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 optimizng 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
characterized by the mean and variance parameters

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

subject to

\[ x_t = \alpha + \beta_t u_t + \gamma x_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \]

\[ \beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2). \]

\( \alpha \) and \( \gamma \) are assumed to be known. Shock variances are known as well. Naturally, \( \omega \geq 0, \) \( \delta \in [0, 1]. \) Initial belief about \( \beta_0 \) is Gaussian with mean \( \mu_0 \) and variance \( \Sigma_0. \) 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 \( \beta_t, \) 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 \( \beta_t \) are normally distributed, because the prior distribution and likelihood functions for each \( t \) are all Gaussian. Hence, the beliefs about \( \beta_t \) are completely characterized by the mean and variance parameters \( \mu_t, \) and \( \Sigma_t \) conditional on current information set, i.e. following the choice of control \( u_t \) and realization of the shock \( \epsilon_t. \) By applying recursive Bayesian updating in the linear regression setup, the following updating equations can be derived:

\[ \mu_t = \mu_{t-1} + \Sigma_{t|t-1} u_t \left( u_t^2 \Sigma_{t|t-1} + \sigma^2 \right)^{-1} \left( x_t - \alpha - \mu_{t-1} u_t - \gamma x_{t-1} \right), \]

\[ \Sigma_t = \Sigma_{t|t-1} - \left( \Sigma_{t|t-1} \right)^2 u_t^2 \left( \Sigma_{t|t-1} + \sigma^2 \right)^{-1}. \]

where \( \Sigma_{t|t-1} = \Sigma_{t-1} + \sigma^2 \) is the conditional predictive variance of the hidden state \( \beta_t. \)

Here, learning is equivalent to Kalman filtering. The updating equation for variance is the deterministic process, which would be non-increasing if \( \sigma^2 = 0. \) In other words, if multiplicative policy parameter \( \beta_t \) 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\textsuperscript{1} is
\begin{equation}
S_t = (x_t, \mu_t, \Sigma_t) \in S.
\end{equation}

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 \( u : S \rightarrow \mathbb{R} \) we can compute expectation of future state conditional on the current information state:
\begin{align}
E_t x_{t+1} &= \alpha + \mu_{t+1|t} u_{t+1} + \gamma x_t, \\
E_t \Sigma_{t+1} &= \Sigma_t + \sigma_t^2 - (\Sigma_t + \sigma_t^2)^2 u_t^2 (\Sigma_t + \sigma_t^2) + \sigma_t^2)^{-1}.
\end{align}

By law of iterated expectations, expected evolution of the mean beliefs is trivial
\begin{equation}
E_t \mu_{t+1|t} = \mu_t.
\end{equation}

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
\begin{equation}
V(x_t, \mu_{t+1|t}, \Sigma_{t+1|t})
= \min_{\{u_t\}} \left\{ L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) \right. \\
+ \delta \int V(\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}), \Sigma_{t+2|t+1} (\Sigma_{t+1|t} u_{t+1})) \\
\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. \right\}
\end{equation}

\textsuperscript{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
\[
L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) = \int \left( (\alpha + \beta x_t + \gamma x_t + \epsilon_{t+1} - x^*)^2 + \omega(u_{t+1} - u^*)^2 \right) \\
\times p(\beta_{t+1}|x_t, \mu_{t+1|t}, \Sigma_{t+1|t})q(\epsilon_{t+1})d\beta_{t+1}d\epsilon_{t+1} \\
= \left( \Sigma_{t+1|t} + \mu_{t+1|t}^2/\gamma \right) u_{t+1}^2 + 2(\gamma \mu_{t+1|t}(\alpha - x^*))u_{t+1} + 2\gamma(\alpha - x^*)x_t + \gamma^2 x_t^2 + \sigma^2 + (\alpha - x^*)^2 + \omega(u_{t+1} - u^*)^2,
\]
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 $\mu_t$ when $\Sigma_t$ is fixed at 0.05. The middle panel in the figure represents the policy function in $x_t$ and $\Sigma_t$ variables when $\mu_t = -1.64$. The bottom panel contains the plot of the policy function against $\mu_t$ and $\Sigma_t$ 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:
\[
V_t^* \leq V_t^0 := E_{t-1} \sum_{\tau=0}^{\infty} \delta^\tau \left( (x_{t+\tau} - x^*)^2 + \omega(u^*)^2 \right) \\
= \sum_{\tau=0}^{\infty} \delta^\tau \left( \frac{(\alpha + \gamma x_{t-1} - x^*)^2 - \delta \gamma (x^*)^2 - \alpha^2 - \gamma x^*(2\alpha - x^*) + \gamma x_{t-1}(1 + \gamma) - 2x_{t-1}(x^* - \alpha + \gamma^2 x^*)}{(1 - \delta)(1 - \gamma^2 \delta)} \right) \\
+ \frac{\gamma^2 \delta^2 (x_{t-1} - x^*)^2}{(1 - \delta)(1 - \gamma^2 \delta)} + \frac{\sigma^2}{(1 - \delta)(1 - \gamma^2 \delta)} + \frac{\omega(u^*)^2}{1 - \delta}.
\]
Figure 1: Actively adaptive optimal control. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma^2_t = 1.0$, $\sigma^2_\eta = 0.01$. (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 its 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
(4.2)  
\( V^{pm}_{t}(x_t, \mu_t, \Sigma_{t+1|t}) = \min_{u_{t+1}} \left\{ \mathbb{E}_t \left\{ L(x_{t+1}, u_{t+1}) + \delta V^0(x_{t+1}) \right\} \right\} \)  

\[ = \min_{u_{t+1}} \left\{ \mathbb{E}_t \left[ (\alpha + \beta x_{t+1} + \gamma \epsilon_{t+1} + \epsilon_t + x^*)^2 + \omega (u_{t+1} - u^*)^2 \right] + \delta \mathbb{E}_t V^0 (\alpha + \beta x_{t+1} + \gamma \epsilon_{t+1} + \epsilon_t + x^*) \right\} \]  

\[ = \frac{2\alpha x x_t (1 - \delta) + \alpha^2 (1 + \gamma \delta) - 2x^*(\alpha + x \gamma (1 - \delta))(1 - \gamma^2 \delta)}{(1 - \delta)(1 - \gamma \delta)(1 - \gamma^2 \delta)} + \frac{(x^*)^2 (1 - \gamma \delta)(1 - \gamma^2 \delta) - (1 - \gamma \delta) (\gamma^2 (-1 + \delta) x_t^2 - \sigma^2 - (u^*)^2 (1 - \gamma^2 \delta) \omega)}{(1 - \delta)(1 - \gamma \delta)(1 - \gamma^2 \delta)} - \frac{(1 - \gamma^2 \delta) \left( \frac{\alpha - x + \gamma x - (x - x^*) \gamma^2 \delta}{(1 - \gamma \delta)(1 - \gamma^2 \delta)} \mu_t + u^* \omega \right)^2}{\mu_t + \omega - \gamma^2 \delta \omega + \Sigma_{t+1|t}}. \]

Performance of \( V^{pm}_{t}\) 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 \( V^0_t \).

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

(4.3)  
\( V^{0,pm}_t(x_t, \mu_t, \Sigma_{t+1|t}) = \min \left\{ V^0_t(x_t), V^{pm}_{t}(x_t, \mu_t, \Sigma_{t+1|t}) \right\} \)
This bound is nontrivial improvement since there is no uniform dominance among \( V_{\text{pm}}^t \) and \( V_0^t \). The combined bound could then be used to derive a bound in the policy space, which is given by the following expression.

\[
-\mu_t (\alpha + \gamma x_t - x^*) - \omega u^* - \sqrt{D} \\
\leq u^*_{t+1}
\]

\[
-\mu_t (\alpha + \gamma x_t - x^*) - \omega u^* + \sqrt{D} \\
\leq u^*_{t+1}
\]

where

\[
D = (\mu_t (\alpha + \gamma x_t - x^*))^2 - (\mu_t^2 + \Sigma_{t|t} + \sigma_\eta^2 + \omega) \left( (\alpha + \gamma x_t - x^*)^2 + \sigma_x^2 + \omega (u^*)^2 - V_{\text{pm}}^0(x_t, \mu_t, \Sigma_{t+1|t}) \right).
\]

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
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 $\epsilon_{t+1}$, it turns out that the optimal choice of policy is the same for all $\sigma^2_{\epsilon}$. In particular, it is equal to the control that would obtain under $\sigma^2_{\epsilon} = 0$, 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:

$$V^{CE}(x_t) = \min_{u_{t+1}} \left\{ E_t \left( \alpha + \beta_t u_{t+1} + \gamma x_t + \epsilon_{t+1} - x^* \right)^2 + \omega (u_{t+1} - u^*)^2 + \delta E_{t+1} V^{CE} \left( \alpha + \beta_{t+1} u_{t+1} + \gamma x_t + \epsilon_{t+1} \right) \right\}. \quad (5.1)$$
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: $\alpha = 0$, $\gamma = 0.9$, $\delta = 0.75$, $\omega = 1$, $x^* = 1$, $u^* = 0$, $\sigma_x^2 = 1$, $\sigma_u^2 = 0.04$. Fixed coordinates in the top panel: $\mu_t = -0.5$, $\Sigma_{tt} = 0.25$. Fixed coordinates in the middle panel: $x_t = 0$, $\Sigma_{tt} = 0.25$. 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

$$Ax^2 + 2Bx + C = \min_u \left\{ (\mu^2 + \omega + \delta u^2) u^2 + 2(\mu \gamma x + \mu(\alpha - x^*) - \omega u^*) + \delta \mu \alpha A + \delta \mu B \right\} u$$

$$+ \gamma^2 x^2 + \sigma_x^2 + (\alpha - x^*)^2 + 2\gamma(\alpha - x^*) + \omega(u^*)^2 + \delta A(\alpha + \gamma x)^2 + \delta \sigma_x^2 A + 2\delta B(\alpha + \gamma x) + \delta C \right\}.$$ 

Hence,

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

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

Under (5.2), the value function becomes

$$(5.3) \quad V^{CE}(x) = -\frac{(\mu \gamma (1 + \delta A)x + \mu(\alpha - x^*) - \omega u^* + \delta \alpha A + \delta B)}{\mu^2 (1 + \delta A) + \omega} x$$

$$+ \gamma^2 x^2 + (\alpha - x^*)^2 + (1 + \delta A)\sigma_x^2 + \omega(u^*)^2 + 2\gamma(\alpha - x^*)x$$

$$+ \delta A(\alpha + \gamma x)^2 + 2\delta B(\alpha + \gamma x) + \delta C$$

$$= Ax^2 + 2Bx + C$$
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: $\alpha = 0$, $\gamma = 0.9$, $\delta = 0.75$, $\omega = 1$, $x^* = 1$, $u^* = 0$, $\sigma_x^2 = 1$, $\sigma_u^2 = 0.04$. Fixed coordinates in the top panel: $\mu_t = -0.5$, $\Sigma_{tu} = 0.25$. Fixed coordinates in the middle panel: $x_t = 0$, $\Sigma_{tt} = 0.25$. 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

\begin{equation}
A = -\frac{\mu^2 \gamma^2 (1 + \delta A)^2}{\mu^2 (1 + \delta A) + \omega} + \gamma^2 (1 + \delta A),
\end{equation}

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

\begin{equation}
A = -\frac{\gamma (1 - \mu \gamma)}{\mu (1 - \delta \gamma^2)}
\end{equation}

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

\begin{equation}
B = \frac{\mu \gamma (1 + \delta A) (\mu (x^* - \alpha) + \omega u^* - \delta \alpha \mu A - \delta \mu B)}{\mu^2 (1 + \delta A) + \omega} + \delta \gamma B - \gamma (x^* - \alpha) + \delta \alpha \gamma A
\end{equation}

which can be simplified to

\begin{equation}
B = \frac{\omega \gamma ((1 + \delta A) (\mu u^* + \alpha) - x^*)}{\mu^2 (1 + \delta A) + (1 - \gamma \delta) \omega}.
\end{equation}
Finally, the third equation is obtained by equating constant terms:

\[
(1 - \delta)C = - \frac{\mu(\alpha(1 + \delta A) + \delta B - x^*) - \omega u^*)^2}{\mu^2(1 + \delta A) + \omega}
+ (1 + \delta A)\sigma_t^2 + (\alpha - x^*)^2 + \omega(u^*)^2 + \delta \alpha^2 A + 2\delta \alpha B.
\]

(5.8)

Figure 7 depicts the policy response surface as a function of physical and informational state variables \(x_t\) and \(\mu_t\). Since \(u_{CE, t+1}\) does not depend on \(\Sigma_t\), plotting additional slices is redundant. As the certainty equivalent policy is linear in \(x_t\), the impact of \(\mu_t\) 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 \(\beta \sim N(\mu_{t+1|t}, \Sigma_{t+1|t})\) 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 \(\beta\) will evolve over time is not taken into 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: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma^2_\epsilon = 1.0$, $\sigma^2_\eta = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$.

simplify notation, we omit time subscripts on $\mu$ and $\Sigma' = \Sigma + \sigma^2_\eta$. 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

$$
V(x) = \min_u \left\{ \mathbb{E} \left( \beta^2 u^2 + \gamma^2 x^2 + \sigma^2_\epsilon + (\alpha - x^*)^2 + 2\beta \gamma xu + 2\beta u + 2\beta(\alpha - x^*)u + 2\gamma x \epsilon \\
+ 2\gamma(\alpha - x^*)x + 2(\alpha - x^*)\epsilon + \omega u^2 - 2\omega uu^* + \omega(u^*)^2 + \delta V(\alpha + \mu u + \gamma x + \epsilon) \right) \right\}.
$$

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

$$Ax^2 + 2Bx + C = \min_u \left\{ \left( \Sigma' + \mu^2 \right) u^2 + \gamma^2 x^2 + \sigma^2_\epsilon + (\alpha - x^*)^2 + 2\gamma \mu xu \\
+ 2\mu(\alpha - x^*)u + 2\gamma(\alpha - x^*)x + \omega u^2 - 2\omega uu^* + \omega(u^*)^2 \\
+ \delta A \left( \alpha^2 + (\Sigma' + \mu^2) u^2 + \gamma^2 x^2 + \sigma^2_\epsilon + 2\gamma \mu xu + 2\alpha \mu u + 2\alpha \gamma x \right) \\
+ \delta B \left( \alpha + \mu u + \gamma x \right) + \delta C \right\}.$$
Performing explicit minimization, we get

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

(6.2)

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

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

Substitute (6.2) into the cost to go function:

$$V^{AU}(x) = -\frac{(\gamma \mu (1 + \delta A)x + \mu (\alpha (1 + \delta A) - x^* + \delta B) - \omega u^*))^2}{(\Sigma' + \mu^2)(1 + \delta A) + \omega}$$

$$+ \gamma^2 x^2 + \sigma x + (\alpha - x^*)^2 + 2\gamma (\alpha - x^*) x + \omega (u^*)^2$$

$$+ \delta \alpha^2 A + \delta \gamma^2 Ax^2 + \delta \sigma^2 A + 2\delta \alpha \gamma Ax + \delta B x + \delta \gamma B x + \delta C$$

$$= Ax^2 + 2Bx + C.$$

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

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

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

and

(6.5) $$-\frac{(\mu (x^* - \alpha (1 + \delta A) - \delta B) + \omega u^*))^2}{(\Sigma' + \mu^2)(1 + \delta A) + \omega} + (1 + \delta A) \sigma\gamma^2 + (\alpha - x^*)^2 + \omega (u^*)^2 + \delta \alpha^2 A + \delta \alpha B$$

$$= (1 - \delta) C.$$

Equation for $B$ can be made explicit

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

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. 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

(7.1) $$X_{t+1} = A_{s(t+1)}X_t + B_{s(t+1)} U_{t+1} + C_{s(t+1)} \xi_{t+1},$$
Figure 9: Anticipated utility passively adaptive control. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma^2_\epsilon = 1.0$, $\sigma^2_\eta = 0.01$. (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.\footnote{If regimes were observed, coefficients of the quadratic value function would be directly regime-dependent.} 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.
Figure 10: Volumetric plot of anticipated utility passively adaptive policy function. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_x^2 = 1.0$, $\sigma^2 = 0.01$.

The Bellman equation for MJLQ model becomes

\[
J(X_t, p_{t+1}) = X_t'V(p_{t+1})X_t + w(p_{t+1}) = \min_{U_{t+1}} \left\{ X_t'QX_t + U_{t+1}'R(p_{t+1})U_{t+1} + 2X_t'N(p_{t+1})U_{t+1} + \delta E_t J(X_{t+1}, p_{t+1}) \right\}
\]

\[
= \min_{U_{t+1}} \left\{ X_t'QX_t + U_{t+1}'R(p_{t+1})U_{t+1} + 2X_t'N(p_{t+1})U_{t+1} + \delta \sum_{j,k} p_{t+1,j} P_{jk} (X_{t+1,k}'V(p_{t+2})X_{t+1,k} + w(p_{t+2})) \right\},
\]

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. $R$ 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.

---

3Because of timing differences with Costa, Fragoso, and Marques (2005) and Svensson and Williams (2007), matrices $Q$, $R$ and $N$ encapsulate expected period $t+1$ loss function.
The first-order condition with respect to $U_{t+1}$ is
\begin{equation}
U_{t+1} R(p_{t+1}) + \delta \sum_{j,k} p_{t+1,j} P_{jk} (X_t A_k' V(P'_{p_{t+1}}) B_k + U_{t+1} B_k') = 0
\end{equation}
and can be written as
\begin{equation}
U_{t+1} = -G(p_{t+1})^{-1} K(p_{t+1}) X_t,
\end{equation}
where
\begin{align*}
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.
\end{align*}

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

The scalar $w(p_{t+1})$ is only important for the expected loss function, not for the control. It solves the equation
\begin{equation}
w(p_{t+1}) = \delta \sum_{j,k} p_{t+1,j} P_{jk} \left( \text{tr} \left(V(P'_{p_{t+1}}) C_k C_k'\right) + w(P'_{p_{t+1}}) \right) .
\end{equation}

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.

Figure 11: Phase portrait of expected state dynamics under anticipated utility policy. Parameter values: $\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.75$, $\sigma^2 = 1.0$, $\sigma^2 = 0.04$, $\omega = 1.0$, $x^* = 1.0$, $u^* = 0$. Mean belief: $\mu_t = 0.5$. The first-order condition with respect to $U_{t+1}$ is
\begin{equation}
U_{t+1} R(p_{t+1}) + \delta \sum_{j,k} p_{t+1,j} P_{jk} (X_t A_k' V(P'_{p_{t+1}}) B_k + U_{t+1} B_k') = 0
\end{equation}
and can be written as
\begin{equation}
U_{t+1} = -G(p_{t+1})^{-1} K(p_{t+1}) X_t,
\end{equation}
where
\begin{align*}
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.
\end{align*}

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

The scalar $w(p_{t+1})$ is only important for the expected loss function, not for the control. It solves the equation
\begin{equation}
w(p_{t+1}) = \delta \sum_{j,k} p_{t+1,j} P_{jk} \left( \text{tr} \left(V(P'_{p_{t+1}}) C_k C_k'\right) + w(P'_{p_{t+1}}) \right) .
\end{equation}

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 $N(\mu_{t+1|t}, \Sigma_{t+1|t})$ distribution into $S$ segments of equal probability, and define $S$ states (regimes) as the expected values of truncated normal distributions over each segment. Thus, $p_{t+1} = (1/S, \ldots, 1/S)'$ and $\beta_k = \mathbb{E}(\beta|\beta \in [\Phi^{-1}(k-1)/S, \mu_{t+1|t}, \Sigma_{t+1|t}), \Phi^{-1}(k/S, \mu_{t+1|t}, \Sigma_{t+1|t})], \beta \sim N(\mu_{t+1|t}, \Sigma_{t+1|t})$, where $\Phi^{-1}$ is the inverse cumulative density function of normal distribution. The transition probability matrix is similarly defined by discretizing the probability distribution of $\beta_{t+1}$ conditional on $\beta_t = \beta_k$ for each $k = 1, \ldots, S$. Finally, to link state equation (2.2) with regime switching linear system (7.1), use the following definitions: $X_t = (x_t - x^*)$, $U_t = u_t$, $A_k = \begin{pmatrix} \alpha + (\gamma - 1)x^* + \beta_k u^* \gamma \\ \gamma x^* + \gamma (\alpha - x^*) \gamma^2 \end{pmatrix}$, $B_k = \begin{pmatrix} 0 \\ \beta_k \end{pmatrix}$, $C_k = \begin{pmatrix} 0 \\ \sigma_x \end{pmatrix}$, $Q = \begin{pmatrix} \gamma (\alpha - x^*)^2 + 2\gamma^2 x^* + \omega u^* + \sigma_x^2 \\ \gamma^2 \end{pmatrix}$, $R = \omega + \sum_{j=1}^{S} p_{j,t+1} \beta_j^2$.

Having described the solution method, we present 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 $\beta$), tenable modulation of the state transmission channel (time-varying $\gamma$), potential regimes of high or low shock variance (time-varying $\sigma_x$), regime-switching mean dynamics (time-varying $\alpha$) 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
interested in solving the following Bellman equation

\[
V(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}) = \min_{\{u_{t+1}\}} \left\{ L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) + \delta \int \int V(\alpha + (\beta_{t+1} + \eta_{t+1})u_{t+1} + \gamma x_t + \epsilon_{t+1}, \beta_{t+1} + \eta_{t+1}, \Sigma_{t+2|t}) \times p(\beta_{t+1}|x_t, \mu_{t+1|t}, \Sigma_{t+1|t})p(\eta_{t+1})q(\epsilon_{t+1})d\mu_{t+1|t}d\eta_{t+1}d\epsilon_{t+1} \right\}
\]

\[
= \min_{\{u_{t+1}\}} \left\{ L(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) + \delta \int \int V(\alpha + \mu_{t+2|t}u_{t+1} + \gamma x_t + \epsilon_{t+1}, \mu_{t+2|t}, \Sigma_{t+2|t}) \times p(\mu_{t+2|t})q(\epsilon_{t+1})d\mu_{t+2|t}d\epsilon_{t+1} \right\}.
\]
Figure 13: Volumetric plot of passively adaptive MJLQ(3) policy function. Parameters: $\alpha = \omega = u^* = 0, \gamma = 0.9, \delta = 0.75, x^* = 1, \sigma^2 = 1.0, \sigma^2_\eta = 0.01$.

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 $\Sigma_{t+2|t+1}$. The passively optimal approach recognizes that the regression coefficient $\beta_{t+1}$ is subject to shocks $\eta_{t+1}$ that can stir it away from the current value $\beta_t$. 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 $\eta_{t+1} \sim \mathcal{N}(0, \sigma^2_\eta)$. Forcing $\sigma^2_\eta = 0$ should reduce the control to the anticipated utility policy. 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.

4Such 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 $\tau > 0$.

5We 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.

6This implies that anticipated utility function is a good starting point for the value iteration for small $\sigma^2_\eta$. 

---

---
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.

Figure 15 plots three slices of the passively adaptive optimal policy function. The top slice is a function of $x_t$ and $\mu_t$ when $\Sigma_t$ is fixed at 0.05. The middle surface represents the policy function as a function of $x_t$ and $\Sigma_t$ for $\mu_t = -1.64$. The bottom graph plots the policy function against $\mu_t$ and $\Sigma_t$ 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.

---

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 $\Sigma$ until the solution doesn’t change.
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:  

\[ \text{Objective} = \min \left( \text{Explicit two-period problem} \right) \]

An alternative formulation could use finite-horizon dynamic programming and would be less explicit.

---

\( ^8 \) Alternative formulation could use finite-horizon dynamic programming and would be less explicit.
Figure 16: Volumetric plot of passively adaptive optimal policy function. Parameters: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma^2 = 1.0$, $\sigma^2_0 = 0.01$.

\begin{equation}
\min_{u_{t+1}, u_{t+2}} \mathbb{E}_t \left\{ (\alpha + \beta_{t+1} u_{t+1} + \gamma x_t + \epsilon_{t+1} - x^*)^2 + \omega (u_{t+1} - u^*)^2 \\
+ \delta \left[ (\alpha + \beta_{t+2} u_{t+2} + \gamma x_{t+1} + \epsilon_{t+2} - x^*)^2 + \omega (u_{t+2} - u^*)^2 \right] \right\}
\end{equation}

\begin{align}
= \min_{u_{t+1}, u_{t+2}} \left\{ (\alpha + \gamma x_t - x^*)^2 + \sigma^2 + \mathbb{E}_t (\beta_{t+1})^2 u_{t+1}^2 + 2(\alpha + \gamma x_t - x^*) (\mathbb{E}_t \beta_{t+1}) u_{t+1} \\
+ \delta \mathbb{E}_t \left[ (\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^*) \right] \\
+ \omega (u_{t+1} - u^*)^2 + \delta \omega (u_{t+2} - u^*)^2 \right\}
\end{align}

\begin{align}
= \min_{u_{t+1}, u_{t+2}} \left\{ (\alpha + \gamma x_t - x^*)^2 + \sigma^2 + (\mu + \Sigma + \sigma^2_0) u_{t+1}^2 + 2(\alpha + \gamma x_t - x^*)\mu u_{t+1} \\
+ \delta (\alpha(1 + \gamma) + \gamma^2 x_t - x^*)^2 + \delta(1 + \gamma^2)\sigma^2 + \delta \mathbb{E}_t (\beta_{t+2}^2 | u_{t+1} |) u_{t+2}^2 \\
+ \delta \gamma \mathbb{E}_t \left[ \beta_{t+1}^2 \left| u_{t+1} \right| \right] u_{t+1}^2 + 2\delta(\alpha(1 + \gamma) + \gamma^2 x_t - x^*) \mathbb{E}_t \left[ \beta_{t+1} \left| u_{t+1} \right| \right] u_{t+1} \\
+ 2\delta \gamma (\alpha(1 + \gamma) + \gamma^2 x_t - x^*) \mathbb{E}_t \left[ \beta_{t+2} \left| u_{t+1} \right| \right] u_{t+1} + 2\delta \gamma \mathbb{E}_t \left[ \beta_{t+1} \beta_{t+2} \left| u_{t+1} \right| \right] u_{t+1} u_{t+2} \\
+ \omega (u_{t+1} - u^*)^2 + \delta \omega (u_{t+2} - u^*)^2 \right\}.
\end{align}
Figure 17: Phase portrait of expected state dynamics under passively optimal policy. Parameter values: \( \alpha = 0.1, \gamma = 0.9, \delta = 0.75, \sigma^2_t = 1.0, \sigma^2_\eta = 0.04, \omega = 1.0, x^* = 1.0, u^* = 0 \). 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)

\[
E_t [\beta_{t+2} | u_{t+1}] = E_t [\beta_{t+1} | u_{t+1}] = E_t [\beta_{t+2}] = \mu_t.
\]

Future variances, on the other hand, follow nontrivial dynamics:

\[
E_t [\beta^2_{t+1} | u_{t+1}] = \mu_t^2 + \Sigma_t + \sigma^2_\eta, \tag{9.2}
\]

\[
E_t [\beta^2_{t+1} | u_{t+1}] = \mu_t^2 + \Sigma_t + \sigma^2_\eta \tag{9.3}
\]

\[
E_t [\beta_{t+1}^2 | u_{t+1}] = \mu_t^2 + \Sigma_t + \sigma^2_\eta \tag{9.4}
\]

\[
E_t [\beta_{t+2}^2 | u_{t+1}] = \mu_t^2 + \Sigma_t + \sigma^2_\eta \tag{9.5}
\]

where

\[
\Sigma_{t+1}(u_{t+1}) = \Sigma_t + \sigma^2_\eta - \frac{(\Sigma_t + \sigma^2_\eta)^2 u_{t+1}^2}{\Sigma_t + \sigma^2_\eta + \sigma^2_t} \tag{9.6}
\]

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: $\alpha = 0.01, \gamma = 0.9, \delta = 0.75, \omega = 1.6, x^* = 1, u^* = 0, \sigma_e^2 = 0.2, \sigma_\eta^2 = 0.4$.

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, $\omega$. As $\omega$ 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
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 $\Sigma_t$, 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: $\alpha = \omega = u^* = 0$, $\gamma = 0.9$, $\delta = 0.75$, $x^* = 1$, $\sigma_2 = 1.0$, $\sigma_2^u = 0.01$.

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:

\begin{equation}
V^*(x_t, \mu_{t+1|t}, \Sigma_{t+1|t}, u_{t+1}) = 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(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}))
\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{equation}

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: $u_{t+1} \in \{u^*_{t+1}, u^P_{t+1}, u^LL_{t+1}, u^MJLQ_{t+1}, u^AU_{t+1}, u^CE_{t+1}\}$. 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 $\omega$ 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.