LIMITS OF PASSIVE LEARNING IN THE BAYESIAN DUAL CONTROL OF DRIFTING COEFFICIENT REGRESSION

SERGEI MOROZOV

Abstract. We study the quality of passively adaptive approximations to both passively adaptive optimal and actively adaptive optimal solutions to the Bayesian dual control problem when coefficients of the target state evolution drift continuously as in Beck and Wieland (2002). Amongst the passive learning approaches we compare the performance of certainty equivalent control, anticipated utility policy, limited lookahead and Markov jump-linear-quadratic approximation. Solutions featuring active experimentation are of two kinds - the solution to the original infinite horizon dual control problem found by Dynamic programming algorithm, and its one-period limited lookahead version. Certainty equivalent and actively optimal policies displays the largest amount of experimentation, accidental for the former and intentional for the latter. While we find only modest differences *in expectation* between more advanced passive policies on the one hand and either of the active policies on the other, the fully optimal active policy is the only one robust to unfortunate rare draws and prevents partitioning of the state space into two basins of attraction with escape-like dynamics between the two. In addition, anticipated utility policy and approximating Markov jump-linear-quadratic policy with small number of regimes are hard to distinguish, upholding computational advantages of anticipated utility.

JEL classification: C44; C63; D83; E17; E52

Keywords: Bayesian dual control, certainty equivalence, anticipated utility, Markov Jump Linear Quadratic control, drifting coefficients, passively optimal control, active experimentation, limited lookahead

To me there is something thrilling and exalting in the thought that we are drifting forward into a splendid mystery – into something that no mortal eye hath yet seen, and no intelligence hath yet declared.
–Edward Chapin

1. Introductory Remarks

Imperfect information in the form of model uncertainty in the dynamic intertemporal choice problems makes the optimizing decision-maker confront difficult compromise between simultaneously stabilizing the policy target and estimating the impact of policy action. Simultaneous solution to a combined control and sequential design of experiment problem is known as the dual control and was originally introduced and discussed by A. A. Feldbaum in a sequence of four seminal papers from 1960 and 1961 (Feldbaum, 1960a,b, 1961a,b). Feldbaum was the first to show that, in principle, the optimal solution can be found by dynamic programming, via what later became known as Bellman functional equation. The numerical problems when solving the functional equation are very large and only few simple examples have been solved. More so, it is difficult to state conditions under which the solution to the imperfect information dynamic programming problem actually exists. Accordingly, an entire genres of economic and engineering literatures have been devoted to finding simpler suboptimal solutions and their comparison with dual optimal dual ones when they could be found. This brief note is in the same lineage.

Date: April 23, 2009.
Version 1.03.
Usual disclaimers apply. All remaining errors are the responsibility of the second author.

Specifically, we revisit a problem of controlling a regression with continuously evolving coefficients that was studied in Beck and Wieland (2002). Beck and Wieland allow the parameter that is multiplicative to the decision variable to drift away in a random walk fashion away from its initial value. We expand the number of suboptimal approximate policies to include not just certainty equivalent control but also anticipated utility policy, Markov Jump Linear Quadratic approximation, leading to the limiting case of the optimal passively adaptive control. The development of the passively adaptive optimal policy is new for the class of dual control models with dynamic uncertainty as is the adaptation of Markov jump-linear-quadratic control to the systems with continuous drift. In addition, we include an example of suboptimal actively adaptive policy from a family of limited lookahead controls.

Brief synopsis of the paper is as follows. Section 2 sets the stage by outlining a particular model of Bayesian dual control of drifting coefficient regressions. Section 3 characterizes the actively adaptive optimal control that fully balances the tradeoff between stabilization and experimentation. Section 4 provides new analytic bounds on the optimal cost-to-go function and on the optimal policy function. These could be used to accelerate the dynamic programming algorithm by refining initial guesses. Sections 5 through 9 map out various suboptimal approximations which are made convenient by way of ignoring some aspects of the decision problem. In particular, section 5 develops certainty equivalent adaptive approach that shuts down uncertainty about the coefficients of the state transition equation, setting them equal to the current mean estimate. The policy is adaptive because the parameter estimates are updated, thus adapted, every period. Next, section 6 relaxes the certainty assumption by letting the decision maker surround the policy effectiveness parameter with a cloud of uncertainty while restricting this uncertainty to be both time-invariant and immune to the choice of policy. This is the so called anticipated utility approximation that leaves the policy maker of two minds as the controlled process unfolds over time. On the one hand, the policy is clearly adaptive since the estimates of the uncertain parameter are updated every period. On the other, both the future coefficient drift and the future learning (i.e. future updating of parameter estimates) do not feed back on the current policy choice. Section 7 drives the treatment of uncertain dynamic coefficients one step further by allowing the future parameter dynamics to feature prominently in the mind of optimizing agent, albeit in a different form. The form of evolving coefficient dynamics is given by the Markov jump-linear system where the coefficients transitions are governed by a regime-switching process. Section 8 takes passively adaptive class of solutions to the ultimate limit. In this limit, the future coefficient dynamics is correctly anticipated but is deemed not impacted by the current control. In section 9 we change gears by offering a suboptimal alternative with the full recognition of experimentation incentive and continuous coefficient drift but only looking ahead one period. Section 10 deals with the six-way comparison amongst various alternatives. We compare policy functions as well as expected loss functions, expected state transitions, and expected beliefs. The contrasting features are illustrated with simulated outcomes under different policies. We explore evolving distributions of simulated outcomes as well as persistence properties of simulated time series, and how they are impacted by the model parameters. We diagnose the aspects of the model that influence the differences in outcomes and the size of probing component in particular. In addition, we comment on computational demands of various approximating frameworks. Lastly, section 11 offers concluding remarks and suggests profitable agenda for future research.

2. Dual Control of Drifting Coefficient Regressions

The objective of control is to stabilize the target variable x_t around its target value x^* while exercising control u_t in the sense of minimizing the discounted sum of squared

deviations:

$$\min_{\{u_t\}_{t=0}^{\infty}} \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \delta^t \left((x_t - x^*)^2 + \omega (u_t - u^*)^2 \right) \right], \tag{2.1}$$

subject to

$$x_t = \alpha + \beta_t u_t + \gamma x_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{2.2}$$

$$\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2). \tag{2.3}$$

α and γ are assumed to be known. Shock variances are known as well. Naturally, $\omega \geq 0$, $\delta \in [0, 1)$. Initial belief about β_0 is Gaussian with mean μ_0 and variance Σ_0. The timing assumption is such that, technically speaking, u_{t+1} is measurable with respect to filtration $\mathcal{F}_t$ generated by histories of stochastic process up until time t.

The focus here is on the time-varying uncertainty regarding a parameter that is multiplicative to the decision variable because this type of parameter is crucial for the tradeoff between current control and estimation. Time variation of the impact of policy action encapsulates the idea of the continuously adapting economic environment, driven perhaps by response of economic agents engaged in the larger dynamic game that is abstracted away here. More generally, time-varying parameter uncertainty captures the absence of consensus concerning stability of data generating process over time (including regime switches, threshold effects, or continuous adaptation). This kind of uncertainty has found its way into many recent macroeconomic papers. For instance, Canova (2006) documents the lack of posterior tightening as new data becomes available in the time-invariant small-scale New Keynesian model of the US economy. Cogley and Sargent (2001) detect important departures from time-invariance in the US inflation dynamics as well.

Beck and Wieland (2002) show that optimal control involves a certain degree of active learning (experimentation) but to a lesser extent than in the model without time variation in β_t, and also less aggressive then for a certainty-equivalent rule that completely disregards parameter uncertainty. The reason is similar in both cases. The expected payoff to learning current parameter value is reduced once it is recognized that the parameter will change again, or parameter uncertainty is assumed away altogether. On the other hand, time-variation in the unknown parameter implies also that the incentive to experiment never disappears, and so experimentation will not die out.

The beliefs about β_t are normally distributed, because the prior distribution and likelihood functions for each t are all Gaussian. Hence, the beliefs about β_t are completely characterized by the mean and variance parameters μ_t, and Σ_t conditional on current information set, i.e. following the choice of control u_t and realization of the shock ϵ_t. By applying recursive Bayesian updating in the linear regression setup, the following updating equations can be derived:

$$\mu_t = \mu_{t-1} + \Sigma_{t|t-1} u_t \left(u_t^2 \Sigma_{t|t-1} + \sigma_\epsilon^2 \right)^{-1} \left(x_t - \alpha - \mu_{t-1} u_t - \gamma x_{t-1} \right), \tag{2.4}$$

$$\Sigma_t = \Sigma_{t|t-1} - (\Sigma_{t|t-1})^2 u_t^2 \left(u_t^2 (\Sigma_{t|t-1}) + \sigma_\epsilon^2 \right)^{-1}, \tag{2.5}$$

where $\Sigma_{t|t-1} = \Sigma_{t-1} + \sigma_\eta^2$ is the conditional predictive variance of the hidden state β_t.

Here, learning is equivalent to Kalman filtering. The updating equation for variance is the deterministic process, which would be non-increasing if $\sigma_\eta^2 = 0$. In other words, if multiplicative policy parameter β_t were not time-varying, the learning would eventually converge.

Endowing decision-maker with the knowledge of econometrics sets in motion the dynamic view of the system as one where policy decisions are made on the basis of the current observed physical state and current available information, the stochastic elements are realized, new observations of the physical state are made, beliefs are updated and the process repeats itself. Note that even in the absence of explicit autoregressive dynamics of the physical state, the overall system dynamics is path-dependent through the information accumulation channel. Information becomes new state variable. The combined state which we'll be referring to as

extended state[1] is

$$S_t = (x_t, \mu_t, \Sigma_t) \in \mathcal{S}. \tag{2.6}$$

Keeping track of information state in addition to the physical state with information state is both a major headache and a major conceptual breakthrough. The breakthrough originates in the demonstration of the formal equivalence between the Markovian decision model with extended state and the original non-Markovian formulation by Hinderer (1970). The headache sprouts from the realization that the information state is, in general, infinitely-dimensional, as it is encoded by a continuous distribution. Keeping track of distributions could be exceedingly hard unless the full arsenal of Bayesian tricks is used, such as the adoption of conjugate prior distributions and model likelihoods. This is the assumption we made here so that the information state is captured by the two sufficient statistics, that evolve according to (2.4)-(2.5).

Forward-looking decisions are made in the view of rewards and losses accruing to the future state, including the future information state. In the same manner as future state is manipulated by the use of current control, same control can be used to impact the future information to the policy-maker's advantage. Doing so is the essence of directed or active learning. To the extent that manipulation of future information flows comes at the expense of current stabilization goal, the control has *dual*, conflicting objectives. This makes the two types of state variables hard to disentangle and poses critical computational challenge.

Under arbitrary policy rule $u : \mathcal{S} \to \mathbb{R}$ we can compute expectation of future state conditional on the current information state:

$$\mathbb{E}_t x_{t+1} = \alpha + \mu_{t+1|t} u_{t+1} + \gamma x_t, \tag{2.7}$$

$$\mathbb{E}_t \Sigma_{t+1} = \Sigma_t + \sigma_\eta^2 - (\Sigma_t + \sigma_\eta^2)^2 u_t^2 \left(u_t^2(\Sigma_t + \sigma_\eta^2) + \sigma_\epsilon^2\right)^{-1}. \tag{2.8}$$

By law of iterated expectations, expected evolution of the mean beliefs is trivial

$$\mathbb{E}_t \mu_{t+1|t} = \mu_t. \tag{2.9}$$

The actively adaptive optimal solution to the problem (2.1) that incorporates the experimentation motive will be shown in the section 3. Sections 5 through 8 develop various approximate suboptimal solutions. Section 9 adds one example of actively adaptive suboptimal policy to the jamboree by considering one-period limited lookahead control. Section 10 illustrates the relationships various policies have amongst themselves especially as it pertains to exploration (intentional or not) and simulated losses.

3. Actively Adaptive Optimal Control

The dynamic program (2.1) has three natural state variables - "physical" state variable x_t, and two informational state variables describing beliefs about the impact of the policy choice – mean predictive belief $\mu_{t+1} = \mu_{t+1|t}$ and predictive belief variance $\Sigma_{t+1|t}$. The Bellman equation associated with stationary optimal policy is given by

(3.1)

$$\begin{aligned} &V(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}) \\ &= \min_{\{u_{t+1}\}} \Big\{ L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) \\ &+ \delta \int V\left(\alpha + \beta_{t+1} u_{t+1} + \gamma x_t + \epsilon_{t+1}, u_{t+1}, \mu_{t+2}(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}), \Sigma_{t+2|t+1}(\Sigma_{t+1|t}, u_{t+1})\right) \\ &\times p(\beta_{t+1}|x_t, \mu_{t+1|t}, \Sigma_{t+1|t}) q(\epsilon_{t+1}) d\beta_{t+1} d\epsilon_{t+1} \Big\} \end{aligned}$$

[1] Kumar (1985) refers to this extended state as *hyperstate*.

where $L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_t)$ is expected one-period loss

$$
\begin{aligned}
L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) &= \int \left((\alpha + \beta_{t+1} u_{t+1} + \gamma x_t + \epsilon_{t+1} - x^*)^2 + \omega (u_{t+1} - u^*)^2\right) \\
&\times p(\beta_{t+1}|x_t, \mu_{t+1|t}, \Sigma_{t+1|t}) q(\epsilon_{t+1}) d\beta_{t+1} d\epsilon_{t+1} \\
&= \left(\Sigma_{t+1|t} + \mu_{t+1|t}^2\right) u_{t+1}^2 + 2\left(\gamma \mu_{t+1|t} x_t + \mu_{t+1|t}(\alpha - x^*)\right) u_{t+1} \\
&+ 2\gamma(\alpha - x^*) x_t + \gamma^2 x_t^2 + \sigma_\epsilon^2 + (\alpha - x^*)^2 + \omega (u_{t+1} - u^*)^2,
\end{aligned} \tag{3.2}
$$

and we exploited the fact that variance updating is a deterministic process. In both (3.1) and (3.2) $p(\cdot)$ is a Gaussian density representing posterior beliefs about the drifting parameter, while q similarly describes Gaussian distribution of physical state innovation.

Although the stochastic process under control is linear and the loss function is quadratic, the belief updating equations are non-linear, and hence the dynamic optimization problem is more difficult than those in the class of linear quadratic problems. Following Easley and Kiefer (1988), it could be shown that Bellman functional operator is a contraction and a stationary optimal policy exists such that corresponding value function is continuous and satisfies the above Bellman equation. Accordingly, the optimal policy and value functions can be obtained by numerical dynamic programming methods. In particular, we use a combination of the value and policy iterations on the three-dimensional grid in the state-space with the integration step in (2.1) carried out with the help of Gauss-Hermite quadrature and tri-linear interpolation.

Figure 1 draws three slices of the actively adaptive optimal policy function. The top slice is a function of x_t and μ_t when Σ_t is fixed at 0.05. The middle panel in the figure represents the policy function in x_t and Σ_t variables when $\mu_t = -1.64$. The bottom panel contains the plot of the policy function against μ_t and Σ_t with $x_t = 2.2$. In addition, figure 2 is a volumetric plot that summarizes the policy function against all three dimensions by color coding function values. Areas of rapid change in the shape of the policy function are given by multiple color regions. Log-scale for the variance makes the features stand out more.

Under the actively optimal policy, equations (2.7) and (2.8) can be iterated forward to generate the path of the expected state The path would be realized if all future target state shocks were zero, $\epsilon_{t+\tau} = 0$, controls followed the optimal policy rule, but the unobserved multiplicative policy coefficient continued to drift randomly. The phase portrait for the dynamical systems of state expectations is given in figure 3 together with a representative path. The phase portrait implies convergence towards the target x^* in the long run. The uncertainty about the multiplicative policy coefficient begins to increase once the incremental progress towards the target slows sufficiently. This is because identification/learning needs variability of system inputs and outputs.

4. Useful Bounds on Actively Optimal Policy

In addition to reimplementation of dynamic programming algorithm, I derived new analytic bounds on the optimal cost-to-go function and on the optimal policy function. The bounds could be used to accelerate the dynamic programming algorithm by refining initial guesses. The optimal cost-to-go bound can be derived via analytic q-factor of the *inert* policy ($u_{t+\tau} \equiv 0 \, \forall \tau \geq 0$) and is as follows:

$$
\begin{aligned}
V_t^* \leq V_t^0 &:= \mathbb{E}_{t-1} \sum_{\tau=0}^{\infty} \delta^\tau \left((x_{t+\tau} - x^*)^2 + \omega (u^*)^2\right) \\
&= \frac{(\alpha + \gamma x_{t-1} - x^*)^2 - \delta\gamma\left((x^*)^2 - \alpha^2 - \gamma x^*(2\alpha - x^*) + \gamma x_{t-1}^2(1+\gamma) - 2x_{t-1}(x^* - \alpha + \gamma^2 x^*)\right)}{(1-\delta)(1-\gamma\delta)(1-\gamma^2\delta)} \\
&+ \frac{\gamma^3\delta^2 (x_{t-1} - x^*)^2}{(1-\delta)(1-\gamma\delta)(1-\gamma^2\delta)} + \frac{\sigma_\epsilon^2}{(1-\delta)(1-\gamma^2\delta)} + \frac{\omega (u^*)^2}{1-\delta}.
\end{aligned} \tag{4.1}
$$

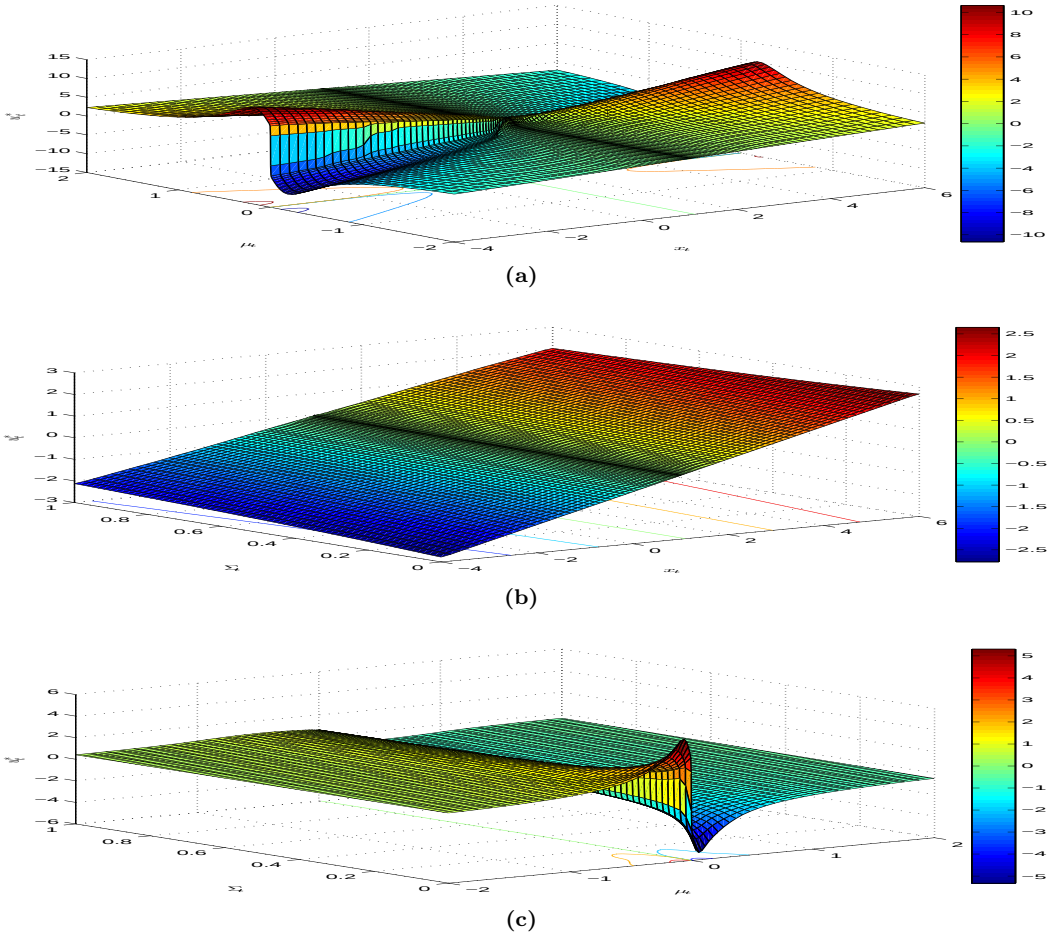


Figure 1: Actively adaptive optimal control. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$. (a) $\Sigma_t = 0.5$ slice; (b) $\mu_t = -1.64$ slice; (c) $x_t = 2.2$ slice.

The bound does not depend on the belief state components. The tightness of the bound is tested with the help of figure 4 which displays both the actively optimal cost-to-go and it's analytical bound. Evidently the bound is not very tight away from the target x^* and from $\mu_t = 0$ belief subspace.

Expression 4.1 is not the only analytic q-factor available. Convenient independence of belief evolution allows us to synthesize an analytic formula for the q-factor of the so-called *pseudo-myopic* policy, which which is the optimal policy choice when the continuation cost-to-go is equal to that of the inert policy. After some tedious algebra best relegated to computer algebra systems, we obtain

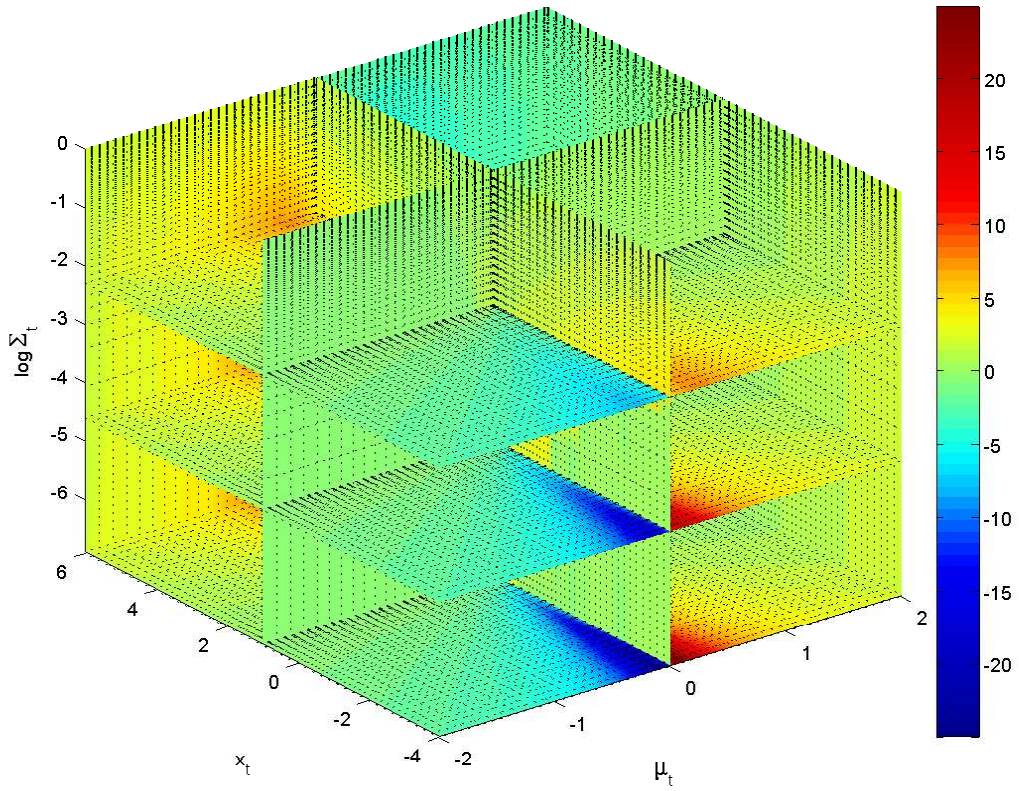


Figure 2: Volumetric plot of actively adaptive optimal policy function. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$.

(4.2)

$$
\begin{aligned}
V_t^{pm}(x_t, \mu_t, \Sigma_{t+1|t}) &= \min_{u_{t+1}} \mathbb{E}_t \left\{ L(x_{t+1}, u_{t+1}) + \delta V^0(x_{t+1}) \right\} \\
&= \min_{u_{t+1}} \Big\{ \mathbb{E}_t \left[(\alpha + \beta_{t+1} u_{t+1} + \gamma x_{t+1} + \epsilon_{t+1} - x^*)^2 + \omega \left(u_{t+1} - u^* \right)^2 \right] \\
&+ \delta \mathbb{E}_t V^0 \left(\alpha + \beta_{t+1} u_{t+1} + \gamma x_{t+1} + \epsilon_{t+1} \right) \Big\} \\
&= \frac{2\alpha\gamma x_t(1-\delta) + \alpha^2(1+\gamma\delta) - 2x^*(\alpha + x_t\gamma(1-\delta))(1-\gamma^2\delta)}{(1-\delta)(1-\gamma\delta)(1-\gamma^2\delta)} \\
&+ \frac{(x^*)^2(1-\gamma\delta)(1-\gamma^2\delta) - (1-\gamma\delta)(\gamma^2(-1+\delta)x_t^2 - \sigma_\epsilon^2 - (u^*)^2(1-\gamma^2\delta)\omega)}{(1-\delta)(1-\gamma\delta)(1-\gamma^2\delta)} \\
&- \frac{(1-\gamma^2\delta)\left(\frac{(\alpha - x^* + \gamma x - (x - x^*)\gamma^2\delta)\mu_t}{(1-\gamma\delta)(1-\gamma^2\delta)} + u^*\omega \right)^2}{\mu_t + \omega - \gamma^2\delta\omega + \Sigma_{t+1|t}}.
\end{aligned}
$$

Performance of V_t^{pm} as a bound is studied in figure 5. Although the new bound seems not very attractive, there are regions in the state space where it outperforms V_t^0.

Since the minimum of the two upper bounds is also an upper bound, we define combined bound

$$
V_t^{0,pm}(x_t, \mu_t, \Sigma_{t+1|t}) = min \left\{ V_t^0(x_t), V_t^{pm}(x_t, \mu_t, \Sigma_{t+1|t}) \right\} \tag{4.3}
$$

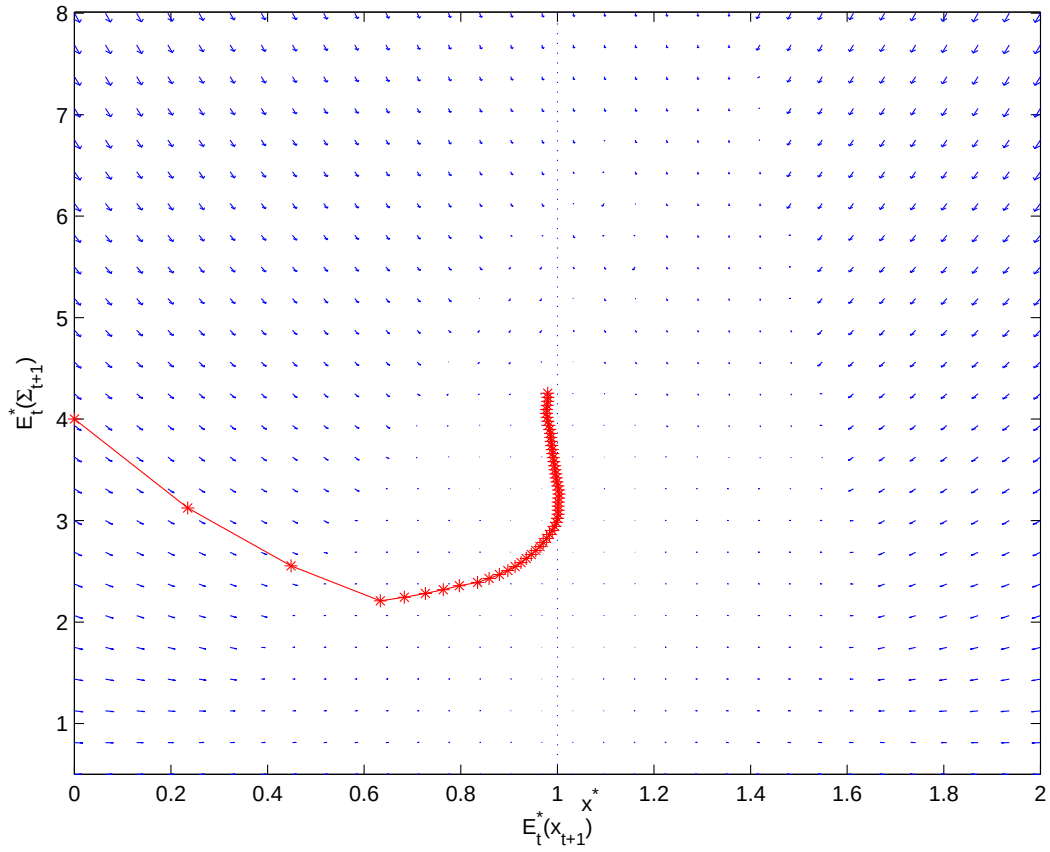


Figure 3: Phase portrait of expected state dynamics under actively optimal policy. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$.

This bound is nontrivial improvement since there is no uniform dominance among V_t^{pm} and V_t^0. The combined bound could then be used to derive a bound in the policy space, which is given by the following expression.

$$\begin{aligned} &\frac{-\mu_t(\alpha + \gamma x_t - x^*) - \omega u^* - \sqrt{D}}{\mu_t^2 + \Sigma_{t|t} + \sigma_\eta^2 + \omega} \\ &\leq u_{t+1}^* \\ &\frac{-\mu_t(\alpha + \gamma x_t - x^*) - \omega u^* + \sqrt{D}}{\mu_t^2 + \Sigma_{t|t} + \sigma_\eta^2 + \omega}, \end{aligned} \tag{4.4}$$

where

$$\begin{aligned} D = &\left(\mu_t(\alpha + \gamma x_t - x^*) - \omega u^*\right)^2 \\ &- \left(\mu_t^2 + \Sigma_{t|t} + \sigma_\eta^2 + \omega\right)\left((\alpha + \gamma x_t - x^*)^2 + \sigma_\epsilon^2 + \omega(u^*)^2 - V_t^{0,pm}(x_t, \mu_t, \Sigma_{t+1|t})\right). \end{aligned}$$

Casual inspection of the bounds' distance to the optimal policy in figure 6 suggests that the midpoint could be a reasonable guess for the optimization steps in the dynamic programming algorithm.

5. Certainty Equivalent Policy

The certainty equivalent policy rule corresponds to the optimal strategy that disregards parameter uncertainty and belief updating. In other words, the decision maker behaves as if

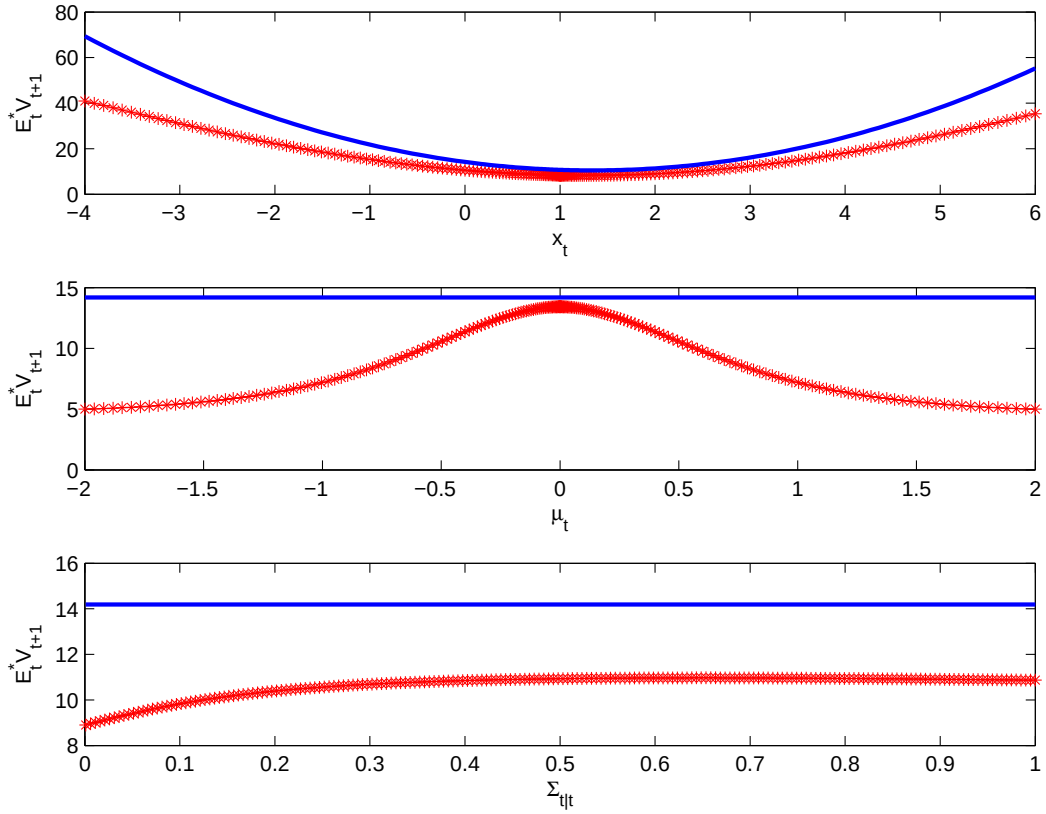


Figure 4: Analytic bound for the actively optimal cost-to-go function in the model with active learning and dynamic model uncertainty. Parameter values: $\alpha = 0$, $\gamma = 0.9$, $\delta = 0.75$, $\omega = 1$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = 1$, $\sigma_\eta^2 = 0.04$. Fixed coordinates in the top panel: $\mu_t = -0.5$, $\Sigma_{t|t} = 0.25$. Fixed coordinates in the middle panel: $x_t = 0$, $\Sigma_{t|t} = 0.25$. Fixed coordinates in the bottom panel: $x_t = 0$, $\mu_t = -0.5$.

he knows the impact of policy action perfectly and assumes that the impact does not change over time. While he does not ignore the state noise ϵ_{t+1}, it turns out that the optimal choice of policy is the same for all σ_ϵ^2. In particular, it is equal to the control that would obtain under $\sigma_\epsilon^2 = 0$, i.e. in the absence of noise altogether.

Certainty equivalent approach is known to be optimal in the standard linear quadratic problems with measurement error (Hansen and Sargent, 2004), but it is definitely not optimal in the case of multiplicative parameter uncertainty. Nevertheless, it constitutes a useful benchmark, and an important competitor among various approximations.

Certainty equivalent policy is a solution of the following stationary Bellman equation:

$$V^{CE}(x_t) = \min_{u_{t+1}} \Big\{ \mathbb{E}_t \left(\alpha + \beta_{t+1} u_{t+1} + \gamma x_t + \epsilon_{t+1} - x^*\right)^2 + \omega (u_{t+1} - u^*)^2 + \delta \mathbb{E}_t V^{CE} \left(\alpha + \beta_{t+1} u_{t+1} + \gamma x_t + \epsilon_{t+1}\right) \Big\}. \tag{5.1}$$

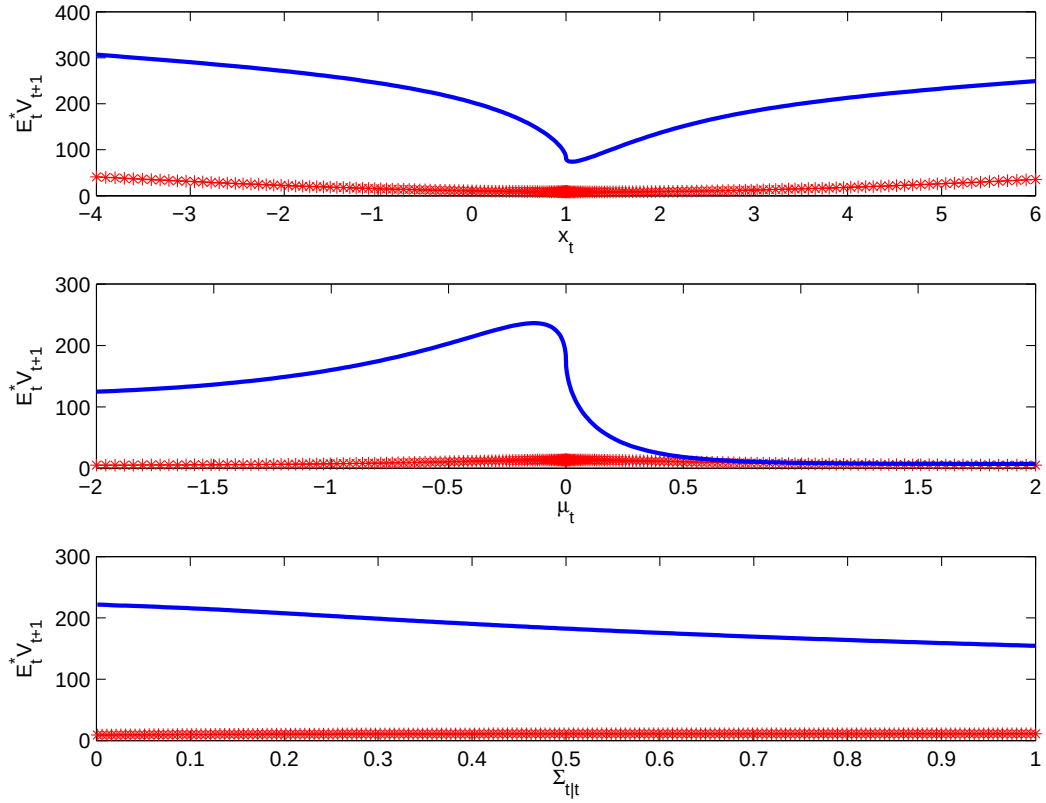


Figure 5: Analytic bound for the actively optimal cost-to-go function based on the pseudo-myopic policy in the model with active learning and dynamic model uncertainty. Parameter values: $\alpha = 0$, $\gamma = 0.9$, $\delta = 0.75$, $\omega = 1$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = 1$, $\sigma_\eta^2 = 0.04$. Fixed coordinates in the top panel: $\mu_t = -0.5$, $\Sigma_{t|t} = 0.25$. Fixed coordinates in the middle panel: $x_t = 0$, $\Sigma_{t|t} = 0.25$. Fixed coordinates in the bottom panel: $x_t = 0$, $\mu_t = -0.5$.

Conjecture that $V^{CE}(x) = Ax^2 + 2Bx + C$ for all x. Then

$$\begin{aligned} Ax^2 + 2Bx + C = \min_u \Big\{ & \left(\mu^2 + \omega + \delta\mu^2 A\right) u^2 \\ & + 2\left(\mu\gamma x + \mu(\alpha - x^*) - \omega u^* + \delta\mu\gamma A x + \delta\alpha\mu A + \delta\mu B\right) u \\ & + \gamma^2 x^2 + \sigma_\epsilon^2 + (\alpha - x^*)^2 + 2\gamma(\alpha - x^*) + \omega(u^*)^2 \\ & + \delta A(\alpha + \gamma x)^2 + \delta\sigma_\epsilon^2 A + 2\delta B(\alpha + \gamma x) + \delta C \Big\}. \end{aligned}$$

Hence,

$$u_{t+1}^{CE} = -\frac{\mu\gamma(1 + \delta A)}{\mu^2(1 + \delta A) + \omega} x_t + \frac{\mu(x^* - \alpha) - \delta\mu(\alpha A + B) + \omega u^*}{\mu^2(1 + \delta A) + \omega}. \tag{5.2}$$

To implement (5.2) we'll need the values of constants A, B, and C.

Under (5.2), the value function becomes

$$\begin{aligned} V^{CE}(x) = & -\frac{\left(\mu\gamma(1 + \delta A)x + \mu(\alpha - x^*) - \omega u^* + \delta\alpha\mu A + \delta\mu B\right)^2}{\mu^2(1 + \delta A) + \omega} \\ & + \gamma^2 x^2 + (\alpha - x^*)^2 + (1 + \delta A)\sigma_\epsilon^2 + \omega(u^*)^2 + 2\gamma(\alpha - x^*)x \\ & + \delta A(\alpha + \gamma x)^2 + 2\delta B(\alpha + \gamma x) + \delta C \\ = & Ax^2 + 2Bx + C \end{aligned} \tag{5.3}$$

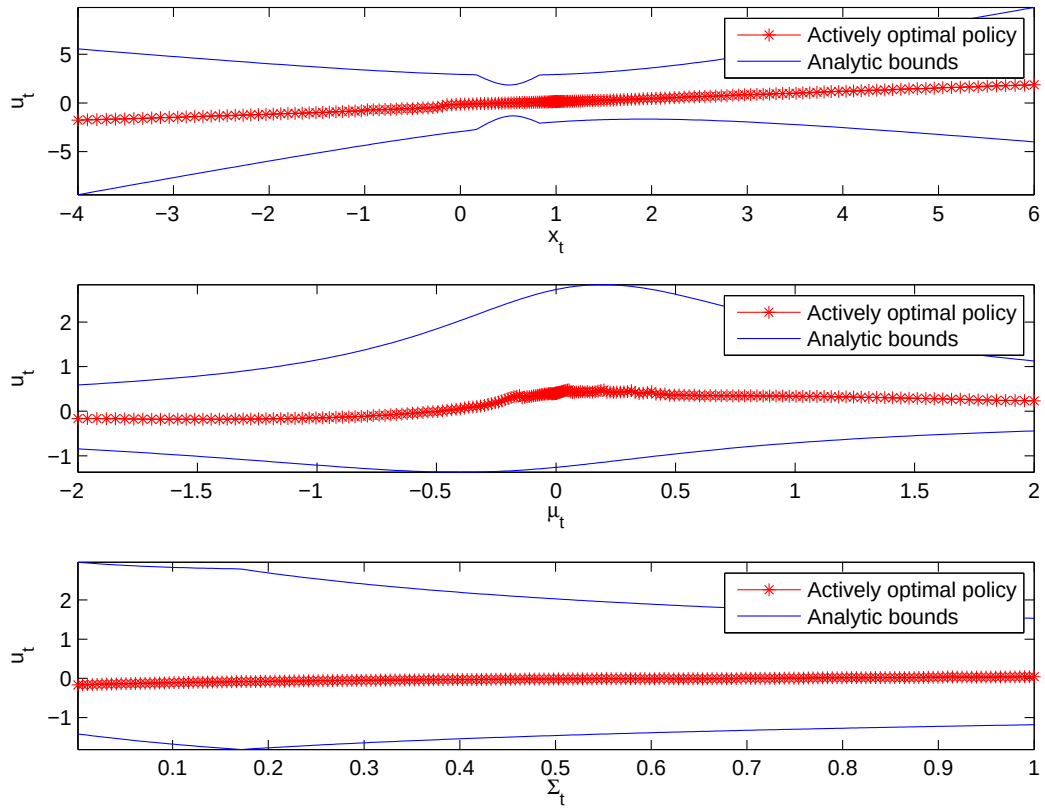


Figure 6: Analytic bound for the actively optimal policy function based on analytic cost-to-go bounds in the model with active learning and dynamic model uncertainty. Parameter values: $\alpha = 0$, $\gamma = 0.9$, $\delta = 0.75$, $\omega = 1$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = 1$, $\sigma_\eta^2 = 0.04$. Fixed coordinates in the top panel: $\mu_t = -0.5$, $\Sigma_{t|t} = 0.25$. Fixed coordinates in the middle panel: $x_t = 0$, $\Sigma_{t|t} = 0.25$. Fixed coordinates in the bottom panel: $x_t = 0$, $\mu_t = -0.5$.

for all $x \in \mathcal{X}$. Equating coefficients on the like powers of x yields three equations in three unknowns. The first one is

$$A = -\frac{\mu^2\gamma^2(1+\delta A)^2}{\mu^2(1+\delta A)+\omega} + \gamma^2(1+\delta A), \tag{5.4}$$

which is a one dimensional version of algebraic Riccati equation. Of the two roots, only one is positive and constitute the limit of time-dependent Riccati recursion associated with finite horizon problem. Notice that it becomes linear when $\omega = 0$ with

$$A = -\frac{\gamma(1-\mu\gamma)}{\mu(1-\delta\gamma^2)} \tag{5.5}$$

as a solution. The second equation is derived by collecting linear terms:

$$B = \frac{\mu\gamma(1+\delta A)\left(\mu(x^*-\alpha)+\omega u^* - \delta\alpha\mu A - \delta\mu B\right)}{\mu^2(1+\delta A)+\omega} + \delta\gamma B - \gamma(x^*-\alpha) + \delta\alpha\gamma A \tag{5.6}$$

which can be simplified to

$$B = \frac{\omega\gamma\left((1+\delta A)(\mu u^*+\alpha) - x^*\right)}{\mu^2(1+\delta A)+(1-\gamma\delta)\omega}. \tag{5.7}$$

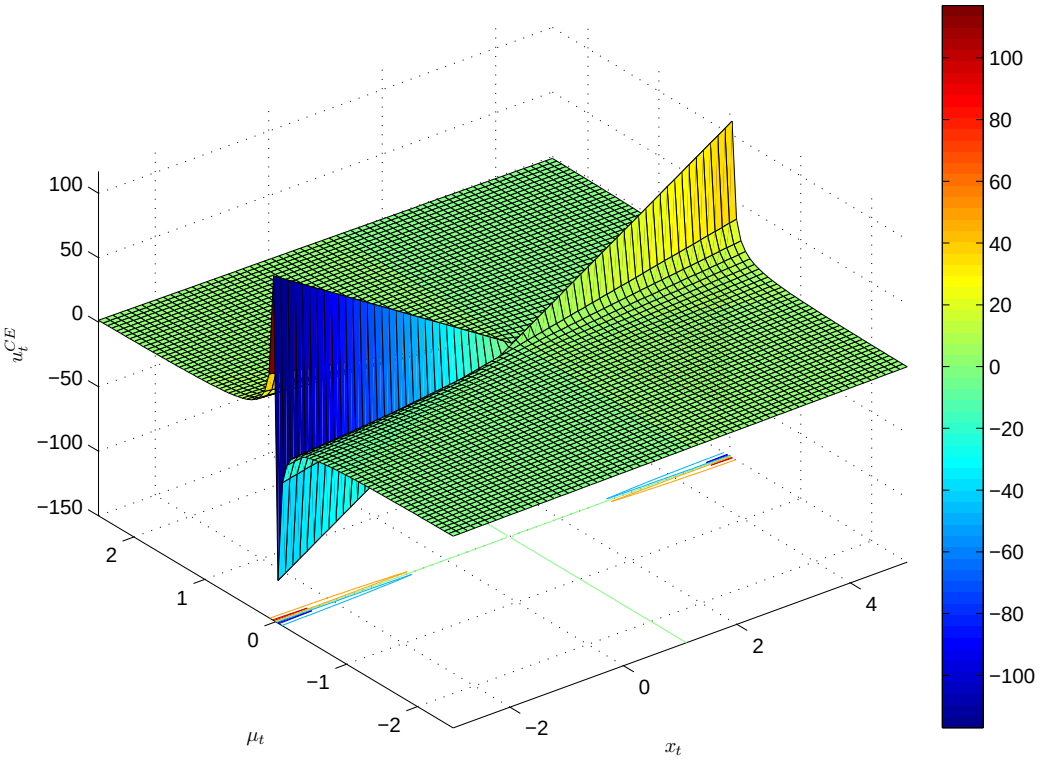


Figure 7: Certainty equivalent policy.

Finally, the third equation is obtained by equating constant terms:

$$(1-\delta)C = -\frac{\left(\mu(\alpha(1+\delta A)+\delta B - x^*) - \omega u^*\right)^2}{\mu^2(1+\delta A)+\omega} + (1+\delta A)\sigma_\epsilon^2 + (\alpha - x^*)^2 + \omega(u^*)^2 + \delta\alpha^2 A + 2\delta\alpha B. \tag{5.8}$$

Figure 7 depicts the policy response surface as function of physical and informational state variables x_t and μ_t. Since u_{t+1}^{CE} does not depend on Σ_t, plotting additional slices is redundant. As the certainty equivalent policy is linear in x_t, the impact of μ_t is to modify the slope and intercept.

The phase portrait for the dynamical systems of state expectations under the certainty equivalent control is given in figure 8 together with a representative path. The phase portrait implies convergence towards the target x^* in the long run. While convergence is not instantaneous with $\omega > 0$, the certainty equivalent policy makes rapid progress towards the target, covering 95% of the distance in only 7 steps. As before, the uncertainty about the multiplicative policy coefficient begins to increase in the vicinity of x^* because the variability of inputs and outputs becomes insufficient for the identification.

6. Anticipated Utility Policy

Anticipated utility problem differs from certainty equivalent formulation in that $\beta \sim \mathcal{N}(\mu_{t+1|t}, \Sigma_{t+1|t})$ conditional on the information at the end of date t, and this uncertainty is taken into account when formulating the decision rule in period t. The problem is known as Bayesian linear regulator (Cogley and Sargent, 2005). The fact that the belief about β will evolve over time is not taken in to account, however. The only natural state variable looking forward is x_t as beliefs are presumed to remain static. For this reason, and to

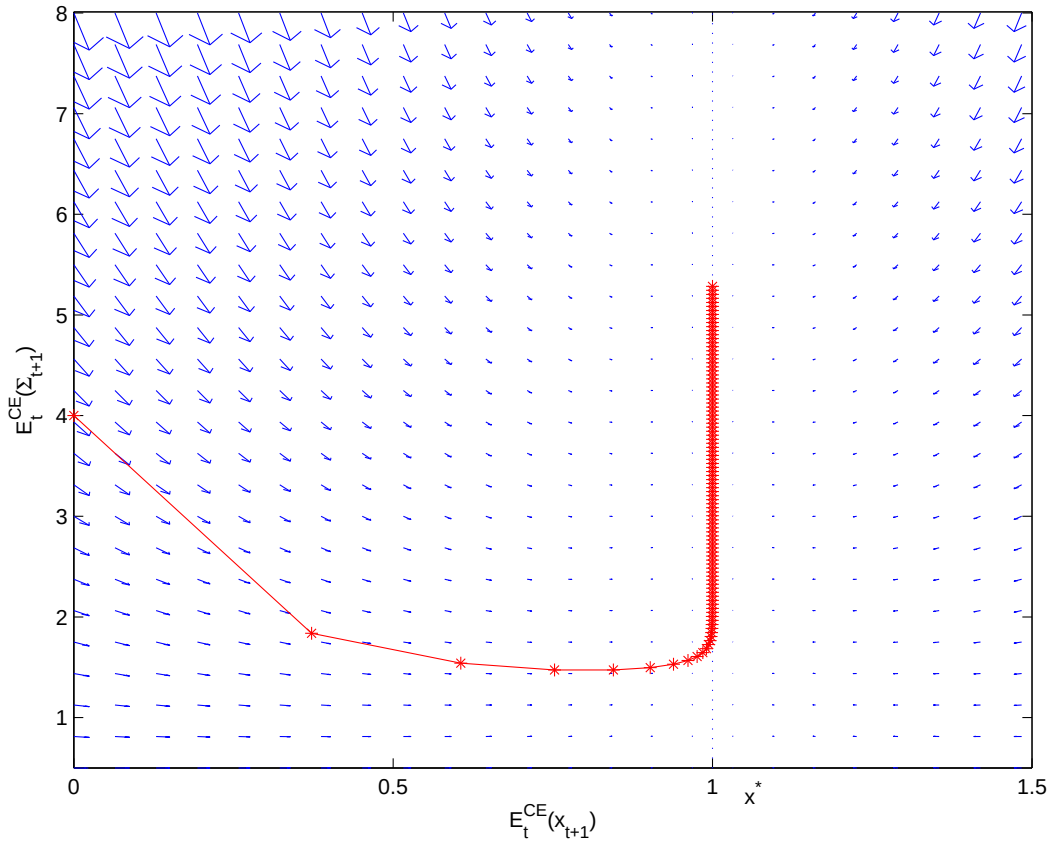


Figure 8: Phase portrait of expected state dynamics under certainty equivalent policy. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$.

simplify notation, we omit time subscripts on μ and $\Sigma' = \Sigma + \sigma_\eta^2$. It should be remembered, however, that these will be updated once x_{t+1} is observed and anticipated utility control will be recalculated.

Bellman equation that anticipated utility decision maker solves is

(6.1)

$$V(x) = \min_u \Big\{ \mathbb{E}\Big(\beta^2 u^2 + \gamma^2 x^2 + \sigma_\epsilon^2 + (\alpha - x^*)^2 + 2\beta\gamma xu + 2\beta u\epsilon + 2\beta(\alpha - x^*)u + 2\gamma x\epsilon + 2\gamma(\alpha - x^*)x + 2(\alpha - x^*)\epsilon + \omega u^2 - 2\omega u u^* + \omega(u^*)^2 + \delta V(\alpha + \beta u + \gamma x + \epsilon) \Big) \Big\}.$$

Conjecture quadratic value function $V(x) = Ax^2 + 2Bx + C$. Then

$$\begin{aligned} Ax^2 + 2Bx + C = \min_u \Big\{ & \left(\Sigma' + \mu^2\right) u^2 + \gamma^2 x^2 + \sigma_\epsilon^2 + (\alpha - x^*)^2 + 2\gamma\mu xu \\ & + 2\mu(\alpha - x^*)u + 2\gamma(\alpha - x^*)x + \omega u^2 - 2\omega u^* u + \omega(u^*)^2 \\ & + \delta A \left(\alpha^2 + (\Sigma' + \mu^2)u^2 + \gamma^2 x^2 + \sigma_\epsilon^2 + 2\gamma\mu xu + 2\alpha\mu u + 2\alpha\gamma x\right) \\ & + \delta B \left(\alpha + \mu u + \gamma x\right) + \delta C \Big\}. \end{aligned}$$

Performing explicit minimization, we get

$$\begin{aligned} u_{t+1}^{AU} &= -\frac{\gamma\mu(1+\delta A)x_t - \mu x^* + \alpha\mu(1+\delta A) - \omega u^* + \delta B\mu}{(\Sigma' + \mu^2)(1+\delta A) + \omega} \\ &= -\frac{\gamma(1+\delta A)\mu}{(\Sigma' + \mu^2)(1+\delta A) + \omega} x_t + \frac{(x^* - \alpha(1+\delta A) - \delta B)\mu + \omega u^*}{(\Sigma' + \mu^2)(1+\delta A) + \omega}, \end{aligned} \tag{6.2}$$

where A and B are yet to be determined. Notice that the slope does not, in fact, depend on α.

Substitute (6.2) into the cost to go function:

$$\begin{aligned} V^{AU}(x) = &-\frac{\left(\gamma\mu(1+\delta A)x + \mu(\alpha(1+\delta A) - x^* + \delta B) - \omega u^*\right)^2}{(\Sigma' + \mu^2)(1+\delta A) + \omega} \\ &+ \gamma^2 x^2 + \sigma_\epsilon + (\alpha - x^*)^2 + 2\gamma(\alpha - x^*)x + \omega(u^*)^2 \\ &+ \delta\alpha^2 A + \delta\gamma^2 A x^2 + \delta\sigma_\epsilon^2 A + 2\delta\alpha\gamma A x + \delta\alpha B + \delta\gamma B x + \delta C \\ &= Ax^2 + 2Bx + C. \end{aligned}$$

Equating like powers of x yields the following equations for the three unknown coefficients A, B, and C:

$$-\frac{-\gamma^2\mu^2(1+\delta A)^2}{(\Sigma' + \mu^2)(1+\delta A) + \omega} + \gamma^2(1+\delta A) = A, \tag{6.3}$$

$$\frac{\gamma\mu(1+\delta A)\left(\mu(x^* - \alpha(1+\delta A) - \delta B) + \omega u^*\right)}{(\Sigma' + \mu^2)(1+\delta A) + \omega} + \gamma(\alpha - x^*) + \delta\alpha\gamma A = (1-\delta\gamma)B, \tag{6.4}$$

and

$$-\frac{\left(\mu(x^* - \alpha(1+\delta A) - \delta B) + \omega u^*\right)^2}{(\Sigma' + \mu^2)(1+\delta A) + \omega} + (1+\delta A)\sigma_\epsilon^2 + (\alpha - x^*)^2 + \omega(u^*)^2 + \delta\alpha^2 A + \delta\alpha B = (1-\delta)C. \tag{6.5}$$

Equation for B can be made explicit

$$B = \frac{\alpha\gamma(1+\delta A)^2\Sigma' + \gamma\omega\left((1+\delta A)(\mu u^* + \alpha) - x^*\right) - \gamma\Sigma'(1+\delta A)x^*}{(1+\delta A)\left(\Sigma'(1-\gamma\delta) + \mu^2\right) + \omega(1-\gamma\delta)}. \tag{6.6}$$

The equation (6.3) that defines A generally has two roots, only one of which could be positive.

Figure 9 provides familiar-looking slices of the anticipated utility passively adaptive policy function that are defined by constraining one of the three state dimensions. Like the certainty equivalent policy, anticipated utility control is linear in x_t but is less aggressively sloped. Figure 10 presents the same information in the form of volumetric plot.

The phase portrait for the dynamical systems of state expectations under the anticipated utility rule is given in figure 11 together with a representative path. The decrease in the uncertainty about the multiplicative policy parameter is inconspicuous and the progress towards the target is at a measured pace. Because of small incremental steps, the learning process reverts relatively farther away from x^*.

7. Markov Jump Linear Quadratic Control

A very explicit but still relatively general form of model uncertainty that remains tractable is given by a so-called Markov jump-linear-quadratic (MJLQ) model, where multiplicative model uncertainty takes the form of different regimes that follow a finite-state Markov chain. Costa, Fragoso, and Marques (2005) devoted entire monograph to filtering, optimal control, partial information control and robust control of discrete time Markov jump linear systems.

As a way of introduction to MJLQ framework, let's assume that the state process takes the form of regime-switching linear system

$$X_{t+1} = A_{s(t+1)}X_t + B_{s(t+1)}U_{t+1} + C_{s(t+1)}\epsilon_{t+1}, \tag{7.1}$$

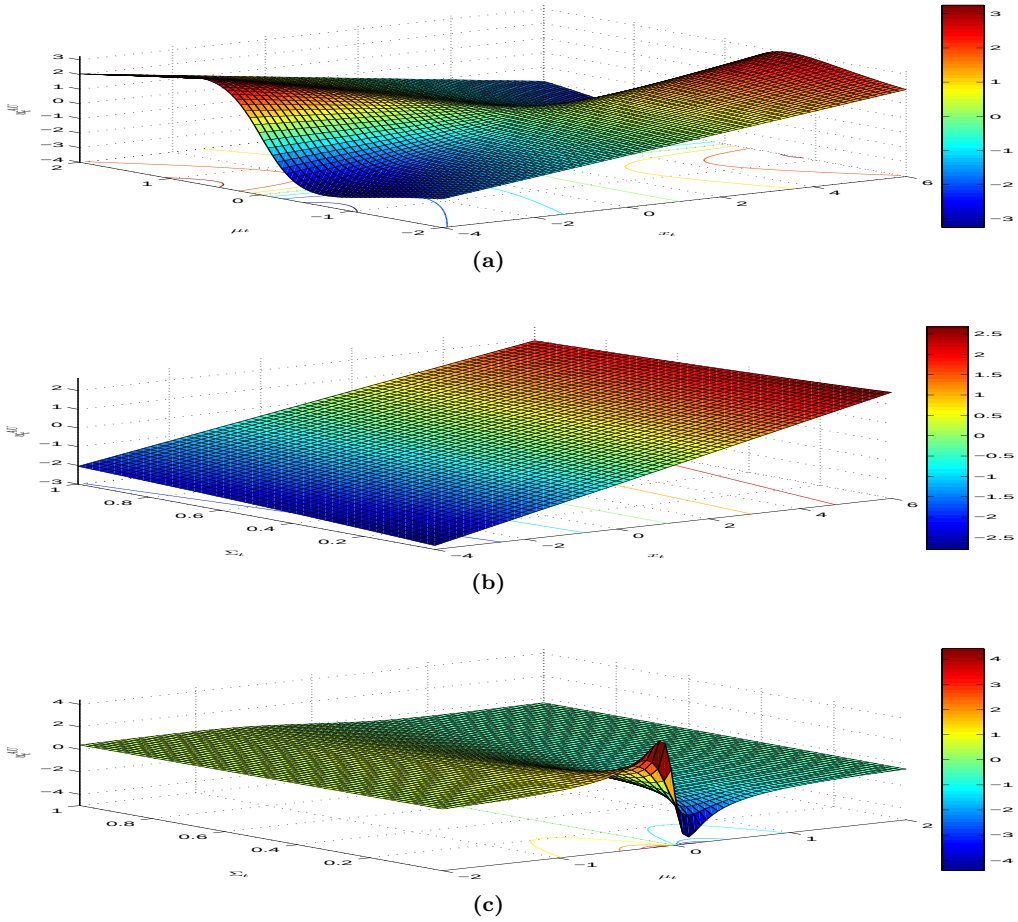


Figure 9: Anticipated utility passively adaptive control. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$. (a) $\Sigma_t = 0.5$ slice; (b) $\mu_t = -1.64$ slice; (c) $x_t = 2.2$ slice.

where coefficient matrices (of conforming dimensions, and with new state variable X_t subsuming the constant term) are random and can take any one of S different values in period $t+1$, corresponding to S regimes $s(t+1) = 1, \ldots, S$. The regimes follow a Markov process with constant transition probabilities,

$$P_{ij} = \Pr\{s(t+1) = j | s(t) = i\}, \quad i, j = 1, \ldots, S,$$

forming the transition probability matrix P. Furthermore, as the regimes are unobserved, the probability distribution over regimes in period t is non-trivial. That distribution, encoded with the vector $p_t = (p_{1t}, \ldots, p_{St})'$ evolves as

$$p_{t+1} = P' p_t.$$

Just like in the anticipated utility case under multiplicative uncertainty in the linear quadratic Gaussian case, the value function stays quadratic in the physical state X_t, but now with coefficients that depend on the distribution of regime probabilities.[2] Solution for the entire simplex of regime probabilities would require function approximation methods, but for any particular probability distribution over regimes, the solution could be obtained easily by using Riccati recursions over receding finite horizon control as we now show.

[2]If regimes were observed, coefficients of the quadratic value function would be directly regime-dependent.

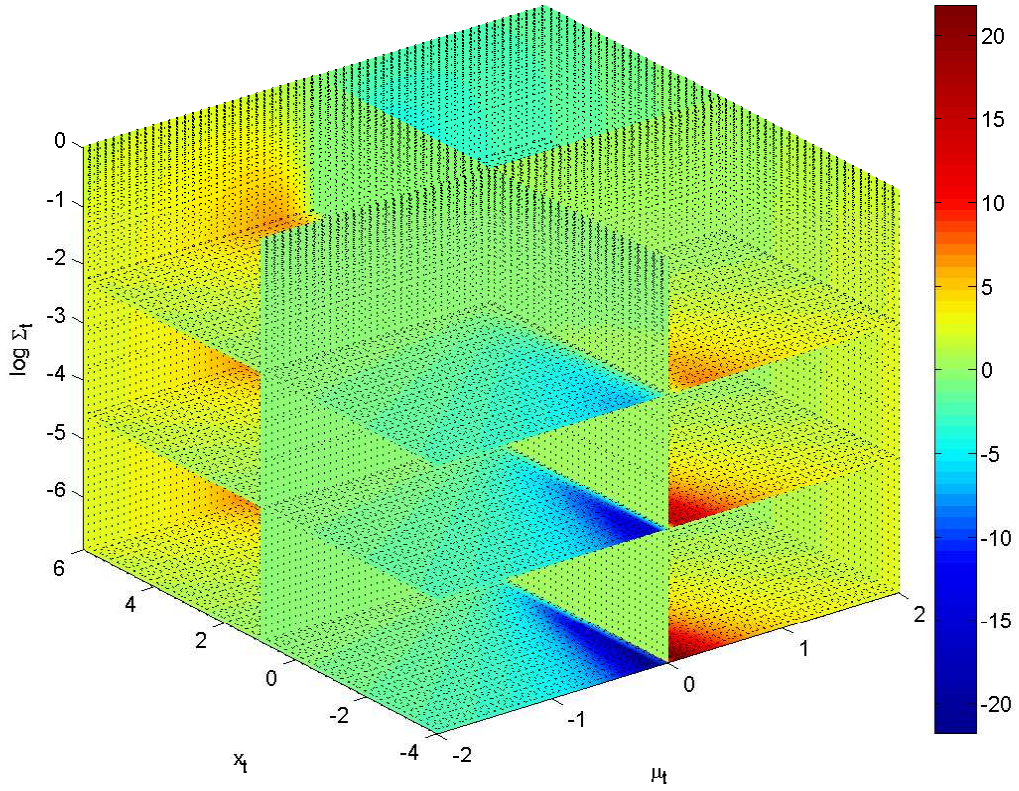


Figure 10: Volumetric plot of anticipated utility passively adaptive policy function. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$.

The Bellman equation for MJLQ model becomes

(7.2)

$$\begin{aligned} J(X_t, p_{t+1}) &= X_t' V(p_{t+1}) X_t + w(p_{t+1}) = \\ &= \min_{U_{t+1}} \left\{ X_t' Q X_t + U_{t+1}' R(p_{t+1}) U_{t+1} + 2 X_t' N(p_{t+1}) U_{t+1} + \delta \mathbb{E}_t J(X_{t+1}, p_{t+1}) \right\} \\ &= \min_{U_{t+1}} \Big\{ X_t' Q X_t + U_{t+1}' R(p_{t+1}) U_{t+1} + 2 X_t' N(p_{t+1}) U_{t+1} \\ &+ \delta \sum_{j,k} p_{t+1,j} P_{jk} \left(X_{t+1,k}' V(p_{t+2}) X_{t+1,k} + w(p_{t+2}) \right) \Big\}, \end{aligned}$$

where $X_{t+1,k} = A_k X_t + B_k U_{t+1} + C_k \epsilon_{t+1}$, $p_{t+2} = P' p_{t+1}$, Q and R positive semi-definite matrices, and N is a vector, altogether defining quadratic period loss function.[3] R and N depend on the regime probability distribution p_{t+1}, while Q, as can be shown by straightforward algebraic manipulation, does not. Analogously to anticipated utility solution, probabilistic structure is assumed known and unchanging in all perpetuity. Unlike anticipated utility, the probabilistic structure is that of dynamic coefficients following a regime-switching process while in the anticipated utility case coefficients are statically uncertain.

[3]Because of timing differences with Costa, Fragoso, and Marques (2005) and Svensson and Willams (2007), matrices Q, R and N encapsulate *expected* period $t+1$ loss function.

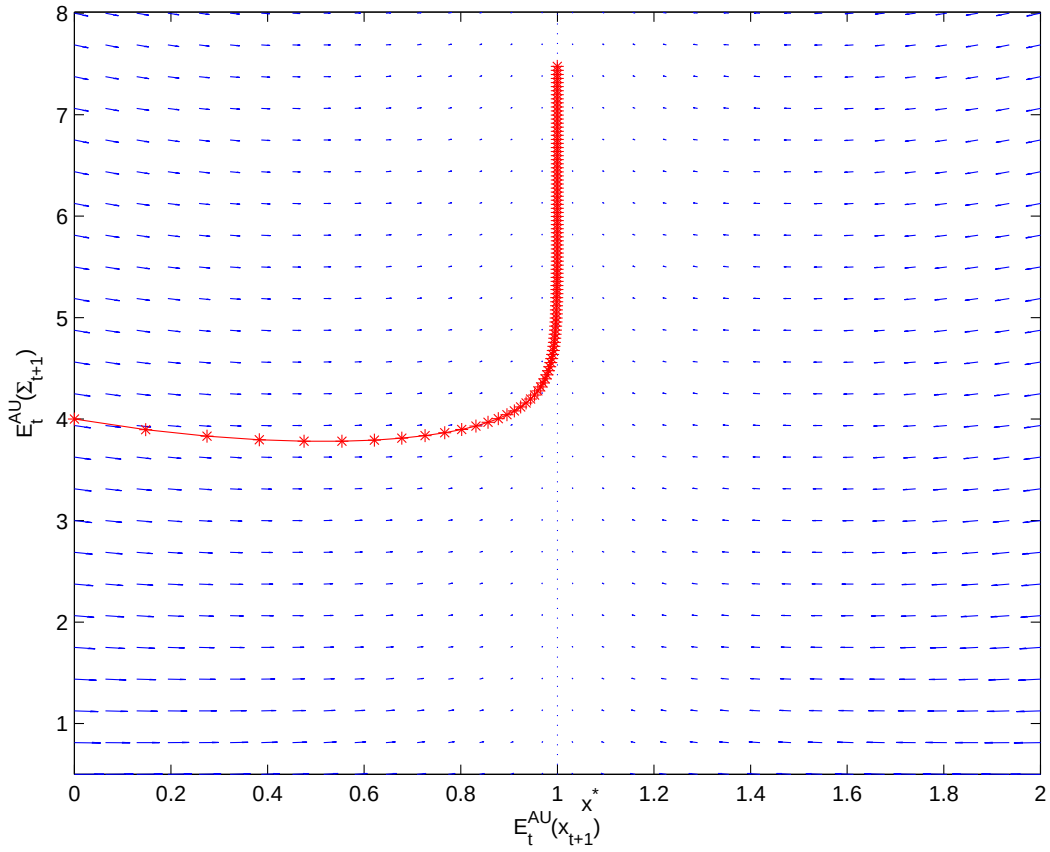


Figure 11: Phase portrait of expected state dynamics under anticipated utility policy. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$.

The first-order condition with respect to U_{t+1} is

$$U'_{t+1}R(p_{t+1}) + \delta \sum_{j,k} p_{t+1,j} P_{jk} \left(X'_t A'_k V(P' p_{t+1}) B_k + U'_{t+1} B'_k \right) = 0 \tag{7.3}$$

and can be written as

$$U_{t+1} = -G(p_{t+1})^{-1} K(p_{t+1}) X_t, \tag{7.4}$$

where

$$G(p_{t+1}) = R(p_{t+1}) + \delta \sum_{j,k} p_{t+1,j} P_{jk} B'_k V(P' p_{t+1}) B_k,$$

$$K(p_{t+1}) = N(p_{t+1})' + \delta \sum_{j,k} p_{t+1,j} P_{jk} B'_k V(P' p_{t+1}) A_k.$$

This leads to the following Riccati equation for the matrix $V(p_{t+1})$:

$$V(p_{t+1}) = Q + \delta \sum_{j,k} p_{t+1,j} P_{jk} A'_k V(P' p_{t+1}) A_k - K(p_{t+1})' G(p_{t+1})^{-1} K(p_{t+1}). \tag{7.5}$$

The scalar $w(p_{t+1})$ is only important for the expected loss function, not for the control. It solves the equation

$$w(p_{t+1}) = \delta \sum_{j,k} p_{t+1,j} P_{jk} \left(\operatorname{tr} \left(V(P' p_{t+1}) C_k C'_k \right) + w(P' p_{t+1}) \right). \tag{7.6}$$

Riccati equation (7.5) can be solved by receding horizon control. Starting with the continuation cost-to-go function at sufficiently distant horizon, the Riccati recursion is rolled

backwards to find the current period expected cost-to-go. The horizon is extended until the difference between the current period value functions is below tolerance threshold. For the continuation cost-to-go function at the receding terminal horizon Svensson and Williams (2005) recommend using the expected value of observed regime control given terminal horizon's regime probabilities. Calculation of the optimal MJLQ control when regimes are observed is similar to the one above, except that instead of one matrix of coefficients, it results in one matrix per regime. Instead of one Riccati equation, we have a system of coupled Riccati equations. The system can be uncoupled by method of do Val, Geromel, and Costa (1998). Once uncoupled, doubling algorithm can be used to solve the resulting optimal linear regulator problem Hansen and Sargent (2004). For additional details and an algorithm, see Svensson and Williams (2005) and Costa, Fragoso, and Marques (2005).

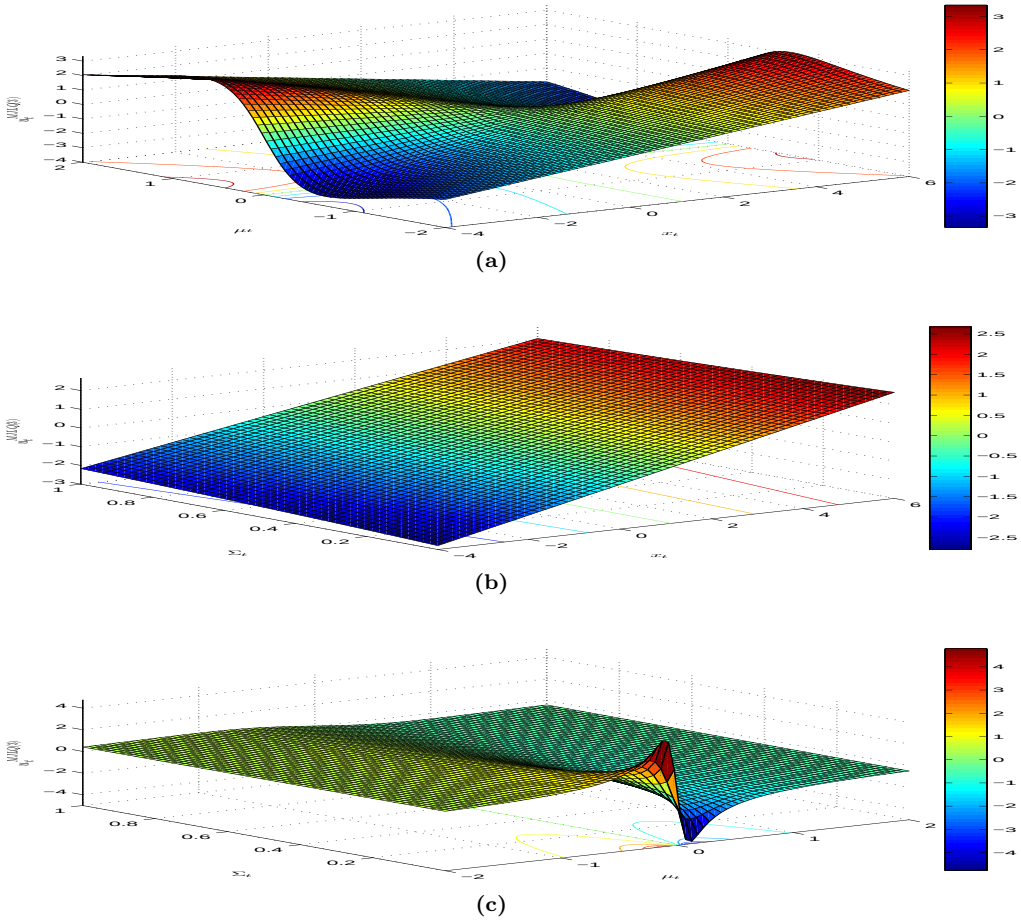
In order to be able to apply MJLQ idea to our setting, we need to map the drifting coefficients specification to the finite-state Markov chain representation. There are many ways to do devise an approximating scheme, none of them perfect because random walk is a non-stationary process whose variance grows over time without bound whereas any finite-state Markov chain is bound to be bounded. As a first rough cut, we envisage the following scheme. Partition the support of $\mathcal{N}(\mu_{t+1|t}, \Sigma_{t+1|t})$ distribution into S segments of equal probability, and define S states (regimes) as the expected values of respective truncated normal distributions over each segment. Thus, $p_{t+1} = (1/S, \dots, 1/S)'$ and $\beta_k = \mathbb{E}(\beta|\beta \in [\Phi^{-1}((k-1)/S, \mu_{t+1|t}, \Sigma_{t+1|t}), \Phi^{-1}(k/S, \mu_{t+1|t}, \Sigma_{t+1|t})], \beta \sim \mathcal{N}(\mu_{t+1|t}, \Sigma_{t+1|t}))$, where Φ^{-1} is the inverse cumulative density function of normal distribution. The transition probability matrix is similarly defined by discretizing the probability distribution of β_{t+2} conditional on $\beta_{t+1} = \beta_k$ for each $k = 1, \dots, S$. Finally, to link state equation (2.2) with regime switching linear system (7.1), use the following definitions: $X_t = \begin{pmatrix} 1 \\ x_t - x^* \end{pmatrix}$, $U_t = u_t$, $A_k = \begin{pmatrix} 1 & 0 \\ \alpha + (\gamma - 1)x^* + \beta_k u^* & \gamma \end{pmatrix}$, $B_k = \begin{pmatrix} 0 \\ \beta_k \end{pmatrix}$, $C_k = \begin{pmatrix} 0 \\ \sigma_\epsilon \end{pmatrix}$, $Q = \begin{pmatrix} \gamma^2(x^*)^2 + (\alpha - x^*)^2 + 2\gamma(\alpha - x^*)x^* + \omega(u^*)^2 + \sigma_\epsilon^2 & \gamma^2 x^* + \gamma(\alpha - x^*) \\ \gamma^2 x^* + \gamma(\alpha - x^*) & \gamma^2 \end{pmatrix}$,
$N = \begin{pmatrix} (\alpha + (\gamma - 1)x^*) - \omega u^*) \left(\sum_{j=1}^S p_{j,t+1}\beta_j\right) \\ \gamma \sum_{j=1}^S p_{j,t+1}\beta_j \end{pmatrix}$, $R = \omega + \sum_{j=1}^S p_{j,t+1}\beta_j^2$.

Having described the solution method, we presents some initial three-regime MJLQ policy calculations in figures 12 and 13. Just like in the case of anticipated utility, acknowledging parameter uncertainty smears the edges of the policy function along $\mu_t = 0$ line in comparison to the corresponding plot for the certainty equivalent policy (figure 7). On the other hand, comparison with the actively adaptive policy function suggests that uncertainty effect results in too much smoothing. In other words, uncertainty causes disquiet but the fact that it could be surmounted by an appropriate policy action is not recognized.

The phase portrait for the dynamical systems of state expectations under MJLQ(3) policy is given in figure 14 together with a representative path. The decrease in the uncertainty about the multiplicative policy parameter is also not very strong and is quickly forsaken.

8. Passively Adaptive Optimal Control

While MJLQ is general and tractable policy that can account for the plausible changes in effectiveness of policy action (time-varying β), tenable modulation of the state transmission channel (time-varying γ), potential regimes of high or low shock variance (time-varying σ_ϵ), regime-switching mean dynamics (time-varying α) or any combination of these features, it does not permit continuous adaptation as in equation (2.3) without proliferating the number of regimes and destroying tractability. It is of interest to solve for a *passively* adaptive policy that is explicit about the random-walk-type coefficient drift. In other words, we are



(a)

(b)

(c)

Figure 12: Passively adaptive MJLQ(3) control. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$. (a) $\Sigma_t = 0.05$ slice; (b) $\mu_t = -1.64$ slice; (c) $x_t = 2.2$ slice.

interested in solving the following Bellman equation

$$
\begin{aligned}
V(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}) = \min_{\{u_{t+1}\}} \Big\{ & L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) \\
& + \delta \int\int\int V\left(\alpha + (\beta_{t+1} + \eta_{t+1})u_{t+1} + \gamma x_t + \epsilon_{t+1}, \beta_{t+1} + \eta_{t+1}, \Sigma_{t+2|t}\right) \\
& \times p(\beta_{t+1}|x_t, \mu_{t+1|t}, \Sigma_{t+1|t}) p(\eta_{t+1}) q(\epsilon_{t+1}) d\mu_{t+1|t} d\eta_{t+1} d\epsilon_{t+1} \Big\} \\
= \min_{\{u_{t+1}\}} \Big\{ & L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) \\
& + \delta \int\int V\left(\alpha + \mu_{t+2|t} u_{t+1} + \gamma x_t + \epsilon_{t+1}, \mu_{t+2|t}, \Sigma_{t+2|t}\right) \\
& \times p(\mu_{t+2|t}) q(\epsilon_{t+1}) d\mu_{t+2|t} d\epsilon_{t+1} \Big\}.
\end{aligned}
\tag{8.1}
$$

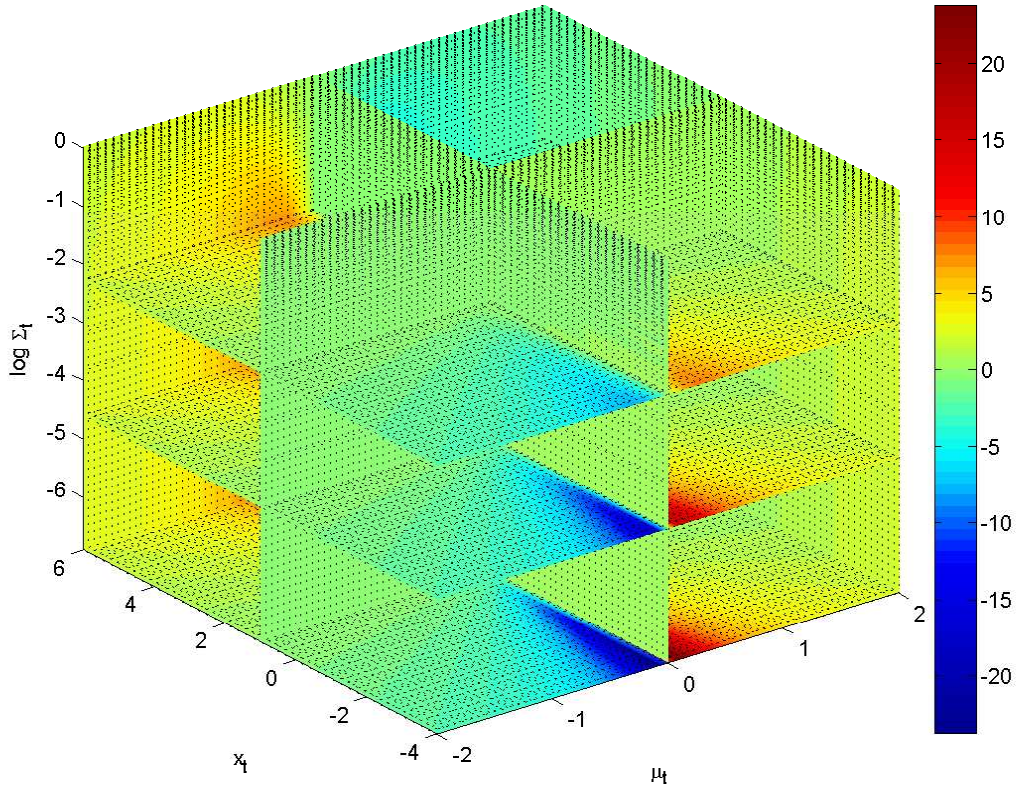


Figure 13: Volumetric plot of passively adaptive MJLQ(3) policy function. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$.

It differs from (3.1) by ignoring the impact of the policy choice, u_{t+1} on subsequent informational state variables $\mu_{t+2|t+1}$ and $\Sigma_{t+2|t+1}$.[4] The passively optimal approach recognizes that the regression coefficient β_{t+1} is subject to shocks η_{t+1} that can stir it away from the current value β_t. The second equality in (8.1) is motivated by the martingale property of conditional expectations by setting $\mu_{t+2|t} = \mu_{t+1|t} + \eta_{t+1}$, with $\eta_{t+1} \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$.[5] Forcing $\sigma_\eta^2 = 0$ should reduce the control to the anticipated utility policy.[6] If, in addition, $\Sigma_{t+1|t} = 0$, we obtain certainty equivalent control. Finally, if $\Sigma_{t+1|t} = 0$, but $\sigma_\eta > 0$, we obtain different generalization of certainty equivalence where the decision maker is certain about the current value of the policy effectiveness but expects that value to drift continuously away in the future periods. Obviously, this kind of policy is only of interest in the drifting coefficient case.

[4]Such ignorance is identical to the assumption of no control in the future periods. Applying Bayes rule with $u_{t+\tau} = 0$ results in the same $\Sigma_{t+\tau|t}$ for $\tau > 0$.

[5]We should note that the latter assumption is not in itself equivalent to the martingale property if $\mu_{t+\tau|t}$ are indeed treated as conditional expectations. Under such interpretation, the decision-maker contemplates the future drift of the current belief consistently with stochastic process for the multiplicative parameter, which is not observed but whose form is known. Meticulous adherence with conditional expectations interpretation of $\mu_{t+\tau|t}$ and the accompanying martingale property is best admitted with $\mu_{t+\tau|t} = \mu_{t+1|t}$ for all $\tau > 0$ assumption. Limited experimentation with this alternative assumption indicates that it results in a policy that is intermediate between the passively optimal policy studied here and MJLQ(S) family of policies.

[6]This implies that anticipated utility function is a good starting point for the value iteration for small σ_η^2.

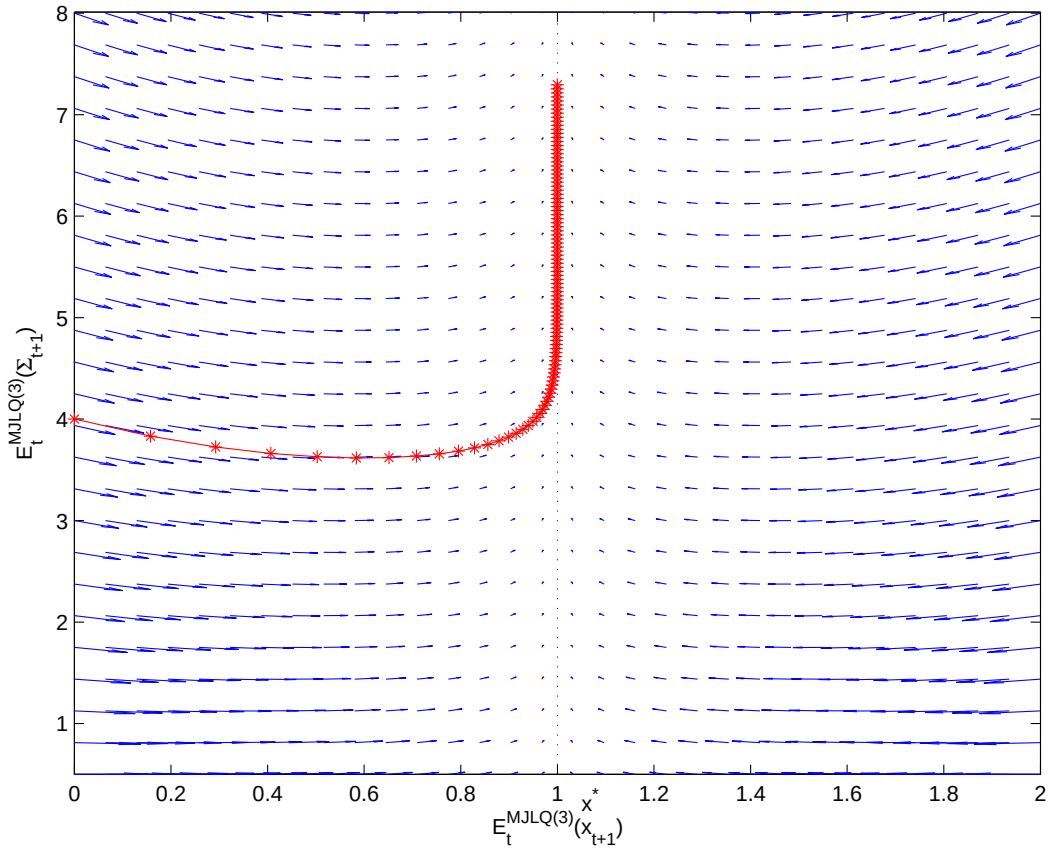


Figure 14: Phase portrait of expected state dynamics under MJLQ(3) policy. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$.

It is equally difficult to compute as it involves the same double integration with respect to slightly different predictive distribution.

Recursion (8.1) can be solved numerically by the same kind of state-space discretization and a combination of value and policy iterations as for the actively optimal control. The presence of the double integration in the equation (8.1) makes doing so computationally more challenging, though still well within reach of modern computers.[7]

Figure 15 plots three slices of the passively adaptive optimal policy function. The top slice is a function of x_t and μ_t when Σ_t is fixed at 0.05. The middle surface represents the policy function as a function of x_t and Σ_t for $\mu_t = -1.64$. The bottom graph plots the policy function against μ_t and Σ_t with $x_t = 2.2$. Volumetric plot in figure 16 summarizes the policy function against all three dimensions by color coding function values. It is harder to read, however.

The phase portrait for the dynamical systems of state expectations under the passively optimal policy is given in figure 17 together with a representative path starting at $\Sigma_t = 4.0$, $x_t = 0$. The decrease in the uncertainty about the multiplicative policy parameter is arrested in about five steps.

[7]Additional complication arises due to the unbounded drift of predictive variance $\Sigma_{t+\tau|t}$ in the absence of new observations which results in the non-stationarity of the problem. Fortunately, the problem can be cured with the technique similar to the receding control by extending the finite absorbing boundary at $\bar{\Sigma}$ until the solution doesn't change.

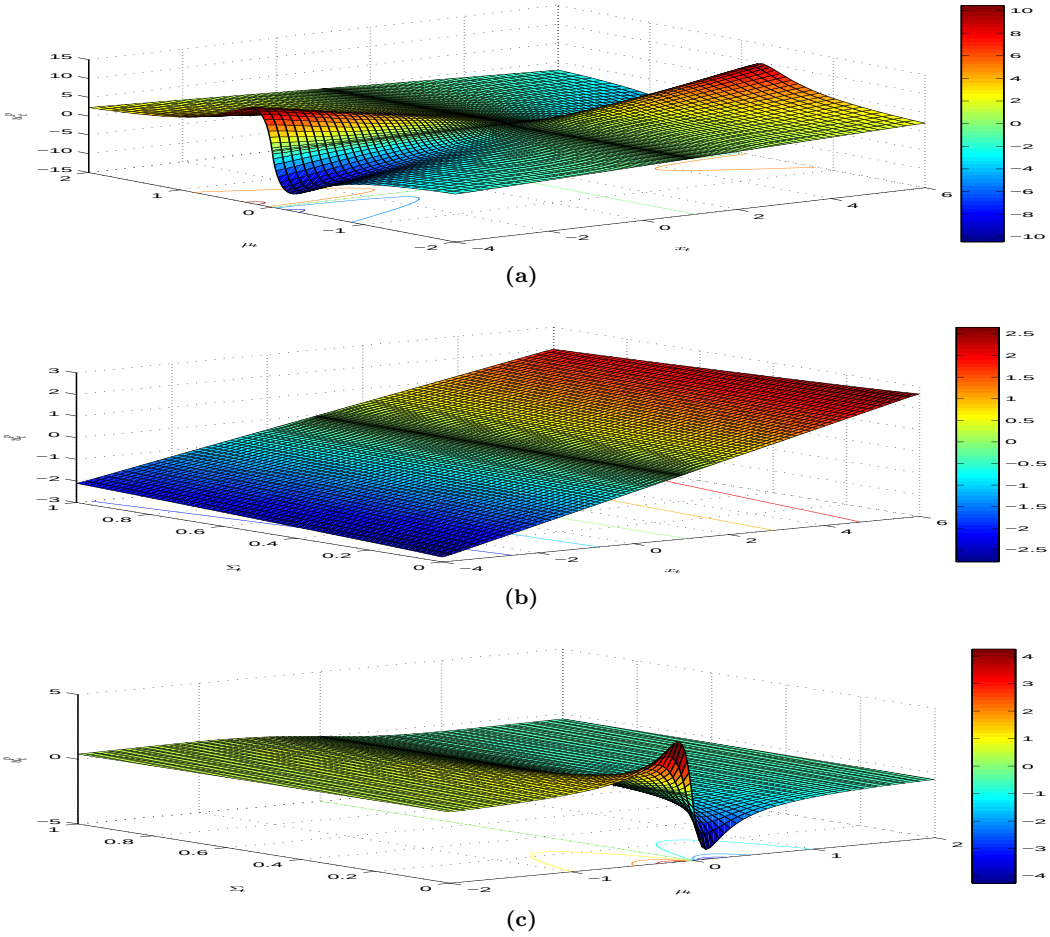


Figure 15: Passively adaptive optimal control. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$. (a) $\Sigma_t = 0.05$ slice; (b) $\mu_t = -1.64$ slice; (c) $x_t = 2.2$ slice.

9. One-Period Limited Lookahead Policy

One-period limited lookahead policy is a solution to one-period-ahead finite horizon version of the original problem. It is *actively* adaptive in the sense that the impact of the policy choice on the next period beliefs is explicitly accounted for. *Adaptive* nomen reflects the fact that even though the solution provides two controls - one for the current period $(t+1)$ and one for the next period $(t+2)$, the decision maker is only committed to implementing the current period control, discarding u_{t+2} at the beginning of the next period and recalculating the solution to the limited lookahead problem anew. At the same time, limited lookahead policy is suboptimal since it disregards any losses that policy maker incurs in periods beyond the lookahead horizon as well as associated future beliefs. In this sense, limited lookahead policy is a generalization of the myopic rule that only minimizes expected one-period loss given current beliefs. This kind of sub-optimality is in stark contrast with all the other policies considered here as they solve respective infinite-horizon problems.

The objective of one-period lookahead policy is to minimize explicit two-period problem:[8]

[8]Alternative formulation could use finite-horizon dynamic programming and would be less explicit.

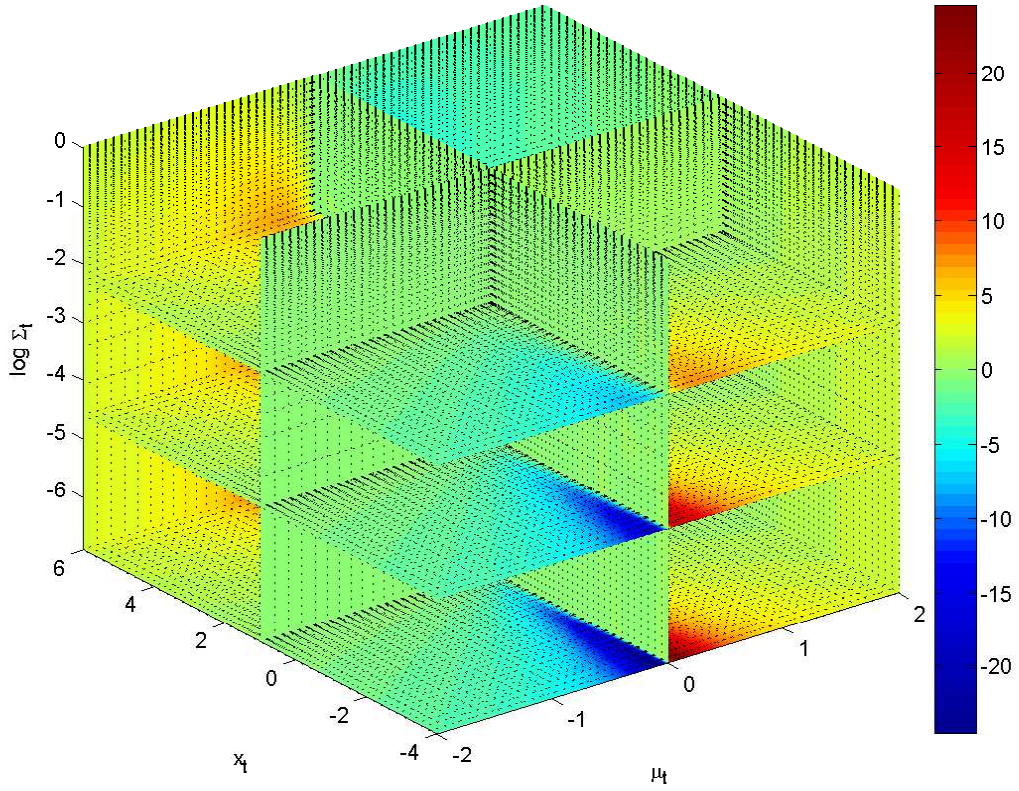


Figure 16: Volumetric plot of passively adaptive optimal policy function. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$.

(9.1)

$$
\begin{aligned}
&\min_{u_{t+1},u_{t+2}} \mathbb{E}_t \Big\{ (\alpha + \beta_{t+1}u_{t+1} + \gamma x_t + \epsilon_{t+1} - x^*)^2 + \omega(u_{t+1} - u^*)^2 \\
&+ \delta \Big[(\alpha + \beta_{t+2}u_{t+2} + \gamma x_{t+1} + \epsilon_{t+2} - x^*)^2 + \omega(u_{t+2} - u^*)^2 \Big] \Big\} \\
= &\min_{u_{t+1},u_{t+2}} \Big\{ (\alpha + \gamma x_t - x^*)^2 + \sigma_\epsilon^2 + \mathbb{E}_t \left(\beta_{t+1}\right)^2 u_{t+1}^2 + 2(\alpha + \gamma x_t - x^*) \left(\mathbb{E}_t \beta_{t+1}\right) u_{t+1} \\
&+ \delta \mathbb{E}_t \Big[(\alpha + \beta_{t+2}u_{t+2} + \gamma(\alpha + \beta_{t+1}u_{t+1} + \gamma x_t + \epsilon_{t+1}) + \epsilon_{t+2} - x^*)^2 \Big] \\
&+ \omega(u_{t+1} - u^*)^2 + \delta\omega(u_{t+2} - u^*)^2 \Big\} \\
= &\min_{u_{t+1},u_{t+2}} \Big\{ (\alpha + \gamma x_t - x^*)^2 + \sigma_\epsilon^2 + \left(\mu_t + \Sigma_t + \sigma_\eta^2\right) u_{t+1}^2 + 2(\alpha + \gamma x_t - x^*)\mu_t u_{t+1} \\
&+ \delta \left(\alpha(1+\gamma) + \gamma^2 x_t - x^*\right)^2 + \delta(1+\gamma^2)\sigma_\epsilon^2 + \delta\mathbb{E}_t \left[\beta_{t+2}^2 \middle| u_{t+1}\right] u_{t+2}^2 \\
&+ \delta\gamma^2 \mathbb{E}_t \left[\beta_{t+1}^2 \middle| u_{t+1}\right] u_{t+1}^2 + 2\delta(\alpha(1+\gamma) + \gamma^2 x_t - x^*)\mathbb{E}_t \left[\beta_{t+2} \middle| u_{t+1}\right] u_{t+2} \\
&+ 2\delta\gamma(\alpha(1+\gamma) + \gamma^2 x_t - x^*)\mathbb{E}_t \left[\beta_{t+1} \middle| u_{t+1}\right] u_{t+1} + 2\delta\gamma\mathbb{E}_t \left[\beta_{t+1}\beta_{t+2} \middle| u_{t+1}\right] u_{t+1}u_{t+2} \\
&+ \omega(u_{t+1} - u^*)^2 + \delta\omega(u_{t+2} - u^*)^2 \Big\}.
\end{aligned}
$$

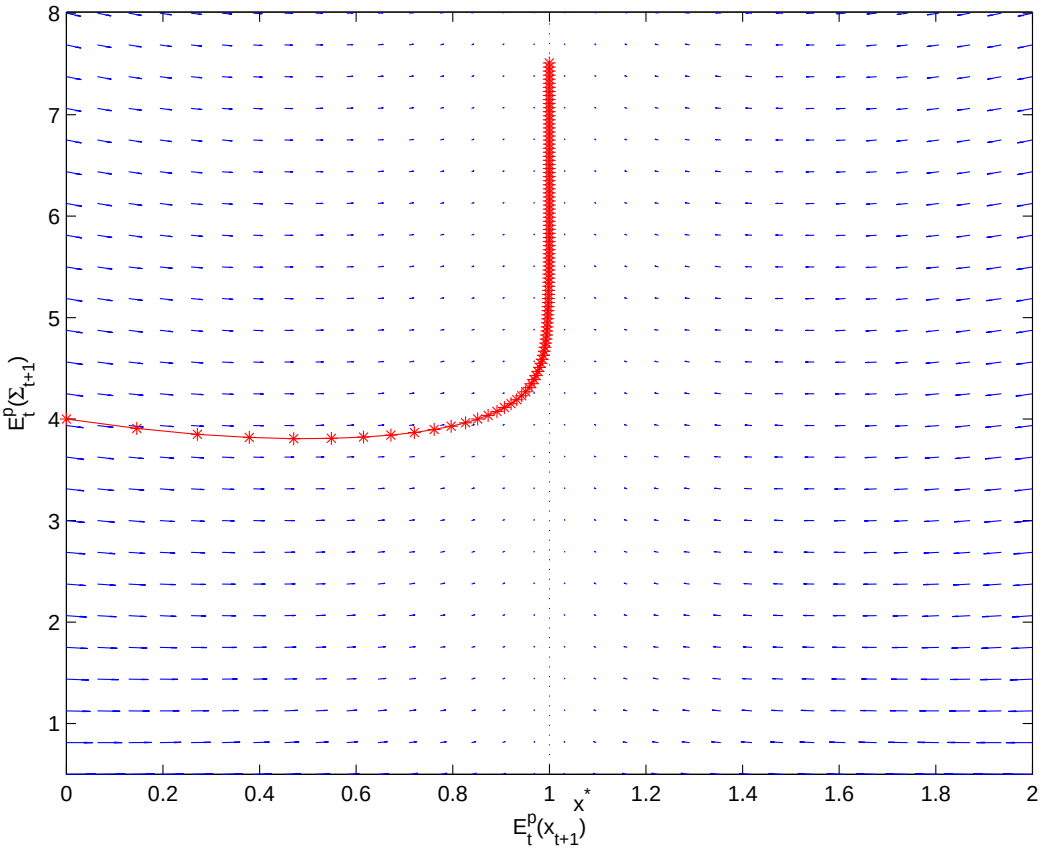


Figure 17: Phase portrait of expected state dynamics under passively optimal policy. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$

Notice that the period $t+2$ part of the objective function involves date t expectations of random variable conditional on future control u_{t+1}. By the updating equation for the mean (or, alternatively, by the law of iterated expectations) $\mathbb{E}_t\left[\beta_{t+2}\middle|u_{t+1}\right] = \mathbb{E}_t\left[\beta_{t+1}\middle|u_{t+1}\right] = \mathbb{E}_t\left[\beta_{t+2}\right] = \mu_t$. Future variances, on the other hand, follow nontrivial dynamics:

$$\mathbb{E}_t\beta_{t+1}^2 = \mu_t^2 + \Sigma_t + \sigma_\eta^2, \tag{9.2}$$

$$\mathbb{E}_t\left[\beta_{t+1}^2\middle|u_{t+1}\right] = \mu_t^2 + \Sigma_{t+1}(u_{t+1}), \tag{9.3}$$

$$\mathbb{E}_t\left[\beta_{t+1}\beta_{t+2}\middle|u_{t+1}\right] = \mu_t^2 + \Sigma_{t+1}(u_{t+1}), \tag{9.4}$$

$$\mathbb{E}_t\left[\beta_{t+2}^2\middle|u_{t+1}\right] = \mu_t^2 + \Sigma_{t+1}(u_{t+1}) + \sigma_\eta^2, \tag{9.5}$$

where

$$\Sigma_{t+1}(u_{t+1}) = \Sigma_t + \sigma_\eta^2 - \frac{\left(\Sigma_t + \sigma_\eta^2\right)^2 u_{t+1}^2}{\left(\Sigma_t + \sigma_\eta^2\right) u_{t+1} + \sigma_\epsilon^2} \tag{9.6}$$

is belief variance that would obtain at the end of period $t+1$ if control u_{t+1} were chosen at its beginning. Upon substituting the above relationships into (9.1) the resulting objective function is no longer quadratic. It is not even globally convex. Figure 18 shows typical behavior with a kink developing away from the minimum, which appears unique.

Figure 19 displays the one-period lookahead policy function along the three orthogonal subspaces in the state space, while figure 20 renders the policy function via a volumetric plot. The shape is by now habitual.

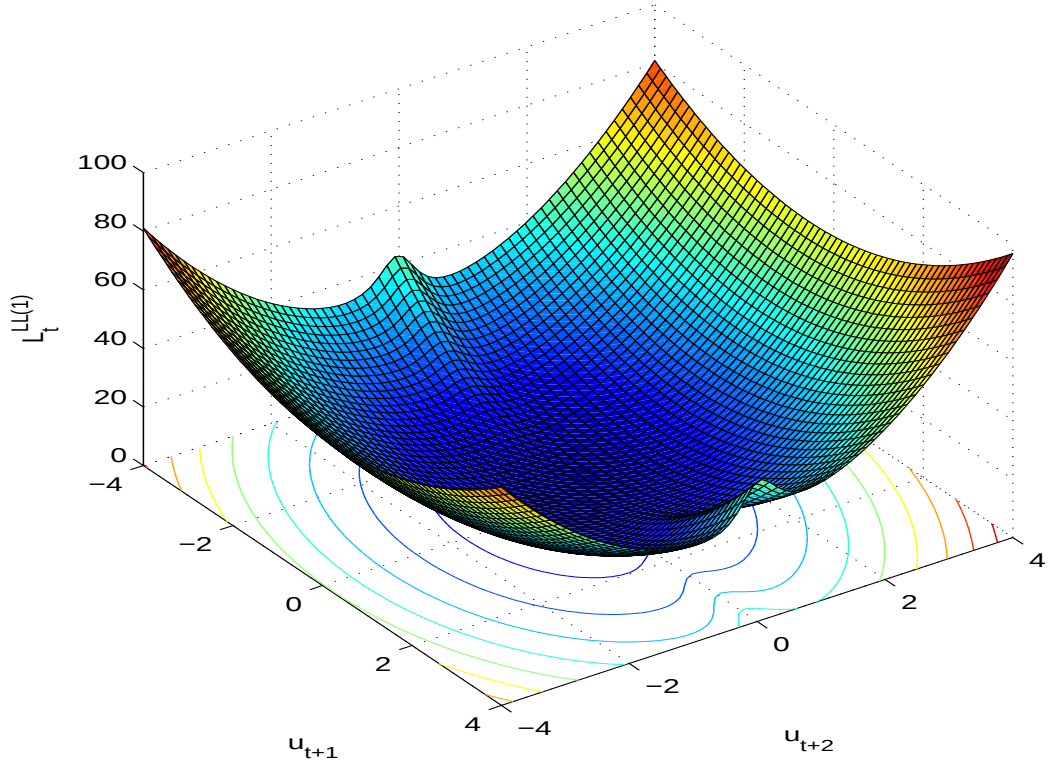


Figure 18: One period limited lookahead loss function. State coordinates: $x_t = 0$, $\mu_t = -0.5$, $\Sigma_t = 1.0$. Parameter values: $\alpha = 0.01$,$\gamma = 0.9$, $\delta = 0.75$, $\omega = 1.6$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = 0.2$, $\sigma_\eta^2 = 0.4$.

The phase portrait for the dynamical systems of state expectations under one-period lookahead policy is given in figure 21 together with a representative path emanating from $\Sigma_t = 4.0$, $x_t = 0$. The shape of the path is the same as for other policies.

10. Comparison

10.1. **Controls.** Figure 22 provides comparison of various alternative policies as functions of the physical state x_t. Certainty equivalent equivalent policy function certainly stands out, displaying much more aggressive reaction to the deviation of the physical state from its target x^*. In contrast, the remaining five policy functions take uncertainty into account by responding in a more gradual manner. Once parameter uncertainty is acknowledged, however, the contributions of other solution elements, such as active experimentation, coefficient drift, or infinite horizon, are of the second order of importance. Accordingly, the policy functions for actively optimal, passively optimal, MJLQ(3)-approximated passively optimal, one-period limited lookahead, and anticipated utility solutions are all very close to each other and hard to distinguish visually. The two panels are helpful in ascertaining the generality of this finding with respect to the weight on control in the period loss function, ω. As ω increases, all policy functions are rotated clockwise resulting in more cautious policy. The drive towards caution is strongest for the certainty equivalent solution which nonetheless remains the most vigorous of the group. In terms of policy's ranking with respect to the gradualism, increase in control cost introduces slight alterations. MJLQ policy stays the second least aggressive but the most hesitant policy award is transferred from the anticipated utility policy to the passively optimal policy. Last thing worth noticing in both

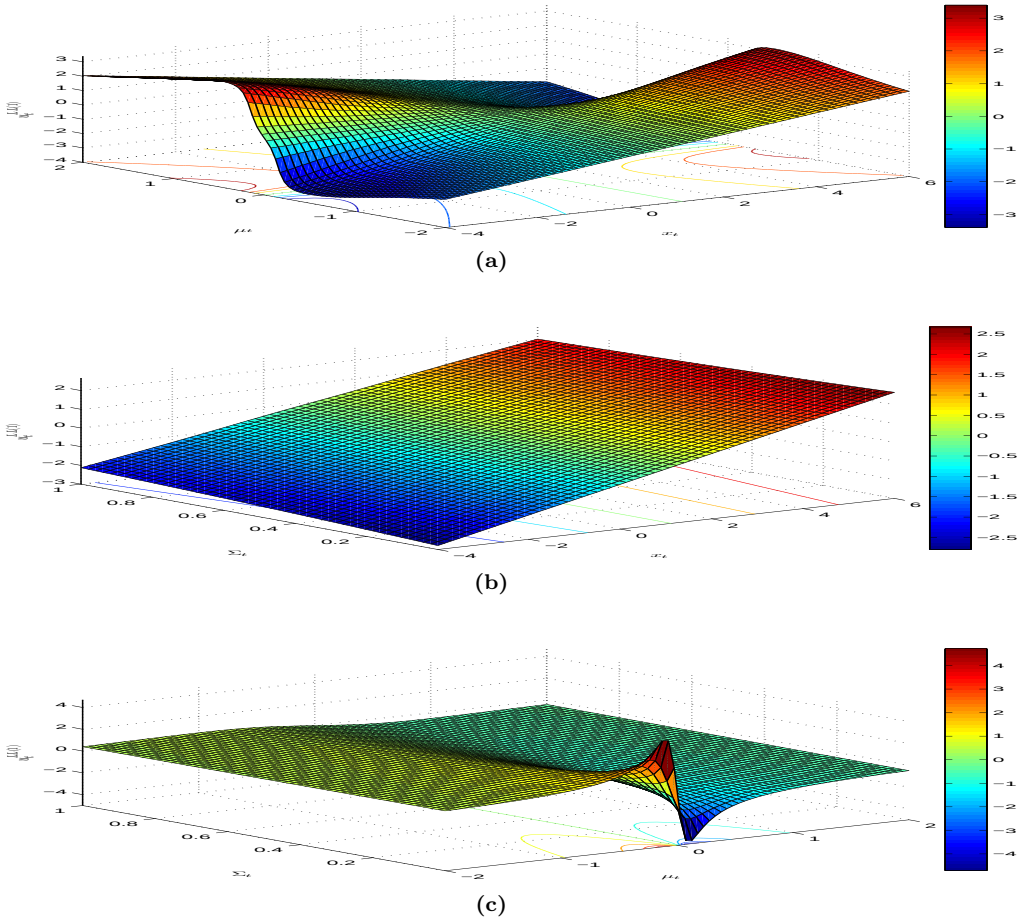


Figure 19: One-period lookahead control. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$. (a) $\Sigma_t = 0.05$ slice; (b) $\mu_t = -1.64$ slice; (c) $x_t = 2.2$ slice.

panels is that the two active policies are only very mildly nonlinear, pointing to the relative unimportance of active experimentation.

However, figure 23 suggests that the difference among policies is deepened as parameter uncertainty is heightened. In both panels, the certainty equivalent policy is the most aggressive, differing significantly from the group of policies that recognize uncertainty. Since the certainty equivalent policy does not depend on Σ_t, the gulf between it and the group of other policies widens as uncertainty mounts. Of the remaining five policy rules, the active optimal policy consistently displays the largest amount of exploration in the outlying regions of belief space. We shall investigate how much this difference matters in the cost-to-go space in section 10.2. The relative rankings of policies other than certainty equivalent one can vary over the belief space and depend on parameter values.

10.2. **Cost-to-go Functions.** We evaluate different expected cost-to-go function from the perspective of fully optimizing decision maker. In other words, we compare not the the

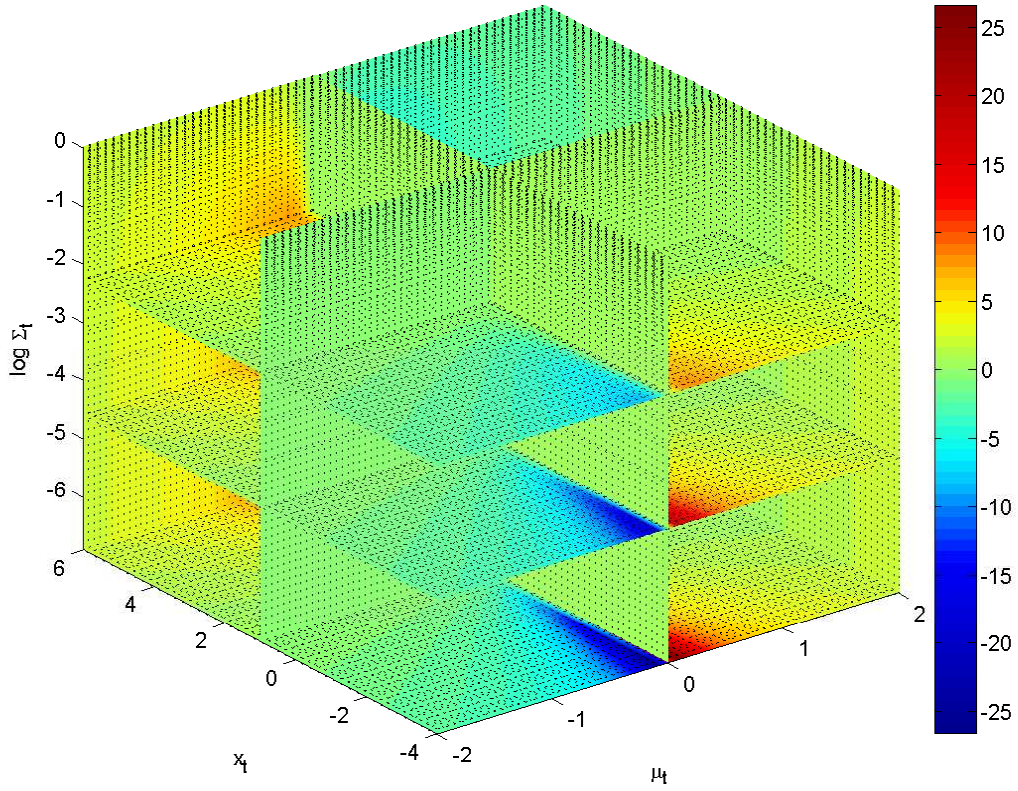


Figure 20: Volumetric plot of one-period limited lookahead policy function. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.01$.

different cost-to-go functions that are minimized by their respective policies, but the Q-factor (Bertsekas, 2005) of the active learning Bellman equation under various policies:

(10.1)

$$
\begin{aligned}
V^*(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) &= \mathbb{E}^* V(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) \\
&= L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) \\
&+ \delta \int V\left(x_{t+1}(\beta_{t+1}, u_{t+1}, x_t, \epsilon_{t+1}), u_{t+1}, \mu_{t+2}(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}), \Sigma_{t+2|t+1}(\Sigma_{t+1|t}, u_{t+1})\right) \\
&\times p(\beta_{t+1}|x_t, \mu_{t+1|t}, \Sigma_{t+1|t}) q(\epsilon_{t+1}) d\beta_{t+1} d\epsilon_{t+1},
\end{aligned}
$$

where $x_{t+1}(\beta_{t+1}, u_{t+1}, x_t, \epsilon_{t+1})$ is a short-hand for the right hand side of (2.2), u_{t+1} is one of the policies under consideration: $u_{t+1} \in \left\{u_{t+1}^*, u_{t+1}^p, u_{t+1}^{LL(1)}, u_{t+1}^{MJLQ(3)}, u_{t+1}^{AU}, u_{t+1}^{CE}\right\}$. Of course, when $u_{t+1} = u_{t+1}^*$, the actively adaptive optimal policy is recovered. To evaluate these value functions, we employ the policy iteration algorithm, now that all six policies are already available on the grid.

The results are shown in figure 24 against the current physical state variable for the two alternative values of parameter ω that controls the balance between intentional and accidental experimentation. The evidence of the figure conforms with the earlier findings. Since it is only the certainty equivalent policy that stands out from the crowd, its value is the only one that differs notably. The benefit to experimentation is virtually negligible, completely overwhelmed by the benefit of simply recognizing parameter uncertainty. Most bang for the buck comes from simply recognizing parameter uncertainty as in the anticipated utility case. This is the same conclusion as the one reached in Cogley, Colacito, and Sargent

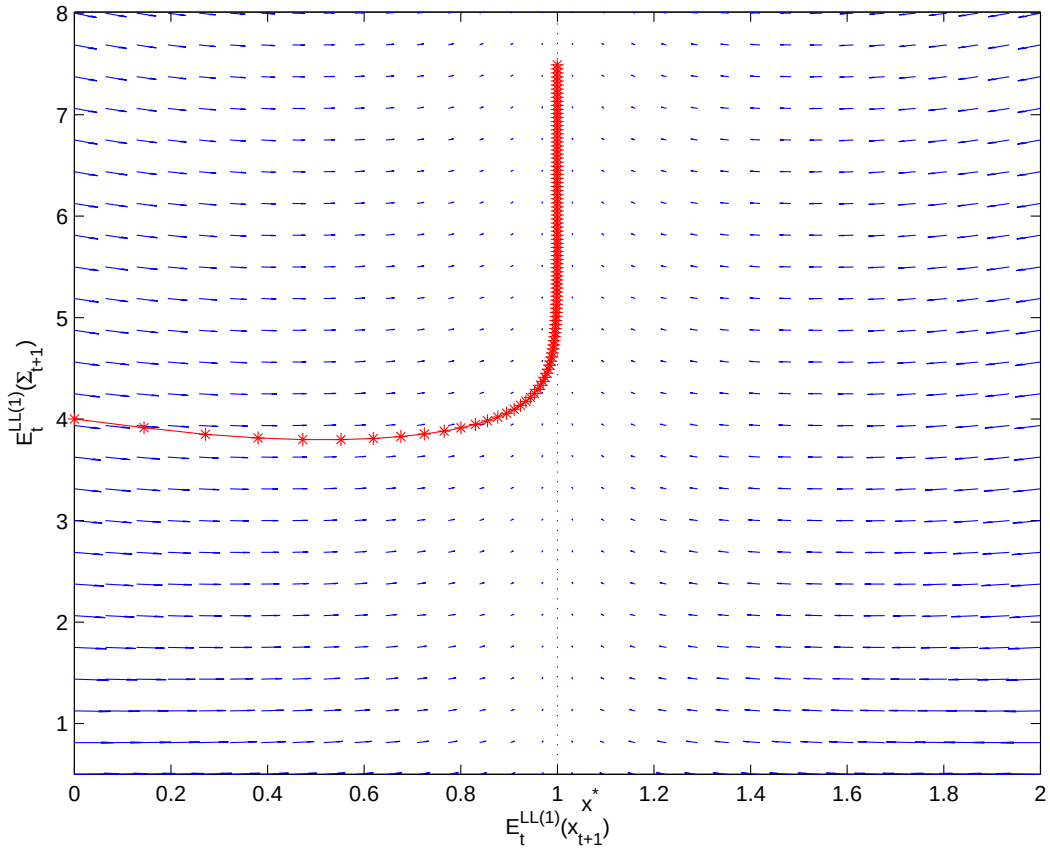
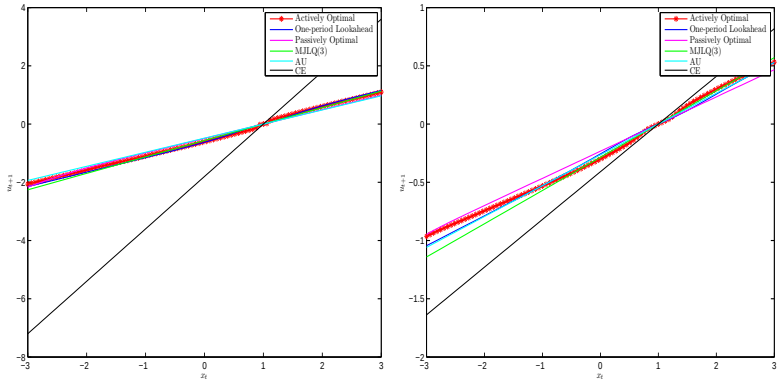
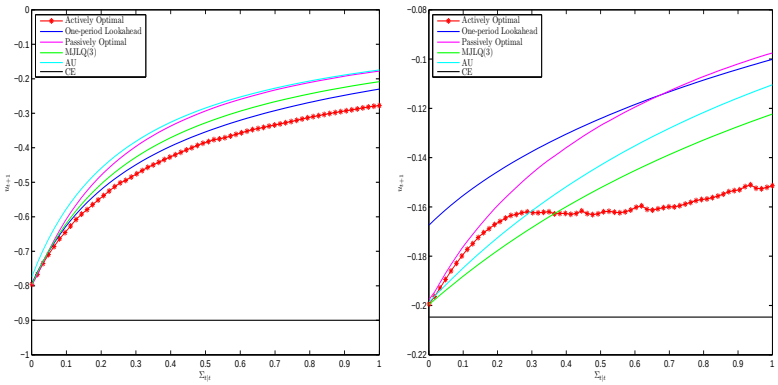


Figure 21: Phase portrait of expected state dynamics under one-period lookahead policy. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$.



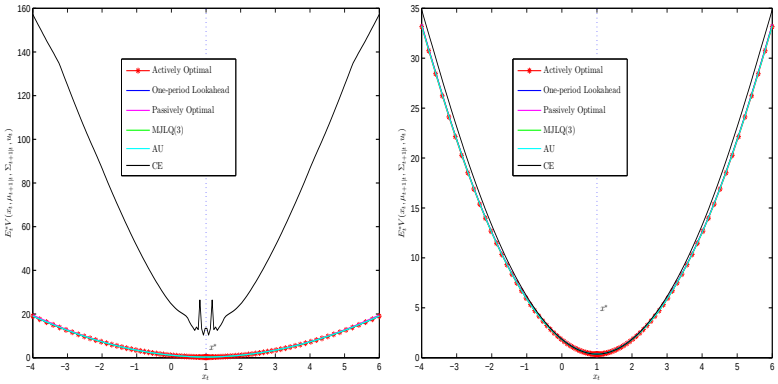
(a) **(b)**

Figure 22: Policy functions under alternative policies. State coordinates: $\mu_t = -0.5$, $\Sigma_t = 0.64$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;



(a) **(b)**

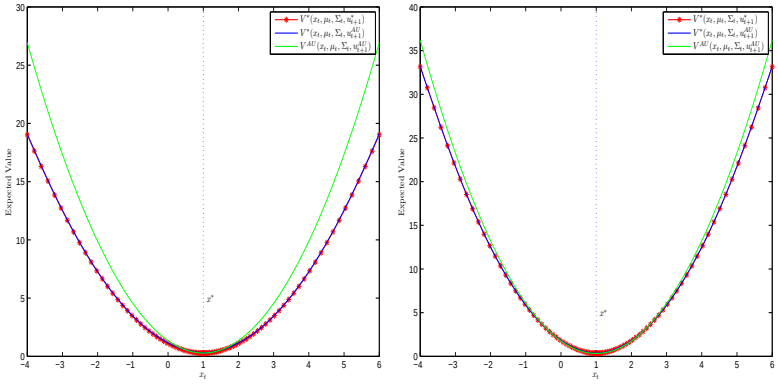
Figure 23: Policy functions under alternative policies. State coordinates: $\mu_t = -0.5$, $x_t = 0.5$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;



(a) **(b)**

Figure 24: Actively adaptive value function under alternative policies. State coordinates: $\mu_t = -0.5$, $\Sigma_t = 0.64$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;

(2007). Following these authors, we caution that the value function that defines anticipated utility control $V^{AU}(S, u^{AU})$ is not the same as $V^*(S, u^{AU})$. The former doesn't allow the actively experimenting decision maker to assess a "what-if" scenario of using the anticipated utility alternative, but instead blinds him to the future dynamics of hidden state and future learning when evaluating a said alternative. The distinction is of minor importance except for the case of extreme uncertainty. For moderate values of Σ_t as in figure 25 we confirm by plotting the three value functions against x_t for two competing values of ω. In both cases using $V^{AU}(S, u^{AU})$ exaggerates the loss under the anticipated utility, while $V^*(S, u^{AU})$ and $V^*(S, u^*)$ are virtually indistinguishable. It should be noted however that, in general, it is not even true that $L(S, u^*) \leq V^{AU}(S, u^{AU})$, where $L(S, u^*)$ is expected one-period loss under the actively optimal policy. In some outlying regions in the state space (especially in the direction of increasing uncertainty) the reverse could be true. Figure 26 gives three slices of Q-factor functions that demonstrate the issue and its potential magnitude. Figure 27 displays the difference $V^*(S, u^{AU}) - V^{AU}(S, u^{AU})$ as a volumetric plot. The significant



(a) **(b)**

Figure 25: Actively adaptive and anticipated utility value functions under actively optimal and anticipated utility policies. State coordinates: $\mu_t = -0.5$, $\Sigma_t = 0.64$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;

regions of the state space are colored in various shades of blue indicating $V^*(S, u^{AU}) \geq V^{AU}(S, u^{AU})$, i.e. where using V^{AU} function would have predicted smaller loss to the anticipated utility policy than if the future learning were taken into account in computing the expectations.

Continuing with optimal Q-factor discussion, we further remark that the differences among the three policies are further tightened with ω. Larger differences arise along Σ_t dimension as shown in figure 28. The first panel is for the case $\omega = 0$ where the optimal Q-factor of the certainty equivalent policy is completely off the scale and is omitted. In both panels the gap grows with Σ_t as intentional experimentation becomes more beneficial. This is despite the fact that extreme uncertainty about the multiplicative policy parameter is not going to last long for the relatively small shock variances used in calculation.

10.3. **Expected Evolution of Observed State.** Figure 29 translates the differences amongst various policies into the differences in the expected transition of the physical state. While it is no more than simply an affine transformation of the policy rule, it lends interpretation to the features of the policy rule and could be used to trace out long term dynamics of x_t in expectation with respect to the current information set. All six policies result in very similar state transition dynamics, except, again, under the certainty equivalent rule. If $\omega = 0$, the certainty equivalent policy leads to complete adjustment to target in one-step, $\mathbb{E}_t^{CE} x_{t+1} = x^*$. If $\omega > 0$, the state evolution resembles the other five much closer. As ω increases, the speed of adjustment wanes.[9] The actively optimal policy is most inertial when the physical state is far away from target, more so than any other policy, albeit not by a large margin.

10.4. **Expected Beliefs.** Figures 30, 31 and 32 elucidate the evolution of beliefs under alternative policies. Because beliefs are completely characterized by the mean and variance and because the law of iterated expectations holds, we only need to predict the evolution of the variance of the belief about slope of the state equation.

Figure 30 argues several points. One is that the quality of all approximations deteriorates with larger uncertainty as the learning outcomes of several policies drift further apart with the variance of the policy parameter. These large variance regions of the state space are

[9]It could also be shown that the slope of the perceived state transition function under the certainty equivalent control is of the same sign as the persistence parameter γ, and it provides the bound on all the other policies. Therefore, in the more natural case of $\gamma > 0$, the convergence is monotone, and most policies display inertia.

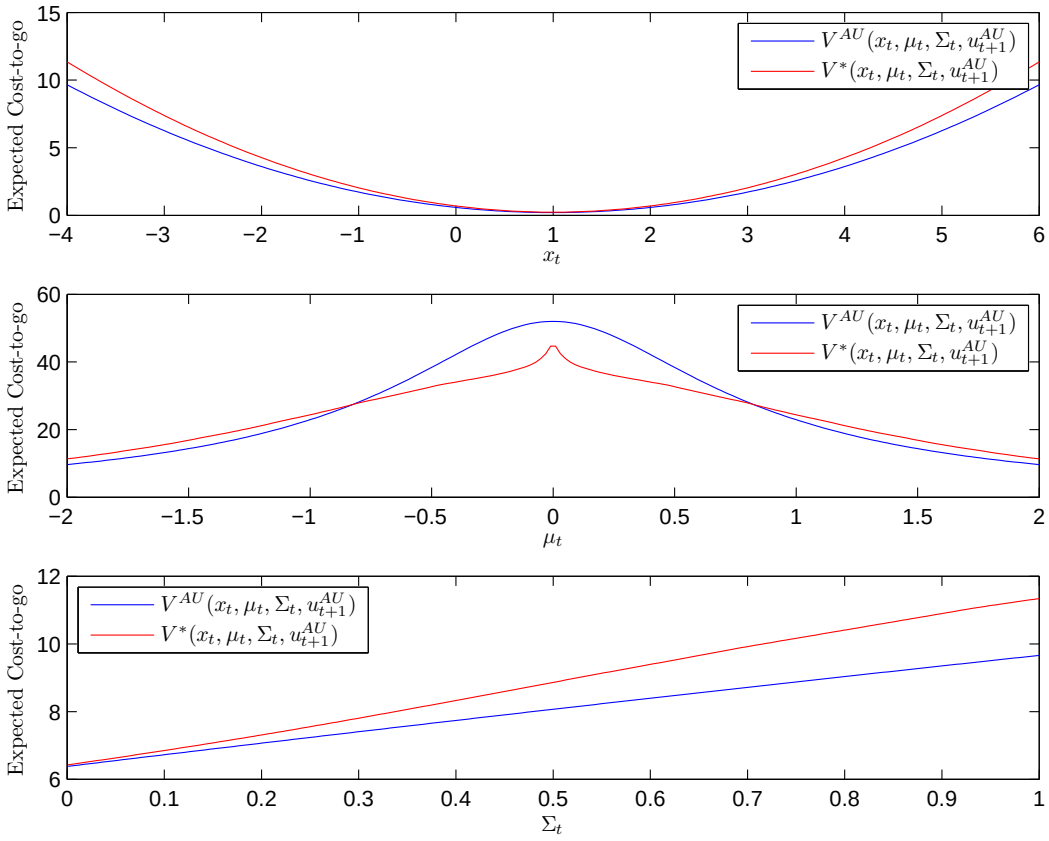


Figure 26: Example where $V^*(S, u^{AU}) \leq V^{AU}(S, u^{AU})$ is violated. Parameters: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\omega = 1.6$, $x^* = 1$, $u^* = 0$,$\sigma^2_\epsilon = \sigma^2_\eta = 0.04$. State coordinates in the top panel: $\mu_t = -2$, $\Sigma_t = 1$. State coordinates in the middle panel: $x_t = -4$, $\Sigma_t = 1$. State coordinates in the bottom panel: $x_t = -4$, $\mu_t = -2$.

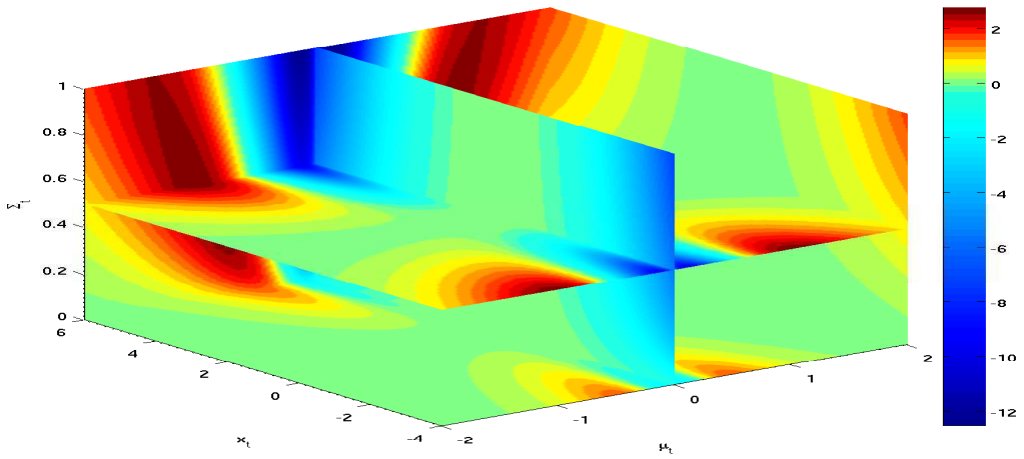
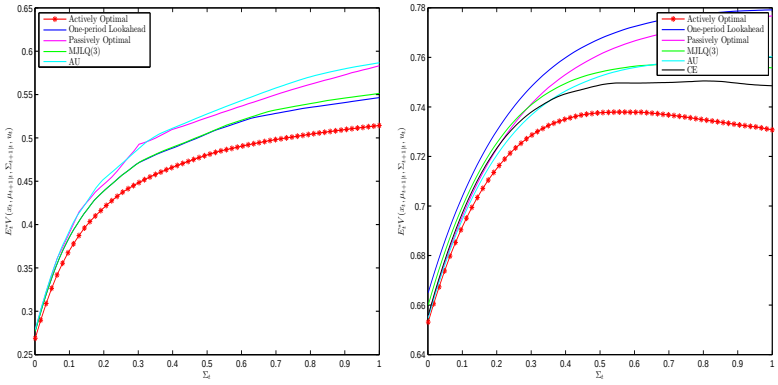
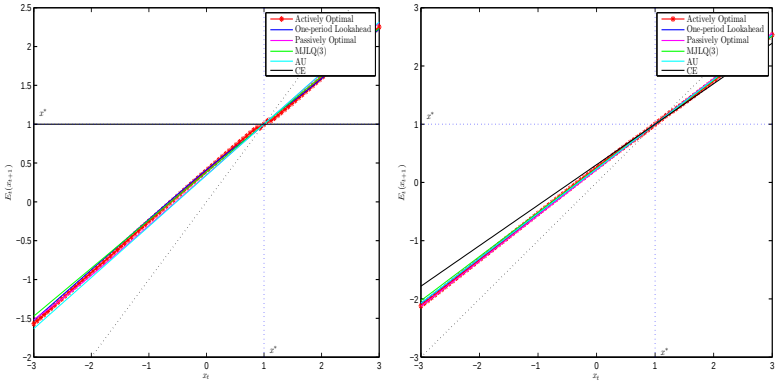


Figure 27: Volumetric plot of $V^*(S, u^{AU}) - V^{AU}(S, u^{AU})$. Parameters: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\omega = 1.6$, $x^* = 1$, $u^* = 0$,$\sigma^2_\epsilon = \sigma^2_\eta = 0.04$.



(a) **(b)**

Figure 28: Actively adaptive value function under alternative policies. State coordinates: $\mu_t = -0.5$, $x_t = 0.5$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;

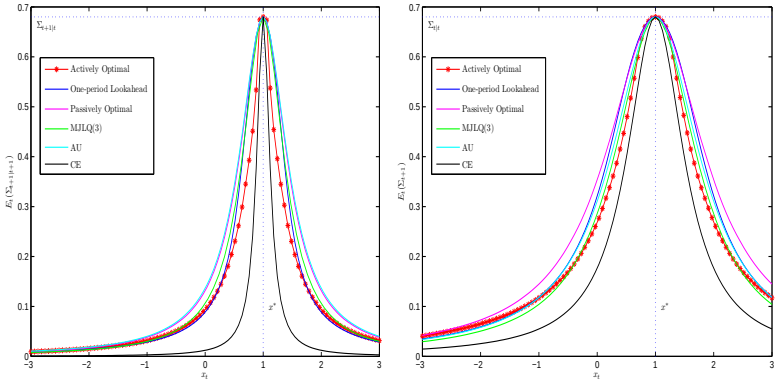


(a) **(b)**

Figure 29: Expected target state under alternative policies. State coordinates: $\mu_t = -0.5$, $\Sigma_t = 0.64$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;

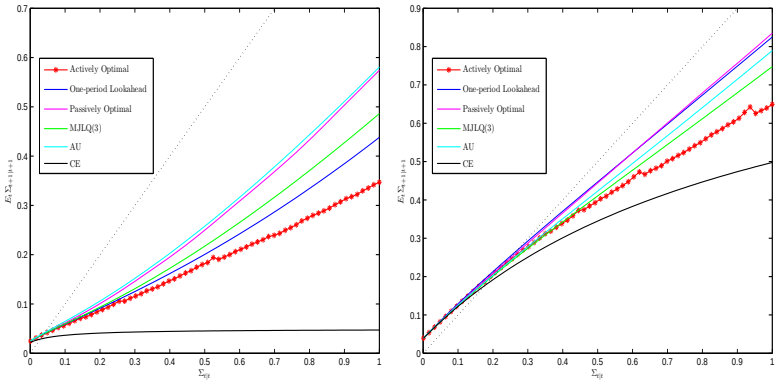
transitory and do not contribute significantly to the overall cost so that in the cost space the approximation remains good. Second observation confirms intuition of Kendrick (2005) that suggests larger role for policy activism with heightened model transmission channel uncertainty. Indeed, apart from the very aggressive certainty equivalent policy, the actively adaptive policy induces the fastest rate of learning by way of the biggest reduction in the variance of the policy parameter (in expectation). The passively optimal policy, in contrast, tends to discourage learning, being consistently one of the least aggressive policies. This is because the passively optimal policy acknowledges the most amount of uncertainty, not only contemporaneously but also its future dynamic evolution, and yet it ignores completely any uncertainty reductions that may stem from future learning.[10] Anticipated utility policy is

[10]This distinction becomes a little more blurred with larger discount factor δ as the passively optimal policy then exhibits larger amount of accidental experimentation. For reasons of space, we do not elaborate on sensitivities of our conclusions to changes in model parameters.



(a) **(b)**

Figure 30: Expected belief variance under alternative policies. State coordinates: $\mu_t = -0.5$, $\Sigma_t = 0.64$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;



(a) **(b)**

Figure 31: Expected belief variance under alternative policies. State coordinates: $\mu_t = -0.5$, $x_t = 0.5$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;

typically adjacent to it, indicating that most of unwillingness to learn originates in the uncertainty about the contemporaneous value of the policy parameter. Passive policy derived from MJLQ(3) tends to inhabit the area midway between actively and passively optimal learning curves. This connotes that the approximation error is inadvertently conducive to quicker learning. Third informal finding concerns convergence of beliefs in expectation. Holding the physical state constant, the variance of the policy parameter will converge to the intersection of the learning curve with the 45-degree line. Allowing the drift in the physical state will complicate the convergence somewhat but since the expectation of the physical state itself converges, the system dynamics will take approximately the same shape as in figure 30. Here, the dynamics of beliefs is stable under all alternative policies. The limit point of the dynamics is at some non-zero value Σ_∞^u that depends on the type of policy under consideration. While more aggressive policies such as the certainty equivalent rule lead to somewhat lower limit variance of belief about the policy parameter and faster

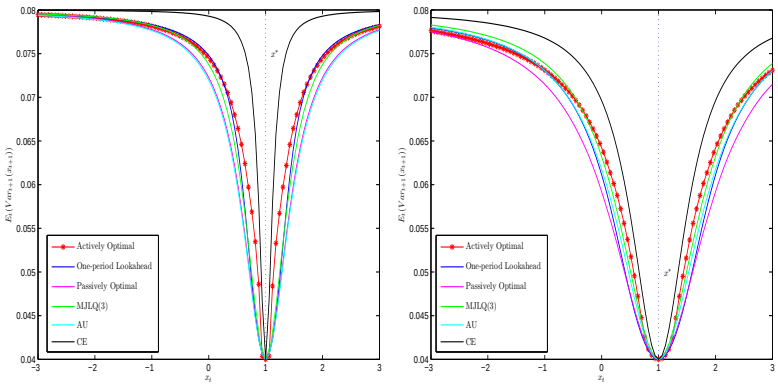


Figure 32: Expected state variance under alternative policies. State coordinates: $\mu_t = -0.5$, $\Sigma_t = 0.64$. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $u^* = 0$, $\sigma_\epsilon^2 = \sigma_\eta^2 = 0.04$. (a) $\omega = 0$; (b) $\omega = 1.6$;

convergence, complete elimination of uncertainty is impossible due to the dynamic coefficient drift. Fourth element of note is that the learning outcomes become nearly identical when $\Sigma_t = 0$. Indeed, anticipated utility policy coincides with certainty equivalent rule exactly, and the other policies only differ through the differences in future states (physical state only for passive policies and both physical and information states for active policies). From the belief updating equation (2.5) it becomes clear that the height of the intercept with $\mathbb{E}(\Sigma)$ axis is roughly proportionate to σ_η^2. Accordingly, we make the fifth observation that the learning curves for all policies are adjusted upward in response to larger dynamic uncertainty. Last ramification of studying figure 30 is dependence of learning curves on the relative cost of control ω. As ω increases learning gets slower,especially for the two active policies and the certainty equivalent rule. The region of convexity for the learning curves that is apparent for $\omega = 0$, disappears when $\omega = 1.6$ Interestingly as well, the one-period lookahead's quality of learning suffers more than for the other policies, making the agent slowest to learn in the region of moderate uncertainty under that policy.

Figure 32 interprets the inferences about $\mathbb{E}_t\Sigma_{t+1|t+1}$ in the space of variance of the future target state by using the relationship

$$\mathbb{E}_t(Var_{t+1}(x_{t+1})|u_{t+1}) = u_{t+1}^2\mathbb{E}_t\left(\Sigma_{t+1|t+1}|u_{t+1}\right) + \sigma_\epsilon^2.$$

This predictive variance of the physical state is shaped by the two offsetting influences, in line with the general philosophy of the dual control. On one hand, it grows with the size of control impulse. On the other, it is tempered by a reduction in variance of the policy parameter due to learning. In consequence, the predictive variance plotted against the physical state x_t take characteristic shape of a potential well, with a sharp minimum at x^*. As figure 32 demonstrates, the passively optimal policy resolves stabilization-learning dilemma entirely in favor of stabilization by yielding consistently lower predictive variance, especially as control becomes relatively more costly. At the other extreme lies the outcome of the certainty equivalent policy rule which manipulates the control without regard to the parameter uncertainty and could lead the target state potentially astray. The actively optimal policy tends to generate more volatile x_{t+1} in the vicinity of the target than other non-certainty-equivalent policies. One-period lookahead and MJLQ-based policies appear to be the closest approximations to the actively optimal policy by balancing the two forces impinging on the predictive state variance in the similar proportions.

10.5. **Expected Dynamics of Extended State.** Here we put together the phase diagrams displayed for each policy individually and the discussion of the preceding two sections.

Figure 33 contrasts the paths of the expected extended state (apart from mean which is constant) under alternative policies for the same parameter configuration and originating from the same starting point in the state space. It is clear that under all policies one would expect eventual convergence of the physical state to target. It is also evident that the learning process will not converge in the sense that limiting value of the belief variance is non-zero unless $\sigma_\eta^2 = 0$.[11] Of the six policies, the certainty equivalent policy exhibits excess accidental experimentation that makes it the only policy with $\mathbb{E}_t\Sigma_{t+\tau|t}$ falling below 2. The first step it makes is also the largest. The actively optimal policy ranks second in the amount of experimentation. It also display intriguing behavior in the vicinity of x^* where for moderate uncertainty levels the actively optimal policy would cause x^* to repel the expected target state. We can speculate that the anomaly is driven by the incentive to explore the state-space in order to slow down the creeping uncertainty which can in turn lead to disastrous future decisions. At the other extreme, passively optimal policy shows almost no learning along the expected state, the least amount of progress towards the target at every step and the highest uncertainty about the multiplicative policy parameter at the and of the path. Anticipated utility policy is very close and so is one-period lookahead policy. MJLQ-type policy displays intermediate degree of gradualism, most likely because treating drifting coefficient as if generated by *stationary* process would tend to understate the true uncertainty in the mind of a decision-maker. One-period lookahead, on the other hand, trades off a different kind of an approximation. By ignoring learning beyond that in the immediate future it understates the benefit to the active probing. The result is that the path nearly coincides with passively optimal one.

10.6. **Simulated Outcomes.** In this section we inquire into the shapes of outcomes that can arise in Monte Carlo simulations under alternative policies. This is useful for a number of reasons. Foremost of these is that it can help uncover certain peculiarities of outcomes given policy rule and sensitivity of these feature to the details of the economic environment in the background. In context of the drifting coefficient regression, and if the drifting coefficients are thought to represent the model uncertainty, study of diversity of simulated outcomes goes beyond the simple duty to report expected losses under different policy as a guide to policy evaluation under model uncertainty.[12] It serves as an additional quantitative and visual aid to communicate how model uncertainty enters the policy evaluation. In doing so it accords with the spirit of Brock, Durlauf, and West's 2007 encouragement to explore the degree of *outcome dispersion* and *action dispersion*.

10.6.1. *Control Sequences.* First, we investigate the shape of the dynamics of the policy choice. Figure 34 shows 400 simulated time-series of length $T = 100$ under different policies, all emerging from identical initial condition $x_0 = 0$, $\mu_0 = -0.1$, $\Sigma_0 = 0.04$. [13] For consistency of comparison across alternative policies the random disturbances were held the same by seeding the random number generator with the same value prior to simulating given policy variety. Several observations emerge. One, the actively optimal policy is the only

[11]The learning process could even diverge to infinity, but since the paths are calculated under the assumption that $\epsilon_{t+\tau} = 0$ for all τ, the probability of divergence is vanishing. Noise in the state process will increase the amount of unintended probing to prevent divergence.

[12]The calculation of these expectations should account for future learning. Thus, Q-factors that we discussed earlier are the most appropriate reporting tools here.

[13]Probability bands for evolving distributions are commonly constructed by connecting the points, such as multiples of standard error or certain percentiles, at each horizon. Resulting objects are used in fan charts in the context of multi-horizon forecasting (Canova, 2007; Cogley, Morozov, and Sargent, 2005), in the standard error bands in reporting of the estimated VAR impulse responses. These evolving distributions are correlated if they are constructed from recursive simulations, and hence plots connecting the points at each horizon are likely to misrepresent the true uncertainty. Sims and Zha (1999) propose an orthogonalization which eliminates this correlation. We sidestep the issue by providing plots with large number of simulated paths. We can simultaneously discern the typical shapes of the time-paths and the frequency of time-path visits to regions in the state space. Fan charts are reported as well.

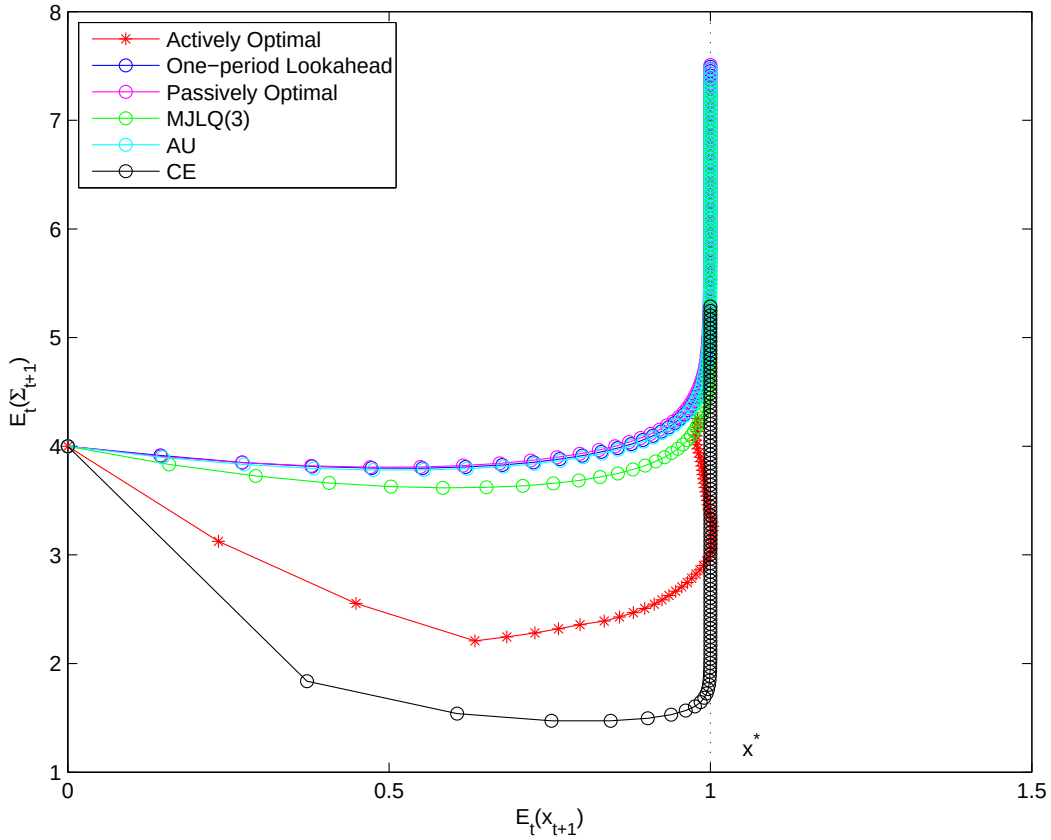


Figure 33: Expected dynamics of extended state under different policies. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 1.0$, $\sigma_\eta^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$. Starting values: $x_t = 0$, $\Sigma_t = 4$.

one without paths gravitating to zero policy intervention line $u_t = 0$, staying above it after the initial exploration phase. Two, most of the actively optimal policy paths start relatively far in the negative territory (given that initial belief is centered around $\mu_0 = -0.1$ which is of the opposite sign to $\beta_1 = 0.5$), switch quickly to large positive values of control, and then settle into the relatively narrow range of policy actions. Three, one period lookahead policy oscillates in the tightest ambit. The three passive policies that recognize uncertainty (passively optimal, MJLQ(3) and anticipated utility) are also the ones with the largest share of occasional outliers.

Figure 35 puts the policy differences under alternative assumptions under the microscope by concentrating on a single time-series realization with exactly equal random elements across policies. Certainty equivalent policy contrasts starkly with others as it quickly tapers off towards $u_t = 0$. One period lookahead seem to involve the least amount of probing and activism of the five non-certainty-equivalent policies. Actively optimal policy continues to display the greatest amount of experimentation during the early stages of the simulation, in both positive and negative directions. After about 15 time steps the five non-certainty-equivalent policies lie virtually on top of each other.

Figure 36 complements the story of figure 34 by representing the policy dynamics by means of a fanchart. Differently shaded bands of that plot indicate the probability ranges of time-varying policy choice. The darkest band encodes the range of $\pm 5\%$ probability around the median. This figure confirms that actively optimal policy varies within notably more constricted range than other policies after the initial outburst of activity. One period

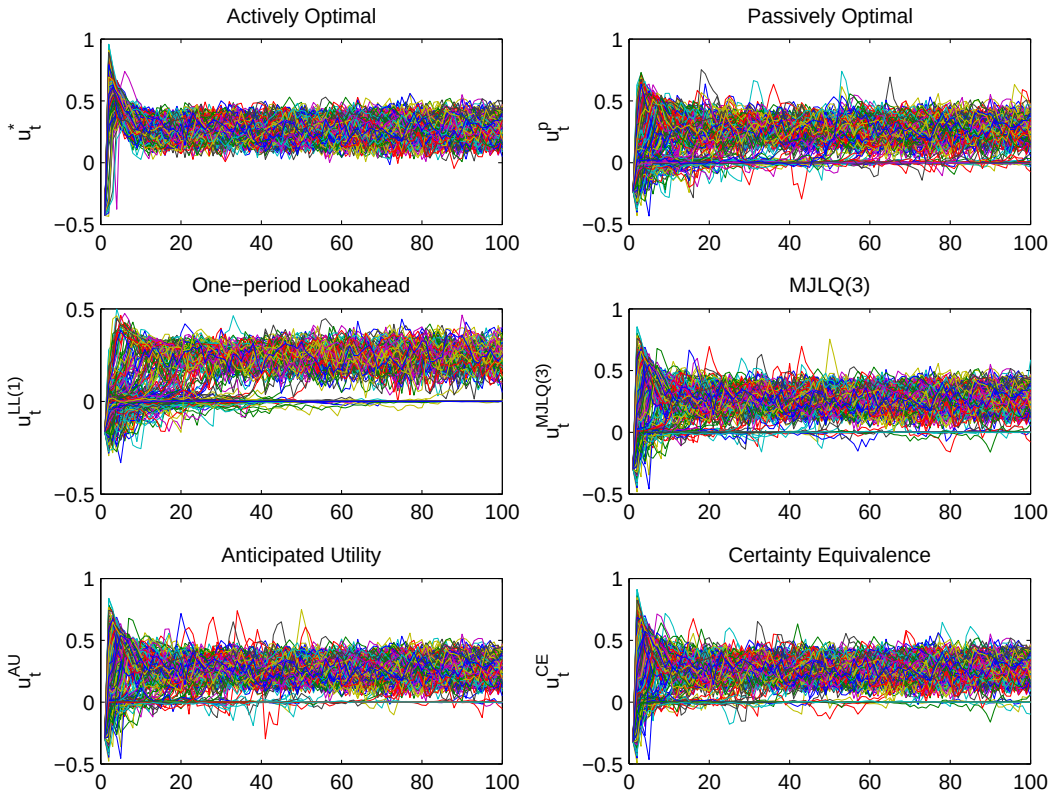


Figure 34: Simulated multiple time-series of control under different policies. Parameter values: $\alpha = -0.05$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 0.01$, $\sigma_\eta^2 = 0.0001$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Starting values: $x_0 = 0$, $\mu_0 = -0.1$, $\Sigma_t = 0.04$. True initial slope: $\beta_1 = 0.5$. Number of time periods: $T = 100$. Number of simulations: $NMC = 400$.

lookahead is also remarkable in that fully 50% of the paths are attracted by $u_t = 0$ line. Distributions of certainty equivalent, anticipated utility and MJLQ(3) controls are also markedly similar at every point in time. Passively optimal policy, in contrast, is dispersed notably more uniformly within the band as indicated by the lack of concentration around the median path.

Another way to look at the differences among policies is with the help of simulation-based pairwise scatter plots. These are shown in figure 37 for a single simulated time-series and in figure 38 for multiple simulations. Similarity is affirmed when the scatterplot is bundled tightly around 45-degree line. From that perspective, anticipated utility and MJLQ(3) policies are remarkably alike. Certainty equivalent and anticipated utility policies are not too different, apart form the incidence of occasional clusters at $u_t = 0$. On the contrary, one period lookahead policy is unlike any other. Also, taking account of additional uncertainty due to multiple simulations is seemingly important. For example, a single simulation scatter plot appears to indicate close affinity of actively optimal policy with certainty equivalent, anticipated utility and MJLQ(3) policies, while the multiple simulation scatter plot bespeaks considerable uncertainty surrounding this relationship. As a results, there is only moderate correlation between the actively optimal policy on the one hand and anticipated utility, MJLQ(3) and certainty equivalent approaches. The relationship between actively and passively optimal policies is also not entirely unequivocal.

10.6.2. *State Sequences.* Corresponding to the dynamics of control sequences we present the dynamic simulations of the physical state x_t and information state (μ_t, Σ_t). Figures 39

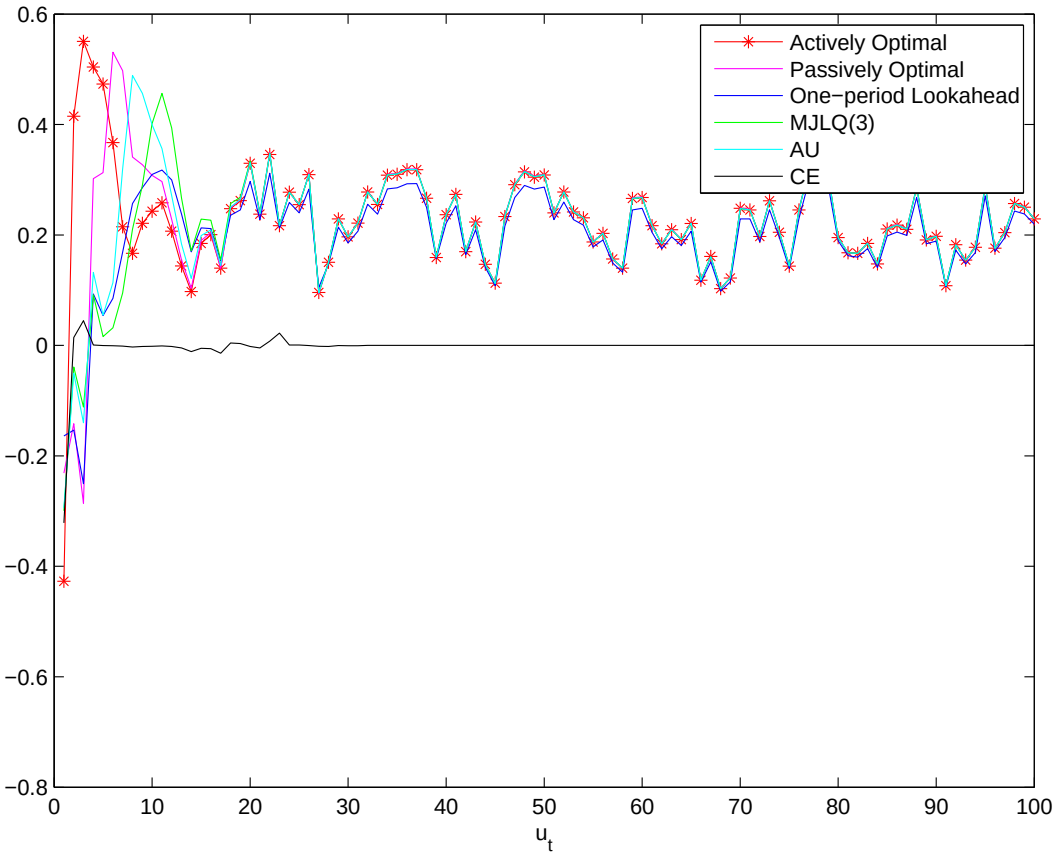


Figure 35: Simulated single time-series of control under different policies. Parameter values: $\alpha = -0.05$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 0.01$, $\sigma_\eta^2 = 0.0001$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Starting values: $x_0 = 0$, $\mu_0 = -0.1$, $\Sigma_t = 0.04$. True initial slope: $\beta_1 = 0.5$. Number of time periods: $T = 100$.

through 43 are dedicated to the physical state x_t. The most overt feature is the bifurcation of the state sequences into two basins of attraction for all suboptimal policies. The gap between the two branches is largest for the certainty equivalent policy, and the smallest for the one period lookahead. Figure 40 concentrates on a single simulation and supports findings of figure 35 in terms of the shape of the paths and the distance from target x^*.

Figure 41 establishes close similarity of evolving distributions for outcomes under anticipated utility, MJLQ(3), and to a lesser extent certainty equivalent policies. The range of state fluctuation is most narrow for actively optimal policy and widest for passively optimal policy, in accordance with the characteristics of policy functions.

Study of figures 42 and 43 confirms that outcomes under anticipated utility and MJLQ(3) policies are essentially identical. It also emphasizes distinctiveness of one period lookahead. Finally, it makes clear that the quality of the approximation of the actively optimal policy by any passive approximation depends on whether under an approximate policy the state stays within the branch that is closer to x^*.

Figures 44 through 48 depict the evolution of mean beliefs in simulations. A fair number of paths under each suboptimal policy remain attracted to $\mu_t = 0$ whereas the average of the unobserved policy slope is 0.5 (at every point in time). For example, the mean belief under the certainty equivalent policy in a particular simulation chosen for the plot in figure 45 is quickly stuck at zero. Such tendency induces slower learning and inhibits the progress

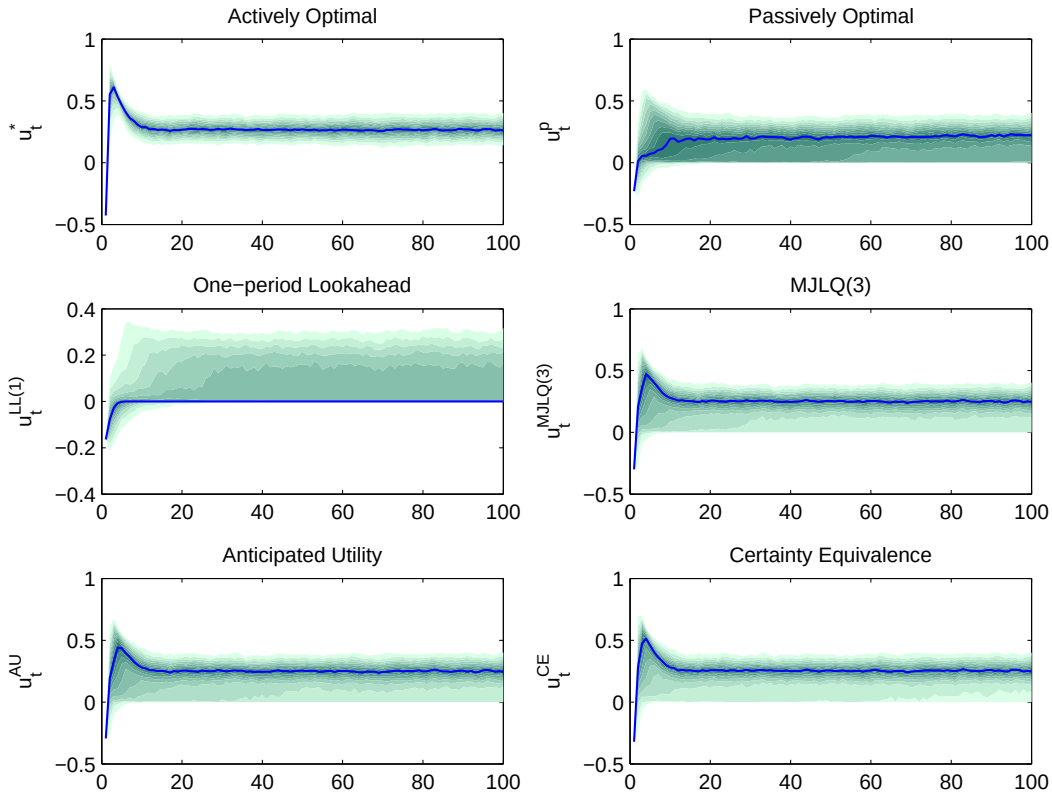


Figure 36: Evolving distribution of simulated time-series of control under different policies. Parameter values: $\alpha = -0.05$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 0.01$, $\sigma_\eta^2 = 0.0001$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Starting values: $x_0 = 0$, $\mu_0 = -0.1$, $\Sigma_t = 0.04$. True initial slope: $\beta_1 = 0.5$. Number of time periods: $T = 100$. Number of simulations: $NMC = 400$.

of the mean belief towards the latent slope process. The actively optimal policy learns the latent slope process the quickest of all as can be seen in figure 46. Again, mean beliefs are most highly correlated between anticipated utility and MJLQ(3) policies. Other pairs could be rather different on occasion, see figures 47 and 48.

Figure 49 demonstrates the superiority of the actively optimal strategy once more. Actively optimal policy is the only one with uniformly declining belief variance. Because of the permanent coefficient drift, the limit belief variance, if it exists, is strictly above zero. The single time-series simulation in figure 50 shows clearly that the belief variance is the lowest for the actively optimal policy, whereas that for certainty equivalent policy diverges. The entire evolving distribution of variance of belief is most narrow for the actively optimal policy. Pairwise scatter plots of belief variances indicate close similarity of anticipated utility and MJLQ(3), and clear supremacy of actively optimal policy. It also appears that the passively optimal policy is no better an approximation than other, simpler to compute, policies.

Since $\mathbb{E}_t x_{t+1}$ is just a linear transformation of x_t and u_{t+1}, most of the conclusions about simulated outcomes for x_t and u_t remain valid for $\mathbb{E}_t x_{t+1}$. Indeed, 54 is a virtual carbon copy of 39. To economize on space we omit the corresponding fancharts and scatter plots.

10.6.3. *Persistence.* A different perspective on the simulated outcomes is afforded by studying persistence of the realized simulated state. That persistence comes from two sources. The first source is the autoregressive dependence in the state equation. The second is the state dependence of the policy choice that feeds back on the subsequent realizations of the

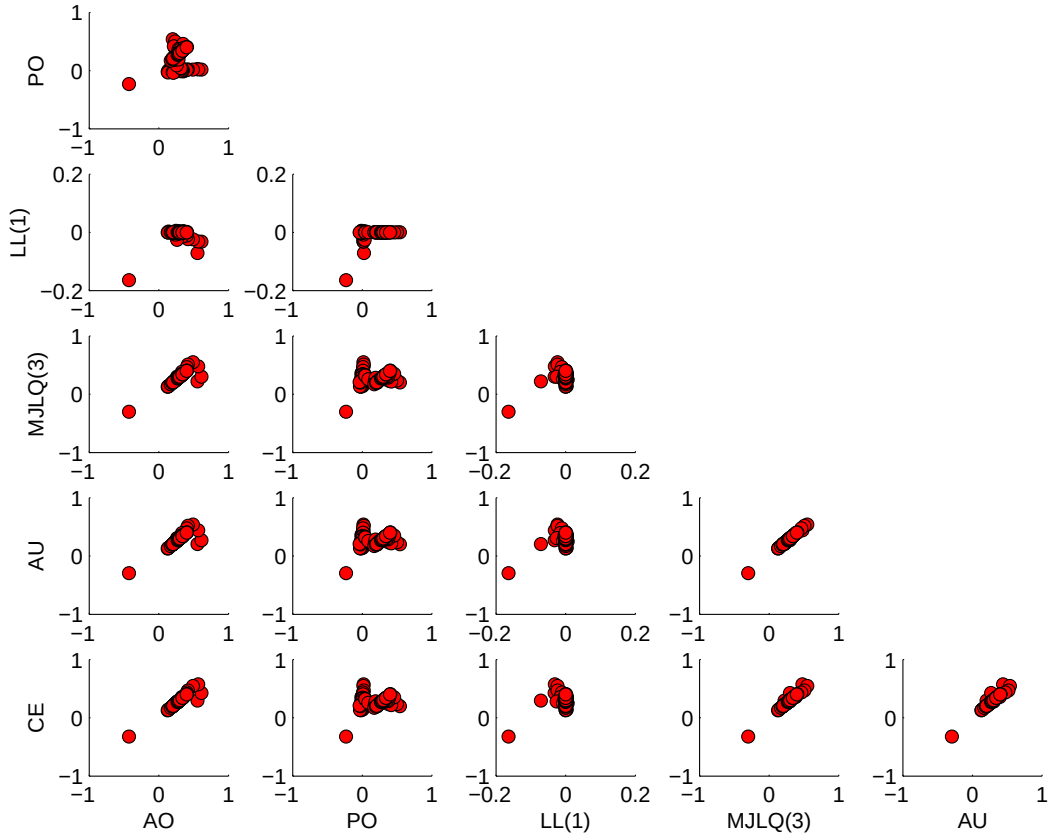


Figure 37: Pairwise comparisons of single simulated time-series of control under different policies. Parameter values: $\alpha = -0.05$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 0.01$, $\sigma_\eta^2 = 0.0001$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Starting values: $x_0 = 0$, $\mu_0 = -0.1$, $\Sigma_t = 0.04$. True initial slope: $\beta_1 = 0.5$. Number of time periods: $T = 100$.

physical state. We use simple first order sample autocorrelation as a linear persistence estimate in place of more complex spectral measures (Hamilton, 1994) or nonlinear measures (Gourieroux and Jasiak, 1999). While this is imperfect and misspecified measures, it is a common lens to study history dependence in time-series. All the results reported here are specific to simulations with direct autocorrelation $\gamma = 0.9$.

First, we report the observations on persistence that are common to different policy rules. When $\omega = 1$, i.e. when state and control deviations are perfectly balanced in the period loss function, we find that the state persistence is mostly driven by the assumed autoregressive dynamics, and not by the state dependence of the policy rule. At the same time, the contribution of control to persistence depends very much on parameters of the dual control problem. For example, the state persistence depends positively on the the weight ω by which control deviations are penalized in the period loss function. With $\omega = 0$, the average state persistence under any of the six policies is at most 0.6, well below the assumed autoregressive coefficient $\gamma = 0.9$ in the state equation. This could be seen in figure 55 where we also investigate the parametric dependence with respect to σ_ϵ^2 and σ_η^2. For the former, we observe that the state persistence tends to decline with larger state uncertainty, regardless of the policy, although there's also large sample variability. For the latter, the results belong to the pool of features that are distinct across policy rules.

Turning to the attributes that are dissimilar, we find the smallest amount of uncertainty about persistence for the actively optimal and certainty equivalent policies across the range

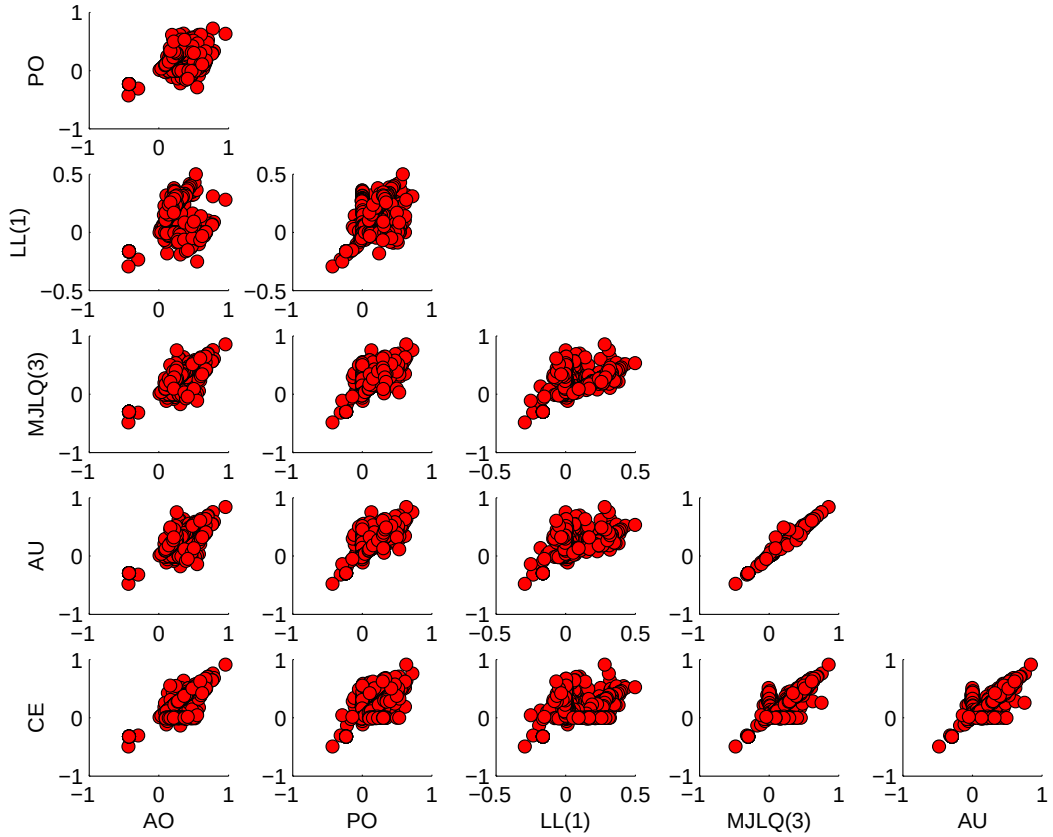


Figure 38: Pairwise comparisons of multiple simulated time-series of control under different policies. Parameter values: $\alpha = -0.05$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 0.01$, $\sigma_\eta^2 = 0.0001$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Starting values: $x_0 = 0$, $\mu_0 = -0.1$, $\Sigma_t = 0.04$. True initial slope: $\beta_1 = 0.5$. Number of time periods: $T = 100$. Number of simulations: $NMC = 400$.

of ω and σ_η^2 values. As σ_η^2 increases, the certainty equivalent control results in by far the largest reduction of the state persistence. The actively optimal and the passively optimal policies are the only ones displaying increase in persistence as σ_η^2 increases. Increasing σ_η^2 also drives the wedge between persistence under the passively optimal policy rule and persistence under anticipated utility and MJLQ(3) approximations. This makes sense as the quality of approximation provided by the latter two policies should necessarily deteriorate with faster parameter drift. In contrast, the distinction between anticipated utility and MJLQ(3) control is so insignificant that it doesn't show across different parameters. One-period lookahead is largely similar to them with only slightly higher persistence across parameter values.

Doctrinaire priors can have long-lasting impact on the persistence. This is shown in figure 56 that plots the average sample autocorrelation of the physical state against prior beliefs. Simulations underlying the construction only differ by initial values as well as policy rules but not in the way of random shocks. With Σ_0 in the vicinity of zero, $T = 100$ time periods is not enough to dissipate the impact of the prior. The direction of the impact depends on the sign of μ_0. If $\mu_0 < 0$, the persistence is amplified. Otherwise, it is dampened. Among the different policies, the optimal and certainty equivalent policies show the quickest dispersion of the impact of the prior beliefs. For the former policy, this is due to significant active experimentation components. For the latter policy, it is due to

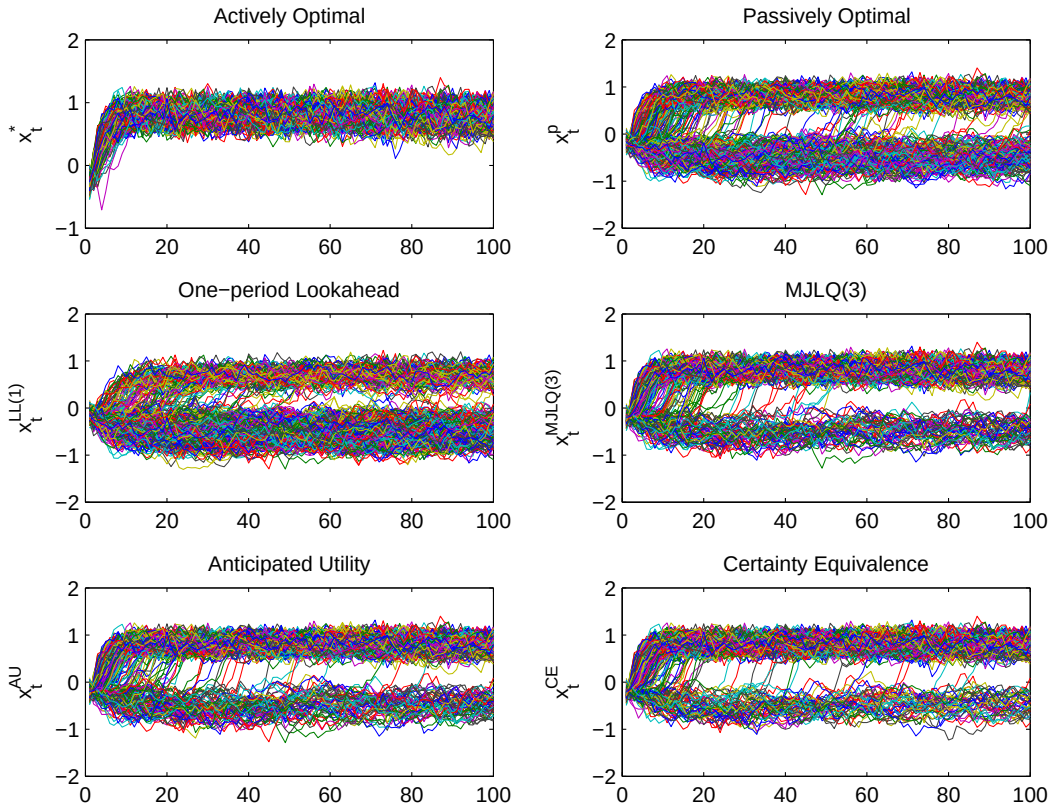


Figure 39: Simulated multiple time-series of target state x_t under different policies. Parameter values: $\alpha = -0.05$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma_\epsilon^2 = 0.01$, $\sigma_\eta^2 = 0.0001$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Starting values: $x_0 = 0$, $\mu_0 = -0.1$, $\Sigma_t = 0.04$. True initial slope: $\beta_1 = 0.5$. Number of time periods: $T = 100$. Number of simulations: $NMC = 400$.

accidental experimentation. For passive policies with uncertainty, the impact of priors on the persistence could be somewhat enduring.

10.6.4. *Regret Function.* The idea of regret goes back to Savage who argued for a decision-making based on the difference between the consequences of the best decision that could have been taken had the underlying circumstances been known and the decision that was in fact taken before they were known. Here we associate regret function in simulations simply with simulated cumulative loss function. This allows us to assess the performance of the given policy in terms of original intertemporal objective function. Indeed, acquiring better knowledge of the unobserved multiplicative policy coefficient has any worth only in the context of the performance index (2.1). Figures 57 through 61 explore performance of various policies via the following version of regret function:

$$C_t = \sum_{\tau=0}^{t} \delta^\tau \left((x_\tau - x^*)^2 + \omega (u_\tau - u^*)^2 \right). \tag{10.2}$$

Figures 57 and 59 demonstrate dominance of the actively optimal policy in terms of simulated regrets. For fairly low discount factor $\delta = 0.75$ used in simulations, convergence of the regret function to its long run value is quick, and there is no discernable time variation in the regret distribution after just a handful of periods. Figure 58 confirms quick convergence and also pinpoints the source of the actively optimal advantage – the loss is larger in the first two periods while the decision maker experiments in order to zoom in his beliefs about