BAYESIAN DUAL CONTROL: REVIEW OF THE LITERATURE

SERGEI MOROZOV

ABSTRACT. Imperfect information in the form of model uncertainty in the dynamic intertemporal choice problems makes the optimizing decision maker face difficult tradeoff between simultaneously controlling the policy target and estimating the impact of policy action, a situation known as dual control. Extant literature on Bayesian dual control is reviewed here. *JEL classification*: C44; C63; D83; E17; E52

Keywords: Bayesian dual control, adaptive control, certainty equivalent control, cautionary control, myopic control, probing, experimentation, anticipated utility, actively adaptive prediction of posterior distributions, limited lookahead, limit beliefs, jump-linear-quadratic model, coefficient augmentation, multi-armed bandit models, numerical dynamic programming, linearization, perturbation

Date: April 24, 2008.

All errors are due entirely to sunspots.

Contents

1. Introduction

The problem of conflict between information gathering and control quality 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). The compromise between probing and control, or in Feldbaum's terminology investigating and directing, lead to the concept of dual control. 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, many attempts have been made to find simpler suboptimal solutions and compare with optimal dual control solutions when the latter could be found. While the main development effort originated in the engineering literature, the idea of the tradeoff between caution and probing impressed economists as well. Indeed, control with dual features is most advantageous when it is necessary to rapidly find good estimates, when the initial estimates are of poor quality, and when the parameters of the process evolve rapidly or in the manner not well understood (Wittenmark, 1995). That these are precisely the characteristics of the realm of economics was argued in Bar-Shalom and Wall (1980); Kendrick (1982); Wieland (2000b).

2. Stabilization and Probing Tradeoff

In the dual control, whether optimal or suboptimal, it is necessary to have both stabilizing and probing features. Both parts of the control law can be obtained in a variety of ways. Stabilizing part of the control action could be further classified in accordance with the amount of uncertainty it is willing to tackle. At one extreme, there's certainty equivalence that disregards the parameter uncertainty altogether. At the other extreme are various cautious policies, so called because they hedge against poor process knowledge. Accordingly, the extant literature, both in economics and engineering, explored above features in some detail.

3. Passive Learning Controls

3.1. **Certainty Equivalent Control.** Early certainty equivalent tradition in economics is rooted in the work of Theil (1957), predating the discovery of the dual control. One of the first application of this approach to macroeconomic stabilization was Holt (1962).

3.2. **Parameter Uncertainty and Cautionary Control.** Nearly simultaneously, the new genre of literature appeared that took cautious feature seriously. Indeed, an optimal adaptive control under model uncertainty, dual or not, should also take the quality of the parameter estimates into consideration. A first example of this kind of control is cautionary myopic control, which is short-sighted and looks only one step ahead. As such it optimizes expected one-stage cost-to-go criterion. Brainard (1967) is one of the first macroeconomic applications in the new genre. A remarkable finding of that paper was that with multiple control instruments it is optimal to employ all of them. The engineering literature that was drawn on by economists near the start of this work was the paper by Farison, Graham, and Shelton (1967) and the book by Aoki (1967). Other early contributions include Shupp (1972) venturing outside linear-quadratic specification, Henderson and Turnovsky (1972) extending cautionary feature to dynamic setting and continuous time, Chow (1973) generalizing the state process to the general linear ARMA type, Kendrick and Majors (1974) using state augmentation to derive cautious control when unknown multiplicative parameters drift over time, Turnovsky (1975) studying the choice of policy instruments in the advanced dynamic monetary equilibrium model, Chow (1978) evaluating the outcomes of the macroeconomic policies, Craine (1979) analyzing the differences in the optimal monetary policy depending on whether the dynamic parameter uncertainty is driven by direct policy impact uncertainty or by the uncertainty about the transition dynamics, etc. An undesired consequence of caution introduced by taking account of poor process knowledge was found that the gain in the controller could decrease. With subdued control signals less information will be gained about the process. If system's parameters are sufficiently strongly time-varying, the parameter uncertainties will increase and even smaller control signals will be generated, creating a vicious circle and eventual *turn-off* of the control. The problem has been reported in Åström and Wittenmark (1971); Wieslander and Wittenmark (1971). It was then recognized that some disregard of parameter uncertainty may be beneficial to move closer to the dual control action.

4. Dual Control by Dynamic Programming

As for the dual control, an early paper by Prescott (1972) considered a toy multi-period control problem with the data generated by the simple regression with an unknown slope, $x_t = \beta u_t + \epsilon_t$. Assuming linear quadratic Bayes risk with the entire weight given to deviation of y_t from its target, he solved for the optimal learning policies as functions of beliefs for several small values of the planning horizon (up to 6). The results showed little difference between the myopic and optimal policies except under very large parameter uncertainty. His numerical study also showed the myopic policy to be superior to the certainty equivalent policy in approximating the optimal policy. Computing optimal solution was not difficult even for fledgling computer technology of the era.

5. Certainty Equivalence Research Continued

In the meantime, the certainty equivalence alternative continued to generate insightful results, largely due to its tractability. We mention detailed study of Pindyck (1973a,b) on the 10-equation macroeconomic model for the U.S., Abel (1975) on application to "monetarist-fiscalist" debate, and Fair (1978) on the assessment of macroeconomic performance of U.S. presidential administrations. Taylor (1974) demonstrated asymptotic unbiasedness and efficiency of least-squares certainty equivalent rules in the Prescott's (1972) model. Taylor (1974) also indicates analytically, albeit for a rather simple model, why the certainty equivalence method does about as well as cautionary myopic and dual methods. Taylor's results were confirmed experimentally by Anderson and Taylor (1976) in a more general setting without lagged dependent variables. The certainty equivalent had also shown some improvement over cautionary myopic approach, especially in circumstances where some disregard of caution could be beneficial. In particular, Åström (1983), extending Prescott's analysis to include an autoregressive term, $x_t = \beta u_t + x_{t-1} + \epsilon_t$, showed that the optimal policy can take relatively large and irregular control actions u_t to probe the system when the Bayes estimate μ_{t-1} of β has poor precision. He also found that the optimal solution is well approximated by the certainty equivalent rule $u_t = -x_{t-1}/\mu_{t-1}$. The burden of computation mounted as it took over a week to solve two-dimensional backward induction recursion for $T = 30$ periods. While the unintended probing signal injected by the certainty equivalent control in the system with autoregressive dynamics, and hence long-lived impulse response, seemed to reinforce learning, it also became clear that the certainty equivalence principle suffers from a general identifiability problem, namely the parameter estimates may converge with positive probability to a false value (Goodwin and Sin, 1984; Becker, Kumar, and Wei, 1985; Ljung, 1987; Kumar, 1990). Taylor's result only applies when a cost criterion is output variance and cannot be extended to general control laws. Indeed, a fundamental result of system identification theory is that the input signal to the process must be persistently exciting or sufficiently rich to achieve consistency in the parameter estimates (Ljung, 1987). In adaptive systems, the input signal is driven by feedback, and in this case there's no guarantee that the process will be properly excited. Moreover, the fact that not even the estimated model stabilizability is ensured may cause a paralysis in the certainty equivalent control law selection (de Larminat, 1984).

6. Limit Beliefs

Even so, both cautionary myopic and optimal dual controls may lead to incorrect limit beliefs. Conversely, the convergence of beliefs does not imply optimality in control. This curious finding was reported in Easley and Kiefer (1988). The reason is virtually the same. Convergence to the correct limit beliefs could fail, unlike in the standard consistency proofs in econometrics, because along any sample path for which beliefs converge, the sequence of actions may also be converging. If actions converge too fast, they may fail to generate enough information to identify the parameter. Examples are given in McLellan (1984); Kihlstrom, Mirman, and Postlewaite (1984). Easley and Kiefer (1988) also prove that limit beliefs are a subset of beliefs that are invariant under Bayesian learning with cautionary myopic action. In most one parameter problems the decision maker will learn the correct value of parameter because any action is informative. They also conclude that the likelihood of learning the truth is higher the more patient the decision-maker.

7. Anticipating Utility

More advanced strategies for cautionary control have also been developed, sometimes under the general rubric of *expected optimal feedback* (Kendrick and Amman, 2006). A particular multi-period extension of cautionary myopic policy could be associated with what is known elsewhere as *anticipated utility* (Quiggin, 1982; Sargent, 1993; Kreps, 1998; Sargent, 1999; Evans and Honkapohja, 2001). In this approach, at each date the decision-maker solves the dynamic program pretending that the model uncertainty is time-invariant. Ironically, this decision delivers a time-varying decision rule that depends on that date beliefs about model parameters. The irony stems from a delusion on the part of the controller of time-invariant beliefs about the model uncertainty which is invalidated every subsequent period as beliefs are, in fact, updated. The anticipated utility approach is broadly recommended by Kreps (1998) for use in games and dynamic economic models. Cogley and Sargent (2006) and Cogley, Colacito, and Sargent (2005) justify the anticipated utility models on the basis of the excellent approximation that it provides to fully rational Bayesian decisions that do anticipate that beliefs will continue to evolve going forward in time, while significantly reducing the computational complexity, at least in the permanent income economy alternating between two regimes and when the precautionary motives are muted. Whether the results extend to other settings is one of the prime stimuli for further research. Another is that anticipated utility modeling strategy is popular in much of the macroeconomic literature on learning. See, for example, Chow (1973); Craine (1979); Amman and Kendrick (1999).

Chow (1978) is one of the first to derive and use anticipating utility controls, which at the time he labeled as optimal feedback control. Chow was keenly aware that, by ignoring the possibility of reducing uncertainty through observations during the control process, the anticipated utility cost-to-go function used to derive the anticipated utility control is not what the decision maker should really expect. Chow argues that the anticipated utility control exaggerates, and thus provides an upper limit to the measure of the effect of uncertainty on the optimal control policy and the associated welfare cost. Surely, a control policy that utilizes additional observations during the control process will be able to improve the value of optimal control in the face of uncertainty. Correctly factoring the anticipation of continuous modification of the joint density of unknown parameters into the decision rule, with or without active experimentation, could only do better.

In Craine (1979), the unobserved coefficients are realizations from a serially independent stationary stochastic process. The assumption of the independent identically distributed draws of coefficients from the known joint distribution makes active component of control redundant. The anticipating utility solution is passively optimal. The physical state state equation includes autoregressive term and a control term. It is then shown that when the uncertainty about the impact of policy is dominant, the anticipating utility is relatively dormant, but when uncertainty about the transition dynamics prevails, a very active counter-cyclical policy results. The divergence of outcomes depending on which transmission channel contains more uncertainty raises an interesting question of whether it is preserved in the models with active experimentation motive.

Relative to the certainty equivalent benchmark, Amman and Kendrick (1999) find that the potential damage from ignoring variances and covariances of parameter estimates could be substantial. Their simple example contrasts the certainty equivalent solution with the anticipated utility control. The certainty equivalent control, being more impulsive, outperforms in the thin majority of simulated dynamic paths. Yet there are some dynamic paths along which faulty parameter estimates lead the certainty equivalent solution seriously astray. It is then prudent to guard against poor initial estimates by factoring the degree of uncertainty about parameter estimates into a policy calculations. Terminology 'anticipated utility' is somewhat recent and is not widely accepted. Rausser (1978) classified anticipating utility approach as sequential stochastic control, while Chow (1975) had referred to this approach as control without learning. 'Sequential stochastic control' label is justified on the grounds that observations will indeed be available in the future but will not be used to adapt the probability distributions on the parameters. Rausser distinguishes this approach from the more general open loop feedback class because the unknown parameters are treated as independent identically distributed random variables for each period of the planning horizon. In contrast, the open loop feedback rule does not necessarily restrict the joint distribution of the future states and parameters. Open loop feedback may therefore encounter the laborious task of determining the expectation of the product of the parameters and the states under multiplicative uncertainty and autoregressive state dependence. Tse and Athans (1972), for example, sidestep the problem by assuming that the parameter multiplying the state is non-random, while Ku and Athans (1973) assume that the expected product of the parameter and the state is equal to the product of the expectations.[1]

Cogley, Colacito, and Sargent (2005) is a recent application of the anticipating utility to the debate on macroeconomic stabilization and learning about the inflation-unemployment tradeoff. Their policy-makers's prior probability over two drastically different prescriptions of the inflation-unemployment tradeoff is one part of his state vector, with the lag of target state variable (unemployment) being the other. For two models calibrated to U.S. data through the early 1960s, they isolate the difference the anticipated utility cost-to-go and dynamically optimal cost-to-go with active experimentation. The discrepancy is ascribed to the benefit of experimentation and is shown to be small. In this setup, the passive anticipated utility learner observes enough variation in the data that he is able to discriminate between the two models almost as fast as the active experimenter. In our view, Cogley, Colacito, and Sargent (2005) also addresses a common criticism of adaptive control that significant parameter uncertainty may indicate not just the potential value of an adaptive control scheme but that the model may be misspecified. Indeed, the decision-maker's experimentation in Cogley, Colacito, and Sargent (2005) is directed not towards learning the parameters of the two competing models but toward the discrimination between them. A follow-up study Cogley, Colacito, Hansen, and Sargent (2008) explores how concerns for robustness alter the costs and benefits of experimentation, and find examples where robustness can either enhance or impede active experimentation.

A form of suboptimal control that is intermediate between the dual optimal control and anticipating utility control could be obtained if we assume that *some* but not all of the measurements will in fact be taken in the future, and the remaining measurements will not be taken. This method allows one to deal with those measurements that are troublesome

[1]In problems with many future periods, the divergence of the two due to the missing covariance term can seriously skew the prediction. This is because higher order moments of the joint normal distribution grow progressively larger.

and complicate the solution. For example, in the limited lookahead framework, losses that are far into the future could be evaluated under beliefs that are anticipated to persist at some intermediate date, not necessarily immediately preceding the distant date. Doing so can help economize the number of anticipated posterior distributions to keep the track of. Bertsekas (2005) refers to this concept as *partial open loop feedback control.* We'll call it *partially optimal anticipating utility policy.*

8. Markov Jump-Linear Quadratic Model

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. MJLQ models have been widely studied in the engineering control literature and found some economic applications (Costa and Fragoso, 1995; do Val, Geromel, and Costa, 1998; do Val and Başar, 1999). 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. Just like in the anticipated utility case under multiplicative uncertainty in the linear quadratic Gaussian framework, the value function stays quadratic in the physical state, but now with coefficients that depend on the regime. These coefficients satisfy coupled Riccati recursions and could be solved quickly by doubling algorithm after uncoupling (Hansen and Sargent, 2004; do Val, Geromel, and Costa, 1998). If regimes are not observable, the optimal policy will depend on the probability distribution over possible regimes. In this case, the value function of the expected optimal feedback remains quadratic in the physical state, but with weights that depend on the probability distribution over potential regimes. Solution for the entire simplex of regime probabilities would require function approximation methods, but for any particular probability distribution over regimes, the MJLQ solution could be computed easily by using Riccati recursions over receding finite horizon control. MJLQ framework appeared in some recent economic applications. Zampolli (2005) uses an MJLQ model to investigate monetary policy under transitions between regimes with and without asset-market bubble. Blake and Zampolli (2005) extend the model with the inclusion of the forward-looking variables and discretion equilibria. do Val and Başar (1999) study macroeconomic stabilization in MJLQ framework with unobserved regimes. Svensson and Williams (2005) combine the two and study equilibria with commitment in timeless perspective of Svensson and Woodford (2005). All these applications abstract entirely from the probing component. Svensson and Willams (2006) is the first laudable study of active learning and experimentation in a relatively small-scale MJLQ framework that also includes forward-looking variables. Results of their study indicate that while the benefits of learning are of first order of importance, the gains from active experimentation are of secondary importance or even insignificant. A new feature of the tradeoff between experimentation and control that is absent in the backward-looking case is a subtle interaction with the forward-looking commitment constraint. With negative Lagrange multiplier on the forward-looking constraint there is larger loss penalty for observed state and control in some outlying regions with large probing component. In those regions, strong desire to experiment is dampened. Yet in other regions of high experimentation motive the Lagrange multiplier on the forward-looking constraint is positive, and so the probing effect is reinforced. As a result, the policies acquire notable asymmetry. In economic terms, the pivotal concerns are not solely emending inference versus injecting destabilizing impulse, but also upsetting the expectations of future variables. Resulting balance amongst these factors as embodied in the dual optimal policy could be rather gordian.

MJLQ approach is ideally suited to the cases when unknown coefficients change over time. In these cases, the naïve anticipating utility control is more ignorant, and likely to concede further advantage to the fully optimal policy. Models with continuously adapting drifting coefficients such random walk or ARMA processes are not tractable analytically because of continuous dependence of the optimal policy on the probability distribution over possible continuum of regimes. However, they can be well approximated by the discrete finite regime Markov chain with the optimal solution admitting explicit form.

9. Coefficient Augmentation and Taylor Series Expansions for Dual Control

So called coefficient-augmented closed loop policy of Kendrick and Majors (1974) is worth mentioning. For one, the paper tackled more difficult case where unknown coefficient drifted over time. More important novelty was that they augmented the physical state vector with the vector of random coefficients, linearized the resulting problem, and obtained a feedback rule for the augmented problem. Consequently, an adaptive feedback solution of this kind made policy a function of both the coefficients and the usual state variables. Since the coefficients are not observed but are instead estimated an expectation of the adaptive feedback policy can be derived and used. The new policy does not coincide with the certainty equivalent control and there could be substantial difference between the two. To be sure, certainty equivalent control treats uncertainty as if it were an additive noise and thus does not take an explicit account of the source of uncertainty. The augmented closed loop rule takes account of the source of uncertainty through an approximation scheme. One of the consequences is that the variance of random parameter innovations matters.

The use of Taylor series expansions of the cost-to-go functions introduces non-trivial choice of the path to linearize around. Kendrick and Majors (1974) were among the first to appreciate this situation. Is it better to linearize around the certainty equivalent solution or around the desired path $x_t = x^*$? Kendrick and Majors (1974), on the basis of a small number of Monte Carlo runs, find that if one linearizes and penalizes about the desired path instead of the certainty equivalent solution, the quality of the linear approximation may not be as good but the adherence to the desired path, which is the ultimate goal, would usually be better. As a result, they recommend that in any procedure involving linearization one should experiment with various candidate paths, with the choice depending on the degree of nonlinearity in the model, the degree of separation between candidate loci of the expansion as well as on the particular set of Monte Carlo runs. Alternatively, a more comprehensive approach that is suggested is to choose the nominal path, about which one linearizes, as part of the optimization procedure. Denham (1964) shows how this can be done. In the context of synthesizing nonlinear control, the optimal choice of the nominal trajectory to minimize state-variable estimation errors propagating along the nominal trajectory, is known as *trajectory shaping* (Van der Stoep, 1968).

While instructive in its own right, Kendrick and Majors (1974) is an important prelude to the use of the second order expansions for the dual control in a series of papers by David Kendrick (Kendrick, 1978, 1979, 1982; Amman and Kendrick, 1999; Kendrick, 2002). They could also be traced back to the contributions by Tse, Bar-Shalom, and Meier (1973); Bar-Shalom, Tse, and Larson (1974); Bar-Shalom and Tse (1976) to the engineering literature. In these papers, the state is also augmented to include uncertain multiplicative coefficients. In addition, the complete adaptive dual control problem is decomposed explicitly into four components – current control, future deterministic control, future cautionary control and future probing control. The last two components comprise future *perturbation* control. The key distinguishing feature of many approximate actively adaptive schemes is how they handle the dependence of future information on present control. In this regard, it should be noted that the augmentation of the state results in the information state (the joint distribution of states and parameters) to be either infinite-dimensional or grow polynomially with the planning horizon. Tse, Bar-Shalom, and Meier (1973) deal with this quandary by maintaining only the first two moments of augmented state estimate. These updated estimates can be computed by any one of the number of methods including the second order extended Kalman filter. The optimal cost-to-go function also undergoes second order expansion in the augmented state around some tentative nominal trajectory (typically trajectory that is expected to occur under the certainty equivalent control). Bar-Shalom and Tse (1976) found that the more sophisticated actively adaptive schemes did not always perform better than the open loop feedback or even the certainty equivalent rule. Norman's 1976 results, using a first-order approximation, also indicate that the performance of alternative schemes is problem-specific. When computational cost is explicitly considered, Norman (1976) isolates some special cases for which the anticipated utility as well as the certainty equivalent rules appear to be the most desirable schemes. More generally, Norman and Jung (1977); Norman (1981, 1994) study the computational complexity of alternative dual control formulations for linear quadratic models, possibly with long lags, and possibly structural. Amman and Kendrick (1994, 1997) is a Monte Carlo based comparative study of second order linearized dual control versus anticipated utility (under the label of optimal feedback with parameter uncertainty) on a set of small macroeconometric models. Dual control performance was found to prevail in most models.

A related actively adaptive control procedure has also been advanced by Chow (1975) for the case of a quadratic but non-additive criteria function, and linear or nonlinear state evolution equation. Chow's method applies a second order Taylor series expansion in perfectly measured states to the relevant cost-to-go function around exogenously chosen nominal path using numerical differentiation. Chow's method differs from that of Tse, Bar-Shalom, and Meier (1973) and Kendrick (2002) in some important respects. First, there's no augmentation of unknown coefficients to the state vector. Second, Chow first takes the expectation of the cost-to-go function and then applies the second order approximation, whilst Tse, Bar-Shalom, and Meier (1973) reverse these steps. As a result, it is no longer possible to isolate the costs of perturbation control explicitly.

Another linearization approach is possible in the model with linear evolution equation with unknown but constant coefficients, quadratic criterion and in the absence of the measurement noise (MacRae, 1972, 1975). Given such specification, the conditional mean and covariance of the parameter estimates completely characterize the relevant information state. MacRae simplifies the problem by replacing unknown future realizations by their conditional means in the covariance update equation, while using martingale property of conditional expectations to nullify the expected mean dynamics. This results in the deterministic covariance update equation which is introduced into the cost-to-go function as a deterministic constraint along with an associated matrix of Lagrange multipliers. The augmented criterion function remains quadratic in control and can easily be computed analytically given its coefficients. However, the derivation of these coefficients requires solving two-point boundary value problem in order to satisfy the constraint on the covariance evolution. Essentially similar suggestion is given by Pronzato, Kulcsár, and Walter (1996) and Kulcsár, Pronzato, and Walter (1996) relying on numerical minimization and short lookahead horizons. Another variation here could be to reduce the dimensionality of the optimization problem by setting controls that are further away to either predetermined values or to outcomes of simpler rules such as cautionary myopic, anticipating utility, etc. This way learning will be anticipated but not fully exploited for the current decision-making.

10. Limited Lookahead Approaches

In the face of intensifying computational difficulties in approaching the dual optimal control and certain inadequacies of the non-dual solutions, the search for simpler solutions with both cautious and probing features continued in the 1980s. A natural next step was to try to solve the two-period dynamic program. Bar-Shalom, Mookerjee, and Molusis (1983) presented adaptive dual controller for a multi-input multi-output autoregressive moving average system by solving the corresponding two-period dynamic programming problem. Reportedly, two-step control showed good performance. In particular, it avoided the turn-off malady and accelerated the convergence of learning process. A two-step-ahead cost-to-go function is also considered in Sternby (1977). The two-step problem gives useful clues how to make sensible approximations that retain the dual features. It may be useful for the infinite-horizon dynamic programming formulations by mapping the regions with high curvature of the cost-to-go function which is exactly where the probing component is most active and where the discontinuities in control are most likely to appear. The simple limited lookahead examples give some useful indications how suboptimal dual controllers could be constructed. For example, to illustrate how the optimal dual control can switch between probing and regulating consider hypothetical figure 1. In the figure, the expected loss function is given for three possible values of a scalar state, and has several minima. For the dashed curve local minimum to the left gives the absolute minimum, while for the full line case the two local minima have the same value. Finally, for the dash-dotted curve the local minimum to the right represents the global minimum. The control action will thus switch in character when state is variable changes. This can be interpreted as that the control action is switching between probing and control intentions.

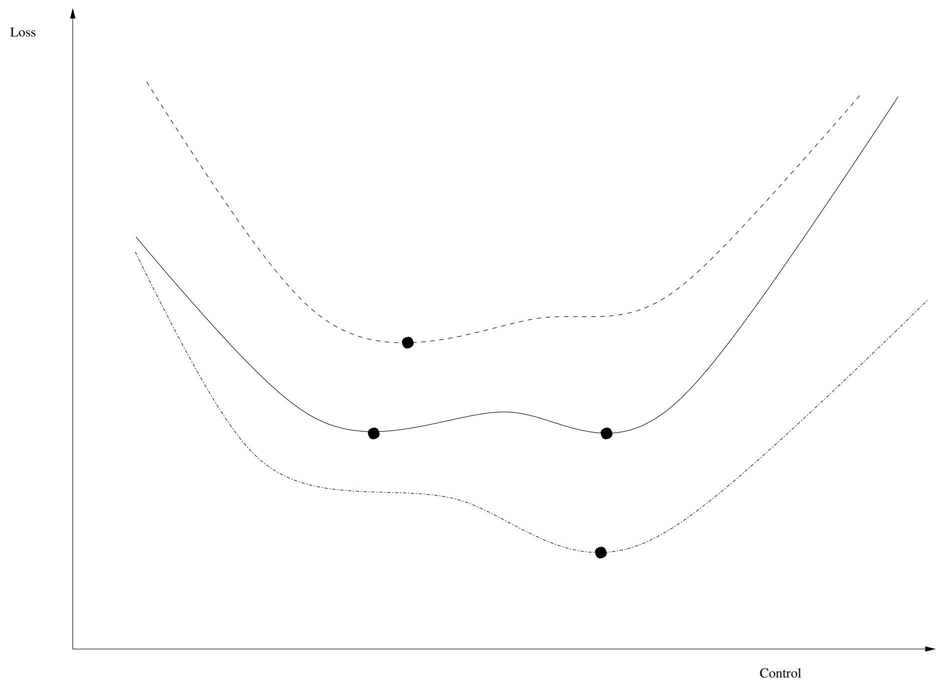

Figure 1: Possible shape of the expected loss function as a function of control signal for three close-by different values of state variable. The global minima for the three cases are marked by dots.

The multi-step limited lookahead approaches can be also developed. The minimization over several steps to obtain limited lookahead dual controller makes it possible to introduce probing in the beginning or when the information about the process is poor and still gaining by being able to make a better control towards the end of control horizon. From a dynamic programming perspective, all that is required is to execute a limited number of cost-to-go updates instead of iterating to convergence. Since in the dynamic programming approach, the optimal cost-to-go function is implicit, curse of dimensionality in storage requirements is not lifted, even though the computing time is reduced by orders of magnitude.

A key issue in implementing a limited lookahead policy is the selection of the cost-to-go approximation at the final step. Limited lookahead policy with zero terminal cost is simply the solution to the finite horizon dynamic subprogram within the infinite horizon problem but it could still be difficult to obtain. One possibility is to trade off accuracy of the cost-to-go approximation with lookahead horizon. In other words, by looking additional periods ahead, one can mitigate the effect of errors in the cost-to-go approximation (Bertsekas, 2005). Cost-to-go approximations integral to the limited lookahead calculations could be arrived at in a variety of ways. For example, one can reduce the amount of available information in the evaluation of the cost-to-go function. One way to achieve this is to assume that no further information gathering will be available along the nominal path of beliefs (Pronzato, Kulcsár, and Walter, 1996; Kulcsár, Pronzato, and Walter, 1996; MacRae, 1972, 1975). Putting it differently, the future trajectory of beliefs is explored as if no future shocks perturb the system while the future control does affect parameter uncertainty. The approach could be viewed as an instance of *preposterior analysis* (Raiffa and Schlaifer, 1961), where the value of information is anticipated by using statistics of future measurements via algorithms based on the notion of closed loop control. Such control schemes involve an experimental dimension that probes the system in anticipation of the value of information to be derived from future observations. Pronzato, Kulcsár, and Walter (1996) and Kulcsár, Pronzato, and Walter (1996) motivate the approach from the optimal experimental design perspective as it strives to exploit the future information by controlling the expected Bayesian Information Criterion matrix. Birmiwal and Bar-Shalom (1985) use preposterior analysis to derive approximate prior probability densities to describe future learning, and, hence, approximate Bellman functional equation by evaluating the value of future information gathering. In cases where autoregressive state coefficients are known, if at all present, this approach amounts to propagating forward the *deterministic* controlled path of belief variance, with possible further simplifications as desired, and then solving simplified Bellman functional equation. The minimization step could incorporate the expected variance dynamics by the direct substitution as in Pronzato, Kulcsár, and Walter (1996); Kulcsár, Pronzato, and Walter (1996) or by attaching Lagrange multipliers to constraints on the variance dynamics as in MacRae (1972, 1975). If autoregressive coefficients are themselves uncertain, the evolution of variance is no longer deterministic. In this case, a Taylor series expansion of beliefs around the nominal path could be used. MacRae (1972, 1975) find that for the unknown autoregressive parameter case, it is no longer necessarily the case that adaptive covariance policy is rather inactive when uncertainty is large and compensates with aggressive policy action once the effect of policy action is known with more precision. As parameters associated with current controls and with lagged observed states can be arbitrarily correlated, no general implications about the relative control intensity can be drawn. For multi-coefficient uncertainty specification such as this, although larger variances imply more uncertainty about model parameters, large correlation imply more information, and larger incentive to experiment. Compared to the model of Wieland (2000a), which also features bivariate parameter uncertainty, correlation is even more important, because incorrect beliefs about persistence can have lasting repercussions.

Incorporation of the predictive evolution of belief variance belongs to a more general class of cost-to-go approximations that involves a heuristic embracements of certain features. The manner in which the cost-to-go approximation is selected is very much problem dependent. In some problems good heuristics come from a cost-to-go approximation based on the solution of a simpler problem that is tractable computationally or analytically (Bertsekas, 2005). For example, one can base a cost-to-go approximation on a suboptimal policy such as certainty equivalent or anticipating utility policy. Generally, any reasonably good suboptimal policy can be used to produce a cost-to-go approximation. For the limited lookahead context in an infinite horizon problem, the cost-to-go function of the suboptimal solution can be used to improve the terminal cost-to-go representation. The hybridization approach could be iterated provided it is not too expensive to do so. These ideas have been successfully applied in the industrial engineering (Kimemia, Gershwin, and Bertsekas, 1982; Kimemia, 1982; Tsitsiklis, 1984). No economic applications has used these hybrids yet, to our knowledge.

The idea of the limited lookahead horizon need not be approached from the recursive viewpoint of Bellman equation. For short horizons, it could be dealt with directly, giving up the convenience of temporal separation of controls. Direct formulation results in multivariate optimization of the performance index with controls for all time-periods decided at time zero even though the decisions are acted upon one at a time. When the next period rolls in, the problem is re-optimized again. The direct route is taken by Maitelli and Yoneyama (1994); Lindoff and Holst (1997a,b). While the solution is often not explicit, the advantage is that root-finding and minimization with moderately large number of variables is not subject to the curse of dimensionality.[2] The obvious disadvantage is cumbersome calculations involved in computing higher order moments of the uncertain parameters, or multiple passes of Kalman smoother recursions. If desired, laborious algebra could be further cut down by some kind of approximations while the retaining the dual feature.[3] The study of Maitelli and Yoneyama (1994) does exactly that by approximating high order cross-moments with products of lower order ones. In terms of performance, the simulation study of Lindoff and Holst (1997b) suggests that cautionary myopic policy is able to control only low order systems with slow varying parameters. For higher order systems or faster time-variations, the cautionary myopic policy has the tendency to be locally unstable. An approximate two-step predictive controller of Maitelli and Yoneyama (1994), active suboptimal dual controller (ASOD) of Wittenmark and Elevitch (1985) and exact two-step predictive control of Lindoff and Holst (1997a) are shown to be superior, especially for higher order systems or faster time-variations. Of the three, the exact two-step predictive control is often the best, despite also having the largest variance of the parameter governing the slope of policy response. Indubitably, the main goal is to control the system so the resulting excessive variance of parameter estimates is of minor subsidiary importance as long as the control performance is good. In comparison, the approximate two-step control has a propensity to concentrate on estimating the unknown parameters at the expense of controlling the output. The two-step limited lookahead control was also more robust against mis-specifications in the simulated linear models. The ASOD control of Wittenmark and Elevitch (1985), that we just mentioned, is a different kind of extension to the cautionary myopic control. Its criterion function includes ad hoc term that rewards good parameter estimates. It works well for low order systems but may have local instability issue for more complex systems. Another drawback is the arbitrariness in the weight given to the reward term. The weight must be chosen correctly to obtain good performance, but it is not known how to do so, in general.

11. Dynamic Programming Revisited

Wieland (2000a) revisited the dynamic programming approach in a more complex model featuring bivariate parameter uncertainty – about the intercept and slope of the regression. The result of Prescott on the strong similarity between cautionary myopic and optimal policies is reversed, with sizeable differences detected between the two detected under a non-negligible range of initial beliefs, or in other words a sizeable extent of optimal experimentation. This results accords with the intuition of Kendrick (1979, 1982) predicting larger role for the experimentation with more sources of uncertainty.[4] Additionally, Wieland

[2]It could be shown that complexity of multivariate minimization and root-finding grows polynomially in the number of variables if the objective is mostly smooth.

[3]Rausser (1978) argues that most approximations performed on the original system will lead to the passively adaptive schemes, while those performed in the process of deriving the optimal rule will generally be actively adaptive.

[4]On the other hand, larger uncertainty for a given number of sources does not necessarily imply more experimentation. For example, Kendrick (1982) finds that active probing is most active for the moderate magnitude of parameter uncertainty. Beyond that range, active experimentation is too costly as the system dynamics will eliminate most of it anyway, in due course. At the other extreme, when the parameters are almost certain, active experimentation diminishes due to the lack of benefit.

(2000a) explored possible limit beliefs that could be reinforced by uninformative actions. The optimal policy typically exhibits a discontinuity at such beliefs and therefore avoids uninformative actions that would render incorrect beliefs self-reinforcing. In contrast, myopic behavior, could lead to persistent and frequent bias in beliefs and actions.

Beck and Wieland (2002) develop a version with single unknown impact of the policy action but with known autoregressive dynamics in the observed target state. The focus is on the uncertainty regarding a parameters that is multiplicative to the decision variable is motivated by the crucial nature of the tradeoff between current control and estimation. In addition, the unknown coefficient drifts over time, following univariate random walk with known innovation variance. In this framework, learning never ceases, but instead beliefs continually track the latent state. Again, noticeable levels of experimentation are detected for the moderate to large levels of uncertainty. Except near the deterministic steady state, the optimal decision rule remains less activist than a certainty equivalent rule and induces gradualism. Gradualism disappears in the vicinity of the deterministic steady state, where aggressive experiments are repeatedly undertaken. The extent of experimentation diminishes with the degree of parameter variation. The justification lies in the lower expected payoff to probing and obtaining more precise parameter estimates when the parameter is going to elude precise estimation all over again. At the same time, steady-state fluctuations are tolerated because they provide information about the unknown time-varying parameter in perpetuity. As learning doesn't converge, it never ceases. The possibility of significant parameter uncertainty (current or future) cannot be ruled out at any time. Costly experiments, especially in the neighborhood of the deterministic steady state are a device to guard against such unpleasant contingency.

12. Finite Information State

When the parameter set consists of a finite number of values, it is easier to solve the dual control numerically because taking expectation in the Bellman equation amounts to summation. Casiello and Loparo (1989) apply preposterior analysis to derive approximate optimal dual control. Bernhardsson (1989) solves the case where the gain of a first-order system may take only two values.

13. Multi-Armed Bandit Problems

A version of the finite action set dual control problem is known as multi-armed bandit problem (Robbins, 1952; Gittins, 1989; Whittle, 1982) whose solution is well understood and developed. In the multi-armed bandit problem, the gambler has to decide which arm of several different slot machines to play to maximize discounted expected reward in a sequence of trials. The distinguishing feature of the bandit problems and the key to its tractability is that the distribution of returns from one arm only changes when that arm is chosen. The index theorem due to Gittins and Jones (1974) transforms the problem of finding the optimal policy into a finite collection of stopping problems, one for each arm. The idea is to find for each arm the stopping time that results in the highest discounted expected return per discounted expected number of periods in operation from only playing this arm. The Gittins index is the resultant highest discounted expected return, and the optimal policy is to select in each period the arm with the highest Gittins index. The simplicity of the solution is fragile and approximations are needed for many extensions. For example, Brezzia and Lai (2002) develop an approximation to Gittins indices based on a numerical solution of an optimal stopping problem for a limiting diffusion which is applicable even in cases where the optimal solution does not reduce to an index, e.g., in finite horizon problems or problems with switching costs. Making use of these approximations to the optimal policy, Brezzia and Lai analyze the value of experimentation and show that unless the horizon or the discount factor are large enough, experimentation does not have much value compared to the simple myopic policy. Multi-armed bandits have been generalized to the continuous time setting (Kaspi and Mandelbaum, 1998). Further, non-stationary bandit problem where the agent is faced with the increased complexity of detecting changes in its environment have been examined as well. For example, Koulouriotis and Xanthopoulos (2008) study a non-stationary, discrete-time, finite horizon bandit problem with a finite number of arms and Gaussian rewards and apply a family of suitably chosen ad hoc learning algorithms that offer intuition-based solutions to the exploitationexploration trade-off and so have the advantage of not relying on strong theoretical assumptions while in the same time can be fine-tuned in order to produce near-optimal results. In particular, they present an evolutionary algorithm that was implemented to solve the non-stationary bandit problem along with ad hoc solution algorithms, namely action-value methods with e-greedy and softmax action selection rules, the probability matching method and finally the adaptive pursuit method. Positive probability of incomplete learning in bandit problems is ascertained in Rothschild (1974b); Sundaram (1992) and Brezzi and Lai (2000) as an outcome of existence of a set with positive probability on which an inferior arm can be played forever once it is chosen. Bergemann and Välimäki (2006) survey the literature on multi-armed bandit models, various extensions and applications in economics.

14. Loose Ends

The vast array of suboptimal solutions listed so far is akin to the expansive and perplexing 'wilderness of bounded rationality' (Ireland, 2003) and it does not end here. There are many more approximating ideas and we'll only provide short, perhaps inadequate glimpse at the rest of them. Just as learning models themselves allow agents to entertain *diverse* and *approximate* models, these approximate dual control ideas let us learn when and how to offer practical advice to the policy maker with an interest in the macroeconomic stabilization under model uncertainty. Ad hoc modifications of the loss function is used in Wittenmark and Elevitch (1985). Milito, Padilla, Padilla, and Cadorin (1982) add squared prediction error into the quadratic loss function. A similar approach is used in Yame (1987) from the information theoretic perspective. Alter and Bélanger (1974) introduce constraint on the trace of one step ahead information matrix. Hughes and Jacobs (1974) propose a limitation on the minimum value of control input. Chan and Zarrop (1985) include the variance of an auxiliary output and its prediction error as additional penalty terms in the loss function. Padilla, Padilla, and Cruz Jr. (1980) supplement the loss function with two sensitivity functions designed to capture parameter uncertainty. Cosimano (2003) applies first-order perturbation method in the vicinity of the deterministic augmented linear regulator. This allows to express approximate optimal solution as a combination of the certainty equivalent solution and a term that captures the impact of uncertainty on the decision-maker's value function. Gapen and Cosimano (2005) compare the costs and benefits of this approach relative to the dynamic programming solution in the model economy of Wieland (2000a). Filatov and Unbenhauen (2004) use bi-criterial optimization to devise dual control policies. Han, Lai, and Spivakovsky (2006) enhance one period cautionary myopic control with the use of policy rollout, a Monte Carlo simulation based technique for the evaluation of the performance of a given policy using only states actually visited in simulation. Sarris and Athans (1973) propose two-step adaptive control that takes into account future adaptation of conditional means but not the variance of the unknown parameters. If the underlying distribution is not Gaussian, Alspach (1972) proposed mixture of normals approximation within limited lookahead framework to obtain a solution scheme. Various advanced dynamic programming techniques could also be redirected for use in economic problems. The techniques include temporal difference methods (Sutton, 1984), approximate and optimistic policy iterations (Bertsekas and Tsitsiklis, 1996), Q-learning (Watkins, 1989), sequential approximation in the state space (Bertsekas and Tsitsiklis, 1996), value iteration with state aggregation (Tsitsiklis and Van Roy, 1996), value iteration with representative states (Tsitsiklis and Van Roy, 1996), Bellman error methods (Schweitzer and Seidman, 1985), approximate linear programming (Trick and Zin, 1993, 1997), linear programming with constraint sampling (de Farias and Van Roy, 2004), etc. Some results on the Bayesian control of Markov chains are discussed in Kumar (1985) alongside more general exposé of stochastic adaptive control ideas.

Most of work on active learning studies the single-agent case, but there are modest literatures about multi-agent (game-theoretical) learning in both non-equilibrium setting (Fudenberg and Levine, 1993; Fudenberg and Kreps, 1996; Dubei and Haimanko, 2004; Jehiel and Samet, 2005) and in equilibrium (Bolton and Harris, 1999; Aghion, Bolton, Harris, and Jullien, 1991; Bergemann and Välimäki, 1996; Dekel, Fudenberg, and Levine, 2004; Keller, Rady, and Cripps, 2005).

Non-Bayesian adaptive control that allows more flexibility in the *design* of control laws has also been explored. See Kumar (1985) for an early survey.

15. Applications

The economic applications with passively or actively adaptive control and purposeful experimentation are far too numerous to survey exhaustively. Some of the work is reviewed in early review of Rausser (1978) and recent keynote address to the Society of Computational Economics by Kendrick (2005). With some overlap with the above reviews, we will recount most notable works.

Early applications of dual control include the work of Rausser and Freebairn (1974a,b) on agricultural trade policy, Rothschild (1974a) on monopolistic pricing with unknown demand, Chong and Cheng (1975) on pricing strategies for the introduction of a new product, Rausser and Hochman (1978) in the context of commodity-marketing boards with possibility of inventory accumulation, Pekelman and Tse (1980) on advertising. Along with these applications, Abel (1975) examined small macroeconomic model for the purposes of 'monetarist-fiscalist' debate, Kendrick (1979) explored a small scale macroeconomic model with measurement error while Bertocchi (1993) built a theory of floating public debt issues using "subscription issues" when demand schedule for bonds is unknown. Bertocchi and Spagat (1993) appraised the optimal monetary policy with experimentation when coefficient uncertainty is confined to the choice of the two possible values. Balvers and Cosimano (1990) study the dynamics of price adjustment in the context of the active learning about the demand schedule. Trefler (1993) considers monopolistic pricing and output decisions in more general framework with endogenous demand information. Datta, Mirman, and Schlee (2000) apply active learning to the model of retail clearance sales with signal dependence and noiseless information. Yetman (2000) constructs a model of monetary policy with uncertainty about the level of potential output and examines the relationship between credibility and the optimal probing when planning horizon is finite (up to ten periods). He finds that only for low levels of credibility or unrealistically large levels of uncertainty or volatility does the optimal policy with probing diverge significantly from a policy that ignores learning, and that the optimal amount of probing diminishes as credibility rises. Related paper where unknown parameter is related to the credibility of the regime change in the private sector inflation expectations is Tesfaselassie and Schaling (2007). The twist is the timing protocol that makes inflationary expectations dependent on past policy decisions. The disinflation under actively optimal policy is more than under cautionary policy but less than under the certainty equivalence, irrespective of the initial level of inflation. Moscarini and Smith (2001), in a continuous time setting of control of variance of a diffusion with uncertain two-state mean, show monotonicity of optimal experimentation level with Bellman value and beliefs when the model is interpreted in R&D context. Hong and Rady (2002) introduce experimentation in an asset pricing model with uncertain supply of liquidity, where strategic seller can infer about liquidity from past prices and trading volume. Cosimano, Emmons, Lee, and Sheehan (2002) study the dynamics of loan and deposit rate adjustment when large financial institutions actively learn about the demand for loans and the supply of deposits. Marcoul and Weninger (2004) study dynamic fishing site choice when unknown abundance is correlated across fishing sites and find empirically that mid-Atlantic surf clam fishing vessel's skippers follow site choice patterns that are consistent with a model of rational search and information acquisition. Ellison, Sarno, and Vilmunen (2006) examine the optimal active learning policy in an open-economy macro model and compare coordinated and uncoordinated equilibria.

There is now also a solid strand of literature on structural estimation of models of learning and experimentation, typically in the context of consumer choice. This line of research was pioneered by Erdem and Keane (1996) on the panel of individual purchases of liquid laundry detergents under certain homogeneity assumptions, including price sensitivity. As pointed out by Osborne (2006), if this assumption is violated, the pricing dynamics following product introductions would be attributed entirely to learning even though there could be no learning in the underlying data generating process. It is then important to match features of the data with identifying assumptions about sources of heterogeneity and differential learning. Crawford and Shum (2005) and Ackerberg (2003) explore richer sources of heterogeneity while not accounting for any types of dynamics that are not learning. Osborne (2006) develops a combined structural framework with consumer learning, switching costs and heterogeneity.

16. Concluding Remarks

Despite much work already done, there's much more to be done both in algorithm development and practical applications. Kumar (1985) lamented that efficient computational methods or analytic solutions to new problems are still needed; that in the area of dual control of linear quadratic Gaussian systems, one needs approximations for which rigorous bounds on the quality of the approximations are available as well as studies of rates of convergence. These issues are still problematic today. In terms of economic applications, especially macroeconomic policy analysis and aggregate demand management under the rational expectations, the area that is dear to us, we would argue that it is particularly important to take account of forward-looking variables such as inflationary expectations, term premia, etc. Unfortunately, combination of active learning with forward-looking variables is very difficult to study as even rudimentary macro models with forward-looking variables proliferate state variables all too easily. Furthermore, estimation of models with forward-looking variables requires that policy rule is pre-specified. Resolution of this simultaneity via iterative guess-verify approach forces one to re-optimize dynamic program multiple times. To our knowledge, no attempt has been made in the literature yet to solve this important problem.

References

Abel, A. B. (1975): "A comparison of three control algorithms to the Monetarist-Fiscalist debate," *Annals of Economic and Social Measurement*, 4, 239–252.

Ackerberg, D. (2003): "Advertising, Learning, and Consumer Choice in Experience Goods Markets: A Structural Empirical Examination," *International Economic Review*, 44(3), 1007–1040.

Aghion, P., P. Bolton, C. Harris, and B. Jullien (1991): "Optimal Learning by Experimentation," *Review of Economic Studies*, 58, 621–654.

Alspach, D. (1972): "Dual Control Based on Approximate A Posteriori Density Functions," *IEEE Transactions on Automatic Control*, AC-17(5), 689–693.

Alter, J., and P. R. Bélanger (1974): "A technique for dual adaptive control," *Automatica*, 10(6), 627–634.

Amman, H., and D. A. Kendrick (1994): "Active Learning: Monte Carlo Results," *Journal of Economic Dynamics and Control*, 18(1), 119–124.

Amman, H. M., and D. A. Kendrick (1997): "Erratum to active learning: Monte Carlo results," *Journal of Economic Dynamics and Control*, 21(10), 1613–1614.

——— (1999): "Should Macroeconomic Policy Makers Consider Parameter Covariances?," *Computational Economics*, 14(3), 263–267.

Anderson, T. W., and J. B. Taylor (1976): "Some Experimental Results on the Statistical Properties of Least Squares Estimates in Control Problems," *Econometrica*, 44(6), 1289 – 1302.

Aoki, M. (1967): *Optimization of Stochastic Systems*. Academic Press, New York.

Åström, K. J. (1983): "Theory and Applications of Adaptive Control–A Survey," *Automatica*, 19(5), 471–486.

Åström, K. J., and B. Wittenmark (1971): "Problems of identification and control," *Journal of Mathematical Analysis and Applications*, 34(1), 90–113.

Balvers, R. J., and T. F. Cosimano (1990): "Actively Learning About Demand and the Dynamics of Price Adjustment," *Economic Journal*, 100(402), 882–898.

Bar-Shalom, Y., P. Mookerjee, and J. Molusis (1983): "A linear feedback dual controller for a class of stochastic systems," in *Développement et utilisation d'outils et modèles mathématiques en automatique, analyse de systèmes et traitement du signal*, pp. 759–774. CNRS, Paris, Belle-Ile, France.

Bar-Shalom, Y., and E. Tse (1976): "Caution, Probing and the Value of Information in the Control of Uncertain Systems," *Annals of Economic and Social Measurement*, 5, 323–336.

Bar-Shalom, Y., E. Tse, and R. E. Larson (1974): "Some Recent Advances in the Development of Closed-Loop Stochastic Control and Resource Allocation Algorithms," in *Proceedings of IFAC Symposium on Stochastic Control*, Budapest, Hungary.

Bar-Shalom, Y., and K. D. Wall (1980): "Dual adaptive control and uncertainty effects in macroeconomic systems optimization," *Automatica*, 16(2), 147–156.

Beck, G. W., and V. Wieland (2002): "Learning and control in a changing economic environment," *Journal of Economic Dynamics and Control*, 26(9-10), 1359–1377.

Becker, A., P. R. Kumar, and C. Z. Wei (1985): "Adaptive Control with Stochastic Approximation Algorithm: Geometry and Convergence," *IEEE Transactions on Automatic Control*, AC-30, 330–338.

Bergemann, D., and J. Välimäki (1996): "Learning and Strategic Pricing," *Econometrica*, 64(5), 1125–1149.

——— (2006): "Bandit Problems," Discussion Paper 1551, Cowles Foundation Discussion Paper, Yale University, New Haven, CT.

Bernhardsson, B. (1989): "Dual control of a first-order system with two possible gains," *International Journal of Adaptive Control and Signal Processing*, 3(1), 15–22.

Bertocchi, G. (1993): "Theory of Public Debt Management with Unobservable Demand," *Economic Journal*, 103(419), 960–974.

Bertocchi, G., and M. Spagat (1993): "Learning, experimentation, and monetary policy," *Journal of Monetary Economics*, 32(1), 169–178.

Bertsekas, D. P. (2005): *Dynamic Programming and Optimal Control*, vol. 1. Athena Scientific, Nashua, NH, 3 edn.

Bertsekas, D. P., and J. N. Tsitsiklis (1996): *Neuro-Dynamic Programming*. Athena Scientific, Boston, MA.

Birmiwal, K., and Y. Bar-Shalom (1985): "Dual control for identification and guidance," in *Proceedings 9th IFAC World Congress*, vol. VII, pp. 65–70, Budapest, Hungary.

Blake, A. P., and F. Zampolli (2005): "Time Consistent Policy in Markov Switching Models," Working paper, Bank of England.

Bolton, P., and C. Harris (1999): "Strategic Experimentation," *Econometrica*, 67(2), 349–374.

Brainard, W. (1967): "Uncertainty and the Effectiveness of Policy," *The American Economic Review*, 57(2), 411 – 425.

Brezzi, M., and T. L. Lai (2000): "Incomplete Learning from Endogenous Data in Dynamic Allocations," *Econometrica*, 68(6), 1511–1516.

Brezzia, M., and T. L. Lai (2002): "Optimal learning and experimentation in bandit problems," *Journal of Economic Dynamics and Control*, 27(1), 87–108.

Casiello, F., and K. A. Loparo (1989): "Optimal Control of Unknown Parameter Systems," *IEEE Transactions on Automatic Control*, AC-34(10), 1092–1094.

Chan, S. S., and M. B. Zarrop (1985): "A suboptimal dual controller for stochastic systems with unknown parameter," *International Journal of Control*, 41, 507–524.

Chong, C. Y., and D. C. Cheng (1975): "Multistage pricing under Uncertain Demand," *Annals of Economic and Social Measurement*, 4, 311–324.

Chow, G. C. (1973): "Effects of Uncertainty on Optimal Control Policies," *International Economic Review*, 14(3), 632–645.

——— (1975): *Analysis and Control of Dynamic Economic Systems*. Wiley, New York.

——— (1978): "Evaluation of Macroeconomic Policies by Stochastic Control Techniques," *International Economic Review*, 19(2), 311–319.

Cogley, T., R. Colacito, L. P. Hansen, and T. J. Sargent (2008): "Robustness and U.S. Monetary Policy Experimentation," Unpublished manuscript, University of California, Davis.

Cogley, T., R. Colacito, and T. J. Sargent (2005): "Benefits from U.S. Monetary Policy Experimentation in the Days of Samuelson and Solow and Lucas," Working paper, forthcoming in journal of money, banking and finance, New York University.

Cogley, T., and T. J. Sargent (2006): "Anticipated Utility and Rational Expectations as Approximations of Bayesian Decision Making," mimeo, University of California, Davis and New York University.

Cosimano, T. F. (2003): "Optimal Experimentation and the Perturbation Method in the Neighborhood of the Augmented Linear Regulator Problem," Unpublished manuscript, University of Notre-Dame.

Cosimano, T. F., W. Emmons, B.-J. Lee, and R. G. Sheehan (2002): "Actively Learning about Demand for Loans and the Supply of Deposits by Large Financial Institutions: The Dynamics of Loan and Deposit Rate Adjustment," Working Paper, University of Notre Dame.

Costa, O. L. V., and M. D. Fragoso (1995): "Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems," *IEEE Transactions on Automatic Control*, 40(12), 2076–2088.

Costa, O. L. V., M. D. Fragoso, and R. P. Marques (2005): *Discrete-Time Markov Jump Linear Systems*, Probability and its Applications. Springer-Verlag, London, 1 edn.

Craine, R. (1979): "Optimal Monetary Policy with Uncertainty," *Journal of Economic Dynamics and Control*, 1(1), 59–83.

Crawford, G. S., and M. Shum (2005): "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica*, 73(3), 1137–1173.

Datta, M., L. J. Mirman, and E. E. Schlee (2000): "Learning with noiseless information and payoff-relevant signals," *Economic Theory*, 16(1), 63–75.

de Farias, D. P., and B. Van Roy (2004): "On Constraint Sampling in the Linear Programming Approach to Approximate Dynamic Programming," *Mathematics of Operations Research*, 29(3), 462–478.

de Larminat, P. (1984): "On the Stabilizability Condition in Indirect Adaptive Control," *Automatica*, 32, 849–857.

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

Denham, W. F. (1964): "Choosing the Nominal Path for a Dynamic System with Random Forcing Function to Optimize Statistical Performance," Technical Report 449, Division of Engineering and Applied Physics, Harvard University.

do Val, J. B. R., and T. Başar (1999): "Receding horizon control of jump linear systems and macroeconomic policy problem," *Journal of Economic Dynamics and Control*, 23(8), 1099–1131.

do Val, J. B. R., J. C. Geromel, and O. L. V. Costa (1998): "Uncoupled Riccati iterations for the linear quadratic control problem of discrete-time Markov jump linear systems," *IEEE Transactions on Automatic Control*, 43(12), 1727–1733.

Dubei, P., and O. Haimanko (2004): "Learning with perfect information," *Games and Economic Behavior*, 46(2), 304–324.

Easley, D., and N. M. Kiefer (1988): "Controlling a Stochastic Process with Unknown Parameters," *Econometrica*, 56(5), 1045–1064.

Ellison, M., L. Sarno, and J. Vilmunen (2006): "Caution or Activism? Monetary Policy Strategies in an Open Economy," Unpublished manuscript.

Erdem, T., and M. Keane (1996): "Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets," *Marketing Science*, 15(1), 1–20.

Evans, G. W., and S. Honkapohja (2001): *Learning and Expectations in Macroeconomics*, Frontiers of Economic Research. Princeton University Press, Princeton, NJ.

Fair, R. C. (1978): "The Use of Optimal Control Techniques to Measure Economic Performance," *International Economic Review*, 19(2), 289–309.

Farison, J. B., R. E. Graham, and R. Shelton (1967): "Identification and Control of Linear Discrete Systems," *IEEE Transactions on Automatic Control*, AC-12, 438–442.

Feldbaum, A. A. (1960a): "Theory of dual control (I)," *Automation and Remote Control*, 21(9), 1240–1249.

——— (1960b): "Theory of dual control (II)," *Automation and Remote Control*, 21(11), 1453–1464.

——— (1961a): "Theory of dual control (III)," *Automation and Remote Control*, 22(1), 3–16.

——— (1961b): "Theory of dual control (IV)," *Automation and Remote Control*, 22(2), 129–143.

Filatov, N. M., and H. Unbenhauen (2004): *Adaptive Dual Control: Theory and Applications*, vol. 302 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, New York.

Fudenberg, D., and D. Kreps (1996): "Learning in Extensive Form Games, II: Experimentation and Nash Equilibrium," mimeo.

Fudenberg, D., and D. K. Levine (1993): "Steady State Learning and Nash Equilibrium," *Econometrica*, 61(3), 547–573.

Gapen, M. T., and T. F. Cosimano (2005): "Optimal Experimentation: a Comparison of the Perturbation and Dynamic Programming Algorithms," Unpublished manuscript, University of Notre-Dame.

Gittins, J. (1989): *Allocation Indices for Multi-Armed Bandits*. Wiley, London.

Gittins, J., and D. Jones (1974): "A Dynamic Allocation Index for the Sequential Allocation of Experiments," in *Progress in Statistics*, ed. by J. Gani, pp. 241–266. North-Hollnad.

Goodwin, G. C., and K. S. Sin (1984): *Adaptive Filtering and Control*. Prentice Hall, Englewood Cliffs, NJ.

Han, J., T. L. Lai, and V. Spivakovsky (2006): "Approximate Policy Optimization and Adaptive Control in Regression Models," *Computational Economics*, 27(4), 433–452.

Hansen, L. P., and T. J. Sargent (2004): "Recursive Models of Dynamic Linear Economies," Unpublished Manuscript.

Henderson, D. W., and S. J. Turnovsky (1972): "Optimal macroeconomic policy adjustment under conditions of risk," *Journal of Economic Theory*, 4(1), 58–71.

Holt, C. C. (1962): "Linear Decision Rules for Economics Stabilization and Growth," *The Quarterly Journal of Economics*, 76(1), 20–45.

Hong, H., and S. Rady (2002): "Strategic Trading and Learning about Liquidity," *Journal of Financial Markets*, 5, 419–450.

Hughes, D. J., and O. L. R. Jacobs (1974): "Turn-off, Escape and Probing in Nonlinear Stochastic Control," in *Proceedings of IFAC Symposium on Stochastic Control*, pp. 343–352, Budapest, Hungary.

Ireland, P. N. (2003): "Irrational Expectations and Econometric Practice: Discussion of Orphanides and Williams 'Inflation Scares and Forecast-Based Monetary Policy'," Federal Reserve Bank of Atlanta Working Paper 2003-22.

Jehiel, P., and D. Samet (2005): "Learning to play games in extensive form by valuation," *Journal of Economic Theory*, 124(2), 129–148.

Kaspi, H., and A. Mandelbaum (1998): "Multi-armed bandits in discrete and continuous time," *Annals of Applied Probability*, 8(4), 1270–1290.

Keller, G., S. Rady, and M. Cripps (2005): "Strategic Experimentation with Exponential Bandits," *Econometrica*, 73(1), 39–68.

Kendrick, D. A. (1978): "Non-convexities from probing an adaptive control problem," *Journal of Economic Letters*, 1(4), 347–351.

——— (1979): "Adaptive control of macroeconomic models with measurement error," in *Optimal Control of Econometric Models*, ed. by S. Holly, B. Rustem, and M. B. Zarrop, pp. 204–230. Macmillan, London.

——— (1982): "Caution and Probing in a Macroeconomic Model," *Journal of Economic Dynamics and Control*, 4(4), 149–170.

——— (2002): *Stochastic Control for Economic Models.* http://eco.utexas.edu/faculty/Kendrick, 2 edn.

——— (2005): "Stochastic control for economic models: past, present and the paths ahead," *Journal of Economic Dynamics and Control*, 29(1-2), 3–30.

Kendrick, D. A., and H. M. Amman (2006): "A Classification System for Economic Stochastic Control Models," *Computational Economics*, 27(4), 453–481.

Kendrick, D. A., and J. Majors (1974): "Stochastic Control with Uncertain Macroeconomic Parameters," *Automatica*, 10(6), 587–593.

Kihlstrom, R. E., L. Mirman, and A. Postlewaite (1984): "Experimental Consumption and the 'Rothschild Effect'," in *Bayesian Models of Economic Theory*, ed. by M. Bayer, and R. E. Kihlstrom. Elsevier, Amsterdam.

Kimemia, J. (1982): "Hierarchical Control of Production in Flexible Manufacturing Systems," Ph.D. dissertation, Department of Electical Engineering and Computer Science, Massachusetts Institute of Technology.

Kimemia, J., S. B. Gershwin, and D. P. Bertsekas (1982): "Computation of Production Control Policies by a Dynamic Programming Technique," in *Analysis and Optimization of Systems*, ed. by A. Bensoussan, and J. L. Lions, pp. 243–269. Springer-Verlag, New York.

Koulouriotis, D., and A. Xanthopoulos (2008): "Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems," *Applied Mathematics and Computation*, 196(2), 913–922.

Kreps, D. (1998): "Anticipated Utility and Dynamic Choice, 1997 Schwartz Lecture," in *Frontiers of Research in Economic Theory*, ed. by D. P. Jacobs, E. Kalai, and M. Kamien. Cambridge University Press, Cambridge, England.

Ku, R., and M. Athans (1973): "On the Adaptive Control of Linear Systems using the Open Loop Feedback Optimal Approach," *IEEE Transactions on Automatic Control*, AC-18, 489–493.

Kulcsár, C., L. Pronzato, and E. Walter (1996): "Dual Control of Linearly Parameterised Models via Prediction of Posterior Densities," *European Journal of Control*, 2(2), 135–143.

Kumar, P. R. (1985): "A Survey of Some Results in Stochastic Adaptive Control," *SIAM Journal on Control and Optimization*, 23(3), 329–380.

——— (1990): "Convergence of Adaptive Control Schemes using Least-Squares Parameter Estimates," *IEEE Transactions on Automatic Control*, AC-35, 416–424.

Lindoff, B., and J. Holst (1997a): "Adaptive Predictive Control for Time-Varying Stochastic Systems," mimeo, Department of Mathematical Statistics, Lund Institute of Technology.

——— (1997b): "Suboptimal Dual Control of Stochastic Systems with Time-Varying Parameters," mimeo, Department of Mathematical Statistics, Lund Institute of Technology.

Ljung, L. (1987): *System Identification: Theory for the User.* Prentice Hall, Englewood Cliffs, NJ.

MacRae, E. C. (1972): "Linear decision with experimentation," *Annals of Economic and Social Measurement*, 1, 437–447.

——— (1975): "An Adaptive Learning Role for Multi-Period Decision Problems," *Econometrica*, 43, 893–906.

Maitelli, A. L., and T. Yoneyama (1994): "A Two-stage Dual Suboptimal Controller for Stochastic Systems using Approximate Moments," *Automatica*, 30(12), 1949–1954.

Marcoul, P., and Q. Weninger (2004): "Search and active learning with correlated information: Empirical Evidence from Mid-Atlantic clam fishermen," Working Paper 04010, Iowa State University.

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

Milito, R., C. Padilla, R. A. Padilla, and D. Cadorin (1982): "An Innovations Approach to Dual Control," *IEEE Transactions on Automatic Control*, AC-27(1), 132–137.

Moscarini, G., and L. Smith (2001): "The Optimal Level of Experimentation," *Econometrica*, 69(6), 1629–1644.

Norman, A. (1976): "First order dual control," *Annals of Economic and Social Measurement*, 5(3), 311–321.

Norman, A. L. (1981): "On the control of structural models, and a reply," *Journal of Econometrics*, 15(1), 13–24,29.

——— (1994): "On the complexity of linear quadratic control," *European Journal of Operations Research*, 73(2), 319–330.

Norman, A. L., and W. S. Jung (1977): "Linear Quadratic Theory for Models with Long Lags," *Econometrica*, 45(4), 905–917.

Osborne, M. (2006): "Consumer Learning, Switching Costs, and Heterogeneity: A Structural Examination," Unpublished manuscript.

Padilla, C. S., R. A. Padilla, and J. B. Cruz Jr. (1980): "A simple algorithm for SAFER control (Sensitivity Adaptive Feedback Estimation Redistribution," *International Journal of Control*, 32(6), 1111–1118.

Pekelman, D., and E. Tse (1980): "Experimentation and Budgeting in Advertising: An Adaptive Control Approach," *Operations Research*, 28(2), 321–347.

Pindyck, R. S. (1973a): *Optimal Planning for Economic Stabilization.* North-Holland, Amsterdam.

——— (1973b): "Optimal Policies for Economic Stabilization," *Econometrica*, 41(3), 529–560.

Prescott, E. C. (1972): "The Multi-Period Control Problem Under Uncertainty," *Econometrica*, 40(6), 1043–1058.

Pronzato, L., C. Kulcsár, and E. Walter (1996): "An Actively Adaptive Control Policy for Linear Models," *IEEE Transactions on Automatic Control*, 41(6), 855 – 858.

Quiggin, J. (1982): "A theory of anticipated utility," *Journal of Economic Behavior and Oragnization*, 3(4), 323–343.

Raiffa, H., and R. Schlaifer (1961): *Applied Statistical Decision Theory.* Harvard University Press, Boston, MA.

Rausser, G. C. (1978): "Active Learning, Control Theory, and Agricultural Policy," *American Journal of Agricultural Economics*, 60(3), 476–490.

Rausser, G. C., and J. W. Freebairn (1974a): "Approximate Adaptive Control Solutions to U.S. Beef Trade Policy," *Annals of Economic and Social Measurement*, 3(1), 177–203.

——— (1974b): "Estimation of Policy Preference Function: An Application to U.S. Beef Import Policy," *The Review of Economics and Statistics*, 56(4), 437–449.

Rausser, G. C., and E. Hochman (1978): *Dynamic Agricultural Systems: Economic Prediction and Control*, Dynamic Economic Series. Elsevier North-Holland, Amsterdam.

Robbins, H. (1952): "Some Aspects of the Sequential Design of Experiments," *Bulletin of American Mathematical Society*, 55, 527–535.

Rothschild, M. (1974a): "Searching for Lowest Price When the Distribution of Prices is Uknown," *Journal of Political Economy*, 82(4), 689–711.

——— (1974b): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9(2), 185–202.

Sargent, T. J. (1993): *Bounded Rationality in Macroeconomics.* Clarendon Press, Oxford.

——— (1999): *The Conquest of American Inflation.* Princeton University Press, Princeton, New Jersey.

Sarris, A. H., and M. Athans (1973): "Optimal Adaptive Control Methods for Structurally Varying Systems," Working Paper 24, National Bureaus of Economic Research, Cambridge, MA.

Schweitzer, P. J., and A. Seidman (1985): "Generalized Polynomial Approximations in Markovian Decision Processes," *Journal of Mathematical Analysis and Applications*, 110(2), 568–582.

Shupp, F. R. (1972): "Uncertainty and Stabilization Policies for a Nonlinear Macroeconomic Model," *Quartely Journal of Economics*, 86(1), 94–110.

Sternby, J. (1977): "A regulator for time-varying stochastic systems," in *Preprints of IFAC 7th World Congress*, Helsinki, Finland.

Sundaram, J. S. B. R. K. (1992): "Denumerable-Armed Bandits," *Econometrica*, 60(5), 1071–1096.

Sutton, R. S. (1984): "Temporal Credit Assignement in Reinforcement Learning," Ph.D. dissertation, University of Massachusetts, Amherst, MA.

Svensson, L. E. O., and N. Willams (2006): "Bayesian and Adaptive Optimal Policy under Model Uncertainty," Working Paper, Princeton University.

Svensson, L. E. O., and N. Williams (2005): "Monetary Policy with Model Uncertainty: Distribution Forecast Targeting," Working Paper, Princeton University.

Svensson, L. E. O., and M. Woodford (2005): "Implementing Optimal Policy through Inflation-Forecast Targeting," in *Infaltion Targeting*, ed. by B. S. Bernanke, and M. Woodford. University of Chicago Press.

Taylor, J. B. (1974): "Asymptotic Properties of Multiperiod Control Rules in the Linear Regression Model," *International Economic Review*, 15(2), 472–482.

Tesfaselassie, M. F., and E. Schaling (2007): "Managing Disinflation under Uncertainty," Working paper, Kiel Institute for the World Economy.

Theil, H. (1957): "A Note on Certainty Equivalence in Dynamic Planning," *Econometrica*, 25(2), 346–349.

Trefler, D. (1993): "The Ignorant Monopolist: Optimal Learning with Endogenous Information," *International Economic Review*, 34(3), 565–581.

Trick, M. A., and S. E. Zin (1993): "A Linear Programming Approach to Solving Stochastic Dynamic Programs," Unpublished Manuscript.

——— (1997): "Spline Approximations to Value Functions: A Linear Programming Approach," *Macroeconomic Dynamics*, 1(1), 255–277.

TSE, E., AND M. ATHANS (1972): "Adaptive Stochastic Control for a Class of Linear Systems," *IEEE Transactions on Automatic Control*, AC-17, 38–52.

TSE, E., Y. BAR-SHALOM, AND L. MEIER (1973): "Wide-Sense Adaptive Dual Control of Stochastic Nonlinear System," *IEEE Transactions on Automatic Control*, AC-18, 98–108.

TSITSIKLIS, J. N. (1984): "Convexity and Characterization of Optimal Policies in a Dynamic Routing Problem," *Journal of Optimization Theory and Applications*, 44(1), 105–136.

TSITSIKLIS, J. N., AND B. VAN ROY (1996): "Feature-Based Methods for Large-Scale Dynamic Programming," *Machine Learning*, 22(1-3), 59–94.

TURNOVSKY, S. J. (1975): "Optimal Choice of Monetary Instrument in a Linear Economic Model with Stochastic Coefficients," *Journal of Money, Credit and Banking*, 7(1), 51–80.

VAN DER STOEP, D. R. (1968): "Trajectory-Shaping for the Minimzation of State-Variable Estimation Errors," *IEEE Transactions on Automatic Control*, AC-13(3), 284–286.

WATKINS, C. J. C. H. (1989): "Learning from Delayed Rewards," Ph.D. dissertation, Cambridge University, Cambridge, England.

WHITTLE, P. (1982): *Optimization Over Time*, vol. 1. Wiley, Chichester.

WIELAND, V. (2000a): "Learning by Doing and the Value of Optimal Experimentation," *Journal of Economic Dynamics and Control*, 24(4), 501–535.

——— (2000b): "Monetary Policy, Parameter Uncertainty and Optimal Learning," *Journal of Monetary Economics*, 46(1), 199–228.

WIESLANDER, J., AND B. WITTENMARK (1971): "An Approach to Adaptive Control Using Real Time Identification," *Automatica*, 7(2), 211–217.

WITTENMARK, B. (1995): "Adaptive Dual Control Methods: An Overview," in *5th IFAC symposium on Adaptive Systems in Control and Signal Processing*, pp. 67–73.

WITTENMARK, B., AND C. ELEVITCH (1985): "An Adaptive Control Algorithm with Dual Features," in *7th IFAC/IFORS Symposium on Identification and System Parameter Estimation*, pp. 587–592, York, UK.

YAME, J. J. (1987): "Dual Adaptive Control of Stochastic Systems via Information Theory," in *Proceeding of 26th IEEE Conference on Decision and Control*, pp. 316–320.

YETMAN, J. (2000): "Probing Potential Output: Monetary Policy, Credibility and Optimal Learning under Uncertainty," Working Paper 2000-10, Bank of Canada.

ZAMPOLLI, F. (2005): "Optimal Monetary Policy in a Regime-Switching Economy: The Response to Abrupt Shifts in Exchange-Rate Dynamics," Working paper, Bank of England.

MORGAN STANLEY
E-mail address: Sergei.Morozov@morganstanley.com