi'Q\Phi\Psi - 2W\Phi\Psi + \beta(A - B\Phi\Psi)'P(A - B\Phi\Psi)\right)x \\
&= -x'\left(R + \Psi'\Phi'Q\Phi\Psi - 2W\Phi\Psi\right)x \\
&\quad -x'\left(\beta A'PA + \beta\Psi'\Phi'B'PB\Phi\Psi - \beta A'PB\Phi\Psi - \beta\Psi'\Phi'B'PA\right)x \\
&= -x'\left(R + \beta A'PA\right)x \\
&\quad -x'\left(\Psi'\Phi'\left(Q + \beta B'PB\right)\Phi\Psi - 2W\Phi\Psi - \beta A'PB\Phi\Psi - \beta\Psi'\Phi'B'PA\right)x \\
&= -x'\left(R + \beta A'PA + \Psi'\Phi'\left(\Psi - \beta B'PA\right) - 2W\Phi\Psi - \beta A'PB\Phi\Psi\right)x \\
&= -x'\left(R + \beta A'PA + \Psi'\Phi'W' - \left(2W + \beta A'PB\right)\Phi\Psi\right)x \\
&= -x'\left(R + \beta A'PA - \Psi'\Phi\Psi\right)x,
\end{aligned}$$

where the last step is obtained by noting that $x'\Psi'\Phi'W'x = x'W\Phi\Psi x$.

Proof of Lemma 2: Observe that

$$\begin{aligned}
E\left((Ax + Bu + C\varepsilon)'P(Ax + Bu + C\varepsilon)|x, u\right) &= \beta(Ax + Bu)'P(Ax + Bu) - \beta\delta(P), \\
\text{where } \delta(P) &= -\text{tr}\left(\sigma_\varepsilon^2 PCC'\right).
\end{aligned}$$

This follows since

$$\begin{aligned}
E\varepsilon'C'PC\varepsilon &= \text{tr}\left(E\varepsilon'C'PC\varepsilon\right) = E\text{tr}\left(\varepsilon'C'PC\varepsilon\right) \\
&= E\text{tr}\left(PC\varepsilon\varepsilon'C'\right) = \text{tr}\left(PCE(\varepsilon\varepsilon')C'\right) \\
&= \text{tr}\left(PC\sigma_\varepsilon^2 I_n C'\right) = \text{tr}\left(\sigma_\varepsilon^2 PCC'\right).
\end{aligned}$$

Since $\delta(P)$ does not depend on u, it follows that, given P, the control choice that solves (12) solves (14). Using Lemma 1 this establishes item 1. Noting that the stochastic objective is

the deterministic objective shifted by $\beta\delta(P)$ and inserting $u = -F(P)x$ into the objective of (14) yields

$$\begin{aligned} -x'T^{\varepsilon}(P)x &= -x'T(P)x + \beta\delta(P) \\ &= -x'\left(T(P) - \beta\Delta(P)\right)x, \end{aligned}$$

where the second equality follow from the fact that the first component of the state equals one. This establishes item 2.

To prove the last statement, first note that $\partial\delta(P)/\partial P_{11} = 0$. This follows because the first row of C, and hence the first row of CC' is zero. Now, let $\tilde{P}$ be a fixed point of T, and let $\tilde{P}_{\varepsilon}$ be given by (15). We now compute

$$\begin{aligned} T^{\varepsilon}\left(\tilde{P}_{\varepsilon}\right) &= T\left(\tilde{P} - \frac{\beta}{1-\beta}\Delta(\tilde{P})\right) - \beta\Delta\left(\tilde{P} - \frac{\beta}{1-\beta}\Delta(\tilde{P})\right) \\ &= T\left(\tilde{P} - \frac{\beta}{1-\beta}\Delta(\tilde{P})\right) - \beta\Delta\left(\tilde{P}\right) \\ &= T\left(\tilde{P}\right) - \frac{\beta^2}{1-\beta}\Delta(\tilde{P}) - \beta\Delta\left(\tilde{P}\right) \\ &= \tilde{P} - \frac{\beta}{1-\beta}\Delta(\tilde{P}) = \tilde{P}_{\varepsilon}. \end{aligned}$$

To establish the third equality, we show that for any perceptions $P \in \mathcal{U}$,

$$T(P + \Upsilon) = T(P) + \beta\Upsilon, \tag{73}$$

where $\Upsilon = \upsilon \oplus 0_{n-1\times n-1}$ simply captures a perturbation of the $(1,1)$ entry of P. To establish this equation first note that by Remark 1, $F(P+\Upsilon) = F(P)$. It follows that

$$-x'T(P+\Upsilon)x = -\left(x'Rx + u'Qu + 2x'Wu\right) - \beta(Ax+Bu)'\left(P+\Upsilon\right)(Ax+Bu),$$

where u is the optimal control decision given perceptions P. Straightforward algebra shows that

$$-x'T(P+\Upsilon)x = -x'T(P)x - \beta(Ax+Bu)'\Upsilon(Ax+Bu).$$

By the forms of A and B we have that $B'\Upsilon = 0$, $\Upsilon B = 0$ and $x'A'\Upsilon Ax = \Upsilon$. Thus

$$-x'T(P+\Upsilon)x = -x'\left(T(P) + \beta\Upsilon\right)x.$$

This establishes (73), which completes the proof. ■

Proof of Lemma 3: To establish the first equation, let

$$\begin{aligned} eq1(P) &= R + \beta A'PA - (\beta A'PB + W)\Phi(P)(\beta B'PA + W') \\ eq2(P) &= \hat{R} + \hat{A}'P\hat{A} - \hat{A}'P\hat{B}\Phi(P)\hat{B}'P\hat{A}. \end{aligned}$$

Our goal is to show $eq1(P) = eq2(P)$. First, we expand each equation.

$$\begin{aligned}
eq1(P) &= R + \beta A'PA - \beta A'PB\Phi(P)B'PA\beta - W\Phi(P)W' \\
&\quad -\beta A'PB\Phi(P)W' - W\Phi(P)B'PA\beta \\
eq2(P) &= R - WQ^{-1}QQ^{-1}W' + \beta A'PA + \beta WQ^{-1}B'PBQ^{-1}W' \\
&\quad -\beta A'PBQ^{-1}W' - \beta WQ^{-1}B'PA - \beta A'PB\Phi(P)B'PA\beta \\
&\quad -\beta WQ^{-1}B'PB\Phi(P)B'PBQ^{-1}W'\beta \\
&\quad +\beta A'PB\Phi(P)B'PBQ^{-1}W'\beta + \beta WQ^{-1}B'PB\Phi(P)B'PA\beta.
\end{aligned}$$

For notational convenience, we will number each summand of each equation in the natural way, being sure to incorporate the sign. Thus $eq1.2 = \beta A'PA$, $eq2.2 = -WQ^{-1}QQ^{-1}W'$, and $eq1 = eq1.1 + \ldots + eq1.6$, where we drop the reference to P to simplify notation. Now notice

$$\begin{aligned}
eq2.2 + eq2.4 + eq2.8 &= WQ^{-1}\left(-Q + \beta B'PB - \beta B'PB\Phi(P)\beta B'PB\right)Q^{-1}W' \\
eq2.5 + eq2.9 &= \beta A'PB\left(\Phi(P)\beta B'PB - I_m\right)Q^{-1}W' \qquad (74) \\
eq2.6 + eq2.10 &= WQ^{-1}\left(\beta B'PB\right)\Phi(P) - I_m)\beta B'PA.
\end{aligned}$$

Since

$$\begin{aligned}
-Q + \beta B'PB - \beta B'PB\Phi(P)\beta B'PB &= -Q + \beta B'PB\Phi(P)\left(\Phi(P)^{-1} - \beta B'PB\right) \\
&= -\Phi(P)^{-1}\Phi(P)Q + \beta B'PB\Phi(P)Q \\
&= -Q\Phi(P)Q, \text{ and} \\
\Phi(P)\beta B'PB - I_m &= \Phi(P)(\beta B'PB - \Phi(P)^{-1}) = -\Phi(P)Q \\
\beta B'PB\Phi(P) - I_m &= (\beta B'PB - \Phi(P)^{-1})\Phi(P) = -Q\Phi(P),
\end{aligned}$$

we may write (74) as

$$\begin{aligned}
eq2.2 + eq2.4 + eq2.8 &= -W\Phi(P)W' \\
eq2.5 + eq2.9 &= -\beta A'PB\Phi(P)W' \\
eq2.6 + eq2.10 &= -W\Phi(P)\beta B'PA.
\end{aligned}$$

It follows that

$$\begin{aligned}
eq2 &= eq2.1 + eq2.3 + eq2.7 + (eq2.2 + eq2.4 + eq2.8) \\
&\quad +(eq2.5 + eq2.9) + (eq2.6 + eq2.10) \\
&= eq1.1 + eq1.2 + eq1.3 + eq1.4 + eq1.5 + eq1.6 = eq1,
\end{aligned}$$

thus establishing item 1.

To demonstrate item 2, compute

$$\begin{aligned}
& \Omega(P')'P\Omega(P)+\hat{F}(P')'Q\hat{F}(P)+\hat{R} \\
= \;& \hat{A}'P\hat{A}+\hat{F}(P')'\left(Q+\hat{B}'P\hat{B}\right)\hat{F}(P)-\hat{F}(P')'\hat{B}'P\hat{A}-\hat{A}'P\hat{B}\hat{F}(P)+\hat{R} \\
= \;& \hat{A}'P\hat{A}+\hat{A}'P\hat{B}\left(Q+\hat{B}'P\hat{B}\right)^{-1}\hat{B}'P\hat{A}-\hat{A}'P\hat{B}(Q+\hat{B}'P\hat{B})^{-1}\hat{B}'P\hat{A} \\
& -\hat{A}'P\hat{B}\left(Q+\hat{B}'P\hat{B}\right)^{-1}\hat{B}'P\hat{A}+\hat{R} \\
= \;& \hat{R}+\hat{A}'P\hat{A}-\hat{A}'P\hat{B}\left(Q+B'P\hat{B}\right)^{-1}\hat{B}'P\hat{A}=T(P).
\end{aligned}$$

where the last equality holds by item 1. ■

Proof of Theorem 1. The proof involves connecting the T-map to the maximization problem (7). For notational simplicity, we consider the equivalent minimization problem

$$\begin{aligned}
V^*(x_0)=\min \quad & E_0\sum\beta^t\left(x_t'Rx_t+u_t'Qu_t+2x_t'Wu_t\right) \\
s.t. \quad & x_{t+1}=Ax_t+Bu_t+C\varepsilon_{t+1},
\end{aligned}$$

where we are abusing notation by also denoted by V^* this problem's value function.

We begin with a well-known series computation that will facilitate our work:

Lemma 4 *If M and N are $n\times n$ matrices and if the eigenvalues of M have modulus less than one then there is a matrix S such that*

$$S=\sum_{t\geq 0}\left(M'\right)^t N\left(M\right)^t.$$

Furthermore, S satisfies $S=M'SM+N$.

Proof. Since the eigenvalues of M have modulus less than one, there is a unique solution to the linear system $S=M'SM+N$

$$vec(S)=(I_n-M'\otimes M')^{-1}vec(N).$$

Now let $S_0=N$ and $S_n=M'S_{n-1}M+N$. Then

$$S_n=\sum_{t=0}^{n}\left(M'\right)^t N\left(M\right)^t.$$

We compute

$$\begin{aligned}
S-S_n &= M'SM+N-(M'S_{n-1}M+N) \\
&= M'(S-S_{n-1})M=\ldots=(M')^n(S-N)(M)^n\to 0,
\end{aligned}$$

where the ellipses indicate induction. ■

The now proceed with the proof of Theorem 1, which involves a series of steps.

Step 1: We first consider a finite horizon problem. Recall that for P an $n \times n$, symmetric, positive semi-definite matrix, define

$$V^P(x) = \min_u x'\hat{R}x + u'Qu + (\hat{A}x + \hat{B}u)'P(\hat{A}x + \hat{B}u).$$

By Lemmas 1 and 3, this problem is solved by $u = -\hat{F}(P)x$, so that $V^P(x) = x'T(P)x$. Now consider the finite horizon problem

$$\begin{aligned} V_N(x) &= \min \sum_{t=0}^{N-1} \left(x_t'\hat{R}x_t + u_t'Qu_t \right) \\ x_{t+1} &= \hat{A}x_t + \hat{B}u_t, \ x_0 = x. \end{aligned} \tag{75}$$

We note that (75) is the finite horizon version of the transformed problem (18). Notice that $V_1(x) = x'T(0)x$. Now we induct on N:

$$\begin{aligned} V_N(x) &= \min \sum_{t=0}^{N-1} \left(x_t'\hat{R}x_t + u_t'Qu_t \right) \\ &= \min \left(x_0'\hat{R}x_0 + u_0'Qu_0 + \sum_{t=1}^{N-1} \left(x_t'\hat{R}x_t + u_t'Qu_t \right) \right) \\ &= \min \left(x'\hat{R}x + u_0'Qu_0 + (\hat{A}x + \hat{B}u_0)'T^{N-1}(0)(\hat{A}x + \hat{B}u_0) \right) = x'T^N(0)x, \end{aligned}$$

where here it is implicitly assumed that the transition $x_{t+1} = \hat{A}x_t + \hat{B}u_t$ is satisfied. Thus $T^N(0)$ identifies the value function for the finite horizon problem (75).

Step 2: We claim that there is an $n \times n$, symmetric, positive semi-definite matrix $\hat{P}$ satisfying $x'T^N(0)x \leq x'\hat{P}x$ for all x. To see this, let F be any matrix that stabilizes $\left(\hat{A}, \hat{B}\right)$. Let $\{\tilde{x}_t\}$ be the state sequence generated by the usual transition equation (with $\tilde{x}_0 = x$) and the policy $u = -Fx$. Then

$$\begin{aligned} V_N(x) &= x'T^N(0)x \leq \sum_{t=0}^{N-1} \tilde{x}_t' \left(\hat{R} + F'QF \right) \tilde{x}_t \\ &= x' \left(\sum_{t=0}^{N-1} \left(\hat{A}' - F\hat{B}' \right)^t \left(\hat{R} + F'QF \right) \left(\hat{A} - \hat{B}F \right)^t \right) x \\ &\leq x' \left(\sum_{t=0}^{\infty} \left(\hat{A}' - F\hat{B}' \right)^t \left(\hat{R} + F'QF \right) \left(\hat{A} - \hat{B}F \right)^t \right) x \equiv x'\hat{P}x, \end{aligned}$$

where the convergence of the series is guaranteed by Lemma 4.

Step 3: We claim $x'T^N(0)x \leq x'T^{N+1}(0)x$. To see this, let $\{x_t^N, u_t^N\}$ solve (75). Then

$$\begin{aligned} x'T^N(0)x &\leq \sum_{t=0}^{N-1}\left(\left(x_t^{N+1}\right)'\hat{R}\left(x_t^{N+1}\right)+\left(u_t^{N+1}\right)'Q\left(u_t^{N+1}\right)\right) \\ &\leq \sum_{t=0}^{N}\left(\left(x_t^{N+1}\right)'\hat{R}\left(x_t^{N+1}\right)+\left(u_t^{N+1}\right)'Q\left(u_t^{N+1}\right)\right) = x'T^{N+1}(0)x. \end{aligned}$$

Step 4: We show that there is an $n\times n$, symmetric, positive semi-definite matrix P^* so that $T^N(0)\to P^*$. Let e_i be the usual n-dimensional coordinate vector, and notice that if M is any $n\times n$ matrix then $e_i'Me_i = M_{ii}$. By steps 2 and 3, $e_i'T^N(0)e_i$ is an increasing sequence bounded above by $\hat{P}_{ii}$, and thus converges. More generally, if v is an n-dimensional vector with 1s in the $i_1,\cdots,i_r$ entries and zeros elsewhere then

$$v'Mv = \sum_{j=1}^{r}\sum_{k=1}^{r} M_{i_j i_k}. \tag{76}$$

By steps 2 and 3, $e_i'T^N(0)e_i$ is an increasing sequence bounded above by $\hat{P}_{ii}$, and thus converges. Now let $v=(1,1,0,\cdots,0)$. Then by (76)

$$v'T^N(0)v = T^N(0)_{11}+T^N(0)_{22}+T^N(0)_{12}+T^N(0)_{21} = T^N(0)_{11}+T^N(0)_{22}+2T^N(0)_{12},$$

where the second equality follows from the fact that T preserves symmetry. Since $v'T^N(0)v$ is increasing and bounded above by $v'\hat{P}v$ it follows that the sum $T^N(0)_{11}+T^N(0)_{22}+2T^N(0)_{12}$ converges. Since the diagonal elements have already been shown to converge, we conclude that $T^N(0)_{12}$ converges. Continuing to work in this way with (76) we conclude that $T^N(0)$ converges to a symmetric matrix P^*. That P^* is positive semi-definite follows from the fact that the zero matrix is positive semi-definite, and by part 2 of Lemma 3, the T-map preserves this property: thus $x'P^*x = \lim x'T^N(0)x \geq 0$.

Step 5: To see that $T(P^*)=P^*$, we work as follows: Let

$$T(P^*) = T\left(\lim_{N\to\infty} T^N(0)\right) = \lim_{N\to\infty} T\left(T^N(0)\right) = \lim_{N\to\infty} T^{N+1}(0) = P^*,$$

where the second equality follows from the continuity of T.

Step 6: We claim

$$\Omega(P) = \beta^{1/2}A - \beta^{1/2}B(Q+\beta B'PB)^{-1}(\beta B'PA+W'). \tag{77}$$

To see this, simply recall that $\Omega(P) = \hat{A}-\hat{B}\hat{F}(P)$. The equation (77) follows immediately from the definitions of $\hat{A},\hat{B}$, and $\hat{F}$.

Step 7: We turn to the stability of $D(T_v)(vec(P^*))$. Using

$$T(P) = R + \beta A'PA - \Psi(P')'\Phi(P)\Psi(P),$$

and noting that

$$\begin{aligned} d\Phi(P) &= -\beta\Phi(P)B'dPB\Phi(P) \text{ and} \\ d\left(\Psi(P')'\right) &= \left(d\Psi(P')\right)' = \left(\beta B'dP'A\right)' = \beta A'dPB, \end{aligned}$$

we have

$$\begin{aligned} dT &= \beta^{1/2}A'dPA\beta^{1/2} - \beta^{1/2}A'dPB\Phi(P)\Psi(P)\beta^{1/2} \\ &+ \beta^{1/2}\Psi(P')'\Phi(P)B'dPB\Phi(P)\Psi(P)\beta^{1/2} - \beta^{1/2}\Psi(P')'\Phi(P)B'dPA\beta^{1/2} \\ &= (\beta^{1/2}A' - \beta^{1/2}\Psi(P')'\Phi(P)B')dP(\beta^{1/2}A - \beta^{1/2}B\Phi(P)\Psi(P)) = \Omega(P')'dP\Omega(P), \end{aligned}$$

where the last equality follows from step 6. Using $vec(XYZ) = (Z' \otimes X)\, vec(Y)$ for conformable matrices X, Y, Z it follows that[35] $vec(dT) = (\Omega(P)' \otimes \Omega(P')')\, vec(dP)$ or $D(T_v)(vec(P)) = \Omega(P)' \otimes \Omega(P')'$.

We are interested in computing the eigenvalues of $D(T_v)(vec(P^*))$. Since P^* is symmetric, since the eigenvalues of $\Omega(P^*)'$ are the same as the eigenvalues of $\Omega(P^*)$, and since the eigenvalues of the Kronecker product are the products of the eigenvalues, it suffices to show that any eigenvalue μ of $\Omega(P^*)$ is strictly inside the unit circle. We will follow the elegant proof of Anderson and Moore (1979). We work by contradiction. Let $v \neq 0$ satisfy $\Omega(P^*)v = \mu v$ and assume $|\mu| \geq 1$. Since P^* is a symmetric fixed point of T, we may use item 2 of Lemma 3 to write

$$P^* = \Omega(P^*)'P^*\Omega(P^*) + \hat{F}(P^*)'Q\hat{F}(P^*) + \hat{R}. \tag{78}$$

Acting on the left of (78) by v' and on the right by v, and exploiting $\Omega(P^*)v = \mu v$, we get[36]

$$(1 - |\mu|^2)v'P^*v = v'\hat{F}(P^*)'Q\hat{F}(P^*)v + v'\hat{D}\hat{D}'v,$$

where we recall that $\hat{R} = \hat{D}\hat{D}'$. Since P^* is positive semi-definite, the left-hand-side is non-positive and the right-hand-side is non-negative, so that both sides must be zero, and thus both terms on the right-hand-side must be zero as well. Since Q is positive definite, $\hat{F}(P^*)v = 0$. Since $\Omega(P^*) = \hat{A} - \hat{B}\hat{F}(P^*)$, it follows that $\Omega(P^*)v = \hat{A}v = \mu v$, i.e. v is an eigenvector of $\hat{A}$. Also, since $\hat{D}$ has full rank, $v'\hat{D}\hat{D}'v = 0$ implies $\hat{D}'v = 0$. But by assumption, $\left(\hat{A}, \hat{D}\right)$ is detectable, which means $|\mu| < 1$, thus yielding the desired contradiction, and step 7 is established.

[35] This result, as well as the result that the eigenvalues of $X \otimes Y$ where X and Y are square matrices is equal to the products of the eigenvalues of X and the eigenvalues of Y, can be found in Section 5.7 of Evans and Honkapohja (2001) and Chapters 8 and 9 of Magnus and Neudecker (1988)

[36] If v is complex, then v' is taken to be the conjugate transpose of v.

Step 8: Now let P_0 be an $n \times n$, symmetric, positive semi-definite matrix. We want to show $T^N(P_0) \to P^*$. An argument analogous to that provided in step 1 shows

$$\begin{aligned} x'T^N(P_0)x &= \min\left(x_N' P_0 x_N + \sum_{t=0}^{N-1}\left(x_t'\hat{R}x_t + u_t'Qu_t\right)\right) \\ x_{t+1} &= \hat{A}x_t + \hat{B}u_t,\ x_0 = x \end{aligned}$$

Because $x_N' P_0 x_N \geq 0$, it follows that $x'T^N(0)x \leq x'T^N(P_0)x$.

Next consider the policy function $u = -\hat{F}(P^*)x$. By Remark **??**, the corresponding state dynamics is given by $x_t = \Omega(P^*)x_0$. It follows that $x'T^N(P_0)x \leq x'G_N(P_0)x$, where $G_N(P_0)$ measures the value of the policy $\hat{F}(P^*)$, that is,

$$G_N(P_0) = (\Omega(P^*)')^N P_0 (\Omega(P^*))^N + \sum_{t=0}^{N-1}\left((\Omega(P^*)')^t\right)\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)(\Omega(P^*))^t.$$

Since $\Omega(P^*)$ is stable, $(\Omega(P^*)')^N P_0 (\Omega(P^*))^N \to 0$, and by Lemma 4

$$\sum_{t=0}^{N-1}\left((\Omega(P^*)')^t\right)\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)(\Omega(P^*))^t \to P^*.$$

Since $x'T^N(0)x \leq x'T^N(P_0)x \leq x'G_N(P_0)x$, and since both $G_N(P_0)$ and $T^N(0)$ converge to P^*, step 8 is complete.

Step 9: With this step we show uniqueness, thus completing the proof. Suppose $\tilde{P}$ is a symmetric positive definite fixed point of the T-map. Then $\tilde{P} = T(\tilde{P}) = T^N(\tilde{P}) \to P^*$. Thus $\tilde{P} = P^*$. ■

Proof of Corollary 1: First notice that $T_v^\varepsilon = T_v - \frac{\beta}{1-\beta}\Delta_v$. Since $\Delta(P) = \delta(P) \oplus 0_{n-1\times n-1}$ and $\partial\delta(P)/\partial P_{11} = 0$ it follows that

$$\text{eig} \circ DT_v^\varepsilon\left(vec(P)\right) = \text{eig} \circ DT_v\left(vec(P)\right). \tag{79}$$

Next, notice that, by (73), $\partial T(P)_{11}/\partial P_{11} = \beta$ and $\partial T(P)_{ij}/\partial P_{11} = 0$ for i, j not both one. It follows that for $w \in \mathbb{R}^{n^2}$,

$$DT_v(w)_{i1} = \begin{cases} \beta & i = 1 \\ 0 & i > 1 \end{cases}.$$

We conclude that if $\Upsilon = \upsilon \oplus 0_{n-1\times n-1}$ then

$$\text{eig} \circ DT_v\left(vec(P + \Upsilon)\right) = \text{eig} \circ DT_v\left(vec(P)\right),$$

which shows that

$$\text{eig} \circ DT_v\left(vec(P_\varepsilon^*)\right) = \text{eig} \circ DT_v\left(vec(P^*)\right).$$

The result follows from (79) and Lemma 2. ■

Discussion of Theorem 2: See Appendix B.

Proof of Theorem 3. It is straightforward to show that $T^{SP}(H) = -2T\left(-\frac{H}{2}\right)$. It follows that

$$T_v^{SP}(vec(H)) = -2T_v\left(vec\left(-\frac{H}{2}\right)\right).$$

By the chain rule,

$$D\left(T_v^{SP}\right)(vec(H^*)) = D\left(T_v\right)\left(vec\left(-\frac{H^*}{2}\right)\right) = D\left(T_v\right)(vec(P^*)),$$

and the result follows from Theorem 1. ■

Discussion of LQ.RTL. We recall the statement of LQ.RTL for convenience:

LQ.RTL The eigenvalues of $A+BF(H^*, A, B)$ not corresponding to the constant term have modulus less than one, and the associated asymptotic second-moment matrix for the process $x_t = (A + BF(H^*, A, B))\, x_{t-1} + C\varepsilon_t$ is non-singular.

As noted in the body, the eigenvalue condition, which provides for the asymptotic stationarity of regressors, allows us to avoid the complex issues involving econometric analysis of explosive regressors. The non-singularity of the asymptotic second-moment matrix is needed so that the regressors remain individually informative; otherwise, asymptotic multicollinearity would destabilize the learning algorithm.

It is interesting to note that a natural condition sufficient to guarantee the needed non-singularity may be stated in terms of "controllability" which is dual to stabilizability. To make the connection, consider the stationary process

$$z_t = \Xi z_{t-1} + \vartheta\varepsilon_t,$$

where $\varepsilon_t \in \mathbb{R}^k$ is i.i.d., zero-mean with invertible variance-covariance matrix σ_ε^2, and so that Ξ stable (roots strictly inside unit circle). We may write

$$z_t = \sum_{m\geq 0}^{m} \Xi^m \vartheta \varepsilon_{t-m},$$

so that the variance of z_t is given by

$$\Omega_z = Ez_t z_t' = \sum_{m\geq 0}^{m} \Xi^m \vartheta \sigma_\varepsilon^2 \vartheta' (\Xi')^m.$$

The following definition captures the condition needed for non-singularity of Ω_z.

Definition. The matrix pair (Ξ, ϑ) is *controllable* provided

$$\bigcap_{m \geq 0} \ker \vartheta'(\Xi')^m = \{0\}. \tag{80}$$

It can be shown that (Ξ, ϑ) is controllable provided

$$\operatorname{rank}(\vartheta, \Xi\vartheta, \Xi^2\vartheta, \ldots, \Xi^{n-1}\vartheta) = n.$$

In this way, controllability acts as a mixing condition, guaranteeing that the variation in ε_t is transmitted across the full span of the state space.

Proposition 2 *If* (Ξ, ϑ) *is controllable then* $\det \Omega_z \neq 0$.

Proof. Since Ω_z is symmetric, positive semi-definite it suffices to show that for any non-zero $v \in \mathbb{R}^n$ we have $v'\Omega_z v > 0$. Now notice that for any $p > 0$,

$$v'\Omega_z v \geq \sum_{m=0}^{p} v'\Xi^m \vartheta \sigma_\varepsilon^2 \vartheta'(\Xi')^m v.$$

Since σ_ε^2 is positive definite, to show the RHS is strictly positive, it suffices to show that there is a p so that

$$\sum_{m=0}^{p} \vartheta'(\Xi')^m v \neq 0.$$

Let p be the least positive integer so that $\vartheta'(\Xi')^p v \neq 0$. Such a p exists because (Ξ, ϑ) is controllable. Then

$$\sum_{m=0}^{p} \vartheta'(\Xi')^m v = \vartheta'(\Xi')^p v \neq 0,$$

and the result follows. ■

We remark that the stated condition also guarantees the second moment matrix of the corresponding VAR(1) is non-singular, and further, the result holds in case the VAR includes a constant term.

Proof of Theorem 4. The proof involves using the theory of stochastic recursive algorithms to show that the asymptotic behavior of our system is governed by the Lyapunov stability of the differential system

$$\frac{dH}{d\tau} = T^{SP}(H, A, B) - H.$$

The proof is then completed by appealing to Theorem 3**.**

Recall the dynamic system under consideration:

$$\begin{aligned} x_t &= Ax_{t-1} + Bu_{t-1} + C\varepsilon_t \\ \mathcal{R}_t &= \mathcal{R}_{t-1} + \gamma_t \left(x_t x_t' - \mathcal{R}_{t-1} \right) \\ H_t' &= H_{t-1}' + \gamma_t \mathcal{R}_{t-1}^{-1} x_{t-1} \left(\lambda_{t-1} - H_{t-1} x_{t-1} \right)' \\ A_t' &= A_{t-1}' + \gamma_t \mathcal{R}_{t-1}^{-1} x_{t-1} \left(x_t - Bu_{t-1} - A_{t-1} x_{t-1} \right)' \\ u_t &= F^{SP}(H_t, A_t, B) x_t \\ \lambda_t &= T^{SP}(H_t, A_t, B) x_t \\ \gamma_t &= \kappa (t+N)^{-\vartheta}, \end{aligned} \tag{81}$$

where $0 < \vartheta, \kappa \leq 1$ and N is a non-negative integer. To apply the theory of stochastic recursive algorithms, we must place our system in the following form:

$$\theta_t = \theta_{t-1} + \gamma_t \mathcal{H}(\theta_{t-1}, X_t) \tag{82}$$

$$X_t = \mathcal{A}(\theta_{t-1}) X_{t-1} + \mathcal{B}(\theta_{t-1}) \hat{\varepsilon}_t, \tag{83}$$

where $\hat{\varepsilon}_t$ is *iid*, but not necessarily mean zero, and $\mathcal{A}(\theta)$ must be a stable matrix, i.e. has roots strictly inside the unit circle. Here $\theta \in \mathbb{R}^M$ for some M. For extensive details on the asymptotic theory of recursive algorithms such as this, see Chapter 6 of Evans and Honkapohja (2001).[37] Note that $\mathcal{R}_t$ and A_t' are matrices and θ is a column vector. Therefore, we identify the space of matrices with $\mathbb{R}^M$ for appropriate M in the usual way using the *vec* operator.

To put (81) in the form (82)-(83) we define

$$\theta_t = \begin{pmatrix} \text{vec}(\mathcal{R}_t) \\ \text{vec}(H_t') \\ \text{vec}(A_t') \end{pmatrix}, \; X_t = \begin{pmatrix} x_t \\ x_{t-1} \\ u_{t-1} \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ \varepsilon_t \end{pmatrix}.$$

Note that $\theta \in \mathbb{R}^M$ for $M = 3n^2$.

For matrix N define $(N)^\bigstar$ as the matrix obtained from N by replacing its first row by a row of zeros. Note that, by LQ.RTL, the eigenvalues of $(A + BF(H^*, A, B))^\bigstar$ are all less than one in modulus. Fixing an estimate θ, define the matrices $\mathcal{A}(\theta)$ and $\mathcal{B}(\theta)$ as follows:

$$\mathcal{A}(\theta) = \begin{pmatrix} \left(A + BF(H, \tilde{A}, B) \right)^\bigstar & 0 & 0 \\ I_n & 0 & 0 \\ F(H, \tilde{A}, B) & 0 & 0 \end{pmatrix} \text{ and } \mathcal{B}(\theta) = \begin{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} & C \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

[37] Other key references are Ljung (1977) and Marcet and Sargent (1989).

where $H, \tilde{A}$ are the matrices corresponding to the relevant components of θ. Recall that the first row of C is zeros. Finally let $\mathcal{R}^*$ be the asymptotic second-moment matrix for the state x_t under optimal decision-making, i.e.

$$\mathcal{R}^* = \lim_{t\to\infty} E_0 x_t x_t' \text{ where } x_t = \left(A + BF(H^*, A, B)\right) x_{t-1} + C\varepsilon_t.$$

We now restrict attention to an open set $W \subset \mathbb{R}^M$ such that whenever $\theta \in W$ it follows that T^{SP} is well-defined, i.e. $\det(2Q - \beta B'HB) \neq 0$, $\mathcal{R}^{-1}$ exists, and the eigenvalues of $\mathcal{A}(\theta)$ have modulus strictly less than one. The existence of such a set is guaranteed by: (i) LQ.1 and Theorem 2, which together with $H^* = -2P^*$ imply that $\det(2Q - \beta B'HB) \neq 0$ for H near H^*, and (ii) LQ.RTL, which implies that $\mathcal{R}$ is invertible near $\mathcal{R}^*$.

Given $\theta \in W$, define $\bar{X}_t(\theta)$ as the stochastic process $\bar{X}_t(\theta) = \mathcal{A}(\theta)\bar{X}_{t-1}(\theta) + \mathcal{B}(\theta)\varepsilon_t$. Let $\bar{x}_t(\theta)$ denote the first n components of $\bar{X}_t(\theta)$ and $\bar{u}_{t-1}(\theta)$ denote the last m components of $\bar{X}_t(\theta)$. With the restriction on W the following limit is well defined:

$$\mathcal{N}(H, \tilde{A}, B) = \lim_{t\to\infty} E_0 \bar{x}_t(\theta)\bar{x}_t(\theta)'.$$

Set

$$\theta^* = \begin{pmatrix} \text{vec}(\mathcal{N}(H^*, A, B)) \\ \text{vec}(H^*) \end{pmatrix},$$

where clearly $\theta^* \in W$.

We now write the recursion (81) in the form (82)-(83). To this end, we define the function $\mathcal{H}(\cdot, X) : W \to \mathbb{R}^M$ component-wise as follows:

$$\begin{aligned} \mathcal{H}^1(\theta_{t-1}, X_t) &= \text{vec}\left(x_t x_t' - \mathcal{R}_{t-1}\right) \\ \mathcal{H}^2(\theta_{t-1}, X_t) &= \text{vec}\left(\mathcal{R}_{t-1}^{-1} x_{t-1}\left((T^{SP}(H_{t-1}, A_{t-1}, B) - H_{t-1})x_{t-1}\right)'\right). \end{aligned}$$

The theory of stochastic recursive algorithms tells us to consider the function $h : W \to \mathbb{R}^M$ defined by

$$h(\theta) = \lim_{t\to\infty} E_0 \mathcal{H}(\theta, \bar{X}_t(\theta)),$$

where existence of this limit is guaranteed by our restrictions on W. The function h has components

$$h^1(\theta) = \text{vec}\left(\mathcal{N}(\theta) - \mathcal{R}\right) \tag{84}$$

$$h^2(\theta) = \text{vec}\left(\mathcal{R}^{-1}\mathcal{N}(\theta)\left(T^{SP}(H, \tilde{A}, B) - H\right)'\right). \tag{85}$$

and captures the long-run expected behavior of θ_t.

We will apply Theorem 4 of Ljung (1977), which directs attention to the ordinary differential equation $\dot{\theta} = h(\theta)$, i.e. $d\theta/d\tau = h(\theta)$, where τ denotes notational time. Notice that

$h(\theta^*) = 0$, so that θ^* is a fixed point of this differential equation. Ljung's theorem tell us that, under certain conditions that we will verify, if θ^* is a Lyapunov stable fixed point, then our learning algorithm will converge to it almost surely. The determination of Lyapunov stability for the system $\dot{\theta} = h(\theta)$ involves simply computing the derivative of h and studying its eigenvalues: if the real parts of these eigenvalues are negative then the fixed point is Lyapunov stable. Computation of the derivative of h at θ^* is accomplished by observing that the term multiplying $\mathcal{N}(\theta)$ in equation (85) is zero when evaluated at θ^* so that, by the product rule, the associated derivative is zero. The resulting block diagonal form of the derivative of h yields repeated eigenvalues that are -1 and the eigenvalues of $\partial h^2/\partial \text{vec}(H')'$, which have negative real part by Theorem 3. It follows that θ^* is a Lyapunov stable fixed point of $\dot{\theta} = h(\theta)$.

To complete the proof of Theorem 4, we must verify the conditions of Ljung's Theorem and augment the algorithm (81) with a projection facility. First we address the regularity conditions on the algorithm. Because ε_t has compact support, we apply Theorem 4 of Ljung (1977) using his assumptions A. Let the set D be the intersection of W with the basin of attraction of θ^* under the dynamics $\dot{\theta} = h(\theta)$. Note that D is both open and path-connected. Let D_R be a bounded open connected subset of D containing θ^* such that its closure is also in D. We note that for fixed X, $\mathcal{H}(\theta, X)$ is continuously differentiable (with respect to θ) on D_R, and for fixed $\theta \in D_R$, $\mathcal{H}(\theta, X)$ is continuously differentiable with respect to X. Furthermore, on D_R the matrix functions $\mathcal{A}(\theta)$ and $\mathcal{B}(\theta)$ are continuously differentiable. Since the closure of D_R is compact, it follows from Coddington (1961), Theorem 1 of Ch. 6, that $\mathcal{A}(\theta)$ and $\mathcal{B}(\theta)$ are Lipschitz continuous on D_R. The rest of assumptions A are immediate given the gain sequence γ_t.

We now turn to the projection facility. Because θ^* is Lyapunov stable there exists an associated Lyapunov function $U : D \to \mathbb{R}_+$. (See Theorem 11.1 of Krasovskii (1963) or Proposition 5.9, p. 98, of Evans and Honkapohja (2001).) For $c > 0$, define the notation

$$K(c) = \{\theta \in D : U(\theta) \leq c\}.$$

Pick $c_1 > 0$ such that $K(c_1) \subset D_R$, and let $D_1 = \text{int}\,(K(c_1))$. Pick $c_2 < c_1$, and let $D_2 = K(c_2)$. Let $\bar{\theta}_t \in D_2$. Define a new recursive algorithm for θ_t as follows:

$$\theta_t = \begin{cases} \hat{\theta}_t = \theta_{t-1} + \gamma_t \mathcal{H}(\theta_{t-1}, X_t) & \text{if} \quad \hat{\theta}_t \in D_1 \\ \bar{\theta}_t & \text{if} \quad \hat{\theta}_t \notin D_1 \end{cases}.$$

With these definitions, Theorem 4 of Ljung applies and shows that $\theta_t \to \theta^*$ almost surely. ■

We remark that if ε_t does not have compact support but has finite absolute moments then it is possible to use Ljung's assumptions B or the results presented on p. 123 – 125 of Evans and Honkapohja (2001) and Corollary 6.8 on page 136.

Appendix B: Details on Euler-equation Learning

As noted in Section 4.2.2, if $A_{22} \neq 0$ it may still be possible to derive a one-step-ahead FOC that may interpreted as an Euler equation. The particular example we considered in that section held that $\det B_2 \neq 0$, but a more general result is available through the use of *transformations.* In this Appendix, we begin with a discussion of transformations, and then prove a general result that will yield Theorem 6 as a corollary.

Recall the LQ-problem under consideration:

$$V^*(x_0) = \max \quad -\sum_t \beta^t \left(x_t' R x_t + u_t' Q u_t + 2 x_t' W u_t\right) \tag{86}$$

$$s.t. \quad x_{t+1} = A x_t + B u_t + C \varepsilon_{t+1}. \tag{87}$$

The usual dimensions of controls and states (and corresponding matrices) apply. We say that the 5-tuple (R, Q, W, A, B) satisfies LQ.1 – LQ.3 provided that R is positive semi-definite, Q is positive definite, and the transformed matrices $\hat{A} = \beta^{1/2}(A - BQ^{-1}W')$, $\hat{B} = \beta^{1/2}B$ and $\hat{R} = R - WQ^{-1}W'$ satisfy LQ.1 – LQ.3.

A *transformation* of problem (86) is a linear isomorphism $\mathcal{S} : \mathbb{R}^n \oplus \mathbb{R}^m \to \mathbb{R}^n \oplus \mathbb{R}^m$ of the form

$$\mathcal{S} = \begin{pmatrix} I_n & 0 \\ \mathcal{S}_{21} & \pm I_m \end{pmatrix}.$$

The idea is to use this transformation to change the notions of states and controls:

$$\begin{pmatrix} \breve{x}_t \\ \breve{u}_t \end{pmatrix} = \mathcal{S} \begin{pmatrix} x_t \\ u_t \end{pmatrix}. \tag{88}$$

We require $\mathcal{S}_{12} = 0$ so that the transformed problem remains recursive, i.e. so that the new (pre-determined) state $\breve{x}_t$ does not depend on the new control $\breve{u}_t$.

Now write

$$\begin{aligned} x_t' R x_t + u_t' Q u_t + 2 x_t' W u_t &= \begin{pmatrix} x_t' & u_t' \end{pmatrix} \begin{pmatrix} R & W \\ W' & Q \end{pmatrix} \begin{pmatrix} x_t \\ u_t \end{pmatrix} \\ &= \begin{pmatrix} \breve{x}_t' & \breve{u}_t' \end{pmatrix} \left(\mathcal{S}^{-1}\right)' \begin{pmatrix} R & W \\ W' & Q \end{pmatrix} \mathcal{S}^{-1} \begin{pmatrix} \breve{x}_t \\ \breve{u}_t \end{pmatrix} \\ &= \breve{x}_t' \breve{R} \breve{x}_t + \breve{u}_t' \breve{Q} \breve{u}_t + 2 \breve{x}_t' \breve{W} \breve{u}_t, \end{aligned} \tag{89}$$

and

$$\begin{aligned} \breve{x}_{t+1} &= x_{t+1} = \begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} x_t \\ u_t \end{pmatrix} + C \varepsilon_{t+1} \\ &= \begin{pmatrix} A & B \end{pmatrix} \mathcal{S}^{-1} \begin{pmatrix} \breve{x}_t \\ \breve{u}_t \end{pmatrix} + C \varepsilon_{t+1} = \breve{A} \breve{x}_t + \breve{B} \breve{u}_t + C \varepsilon_{t+1}, \end{aligned} \tag{90}$$

where the last equalities of each displayed equation provide notation.

A transformation $\mathcal{S}$ provides a formal mechanism though which a problem's states and controls may be modified in order to facilitate analysis. The following example helps motivate the general approach. Consider again the original problem (86). Let

$$\tilde{\mathcal{S}}(W,Q) = \begin{pmatrix} I_n & 0_{n\times m} \\ Q^{-1}W' & I_m \end{pmatrix}.$$

This transformation removes the interaction term in the objective: $\breve{W} = 0$. It provides the transformation used in our paper to define LQ.1 – LQ.3, except that here we are not eliminating discounting.

We now use $\mathcal{S}$ to transform problem (86). Specifically, the corresponding transformed problem is given by

$$\breve{V}^*(\breve{x}_0) = \max \qquad -E_0 \sum_t \beta^t \left(\breve{x}_t' \breve{R} \breve{x}_t + \breve{u}_t' \breve{Q} \breve{u}_t + 2\breve{x}_t' \breve{W} \breve{u}_t \right) \tag{91}$$

$$s.t. \qquad \breve{x}_{t+1} = \breve{A}\breve{x}_t + \breve{B}\breve{u}_t + C\varepsilon_{t+1}. \tag{92}$$

We have the following theorem, which provides the sense in which the original problem and its associated transform are equivalent.

Lemma 5 *Given (R,Q,W,A,B) and transform $\mathcal{S}: \mathbb{R}^n \oplus \mathbb{R}^m \to \mathbb{R}^n \oplus \mathbb{R}^m$, let $(\breve{R},\breve{Q},\breve{W},\breve{A},\breve{B})$ be as determined by (89) and (90). Then*

1. *The 5-tuple (R,Q,W,A,B) satisfies LQ.1 – LQ.3 if and only if the 5-tuple $(\breve{R},\breve{Q},\breve{W},\breve{A},\breve{B})$ satisfies LQ.1 – LQ.3.*
2. *The state-control path (x_t,u_t) solves (86) if and only if the state-control path $(\breve{x}_t,\breve{u}_t)$ solves (91), where (x_t,u_t) and $(\breve{x}_t,\breve{u}_t)$ are related by (88).*

Proof of Lemma 5. To prove part 1 notice that we only need to prove one direction because of the invertibility of the transform. Thus we assume (R,Q,W,A,B) satisfy LQ.1 – LQ.3 and we show $(\breve{R},\breve{Q},\breve{W},\breve{A},\breve{B})$ satisfies LQ.1 – LQ.3. Next, notice that we can redefine the control as its negation: $u_t \to -u_t$ by also sending $W \to -W$ and $B \to -B$: it is immediate that LQ.1 – LQ.3 remain satisfied. Thus we may consider only transforms of the form

$$\mathcal{S} = \begin{pmatrix} I_n & 0 \\ \mathcal{S}_{21} & I_m \end{pmatrix}.$$

Using

$$\mathcal{S}^{-1} = \begin{pmatrix} I_n & 0 \\ -\mathcal{S}_{21} & I_m \end{pmatrix},$$

we may compute directly that $\breve{B} = B, \breve{Q} = Q$, and

$$\begin{aligned}\breve{R} &= R - W\mathcal{S}_{21} - \mathcal{S}_{21}'W' + \mathcal{S}_{21}'Q\mathcal{S}_{21} \\ \breve{W} &= W - \mathcal{S}_{21}'Q. \\ \breve{A} &= A - B\mathcal{S}_{21}\end{aligned}$$

Denote with a tilde matrices associated to the transform of the problem (91) used to examine LQ.1 - LQ.3, and recall that hats identify the corresponding matrices of problem (87). Thus, for example, $\hat{R} = R - WQ^{-1}W$ and $\tilde{R} = \breve{R} - \breve{W}\breve{Q}^{-1}\breve{W}'$. We compute

$$\left(\tilde{\mathcal{S}}(\breve{W}, \breve{Q})^{-1}\right)' \begin{pmatrix} \breve{R} & \breve{W} \\ \breve{W}' & \breve{Q} \end{pmatrix} \tilde{\mathcal{S}}(\breve{W}, \breve{Q})^{-1} = \begin{pmatrix} R - WQ^{-1}W' & 0 \\ 0 & Q \end{pmatrix},$$

which shows that $\tilde{R} = \hat{R}$. Next we find that

$$\begin{aligned}\tilde{A} &= \beta^{1/2}(\breve{A} - BQ^{-1}\breve{W}') \\ &= \beta^{1/2}(A - B\mathcal{S}_{21} - BQ^{-1}(W - \mathcal{S}_{21}'Q)') \\ &= \beta^{1/2}(A - BQ^{-1}W') = \hat{A}.\end{aligned}$$

Since it is immediate that $\tilde{B} = \hat{B}$ and $\tilde{Q} = \hat{Q}$, part 1 is established.

To prove part two, we use the invertibility of $\mathcal{S}$ and the definitions (90) to see that the state-control path (x_t, u_t) is feasible under (87) if and only if the state-control path $(\breve{x}_t, \breve{u}_t)$ is feasible under (92). The proof is completed by observing that equation (89) implies the objective of (86) evaluated at (x_t, u_t) is equal to the objective of (91) evaluated at $(\breve{x}_t, \breve{u}_t)$.■

We now use transforms to identify conditions guaranteeing the existence of one-step-ahead Euler equations. Let (R, Q, W, A, B) characterize a problem and suppose $A_{22} \neq 0$. Suppose further that $\mathcal{S}$ is a transform that yields $(\breve{R}, \breve{Q}, \breve{W}, \breve{A}, \breve{B})$. If $\breve{A}_{22} = 0$ then the optimal path $(\widetilde{x}_t, \widetilde{u}_t)$ satisfies

$$\breve{Q}\breve{u}_t + \breve{W}_t'\breve{x} + \beta\breve{B}'E_t(\breve{R}\breve{x}_{t+1} + \breve{W}\breve{u}_{t+1}) = 0. \tag{93}$$

Inverting the transform, it follows that the optimal path (x_t, u_t) satisfies an FOC of the form

$$\begin{pmatrix} \breve{W}' & \breve{Q} \end{pmatrix} \mathcal{S} \begin{pmatrix} x_t \\ u_t \end{pmatrix} + \beta\breve{B}' \begin{pmatrix} \breve{R} & \breve{W} \end{pmatrix} \mathcal{S}E_t \begin{pmatrix} x_{t+1} \\ u_{t+1} \end{pmatrix} = 0. \tag{94}$$

We interpret equation (94) as an Euler equation on which agents may base their boundedly optimal decision making. In particular, given a PLM of the form $u_t = -Fx_t$, equation (94) may be used to identify the corresponding T-map which determines the ALM. We call label this T-map by T^{EL} just as before because it is a direct generalize of the map identified in Section 4.2.2. In particular, if $A_{22} = 0$ then we may take S to be the identity matrix,

and equation (94) reduces to the usual Euler equation (40). Also, if $A_{22} \neq 0, n_2 = m$ and $\det B_2 \neq 0$ then we may form the transform[38]

$$\mathcal{S} = \begin{pmatrix} I_{n_1} & 0_{n_1 \times n_2} & 0_{n_1 \times m} \\ 0_{n_2 \times n_1} & I_{n_2} & 0_{n_2 \times m} \\ 0_{m \times n_1} & B_2^{-1} A_{22} & I_m \end{pmatrix}.$$

The new control becomes $\breve{u}_t = u_t + B_2^{-1} A_{22} x_{2t}$. We find that

$$\breve{A} = \begin{pmatrix} I_{n_1} & 0_{n_1 \times n_2} \\ A_{21} & 0_{n_2 \times n_2} \end{pmatrix}$$

and

$$\breve{B} = B = \begin{pmatrix} 0_{n_1 \times m} \\ B_2 \end{pmatrix},$$

and we have that $\breve{A}_{22} = 0$. Finally, in this case, equation (94) reproduces the Euler equation (40).

The discussion just provided shows that the following theorem holds Theorem 6 as a special case:

Theorem 7 *Assume LQ.1 – LQ.3 are satisfied and that there is a transform S so that $\breve{A}_{22} = 0$. If agents behave as Euler equation learners with perceptions $u_t = -Fx_t$ and use (41) as their behavioral primitive then F^* is a Lyapunov stable fixed point of the differential equation $dF/d\tau = T^{EL}(F) - F$. That is, F^* is stable under stylized learning.*

Proof of Theorem 7. The proof proceeds in two parts. In part I we assume that $A_{22} = 0$, so that no transform is applied to the matrices. In part II the general case is considered.

Part I. To show $\Phi(R + WF^*) = \Phi(P^*)$ and $\Psi(R + WF^*) = \Psi(P^*)$, we use the Riccati equation we compute

$$\begin{aligned} P^* &= R + \beta A' P^* A - \Psi(P^*)' \Phi(P^*) \Psi(P^*) \\ &= R - W \Phi(P^*) \Psi(P^*) + \beta A' P^* A - \beta A' P^* B \Phi(P^*) \Psi(P^*), \end{aligned}$$

so that

$$P^* = R + WF^* + \beta A' P^* (A + BF^*).$$

Noting that $B'A' = 0$ and using the preceding equation, we conclude that

$$\Psi(P^*) \equiv \beta B' P^* A + W' = \Psi(R + WF^*),$$

and similarly for Φ.

[38] We note that the $\breve{R}$ and $\breve{W}$ produced by this transform will not be the same as in equation (41) because this later equation is written in terms of u_t and not $\breve{u}_t$.

We next compute matrix differentials. At arbitrary (appropriate) F, we have that $dT^{EL} = -(d\Phi \cdot \Psi + \Phi \cdot d\Psi)$, and

$$\begin{aligned} d\Phi &= -\left((Q+\beta B'(R+WF)B)^{-1}\beta B'W\right) \cdot dF \cdot \left(B(Q+\beta B'(R+WF)B)^{-1}\right) \\ d\Psi &= (\beta B'W) \cdot dF \cdot (A). \end{aligned}$$

Evaluated at F^*, we obtain

$$dT^{EL} = (\beta\Phi(P^*)B'W) \cdot dF \cdot (B\Phi(P^*)\Psi(P^*) - A).$$

Applying the vec operator to each side, we obtain

$$D\left(T_v^{EL}\right)(F^*) = -\Omega(P^*)' \otimes \beta^{1/2}\Phi(P^*)B'W,$$

where we recall that

$$\begin{aligned} \Omega(P^*) &= \beta^{1/2}A - \beta^{1/2}B(Q+\beta B'P^*B)^{-1}(\beta B'P^*A + W') \\ &= \beta^{1/2}(A - B\Phi(P^*)\Psi(P^*)). \end{aligned}$$

In the proof of Theorem 1 it is shown that $\text{eig} \circ \Omega(P^*)$ have modulus less than one; thus the proof will be complete if we can show that the eigenvalues of $\beta^{1/2}\Phi(P^*)B'W$ are inside the unit circle. This requires three steps.

Step 1. We show that the eigenvalues of $\beta^{1/2}B\Phi(P^*)\Psi(P^*)$ are inside the unit circle. To see this let $y = -\beta^{1/2}B\Phi(P^*)\Psi(P^*)$ and notice that since the first n_1 rows of B are zeros we have

$$y = \left(\begin{array}{c|c} 0 & 0 \\ \hline y_{21} & y_{22} \end{array}\right).$$

Because $A_{22} = 0$ we conclude that

$$\Omega(P^*) = \beta^{1/2}\left(A - B\Phi(P^*)\Psi(P^*)\right) = \left(\begin{array}{c|c} \beta^{1/2}A_{11} & 0 \\ \hline * & y_{22} \end{array}\right),$$

and step 1 is complete by Theorem 1.

Step 2. Now we show that

$$\text{eig} \circ B\Phi(P^*)\Psi(P^*) = \text{eig} \circ B\Phi(P^*)W'.$$

Let $z = \beta B\Phi(P^*)B'P^*A$ and $\zeta = B\Phi(P^*)W'$, so that $B\Phi(P^*)\Psi(P^*) = z + \zeta$. The structures of A and B imply

$$z = \left(\begin{array}{c|c} 0 & 0 \\ \hline z_{21} & 0 \end{array}\right) \text{ and } \zeta = \left(\begin{array}{c|c} 0 & 0 \\ \hline \zeta_{21} & \zeta_{22} \end{array}\right),$$

and step 2 is follows.

Step 3. Finally, we show that

$$\text{eig} \circ \Phi(P^*)B'W \subset \text{eig} \circ B\Phi(P^*)W'.$$

Since eigenvalues are preserved under transposition, it suffices to show that

$$\text{eig} \circ W'B\Phi(P^*) \subset \text{eig} \circ B\Phi(P^*)W'.$$

To this end, let $\xi = B\Phi(P^*)$, and notice $\xi' = \left(\begin{array}{c|c} 0 & {\xi_2}' \end{array} \right)$. Writing the $n \times m$ matrix W as $W' = \left(\begin{array}{c|c} {W_1}' & {W_2}' \end{array} \right)$, we compute $W'\xi = {W_2}'\xi_2$ and

$$\xi W' = \left(\begin{array}{c|c} 0 & 0 \\ \hline \xi_2 {W_1}' & \xi_2 {W_2}' \end{array} \right).$$

Since ξ_2 and W_2 are $m \times m$ matrices, it follows that $\text{eig} \circ \xi_2 {W_2}' = \text{eig} \circ {W_2}'\xi_2$, and step 3 is complete.[39]

By steps 3 and 2 we have

$$\begin{array}{rcl} \text{eig} \circ \beta^{1/2}\Phi(P^*)B'W & \subset & \text{eig} \circ \beta^{1/2}B\Phi(P^*)W' \\ & = & \text{eig} \circ \beta^{1/2}B\Phi(P^*)\Psi(P^*). \end{array}$$

The result then follows from step 1.

Part II. Now we turn to the case the general case in which $A_{22} \neq 0$ and there is a transform S so that $\breve{A}_{22} = 0$. Denote by $\breve{T}^{EL}$ the T-map that obtains from equation (93) using the PLM $\breve{u}_t = -F\breve{x}_t$, and let $\breve{F}^*$ be the corresponding fixed point. By Lemma 5 and Part I of the proof, we know that $D\breve{T}^{EL}\left(\breve{F}^*\right)$ is a stable matrix. We claim that

$$T^{EL}(F) = \breve{T}^{EL}\left(F - S_{21}\right) + S_{21} \tag{95}$$

$$F^* = \breve{F}^* + S_{21}. \tag{96}$$

To see this, notice that the PLM $u_t = -Fx_t$ corresponds to the PLM $\breve{u}_t = -\left(F - \mathcal{S}_{21}\right)x_t$,where we are using $x_t = \breve{x}_t$. Also the ALM $u_t = -T^{EL}(F)x_t$ implies that

$$\breve{u}_t = -\left(T^{EL}(F) - \mathcal{S}_{21}\right)x_t = -\breve{T}^{EL}\left(F - \mathcal{S}_{21}\right)x_t,$$

where the second equality follows from the construction of $\breve{T}^{EL}$. This demonstrates (95) and (96). It follows that $DT^{EL}\left(F^*\right) = D\breve{T}^{EL}\left(\breve{F}^*\right)$ and the result obtains. ■

[39] If C is a $p \times q$ matix and D is a $q \times p$ matrix then every non-zero eigenvalue of CD is an eigennvalue of DC. To see this suppose $CDv = \mu v$, where μ and v are non-zero. Then $DC(Dv) = \mu Dv$, where Dv is non-zero because CDv is non-zero. Thus μ is a non-zero eigenvalue of DC.

If $q > p$ then the null space of C is nontrivial. Thus zero is an eigenvalue of DC, though it may not be an eigenvalue of CD.

Appendix C: Solving the Quadratic Program

While the LQ-framework developed Section 3.1 is well-understood under various sets of assumptions, we include here, for completeness, precise statements and proofs of the results needed for our work. Here we begin with the deterministic case – the stochastic case, as presented in the next section, relies on the same types of arguments but requires considerably more technical machinery to deal with issues involving measurability. The proofs presented here follow the work of Bertsekas (1987) and Bertsekas and Shreve (1978) with modifications as required to account distinct assumptions.

Solving the LQ-Problem: the Deterministic Case.

The problem under consideration is

$$\min \quad \sum_{t\geq 0} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2 x_t' W u_t\right) \tag{97}$$

$$s.t. \quad x_{t+1} = A x_t + B u_t \tag{98}$$

with x_0 given. Notice that we have shut down the stochastic shock. Also, we are considering the equivalent minimization problem: this is only for notational convenience. We assume LQ.1 – LQ.3 are satisfied.

Formalizing the programming problem

To make formal the problem (97) we define a policy π to be a sequence of functions $\pi_t : \mathbb{R}^n \to \mathbb{R}^m$. For the problems under consideration, optimal policy will be stationary policies, that is, $\pi_t = \pi_s$. However, at a key step in the proof of the Principle of Optimality below, it is helpful to allow for time-varying policies. Set $V_\pi : \mathbb{R}^n \to [0, \infty]$ as

$$\begin{aligned} V_\pi(x) &= \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2 x_t' W u_t\right) \\ x_{t+1} &= A x_t + B u_t \text{ and } u_t = \pi_t(x_t). \end{aligned}$$

By equation (16), and LQ.1 - LQ.3, the summands in the above limit are positive:

$$x' R x + u' Q u + 2 x' W u \geq 0.$$

Thus V_π is well defined. Letting Π be the collection of all policies, we define $V^* : \mathbb{R}^n \to [-\infty, \infty]$ by

$$V^*(x) = \inf_{\pi\in\Pi} V_\pi(x). \tag{99}$$

While V^* is now well defined, more can be said. Specifically,

Lemma 6 $V^*(x) \in [0, \infty)$.

Proof: That $V^*(x) \geq 0$ is immediate. To establish that V^* is finite-valued, choose $\hat{F}$ to stabilize $(\hat{A}, \hat{B})$ and set $F = Q^{-1}W' + \hat{F}$. Then

$$\beta^{\frac{1}{2}}(A - BF) = \beta^{\frac{1}{2}}\left(A - B\left(Q^{-1}W' + \hat{F}\right)\right) = \beta^{\frac{1}{2}}\left(A - BQ^{-1}W'\right) - \beta^{\frac{1}{2}}B\hat{F} = \hat{A} - \hat{B}\hat{F},$$

so that $\beta^{\frac{1}{2}}(A - BF)$ is stable. Let the policy π^F be given by $\pi_i^F(x) = -Fx$. It follows that

$$\begin{aligned} V^*(x) &\leq V_{\pi^F}(x) = x'\left(\sum_{t\geq 0}\beta^t\left(A' - F'B'\right)^t\left(R + F'QF - 2WF\right)(A - BF)^t\right)x \\ &= x'\left(\sum_{t\geq 0}\left(\beta^{\frac{1}{2}}\left(A' - F'B'\right)\right)^t\left(R + F'QF - 2WF\right)\left(\beta^{\frac{1}{2}}(A - BF)\right)^t\right)x < \infty, \end{aligned}$$

where the inequality comes from Lemma 4. ■

We conclude that the solution to the sequence problem (99) is a well-defined, non-negative, real-valued function.

The S-map

Characterization of V^* is provided by the Principle of Optimality. To develop this characterization carefully, we work as follows. As noted above, under LQ.1 – LQ.3 we have that

$$x'Rx + u'Qu + 2x'Wu \geq 0.$$

It follows that given any function $V : \mathbb{R}^n \to [0, \infty)$, we may define $S(V) : \mathbb{R}^n \to [0, \infty)$ by

$$S(V)(x) = \inf_{u\in\mathbb{R}^m}\left(x'Rx + u'Qu + 2x'Wu + \beta V\left(Ax + Bu\right)\right).$$

It is helpful to observe that given V and $\varepsilon > 0$ we can find the stationary policy $\pi_\varepsilon : \mathbb{R}^n \to \mathbb{R}^m$ so that for each $x \in \mathbb{R}^n$, the control choice $u = \pi_\varepsilon(x)$ implies objective $x'Rx + u'Qu + 2x'Wu + \beta V\left(Ax + Bu\right)$ is within ε of the infimum, that is,

$$x'Rx + \pi_\varepsilon(x)'Q\pi_\varepsilon(x) + 2x'W\pi_\varepsilon(x) + \beta V(Ax + B\pi_\varepsilon(x)) \leq S(V)(x) + \varepsilon.$$

The following result is the work-horse lemma for dynamic programming in the deterministic case. The proofs, which be provide for completeness, follow closely Bertsekas (1987) and Bertsekas and Shreve (1978).

Lemma 7 *The map S satisfies the following properties:*

1. *(Monotonicity) If $V_1, V_2 : \mathbb{R}^n \to [0, \infty)$ and $V_1 \leq V_2$ then $S(V_1) \leq S(V_2)$.*
2. *(Principle of Optimality) $V^* = S\left(V^*\right)$.*

3. (Minimality of V^*) If $V : \mathbb{R}^n \to [0, \infty)$ and $S(V) = V$ then $V^* \leq V$.

Proof. To establish item 1, let $\varepsilon > 0$ and choose a stationary policy π_ε so that

$$x'Rx + \pi_\varepsilon(x)'Q\pi_\varepsilon(x) + 2x'W\pi(x) + \beta V_2(Ax + B\pi_\varepsilon(x)) \leq S(V_2)(x) + \varepsilon.$$

Then

$$\begin{aligned} S(V_1)(x) &\leq x'Rx + \pi_\varepsilon(x)'Q\pi_\varepsilon(x) + 2x'W\pi_\varepsilon(x) + \beta V_1(Ax + B\pi_\varepsilon(x)) \\ &\leq x'Rx + \pi_\varepsilon(x)'Q\pi_\varepsilon(x) + 2x'W\pi_\varepsilon(x) + \beta V_2(Ax + B\pi_\varepsilon(x)) \leq S(V_2)(x) + \varepsilon. \end{aligned}$$

Turning to item 2, let π be any policy. Then

$$\begin{aligned} V_\pi(x) &= x'Rx + \pi_0(x)'Q\pi_0(x) + 2x'W\pi_0(x) + \beta V_\pi(Ax + B\pi_0(x)) \\ &\geq x'Rx + \pi_0(x)'Q\pi_0(x) + 2x'W\pi_0(x) + \beta V^*(Ax + B\pi_0(x)) \\ &\geq \inf_{u \in \mathbb{R}^m} x'Rx + x'Qu + 2x'Wu + \beta V^*(Ax + Bu) = S(V^*)(x). \end{aligned}$$

It follows that $V^* \geq S(V^*)$. To establish the reverse inequality, define $V^1_\pi(x)$ to be the value at time 1 given policy π. Thus

$$\begin{aligned} V^1_\pi(x) &= \lim_{T \to \infty} \sum_{t=1}^{T} \beta^t \left(x_t'Rx_t + u_t'Qu_t + 2x_t'Wu_t\right) \\ x_{t+1} &= Ax_t + Bu_t \text{ and } u_t = \pi_t(x_t), \text{ with } x_1 = x. \end{aligned}$$

Let $\delta, \varepsilon > 0$, and choose π so that

$$\begin{aligned} x'Rx + \pi_0(x)'Q\pi_0(x) + 2x'W\pi_0(x) + \beta V^*(Ax + B\pi_0(x)) &\leq S(V^*)(x) + \varepsilon, \text{ and} \\ V^1_\pi(Ax + B\pi_0(x)) &\leq V^*(Ax + B\pi_0(x)) + \delta. \end{aligned}$$

We find that

$$\begin{aligned} V^*(x) &\leq V_\pi(x) = x'Rx + \pi_0(x)'Q\pi_0(x) + 2x'W\pi_0(x) + \beta V^1_\pi(Ax + B\pi_0(x)) \\ &\leq S(V^*)(x) + \varepsilon + \delta, \end{aligned}$$

and the result follows.

To establish item 3, let V be a fixed point of S and $\varepsilon > 0$. and let $\delta = 1/2(1 - \beta)\varepsilon > 0$. Choose stationary policy π_δ so that

$$x'Rx + \pi_\delta(x)'Q\pi_\delta(x) + 2x'W\pi_\delta(x) + V(Ax + B\pi_\delta(x)) \leq S(V)(x) + \delta.$$

Now fix $x \in \mathbb{R}^n$, and let (x_t, u_t) be the sequence generated by the policy π_δ, the transition dynamic (98) , and the initial condition $x_0 = x$.Then

$$\begin{aligned}
V^*(x) &\leq \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right) \\
&\leq \lim_{T\to\infty} \left(\beta^{T+1} V\left(x_{T+1}\right) + \sum_{t=0}^{T} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right)\right) \\
&= \lim_{T\to\infty} \left(\beta^{T+1} V\left(A x_T + B u_T\right) + \sum_{t=0}^{T} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right)\right) \\
&= \lim_{T\to\infty} \left(\beta^T \left(x_T' R x_T + u_T' Q u_T + 2x_T' W u_T + \beta V\left(A x_T + B u_T\right)\right)\right. \\
&\quad \left. + \sum_{t=0}^{T-1} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right)\right) \\
&\leq \lim_{T\to\infty} \left(\beta^T \left(S(V)(x_T) + \delta\right) + \sum_{t=0}^{T-1} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right)\right) \\
&= \lim_{T\to\infty} \left(\beta^T V(x_T) + \sum_{t=0}^{T-1} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right) + \beta^T \delta\right) \\
&\quad \vdots \\
&\leq \lim_{T\to\infty} \left(V(x_0) + \delta \sum_{t=0}^{T} \beta^t\right) = V(x) + \delta \sum_{t=0}^{\infty} \beta^t < V(x) + \varepsilon. \blacksquare
\end{aligned}$$

Solving the deterministic LQ-problem

The following lemma relates the S-map to the T-map, and we use this Lemma to prove Theorem 8:

Lemma 8 *If P is symmetric positive semi-definite and $V_P(x) = x'Px$ is the corresponding quadratic form then $S^N(V_P)(x) = x'T^N(P)x$.*

Proof. That $S(V_P)(x) = x'T(P)x$ follows from Lemma 1. Thus $S(V_P)(x) = V_{T(P)}(x)$. Working by induction,

$$\begin{aligned}
S^N(V_P)(x) &= \left(S^{N-1} \circ S\right)(V_P)(x) = S^{N-1}(S(V_P))(x) \\
&= S^{N-1}(V_{T(P)})(x) = x'T^{N-1}(T(P))x \\
&= x'T^N(P)x. \blacksquare
\end{aligned}$$

Theorem 8 *Assume LQ.1 – LQ.3 and let P^* be as in Theorem 1. Then*

1. *$V^*(x) = x'P^*x$.*
2. *If $\pi(x) = -F(P^*)x$ then $V_\pi(x) = V^*(x)$.*

Proof. To demonstrate item 1, we lean heavily on Lemma 1. Let $V_{P^*}(x) = x'P^*x$. Then

$$S\left(V_{P^*}\right)(x) = x'T\left(V_{P^*}\right)x = x'P^*x = V_{P^*}(x),$$

where the first equality follows from Lemma 1 and the second follows from the fact that P^* is a fixed point of T. It follows from item 3 of Lemma (7) that $V^*(x) \leq V_{P^*}(x)$. Next notice that $0 \leq V^*$, so that by items 1 and 2 of Lemma 7 that $S^N(0) \leq V^*$. But by Lemma 8, $S^N(0)(x) = x'T^N(0)x$. Thus $x'T^N(0)x \leq V^*(x)$. Taking limits we have

$$V_{P^*}(x) = x'P^*x = \lim_{N\to\infty} x'T^N(0)x \leq V^*(x),$$

and the item 1 is established.

To determine the optimal policy consider the optimization problem

$$\inf_{u\in\mathbb{R}^m}\left(x'Rx + u'Qu + 2x'Wu + \beta V_{P^*}\left(Ax + Bu\right)\right). \tag{100}$$

By Lemma 1 the unique solution is given by $u = -F(P^*)x$. Now note that

$$\begin{aligned}
\hat{F}(P^*) &= (Q + \beta B'P^*B)^{-1}\hat{B}'P^*\hat{A} \\
&= (Q + \beta B'P^*B)^{-1}\beta B'P^*\left(A - BQ^{-1}W'\right) \\
&= (Q + \beta B'P^*B)^{-1}\left(\beta B'P^*A + W' - QQ^{-1}W' - \beta B'P^*BQ^{-1}W'\right) \\
&= (Q + \beta B'P^*B)^{-1}\left[\left(\beta B'P^*A + W'\right) - \left(Q + \beta B'P^*B\right)Q^{-1}W'\right] \\
&= F\left(P^*\right) - Q^{-1}W'.
\end{aligned}$$

From $\hat{R} = R - WQ^{-1}W'$ we have

$$\begin{aligned}
& x'Rx + u'Qu + 2x'Wu \\
= \;& x'\hat{R}x + (u + Q^{-1}Wx)'Q(u + Q^{-1}Wx).
\end{aligned}$$

It follows that under the policy $u = -F(P^*)x$,

$$\begin{aligned}
x'Rx + u'Qu + 2x'Wu &= x'\left(\hat{R} + (F(P^*) + Q^{-1}W')'Q(F(P^*) + Q^{-1}W')\right)x \\
&= x'\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)x.
\end{aligned}$$

Now let π be given by $\pi(x) = -F(P^*)x$. We claim that $V_\pi(x) = V^*(x)$. that is, π solves the sequence problem (97). If this policy is followed then the state dynamics are given by $x_t = \beta^{-1/2}\Omega(P^*)x_{t-1}$. Thus

$$\begin{aligned} V_\pi(x) &= \sum_{t\geq 0} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right) = \sum_{t\geq 0} \beta^t x_t' \left(\hat{R} + \hat{F}(P^*)' Q \hat{F}(P^*)\right) x_t \\ &= x' \left(\sum_{t\geq 0} \left(\Omega(P^*)'\right)^t \left(\hat{R} + \hat{F}(P^*)' Q \hat{F}(P^*)\right) \left(\Omega(P^*)\right)^t\right) x = x' \hat{S} x, \end{aligned}$$

where the last equality identifies notation. The existence of $\hat{S}$ is guaranteed by Lemma 4. Note also that $\hat{S}$ is symmetric, positive semi-definite. That same Lemma shows

$$\hat{S} = \Omega(P^*)' \hat{S} \Omega(P^*) + \hat{R} + \hat{F}(P^*)' Q \hat{F}(P^*).$$

But by Lemma 3 and Theorem 1, P^* is the unique symmetric, positive semi-definite satisfying this equation. It follows that $\hat{S} = P^*$ and thus $V_\pi(x) = V^*(x)$. ■

Solving the LQ-Problem: the Stochastic Case. We now turn to the stochastic formulation of the quadratic problem. Certainty equivalence will yield that the policy function takes precisely the same form as in the deterministic case: $u = F(P^*)x$; however, establishing this result requires considerably machinery. The difficulty involves measurability of the value function, a property that is needed in order to properly formulate the expectational form of Bellman's functional equation. The development presented here follow the work of Bertsekas (1987), Bertsekas and Shreve (1978) and, at a key point, leans heavily on Theorem 1.

The problem under consideration is

$$\min \quad E \sum_{t\geq 0} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2x_t' W u_t\right) \tag{101}$$

$$s.t. \quad x_{t+1} = Ax_t + Bu_t + C\varepsilon_{t+1}, \tag{102}$$

where $E\varepsilon_t\varepsilon_t' = \sigma_\varepsilon^2 I_k$. To make this problem precise, and in particular to make sense of the expectations operator, we must establish the collection of admissible policies. Because objective involves an integral, it is necessary to restrict attention to measurable policies, and to remain connected to the inherent topology of the state and controls spaces, we follow the literature and require that our policy functions be Borel measurable. Unfortunately, Borel measurability is not sufficiently flexible to allow for integration of all of the types of functions arising from a general (unbounded) DP problem. At issue is the projection of a Borel subset of a Cartesian product onto one of the factors, is not necessarily Borel measurable. Since, as we will emphasize below, the infimum operation involves precisely this type of projection, we are required to expand our notion of measurability to include these projected sets. We turn to this development now, which is somewhat involved.

Preliminaries.

First, we work to make infima well behaved. Let $N > 0$ represent any dimension, and let $\mathcal{B}_N$ be the Borel sets in $\mathbb{R}^N$, that is, the σ-algebra generated by the opens sets of $\mathbb{R}^N$. Let $\mathcal{Y}$ be *any* uncountable Borel space. If $\mathcal{X} \subset \mathbb{R}^N \times \mathcal{Y}$ define the projection of $\mathcal{X}$ onto $\mathbb{R}^N$ in the usual way:

$$\underset{\mathbb{R}^N}{\Pr}(\mathcal{X}) = \left\{x \in \mathbb{R}^N : (x, y) \in \mathcal{X}\right\}.$$

The collection $\mathcal{A}$ of *analytic sets* in $\mathbb{R}^N$ is the set of all subsets of $\mathbb{R}^N$ of the form $\Pr_{\mathbb{R}^N}(\mathcal{X})$ for any Borel set $\mathcal{X} \subset \mathbb{R}^N \times \mathcal{Y}$.[40] Note that $\mathcal{B}_N \subset \mathcal{A}$, but the reverse inclusion does not hold. The function $g : \mathbb{R}^N \to \mathbb{R}^*$ is *lower semi-analytic* provided that its lower contours are analytic, that is, for any $c \in \mathbb{R}^*$,

$$\left\{x \in \mathbb{R}^N : g(x) < c\right\} \in \mathcal{A}.$$

A lower semi-analytic function is not necessarily Borel measurable. The following result is Proposition 7.47 on page 179 of Bertsekas and Shreve (1978)**:**

Lemma 9 *Suppose $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^*$ is lower semi-analytic. Define $f^* : \mathbb{R}^n \to \mathbb{R}^*$ by*

$$f^*(x) = \inf\left\{f(x, y) : (x, y) \in \mathbb{R}^n \times \mathbb{R}^m\right\}.$$

Then f^* is lower semi-analytic.

Next, we extend measurability in the appropriate way. Let $\Lambda\left(\mathbb{R}^N\right)$ to be the collection of all probability measures on $\mathcal{B}_N$. For $\mu \in \Lambda\left(\mathbb{R}^N\right)$ and $E \subset \mathbb{R}^N$ we define the outer measure of E with respect to μ in the usual way:

$$\mu^*(E) = \inf\{\mu(B) : E \subset B \in \mathcal{B}_N\}.$$

Then, using the method of Caratheodory, we define $\mathcal{B}_N(\mu)$ to be the largest σ-algebra on which μ^* is countably additive:

$$\mathcal{B}_N(\mu) = \{E \subset \mathbb{R}^N : \mu^*(E) + \mu^*(E^c) = 1\}.$$

Notice that $\mathcal{B}_N \subset \mathcal{B}_N(\mu)$ and $\mu^*|_{\mathcal{B}_N} = \mu$. Finally, set

$$\mathcal{U}_N = \bigcap_{\mu \in \Lambda(\mathbb{R}^N)} \mathcal{B}_N(\mu).$$

[40] As is implied by this definition, any two distinct uncountable Borel spaces $\mathcal{Y}$ and $\mathcal{Y}'$ identify the same collection of analytic sets. While this does mean we could have simply take $\mathcal{Y}$ to be a familiar set like $\mathbb{R}$, the generality of this definition will be useful.

There is a variety of equivalent definitions of analytic sets: see Bertsekas and Shreve (1978), Proposition 7.41, page 166. The one we chose to emphasize highlights the relationship between analyticity and infima, though other defintions are more useful for establishing the useful properties of analytic sets.

Then $\mathcal{U}_N$ is a σ-algebra containing $\mathcal{B}_N$ onto which all $\mu \in \Lambda\left(\mathbb{R}^N\right)$ extend. Importantly, by Corollary 7.42.1 on page 169 of Bertsekas and Shreve (1978), analytic sets are universally measurable.

If $f : \mathbb{R}^N \to \mathbb{R}^*$ then f is *universally measurable* provided that $f^{-1}(U) \in \mathcal{U}_N$ for any open $U \subset \mathbb{R}^N$. Notice that universally measurable functions may be integrated against any (properly extended) Borel measure $\mu \in \Lambda\left(\mathbb{R}^N\right)$. Also, since analytic sets are universally measurable it follows that lower semi-analytic functions are universally measurable.

To illustrate more concretely the relationships between analyticity, universal measurability, infima and projections, let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^*$ be Borel measurable. Also, for any set $\mathcal{X} \subset \mathbb{R}^n \times \mathbb{R}^m$, let

$$\Pr_{\mathbb{R}^n}(\mathcal{X}) = \{x \in \mathbb{R}^n : (x, y) \in \mathcal{X}\}.$$

Finally, define f^* as in Lemma 9. If $c \in \mathbb{R}$ let $U_c \subset \mathbb{R}^n \times \mathbb{R}^m$ be the pre-image under f of $(-\infty, c)$ and let $U_c^* \subset \mathbb{R}^n$ be the pre-image of $(-\infty, c)$ under f^*. These sets are related by

$$U_c^* = \Pr_{\mathbb{R}^n}(U_c).$$

Since f is Borel measurable it follows that U_c is Borel measurable. However, U_c^* may not be Borel measurable since Borel measurability is not preserved under projection. But if f is lower semi-analytic then by Lemma **??** together with the inclusion $\mathcal{U}_N \subset \mathcal{A}$, we know that U_c^* is universally measurable, and thus f^* may be integrated.

Finally, we connect universal measurability to evaluation of the objective. To allow for the computation of

$$E \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left(x_t' R x_t + u_t' Q u_t + 2 x_t' W u_t\right),$$

the policy and transition dynamics must preserve universal measurability and determine a distribution over the history x^T. The needed result relies on the notion of a *universally (Borel) measurable stochastic kernel*, which is defined as a function $q : \mathcal{U}_n \times \mathbb{R}^N \to [0, 1]$ so that $q(\cdot, y) \in \Lambda\left(\mathbb{R}^N\right)$, properly extended, and for any $E \in \mathcal{U}_n$, the map

$$y \to q(E, y) : \mathbb{R}^N \to [0, 1]$$

is universally (Borel) measurable.The following Lemma, which is Proposition 7.48 on page 180 of Bertsekas and Shreve (1978), provides a needed result.

Lemma 10 *Suppose $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^*$ is lower semi-analytic and q is a Borel measurable stochastic kernel. Define $\hat{f} : \mathbb{R}^n \to \mathbb{R}^*$*

$$\hat{f}(x) = \int f(x, y) q(dy, x).$$

Then $\hat{f}$ is lower semi-analytic.

Formalizing the programming problem

For simplicity, we consider only stationary policies. This is possible because we state without proof the Principle of Optimality, from which it will follow that the optimal policy will be stationary. The set Π of admissible policies is the collection of universally measurable functions $\pi : \mathbb{R}^n \to \mathbb{R}^m$. We note that if π is an admissible policy and $f : \mathbb{R}^n \oplus \mathbb{R}^n \to \mathbb{R}^*$ is universally measurable then so is the function $x \to f(x, \pi(x))$: see Bertsekas and Shreve (1978), Proposition 7.44, page 172. Next, we build the universally measurable stochastic kernel associated to a given policy π and the transition dynamics (102): Let $\mu_{C\varepsilon} \in \Lambda(\mathbb{R}^n)$ be the Borel measure induced by the random vector $C\varepsilon$, appropriately extended to $\mathcal{U}_n$. Define q as follows: $E \in \mathcal{B}_n$ and $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$q(E, (x, u)) = \mu_{C\varepsilon}(E - (Ax + Bu)). \tag{103}$$

At the bottom of page 189, Bertsekas and Shreve (1978) argue that q is a Borel measurable stochastic kernel. Next, for $\pi \in \Pi$, and or $E \in \mathcal{U}_n$ and $x \in \mathbb{R}^n$, define

$$q^\pi(E, x) = \mu_{C\varepsilon}(E - (Ax + B\pi(x))).$$

Because universal measurability is preserved under composition, q_t^π is a universally measurable stochastic kernel.

Given a policy π and associated kernel q^π, we may use Proposition 7.45 on page 175 of Bertsekas and Shreve (1978) to construct a distribution over histories of the state. Let

$$f_T^\pi(x^T) = \sum_{t=0}^{T} \beta^t \left(x_t' R x_t + \pi(x_t)' Q \pi(x_t) + 2x_t' W \pi(x_t)\right).$$

Then given any initial condition x_0, this proposition guarantees the existence of a measure $\mu_\pi^T(\cdot, x_0) \in \Lambda\left(\left(\mathbb{R}^N\right)^T\right)$ so that

$$\int f_T^\pi(x^T)\mu_\pi^T(dx^T, x_0) = \int\int \cdots \int f_T^\pi(x^T) q^\pi(dx_T, x_{T-1}) q^\pi(dx_{T-1}, x_{T-2}) \cdots q^\pi(dx_1, x_0).$$

Finally, for admissible policy π we set

$$V_\pi(x) = \lim_{T\to\infty} \int f_T^\pi(x^T)\mu_\pi^T(dx^T, x).$$

With this notation, we are finally in a position to make formal the decision problem (101):

$$V^*(x) = \inf_{\pi\in\Pi} V_\pi(x). \tag{104}$$

Note that $V^* : \mathbb{R}^n \to [0, \infty]$. We have the following non-trivial result, which is Corollary 9.4.1 on page 221 of Bertsekas and Shreve (1978), and represents more than a simple combination of Lemmas 9 and 10:

Lemma 11 *The map V^* is lower semi-analytic.*

The S-map

As in the deterministic case, we would like to identify V^* as the solution to associated Bellman system. For lower semi-analytic $V : \mathbb{R}^n \to [0, \infty]$, define $S(V) : \mathbb{R}^n \to [0, \infty]$ by

$$S(V)(x) = \inf_{u \in \mathbb{R}^m} x'Rx + u'Qu + 2x'Wu + \beta \int V(\tilde{x}) q(d\tilde{x}, (x, u)),$$

where q is the Borel measurable stochastic kernel defined by (**??**). By Lemmas 9 and 10, $S(V)$ is lower semi-analytic whenever V is. By Lemma 11 we may act by S on V^*. We have the following analog to Lemma 7 from the deterministic case.

Lemma 12 *The map S satisfies the following properties:*

1. *(Monotonicity) If $V_1, V_2 : \mathbb{R}^n \to [0, \infty)$ are lower semi-analytic, and $V_1 \leq V_2$ then $S(V_1) \leq S(V_2)$.*
2. *(Principle of Optimality) $V^* = S(V^*)$.*
3. *(Minimality of V^*) If $V : \mathbb{R}^n \to [0, \infty)$ is lower semi-analytic and $S(V) = V$ then $V^* \leq V$.*

Item 1 is proved just as before and the remaining two results are recorded as Propositions 9.8 and 9.10 on pages 227 and 228 of Bertsekas and Shreve (1978). One more (highly non-trivial) result is needed: Define $V_0 = 0$ and $V_n = S(V_{n-1})$. By monotonicity, V_n is increasing, and so pointwise convergent in $\mathbb{R}^*$. Denote this limit by V_∞. The following Lemma is Proposition 9.16 on pages 232 of Bertsekas and Shreve (1978):

Lemma 13 *If $S(V_\infty) = V_\infty$ then $V_\infty = V^*$.*

Solving the stochastic LQ-problem

As before, we need the following result:

Lemma 14 *If P is symmetric positive semi-definite and $V_P(x) = x'Px$ is the corresponding quadratic form then $S^N(V_P)(x) = x'T_\varepsilon^N(P)x$.*

Proof. That $S(V_P)(x) = x'T_\varepsilon(P)x$ follows from Lemma 2. Thus $S(V_P)(x) = V_{T_\varepsilon(P)}(x)$. Working by induction,

$$\begin{aligned} S^N(V_P)(x) &= \left(S^{N-1} \circ S\right)(V_P)(x) = S^{N-1}(S(V_P))(x) \\ &= S^{N-1}(V_{T_\varepsilon(P)})(x) = x'T_\varepsilon^{N-1}(T_\varepsilon(P))x \\ &= x'T^N(P)x. \blacksquare \end{aligned}$$

.

Theorem 9 *Assume LQ.1 – LQ.3 and let P^* be as in Theorem 1. Then*

1. $V^*(x) = x'P^*x + \frac{\beta}{1-\beta} tr\left(\sigma_\varepsilon^2 P^* CC'\right)$.
2. *If $\pi(x) = -F(P^*)x$ then $V_\pi(x) = V^*(x)$.*

Proof. To demonstrate item 1, note that

$$V_\infty(x) = \lim_{N\to\infty} V_N(x) = \lim_{N\to\infty} S^N(0)(x) = \lim_{N\to\infty} x'T_\varepsilon^N(0)x = x'P_\varepsilon^* x,$$

where the last equality follows from Theorem 1. Thus

$$S(V_\infty)(x) = S(V_{P_\varepsilon^*})(x) = x'T_\varepsilon\left(P_\varepsilon^*\right)x = x'P_\varepsilon^* x = V_\infty(x).$$

By Lemma 13,

$$V^*(x) = V_\infty(x) = x'P_\varepsilon^* x = x'P^*x + \frac{\beta}{1-\beta} tr\left(\sigma_\varepsilon^2 P^* CC'\right),$$

where the last equality follows from Lemma 2. With this, item 1 is established.

To determine the optimal policy consider the optimization problem

$$\inf_{u\in\mathbb{R}^m} x'Rx + u'Qu + 2x'Wu + \beta \int V^*(\tilde{x}) q(d\tilde{x}, (x,u)),$$

with V^* now written as in item 1. By Lemma 2, together with the bulleted comments that follow, the unique solution is given by $u = -F(P^*)x$. The same computation as in the proof of Theorem 8 demonstrates that under the policy $u = -F(P^*)x$,

$$\begin{aligned} x'Rx + u'Qu + 2x'Wu &= x'\left(\hat{R} + (F(P^*) + Q^{-1}W')'Q(F(P^*) + Q^{-1}W')\right)x \\ &= x'\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)x. \end{aligned}$$

Now let π be given by $\pi(x) = -F(P^*)x$. We claim that $V_\pi(x) = V^*(x)$, that is, π solves the sequence problem (104). If this policy is followed, then, by equation (20), the state dynamics are given by

$$\begin{aligned} x_t &= \beta^{-1/2}\Omega(P^*)x_{t-1} + C\varepsilon_t \\ &= \left(\beta^{-1/2}\Omega(P^*)\right)^t x_0 + \sum_{N=0}^{t-1}\left(\beta^{-1/2}\Omega(P^*)\right)^N C\varepsilon_{t-N}. \end{aligned}$$

Thus

$$\begin{aligned} & V_\pi(x) \\ = \; & E_0\sum_{t\geq 0}\beta^t\left(x_t'Rx_t + u_t'Qu_t + 2x_t'Wu_t\right) = E_0\sum_{t\geq 0}\beta^t x_t'\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)x_t \\ = \; & x'\left(\sum_{t\geq 0}\left(\Omega(P^*)'\right)^t\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)\left(\Omega(P^*)\right)^t\right)x \\ & + E_0\sum_{t\geq 1}\beta^t\sum_{N=0}^{t-1}\left(\left(\beta^{-1/2}\Omega(P^*)\right)^N C\varepsilon_{t-N}\right)'\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)\left(\left(\beta^{-1/2}\Omega(P^*)\right)^N C\varepsilon_{t-N}\right) \\ = \; & x'\left(\sum_{t\geq 0}\left(\Omega(P^*)'\right)^t\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)\left(\Omega(P^*)\right)^t\right)x \\ & + \left(\frac{\beta}{1-\beta}\right)E_0\sum_{t\geq 0}\beta^t\left(\left(\beta^{-1/2}\Omega(P^*)\right)^t C\varepsilon_{t+1}\right)'\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)\left(\left(\beta^{-1/2}\Omega(P^*)\right)^t C\varepsilon_{t+1}\right) \\ = \; & x'\left(\sum_{t\geq 0}\left(\Omega(P^*)'\right)^t\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)\left(\Omega(P^*)\right)^t\right)x \\ & + \left(\frac{\beta}{1-\beta}\right)E_0\sum_{t\geq 0}\varepsilon_{t+1}'C'\left(\Omega(P^*)'\right)^t\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)\Omega(P^*)^t C\varepsilon_{t+1} \\ = \; & x'P^*x + \left(\frac{\beta}{1-\beta}\right)E_0\sum_{t\geq 0}\varepsilon_{t+1}'C'\left(\Omega(P^*)'\right)^t\left(\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)\right)\Omega(P^*)^t C\varepsilon_{t+1}, \end{aligned}$$

where the last equality follows from the work done in proving 8. Since $\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*)$ is symmetric, positive semi-definite, we may use the rank decomposition to write

$$\hat{R} + \hat{F}(P^*)'Q\hat{F}(P^*) = D'D.$$

Then

$$
\begin{aligned}
& E_0 \sum_{t\geq 0} \varepsilon'_{t+1} C' \left(\Omega (P^*)'\right)^t D' D \Omega (P^*)^t C \varepsilon_{t+1} \\
= & \sum_{t\geq 0} \mathrm{tr}\left(D\Omega (P^*)^t C E(\varepsilon_{t+1}\varepsilon'_{t+1}) C' \left(\Omega (P^*)'\right)^t D' \right) \\
= & \sum_{t\geq 0} \mathrm{tr}\left(\left(\Omega (P^*)'\right)^t D' D \Omega (P^*)^t C \sigma_\varepsilon^2 I_n C' \right) \\
= & \mathrm{tr}\left[\left(\sum_{t\geq 0} \left(\Omega (P^*)'\right)^t D' D \Omega (P^*)^t \right) C \sigma_\varepsilon^2 I_n C' \right] \\
= & \mathrm{tr}\left[\left(\sum_{t\geq 0} \left(\Omega (P^*)'\right)^t \left(\hat{R} + \hat{F}(P^*)' Q \hat{F}(P^*) \right) \Omega (P^*)^t \right) C \sigma_\varepsilon^2 I_n C' \right] \\
= & \mathrm{tr}\left[P^* C \sigma_\varepsilon^2 I_n C' \right] = \mathrm{tr}\left(\sigma_\varepsilon^2 P^* C C' \right),
\end{aligned}
$$

and the result follows. ■

Appendix D: Robinson Crusoe model details

Proof of Proposition 1. We reproduce the LQ problem (49) here for convenience:

$$
\begin{aligned}
\max \quad & -E \sum_{t\geq 0} \beta^t \left((c_t - b^*)^2 + \phi s_{t-1}^2 \right) \\
\text{s.t.} \quad & s_{t+1} = A_1 s_t + A_2 s_{t-1} - c_t + \mu_{t+1}
\end{aligned}
\tag{105}
$$

To place in standard LQ form (see (7)), we define the state as $x_t = (1, s_t, s_{t-1})'$ and the control as $u_t = c_t$. Note that the intercept is an exogenous state. The key matrices are given by:

$$
R = \begin{pmatrix} (b^*)^2 & 0 & 0 \\ 0 & \phi & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & A_1 & A_2 \\ 0 & 1 & 0 \end{pmatrix},
$$

and $W = (-b^*, 0, 0)', B = (0, -1, 0)'$, and $Q = 1$. The transformed matrices are $\hat{R} = R - WW', \hat{A} = \beta^{\frac{1}{2}} (A - BW')$, and $\hat{B} = \beta^{\frac{1}{2}} B$: thus

$$
\hat{R} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \phi & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \hat{A} = \beta^{\frac{1}{2}} \begin{pmatrix} 1 & 0 & 0 \\ -b^* & A_1 & A_2 \\ 0 & 1 & 0 \end{pmatrix}.
$$

We see immediately that LQ.1 is satisfied: $\hat{R}$ is positive semi-definite and Q is positive definite. Next let $K = (K_1, K_2, K_3)$ be any 1×3 matrix. Then

$$
\hat{A} - \hat{B}K = \beta^{\frac{1}{2}} \begin{pmatrix} 1 & 0 & 0 \\ -b^* + K_1 & A_1 + K_2 & A_2 + K_3 \\ 0 & 1 & 0 \end{pmatrix}.
$$

By choosing $K_2 = -A_1$ and $K_3 = -A_2$, we see that $(\hat{A}, \hat{B})$ is a stabilizable pair: thus LQ.2 is satisfied. Finally, to verify LQ.3 note that

$$\hat{D} = \begin{pmatrix} 0 \\ \sqrt{\phi} \\ 0 \end{pmatrix},$$

and suppose for some $y \neq 0$ that $\hat{A}y = \mu y$ and $\hat{D}'y = 0$. The condition $\hat{D}'y = 0$ implies that $y_2 = 0$. Since the third component of $\hat{A}y$ must then be zero, it then follows from $\hat{A}y = \mu y$ that $y_3 = 0$ or $\mu = 0$. But $A_2 \neq 0$ implies all three eigenvalues of $\hat{A}$ are non-zero; thus $y_3 = 0$. But no non-zero eigenvector of $\hat{A}$ has the form $(y_1, 0, 0)$. Thus if y is an eigenvalue of $\hat{A}$ then $\hat{D}'y \neq 0$. Thus detectability holds vacuously and LQ.3 is satisfied. The proof is completed by application of Theorem 4′. ■

Details of Robinson Crusoe Economy The LQ set-up (105) does not directly impose non-negativity constraints that are present. We show that under suitable assumptions these constraints are never violated. Carefully specified as an economics problem, (105) might include the constraints

$$0 \leq c_t \leq A_1 s_t + A_2 s_{t-1},\ c_t \in (0, b^*), \text{ and } s_{t+1} \geq 0. \tag{106}$$

To demonstrate these hold under our assumption $\beta A_1 + \beta^2 A_2 > 1$, we first study the non-stochastic steady sate. Let c and s be the respective nonstochastic steady-state values of c_t and s_t. The transition equation implies $c = \Theta s$, where $\Theta = A_1 + A_2 - 1$. Inserting this condition into the Euler equation, reproduced here for convenience,

$$c_t - \beta\phi\hat{E}_t s_{t+1} = b^*(1 - \beta A_1 - \beta^2 A_2) + \beta A_1 \hat{E}_t c_{t+1} + \beta^2 A_2 \hat{E}_t c_{t+2}, \tag{107}$$

and solving for s yields

$$s = \frac{b^*(1 - \beta A_1 - \beta^2 A_2)}{\Theta(1 - \beta A_1 - \beta^2 A_2) - \beta\phi}.$$

It follows that if $\beta A_1 + \beta^2 A_2 > 1$ then $s > 0$ and $c \in (0, b^*)$. Under LQ.RTL, and provided suitable initial conditions hold and the support of μ_{t+1} is sufficiently small, it follows that (106) holds and $s_t > 0$ and $c_t \in (0, b^*)$.

It remains to consider LQ.RTL within the context of the Crusoe Economy. In general, the first criterion – asymptotic stationarity of the state under optimal decision making – will not be satisfied. There is no free-disposal in Bob's world: to prevent an accumulation of new trees Bob must consume. Thus, if, for example, new trees are very productive ($A_1 > 1$) and if weeding them is relatively painless ($\phi \approx 0$) it may be optimal to allow the quantity of new trees to grow without bound. However, productive new trees does not preclude the desire to stabilize the state. For example, if $A_1 = 1.1$ and $A_2 = .1$, then even with $\phi = .01$, asymptotic stationarity of the state under optimal decision making obtains.

Satisfaction of the second part of LQ.RTL follows satisfaction of the first part by an argument in Hayashi, p. 394: for an autoregressive process $y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t$, where ε_t is white noise with positive variance, which satisfies the stationarity condition that the roots of $1 - \sum_{i=1}^{p} \phi_i L^i = 0$ are strictly outside the unit circle, the second moment matrix $Ex_t x_t'$, where $x_t' = (1, y_{t-1}, \ldots y_{t-p})$, is non-singular.

Euler Equation Learning in the Robinson Crusoe Economy Turning to Euler equation learning, first we note that if $A_2 = 0$ then the following simple transform allows for the construction of a one-step-ahead Euler equation:

$$\mathcal{S} = \begin{pmatrix} I_2 & 0_{2\times 1} \\ 0 \quad A_1 & -1 \end{pmatrix}.$$

Next, recall the PLM

$$c_t = f_1 + f_2 s_t + f_3 s_{t-1}.$$

Using this PLM, the following expectations may be computed:

$$\begin{aligned} \hat{E}_t c_{t+1} &= f_1 + (f_2 A_1 + f_3) s_t + f_2 A_2 s_{t-1} - f_2 c_t \\ \hat{E}_t c_{t+2} &= f_1(1 - f_2) + ((f_2(A_1 - f_2) + f_3) A_1 + f_2(A_2 - f_3)) s_t \\ &\quad + (f_2(A_1 - f_2) + f_3) A_2 s_{t-1} - (f_2(A_1 - f_2) + f_3) c_t. \end{aligned}$$

Combining these expectations with (107) provides the following T-map:

$$f_1 \rightarrow \frac{\psi + \beta A_1 f_1 + \beta^2 A_2 f_1 (1 - f_2)}{1 + \beta\phi + \beta A_1 f_2 + \beta^2 A_2 (f_2(A_1 - f_2) + f_3)}$$

$$f_2 \rightarrow \frac{\beta\phi A_1 + \beta A_1 (f_2 A_1 + f_3) + \beta^2 A_2 ((f_2(A_1 - f_2) + f_3) A_1 + f_2(A_2 - f_3))}{1 + \beta\phi + \beta A_1 f_2 + \beta^2 A_2 (f_2(A_1 - f_2) + f_3)}$$

$$f_3 \rightarrow \frac{\beta\phi A_2 + \beta A_1 f_2 A_2 + \beta^2 A_2 (f_2(A_1 - f_2) + f_3)}{1 + \beta\phi + \beta A_1 f_2 + \beta^2 A_2 (f_2(A_1 - f_2) + f_3)}.$$

References

Adam, K., and A. Marcet (2011): "Internal Rationality, Imperfect Market Knowledge and Asset Proces," *Journal of Economic Theory*, 146, 1224–1252.

Anderson, B. D. O., and J. P. Moore (1979): *Optimal Filtering*. Prentice Hall, New Jersey.

Bao, T., J. Duffy, and C. Hommes (2013): "Learning, Forecasting and Optimizing: an Experimental Study," *European Economic Review*, 61, 186–204.

Bertsekas, D. P. (1987): *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, NJ.

Bertsekas, D. P., and S. E. Shreve (1978): *Stochastic Optimal Control: the Discrete-Time Case*. Academic Press, New York.

Branch, W. A., G. W. Evans, and B. McGough (2013): "Finite Horizon Learning," in Sargent and Vilmunen (2013), chap. 9.

Bray, M. (1982): "Learning, Estimation, and the Stability of Rational Expectations Equilibria," *Journal of Economic Theory*, 26, 318–339.

Bray, M., and N. Savin (1986): "Rational Expectations Equilibria, Learning, and Model Specification," *Econometrica*, 54, 1129–1160.

Bullard, J., and J. Duffy (1998): "A Model of Learning and Emulation with Artificial Adaptive Agents," *Journal of Economic Dynamics and Control*, 22, 179–207.

Bullard, J., and K. Mitra (2002): "Learning About Monetary Policy Rules," *Journal of Monetary Economics*, 49, 1105–1129.

Cho, I.-K., N. Williams, and T. J. Sargent (2002): "Escaping Nash Inflation," *Review of Economic Studies*, 69, 1–40.

Coddington, E. A. (1961): *An Introduction to Ordinary Differential Equations*. Prentice-Hall, Englewood Cliffs, N.J.

Cogley, T., and T. J. Sargent (2008): "Anticipated Utility and Rational Expectations as Approximations of Bayesian Decision Making," *International Economic Review*, 49, 185–221.

Eusepi, S., and B. Preston (2010): "Central Bank Communication and Expectations Stabilization," *American Economic Journal: Macroeconomics*, 2, 235–271.

Evans, G. W., and S. Honkapohja (1998): "Economic Dynamics with Learning: New Stability Results," *Review of Economic Studies*, 65, 23–44.

——— (2001): *Learning and Expectations in Macroeconomics*. Princeton University Press, Princeton, New Jersey.

——— (2006): "Monetary Policy, Expectations and Commitment," *Scandinavian Journal of Economics*, 108, 15–38.

Evans, G. W., S. Honkapohja, and K. Mitra (2009): "Anticipated Fiscal Policy and Learning," *Journal of Monetary Economics*, 56, 930–953.

Gabaix, X. (2014): "Sparse Dynamic Programming and Aggregate Fluctuations," Working paper.

Hansen, L. P., and T. J. Sargent (2014): *Recursive Models of Dynamic Linear Economies*. Princeton University Press, Princeton, NJ.

Hommes, C. H. (2011): "The Heterogeneous Expectations Hypothesis: Some Evidence from the Lab," *Journal of Economic Dynamics and Control*, 35, 1–24.

Honkapohja, S., K. Mitra, and G. W. Evans (2013): "Notes on Agents' Behavioral Rules Under Adaptive Learning and Studies of Monetary Policy," in Sargent and Vilmunen (2013), chap. 5.

Howitt, P., and O. Özak (2014): "Adaptive Consumption Behavior," *Journal of Economic Dynamics and Control*, 39, 37–61.

Kiyotaki, N., and R. Wright (1989): "On Money as a Medium of Exhange," *Journal of Political Economy*, 97, 927–954.

Krasovskii, N. (1963): *Stability of Motion*. Stanford University Press, Stanford, California.

Kwakernaak, H., and R. Sivan (1972): *Linear Optimal Control Systems*. Wiley-Interscience, New York.

Lancaster, P., and L. Rodman (1995): *Algebraic Riccati Equations*. Prentice-Hall, Oxford University Press, Oxford, UK.

Lettau, M., and H. Uhlig (1999): "Rules of Thumb and Dynamic Programming," *American Economic Review*, 89, 148–174.

Ljung, L. (1977): "Analysis of Recursive Stochastic Algorithms," *IEEE Transactions on Automatic Control*, 22, 551–575.

Lucas, Jr., R. E. (1972): "Expectations and the Neutrality of Money," *Journal of Economic Theory*, 4, 103–124.

——— (1978): "Asset Prices in an Exchange Economy," *Econometrica*, 46, 1429–1445.

Lucas, Jr., R. E., and E. C. Prescott (1971): "Investment under Uncertainty," *Econometrica*, 39, 659–681.

Magnus, J., and H. Neudecker (1988): *Matrix Differential Calculus*. Wiley, New York.

Marcet, A., and T. J. Sargent (1989): "Convergence of Least-Squares Learning Mechanisms in Self-Referential Linear Stochastic Models," *Journal of Economic Theory*, 48, 337–368.

Marimon, R., E. McGrattan, and T. Sargent (1990): "Money as Medium of Exchange with Artificially Intelligent Agents," *Journal of Economic Dynamics and Control*, 14, 329–373.

Marimon, R., and S. Sunder (1993): "Indeterminacy of Equilibria in a Hyperinflationary World: Experimental Evidence," *Econometrica*, 61, 1073–1107.

——— (1994): "Expectations and Learning under Alternative Monetary Regimes: An Experimental Approach," *Economic Theory*, 4, 131–162.

Muth, J. F. (1961): "Rational Expectations and the Theory of Price Movements," *Econometrica*, 29, 315–335.

Preston, B. (2005): "Learning about Monetary Policy Rules when Long-Horizon Expectations Matter," *International Journal of Central Banking*, 1, 81–126.

Sargent, T. J. (1973): "Rational Expectations, the Real Rate of Interest and the Natural Rate of Unemployment," *Brookings Papers on Economic Activity*, 2, 429–472.

——— (1987): *Macroeconomic Theory,* second edition. Academic Press, New York.

——— (1993): *Bounded Rationality in Macroeconomics*. Oxford University Press, Oxford.

——— (1999): *The Conquest of American Inflation*. Princeton University Press, Princeton NJ.

Sargent, T. J., and J. Vilmunen (eds.) (2013): *Macroeconomics at the Service of Public Policy*. Oxford University Press.

Stokey, N., and R. E. Lucas Jr. (1989): *Recursive Methods in Economic Dynamics*. Harvard University Press, Cambridge, Mass.

Watkins, C. (1989): *Learning from Delayed Rewards*. PhD Thesis, University of Cambridge, Cambridge, UK.

Watkins, C., and P. Dayan (1992): "Technical Note: Q-learning," *Machine Learning*, 8, 279–292.

Williams, N. (2014): "Escape Dynamics in Learning Models," Working paper, University of Wisconsin.