APPROXIMATE POLICY OPTIMIZATION AND ADAPTIVE CONTROL IN REGRESSION MODELS

by

Jiarui Han
Tze Leung Lai
Viktor Spivakovsky

April 2005

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
APPROXIMATE POLICY OPTIMIZATION AND ADAPTIVE CONTROL IN REGRESSION MODELS

by

Jiarui Han
Department of Statistics
Stanford University

Tze Leung Lai
Department of Statistics
Stanford University

Viktor Spivakovsky
Citadel Investment Group
Chicago, IL 60603

April 2005

This research was supported in part by the National Science Foundation grant DMS-0305749

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Approximate Policy Optimization and Adaptive Control in Regression Models

JIARUI HAN\textsuperscript{1,*}, TZE LEUNG LAI\textsuperscript{2} and VIKTOR SPIVAKOVSKY\textsuperscript{3}

\textsuperscript{1}Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.;  
E-mail: jrhan@stanford.edu

\textsuperscript{2}Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.;  
E-mail: lait@stat.stanford.edu

\textsuperscript{3}Citadel Investment Group, 131 S. Dearborn, Chicago, IL 60603, U.S.A.;  
E-mail: Viktor.Spivakovsky@citadelgroup.com

Abstract

The "curse of dimensionality" in dynamic programming has hampered its potential applications to a wide range of dynamic optimization problems in economics. In this paper we use recent advances in approximate dynamic programming to develop an approximate policy optimization procedure that uses Monte Carlo simulations to circumvent the curse of dimensionality. We apply the procedure to the classical problem of "learning by doing" in regression models and demonstrate the value and extent of active experimentation in a variety of numerical studies.

Keywords: dynamic programming, policy iteration, rollout, Monte Carlo, learning by doing.

*Corresponding Author
1. Introduction

Stochastic optimization problems arise naturally in many fields, including economics and engineering. These problems have stochastic components and optimization is in the sense of maximizing the expectation of a stochastic payoff (or minimizing expected loss). A further complication arises when optimization is carried out sequentially over time. Although dynamic programming (abbreviated by DP hereafter) provides a standard tool to solve for the optimal policy (cf. Stokey and Lucas (1989)), its practical implementation is limited by the exponential rise in computational complexity as the dimension of the state space increases, which is often called the curse of dimensionality (Bellman, 1957, page ix).

This paper studies a class of stochastic optimization problems in which economic agents operating in uncertain (stochastic) environments face the dilemma between minimizing current-period loss and information generation in the sequential choice of actions. We consider both the infinite-horizon problem where the objective is to minimize a discounted sum of an infinite stream of expected losses, and the multiperiod (finite-horizon) problem of minimizing the cumulative sum of the expected losses over \( N \) periods. There is an extensive literature on Bayesian learning that assumes an economic agent to have a prior distribution over the set of environments and to update the distribution over successive time periods via the Bayes rule; see e.g. Blume and Easley (1984), Easley and Kiefer (1988, 1989), Kiefer (1989) and Aghion et al. (1991). The main ideas of these works are (i) to apply the martingale convergence theorem to show that the posterior distributions converge to a limit, call the “limit belief”, when actions are non-anticipating (in the sense that the choice of the action at time \( t \) depends on the observations up to time \( t - 1 \)), and (ii) to use the limit belief in conjunction with Bellman’s DP equation to derive the asymptotic properties of an optimal sequence of actions (“controls”). In principle one can apply DP directly to compute the optimal control rule in this Bayesian context and see how it strikes the balance between information generation for future actions and minimal current-period loss. The state space in this Bayesian optimal control problem is the space of probability measures on the unknown parameters describing the economic environment, and therefore the state dynamics are governed by how a prior distribution is transformed into a posterior distribution after the arrival of a new observation.

In particular, Åström (1990, p. 478) used the backward induction algorithm in DP to compute the optimal policy that minimizes

\[
\int E_\beta \left\{ \sum_{t=2}^{T} (y_t - y_t^*)^2 \right\} d\pi(\beta)
\]

in the regression model \( y_t = y_{t-1} + \beta u_t + \epsilon_t \), where \( \epsilon_t \sim N(0, 1) \), \( y_t^* \equiv 0 \) and \( \pi \) is a normal prior distribution on \( \beta \). With \( T = 30 \), the numerical computations took 180 CPU hours on a VAX
11/780 computer and showed that the optimal policy takes relatively large and irregular control actions $u_t$ to probe the system when the Bayes estimate $\beta_{t-1}$ of $\beta$ has poor precision, but is well approximated by the certainty equivalence rule $\beta_{t-1}u_t = -y_{t-1}$. Prescott (1972) considered the Bayes rule minimizing (1) in the simpler model $y_t = \beta u_t + \epsilon_t$. His numerical computations for $T = 2, 3, 4, 6$ and $y^*_t \equiv y^* = 1, 4, 16$ showed little difference between the Bayes rule and the myopic rule that minimizes the current-period loss, except when there is large uncertainty in the estimate of $y^*/\beta$. His numerical study also showed the myopic policy to be superior to the certainty equivalence policy in approximating the optimal policy.

Wieland (1995, 2000) added an intercept term to the regression model, i.e.,

$$y_t = \alpha + \beta u_t + \epsilon_t,$$

and considered the infinite-horizon problem of minimizing

$$\int E_{\alpha, \beta} \left\{ \sum_{t=1}^{\infty} \delta^{t-1} \ell(y_t, u_t) \right\} d\pi(\alpha, \beta),$$

where $0 < \delta < 1$ is a discount factor, $\ell$ is a loss function and $\pi$ is a bivariate normal distribution. Since the posterior distribution of $(\alpha, \beta)$ given $\{(u_i, y_i), i \leq t\}$ is also normal, the state space of the Bayesian optimal control problem is a 5-dimensional space involving the first and second moments of $\alpha$ and $\beta$ and the expectation of $\alpha \beta$. Wieland’s (1995) numerical study is based on approximate solution of Bellman’s DP equation via an iterative scheme that uses (i) Gaussian quadrature to compute integrals (over $\alpha, \beta$) in the Bellman equation and (ii) policy iteration that describes the policy by using a finite grid of points in the 5-dimensional state space. Besides the “curse of dimensionality” in this finite grid approximation, there are additional difficulties due to apparent discontinuities in the policy function and multiple local minima that require expensive grid search for optimum. Wieland (1995, page 97) indicates the large computational cost of this scheme that requires 5 days on a SPARC 2 workstation to compute an approximation to the optimal policy in the case $\delta = 0.95$. His results show that the optimal policy differs substantially from the myopic one, which he attributes to the existence of incorrect limit beliefs that may be reinforced under passive learning but that can be avoided by optimal experimentation.

Because of the computational complexity of DP, no simulation studies have been performed on the Bayes risk (1) or (3) of the computationally formidable optimal policy and to compare it with that of the much simpler myopic or certainty equivalence rule. In addition, instead of Bayes risks, frequentists may want to consider risk functions evaluated at different values of $\alpha$ and $\beta$. In particular, Anderson and Taylor (1976) introduced least-squares certainty equivalence rules and Kendrick (1981) considered dual control rules, whose risk functions they studied by Monte Carlo simulations. Simulation studies are also important
for assessing the robustness properties of the proposed control rules to model misspecification.

During the past decade there have been important advances in approximate dynamic programming in the machine learning/artificial intelligence literature, motivated by large combinatorial optimization problems in traffic and communications networks and computer algorithms to play games such as backgammon and solitaire. Sections 2 and 5 review these advances and modify them for applications to adaptive control of regression models and to other stochastic optimization problems in economics. Whereas Section 2 considers the infinite-horizon discounted problem (3) for which the optimal policy is stationary Markov (in the sense that it only depends on the state) and is approximated by a rollout scheme that mimics conventional policy iteration in dynamic programming, Section 3 shows how the scheme can be modified for finite-horizon problems. Simulation studies of the performance of the proposed approximations to the optimal rules are given in Section 4 where the myopic and other policies are also studied for comparison to assess the value and extent of active learning in adaptive control. Some concluding remarks are given in Section 5.

2. Rollout Algorithms for Approximate Policy Iteration

2.1. POLICY ITERATION IN INFINITE-HORIZON MARKOV DECISION PROBLEMS

Tesauro and Galperin (1996) suggest that if there is a good heuristic policy for a finite-state Markov decision problem, with an absorbing state and a finite action space, then it can be “rolled out” to compute controls that are improvements over those of the heuristic policy. The “rollout” corresponds to executing a single policy iteration step (Bertsekas (2000), page 374) on the states visited during execution of the policy; the terminology arises from the game of backgammon. Subsequently Bertsekas, Tsitsiklis and Wu (1997) have proposed rollout algorithms for combinatorial optimization, Secomandi (2001, 2003) and MacGovern, Moss and Barto (2002) have developed rollout schemes for stochastic routing problems; see also Section 6.4 of Bertsekas (2000). Recently Yan et al. (2005) have introduced iterated rollout schemes in programming a machine to play solitaire so that it can win more games on average than an expert human player.

To begin with, consider a controlled Markov chain \( \{X_t\} \) on a finite state space \( \mathcal{X} \), whose control \( u_t \) at time \( t \) belongs to a finite set \( \mathcal{U} \). Let \( (p_{xy}(u))_{x,y \in \mathcal{X}} \) denote the transition probability matrix corresponding to control \( u \in \mathcal{U} \). An infinite-horizon Markov decision problem is an optimization problem that chooses a sequence \( U = (u_1, u_2, \ldots) \) of controls to minimize the total expected discounted cost

\[
c_U(x) = E \left\{ \sum_{t=1}^{\infty} \delta^{t-1} g(X_t, u_t) | X_1 = x \right\},
\]

(4)
where $0 < \delta < 1$ is a discount factor and $g$ is a cost function. The minimum $V(x)$ of $c_U(x)$ over all control policies $U$ satisfies the Bellman equation

$$V(x) = \inf_{u \in U} \left\{ g(x, u) + \delta \sum_{y \in \mathcal{X}} p_{xy}(u) V(y) \right\}.$$  \hspace{1cm} (5)

The minimum is attained by a stationary policy, i.e., $u_t = \mu(x_t)$, where $\mu(x)$ is the minimizer $u$ of the right-hand side of (5). Two commonly used algorithms to solve (5) are value iteration and policy iteration (Bertsekas (2000), pages 373-376). In particular, the latter has at the $k$th iteration a “policy evaluation” step that computes the cost $c_{\mu^{(k)}}(x)$ of a stationary policy $U^{(k)}$ (associated with $\mu^{(k)} : \mathcal{X} \to \mathcal{U}$) and a “policy improvement” step that defines $\mu^{(k+1)}(x)$ as the minimizer $u$ of

$$g(x, u) + \delta \sum_{y \in \mathcal{X}} p_{xy}(u) c_{\mu^{(k)}}(y).$$ \hspace{1cm} (6)

The cost-to-go function $c_{\mu}$ can be computed by solving the linear system

$$c_{\mu}(x) = g(x, \mu(x)) + \delta \sum_{y \in \mathcal{X}} p_{xy}(\mu(x)) c_{\mu}(y), \ x, y \in \mathcal{X}. \hspace{1cm} (7)$$

Note that (6) in the policy improvement step can be expressed as

$$g(x, u) + E \left\{ \sum_{t=2}^{\infty} \delta^{t-1} g(X_t, u_t^{(k)}) | X_1 = x, u_1 = u \right\}. \hspace{1cm} (8)$$

When the state space $\mathcal{X}$ is large, solving the linear system (7) becomes computationally expensive. Also storing $\mu(x), x \in \mathcal{X}$, for a stationary policy becomes expensive. Instead of solving (7), one can use Monte Carlo simulations to compute the cost-to-go function of a stationary policy $U$ for those states that are visited during the execution of the policy, as is done in a “rollout”, which corresponds to a single iteration of policy improvement via minimization of (8) with $k = 0$ and which uses a good heuristic policy for $u_t^{(0)}$. An “iterated rollout” involves $k \geq 1$, where $k = 1$ corresponds to the rollout of a heuristic policy and $k = 2$ corresponds to the rollout of that rollout policy, etc. Since the calculation of $u_t^{(k)}$ in (8) requires optimization of the cost-to-go that is computed by Monte Carlo simulations, an iterated rollout requires layers of Monte Carlo simulations, and it is therefore important to find a good heuristic policy to initialize because at most a few iterations are computationally tractable. To compute the cost-to-go by Monte Carlo, the infinite series in (8) has to be replaced by a finite sum over $2 \leq t \leq n_\delta$. Unless $\delta$ is very near 1, $n_\delta$ can be of moderate size ($< 100$) for $\delta^{n_\delta}$ to be reasonably small. Details of the Monte Carlo procedure are given in Section 4.

2.2. ROLLOUT IN LINEAR REGRESSION MODELS
Consider the simple linear regression model \( y_t = \alpha + \beta u_t + \epsilon_t \) with independent and identically distributed normal errors \( \epsilon_t \) having mean 0. We can express the adaptive control problem of choosing inputs \( u_t \) sequentially to minimize (3) as a Markov decision problem whose states are

\[
X_t = (E_{t-1}\alpha, E_{t-1}\beta, E_{t-1}\alpha^2, E_{t-1}\beta^2, E_{t-1}\alpha\beta),
\]

in which \( E_{t-1} \) denotes conditional expectation given \( \{(y_i, u_i), \ i \leq t - 1\} \) and \( E_0 \) denotes expectation under the prior distribution. Although the state space is not finite but is a region in \( \mathbb{R}^5 \) satisfying certain constraints (on first and second moments and the cross-product moment), the same ideas described above for rollouts can still be applied. In fact, the Monte Carlo approach to evaluate the cost-to-go is particularly effective in dealing with the curse of dimensionality in multidimensional state spaces; see Rust (1997) who considers continuous state spaces \( \mathcal{X} \) but finite control sets \( \mathcal{U} \). Here the control set \( \mathcal{U} = \mathbb{R} \) is also continuous and unbounded, and the cost function \( g \) involves an integral with respect to a Gaussian measure on \( \mathbb{R} \). Specifically, since \( y_t = \alpha + \beta u_t + \epsilon_t \) and \( u_t \) depends only on \( \{(y_i, u_i), \ i \leq t - 1\} \), we can express (3) in the form of (4) with

\[
g(X_t, u) = \int \ell(E_{t-1}\alpha + uE_{t-1}\beta + z, u) \ d\Psi_t(z),
\]

(10)

where \( \Psi_t \) is the normal distribution with mean 0 and variance

\[
\text{Var}_{t-1}(\alpha) + u^2\text{Var}_{t-1}(\beta) + 2u\text{Cov}_{t-1}(\alpha, \beta) + \text{Var}(\epsilon),
\]

(11)

which is the conditional variance of \( (\alpha - E_{t-1}\alpha) + (\beta - E_{t-1}\beta)u + \epsilon_t \) given \( \{(y_i, u_i), \ i \leq t - 1\} \) and which is a function of the 5-dimensional state \( X_t \). The integral in (10) can be computed by Gaussian quadrature, and has a simple explicit formula for quadratic loss functions \( \ell; \) see (13) below.

Besides Monte Carlo computation of the cost-to-go function evaluated only at the states visited during execution of the policy, another important ingredient of the rollout algorithm is the so-called base policy \( u_t^{(0)} \), which is the heuristic policy whose cost-to-go function is to be evaluated by Monte Carlo. For the control problem (3) in simple linear regression, we take \( u_t^{(0)} \) to be the myopic rule \( u_t^{\text{myop}} \) that minimizes the current-period cost and ignores active experimentation to generate information for future cost reduction, i.e.,

\[
u_t^{\text{myop}} = \text{Arg min}_u g(X_t, u).
\]

(12)

For the quadratic loss function \( \ell(y_t, u_t) = (y_t - y^*)^2 + \omega u_t^2 \) considered by Wieland (1995, 2000) and in the numerical studies of Section 4, (10) and (12) have the following explicit formulas. Let \( \sigma^2 = \text{Var}(\epsilon) \),

\[
a = E_{t-1}\alpha - y^*, \ b = E_{t-1}\beta, \ v_\alpha = \text{Var}_{t-1}(\alpha), \ v_\beta = \text{Var}_{t-1}(\beta), \ v_{\alpha\beta} = \text{Cov}_{t-1}(\alpha, \beta).
\]
Then
\[ g(X_t, u) = \sigma_x^2 + v_\alpha + a^2 + u^2(v_\beta + b^2 + \omega) + 2u(v_\alpha + ab), \tag{13} \]
\[ u_t^{mpp} = -(v_\alpha + ab)/(v_\beta + b^2 + \omega) = -E_{t-1}\{(\alpha - y^*)\beta\}/\{\omega + E_{t-1}\beta^2\}, \tag{14} \]
\[ g(X_t, u_t^{mpp}) = \sigma_x^2 + E_{t-1}[\{(\alpha - y^*)^2\} - \{E_{t-1}[\{(\alpha - y^*)\beta\}]^2/\{\omega + E_{t-1}\beta^2\}. \tag{15} \]

There are three major reasons why we choose the myopic policy to be the base policy. First a major thrust of the numerical studies in the literature has been on whether the optimal policy differs significantly from the myopic policy, thereby revealing the value of optimal experimentation; see Prescott (1972), Kiefer (1989) and Wieland (1995, 2000). Although the rollout with the myopic policy as the base policy is suboptimal, if it can be demonstrated that the rollout already improves the myopic policy substantially, then the same can be said for the optimal policy which is prohibitively difficult to compute and has to rely on some form of approximation. Secondly the myopic policy is relatively easy to implement. Finally note from (8) that like the optimal policy, the rollout also tries to strike a balance between the current-period cost \( g(x, u) \) and the cost-to-go, except that the cost-to-go is based on using the myopic policy for future periods. Since it anticipates no active experimentation in the future, it may tend to over-perturb the myopic rule in the current period. On the other hand, if the experimental design in \( \{(y_i, u_i), \ i \leq t-1\} \) does not yet provide adequate information about the unknown parameters \( \alpha \) and \( \beta \), then one should have more active learning at current time \( t \) than delaying it to the future, for which the benefit of current active learning is discounted by some power of \( \delta \). These heuristics suggest that unless the discount factor is very near 1 (which will be discussed in Section 3), the rollout may already have incorporated nearly optimal experimentation. Although it is difficult to demonstrate this directly because of the computational difficulties of the optimal policy, we shall introduce a tractable super-optimal policy that assumes \( \beta \) to become completely known after a few periods in the future, and compare the performance of the rollout with the super-optimal policy in Section 4.

We have assumed \( \sigma_x \) to be known in the Bayesian regression model. When \( \sigma_x \) is unspecified, we can augment the Bayesian model by putting a gamma prior on \( (2\sigma_x^2)^{-1} \) and extend the rollout procedure together with the formulas (13)-(15) to handle this case. Details are given in Section 4.4.

2.3. AN ORACLE POLICY WITH COMPLETE KNOWLEDGE OF \( \beta \)

We next describe this super-optimal policy in the context of a quadratic loss function \[ \ell(y_t, u_t) = (y_t - y^*)^2 + \omega u_t^2. \] The basic idea is to bound the complicated cost-to-go function for the optimal policy by that of an “oracle” policy that assumes the value of \( \beta \) to be revealed after \( k \) periods in the future. We denote this policy by SP\((k)\), where SP stands for “super-optimal” and \( k \) is chosen to be some small number for computational tractability. To begin
with, note that when \( \beta \) is known, the model becomes that of a location model \( y_t - \beta u_t = \alpha + \epsilon_t \) and the maximum likelihood estimate of \( \alpha \) at time \( n \) is the sample mean \( n^{-1} \sum_{t=1}^{n} (y_t - \beta u_t) \), which has mean 0 and variance \( \sigma^2 / n \), irrespective of how the \( u_t \) are chosen. Similarly the posterior variance of the Bayes estimate of \( \alpha \) does not depend on how the \( u_t \) are chosen, and therefore the myopic rule

\[
u_t^* = -(E_{t-1} \alpha - y^*)/(\beta + \beta^{-1} \omega)
\]

is optimal. Moreover, (15) now reduces to the simpler form

\[
g(X_t, u_t^*) = \sigma^2 + \text{Var}_{t-1} \alpha + \beta^{-2} \omega (E_{t-1} \alpha - y^*)^2/(1 + \beta^{-2} \omega).
\]

Of particular importance are the following recursive formulas for \( a_t = E_t \alpha \) and \( \nu_t = \text{Var}_t \alpha \) when \( \beta \) is known:

\[
u_t^{-1} = \nu_{t-1}^{-1} + \sigma^2,
\]

\[
a_t = \left( \frac{a_{t-1} + \nu_{t-1} - \beta u_t}{\nu_{t-1}} \right) \left( \frac{1}{\nu_{t-1}} + \frac{1}{\sigma^2} \right).
\]

Since \( E_{t-1} (y_t - \beta u_t) = a_{t-1} \) and \( \text{Var}_{t-1} (y_t - \beta u_t) = \nu_{t-1} + \sigma^2 \), it follows from (19) that

\[
E_{t-1} (a_t - y^*)^2 = (a_{t-1} - y^*)^2 + \nu_{t-1} + \sigma^2,
\]

which yields by induction

\[
E_0 (a_t - y^*)^2 = (E_0 \alpha - y^*)^2 + \sum_{i=1}^{t} \nu_{i-1} + t \sigma^2.
\]

The difference equation (18) has an explicit solution \( \nu_j = \nu_0 / (1 + j \nu_0 / \sigma^2) \). Putting this and (20) into (17) then yields

\[
E \left\{ \sum_{t=1}^{\infty} \delta^{t-1} g(X_t, u_t^*) | X_0 = x \right\} = \frac{\sigma^2}{1 - \delta} + \frac{\beta^{-2} \omega (E_0 \alpha - y^*)^2}{(1 + \beta^{-2} \omega)(1 - \delta)}
\]

\[
+ \left( 1 + \frac{\beta^{-2} \omega}{(1 + \beta^{-2} \omega)(1 - \delta)} \right) \sum_{j=0}^{\infty} \frac{\delta^j \nu_0}{1 + j \nu_0 / \sigma^2} + \frac{\beta^{-2} \omega \sigma^2}{1 + \beta^{-2} \omega} \sum_{t=0}^{\infty} t \delta^t.
\]

The preceding argument assumes that \( \beta \) is completely known at all times, as in the SP(1) policy. For SP(k) with small \( k \), say \( k \leq 10 \), a rollout policy with the base policy given by (16) for \( t > k \) and by the myopic rule (14) for \( 1 < t \leq k \) should provide a good and easily implementable approximation since \( E \{ \Sigma_{t=2}^{k} \delta^{t-1} g(X_t, u_t^{\text{my}}) | X_1 \} \) approximates well the optimal finite-horizon cost-to-go for small horizon \( k \) and can be conveniently computed by Monte Carlo, and since \( E \{ \Sigma_{t=k+1}^{\infty} \delta^{t-1} g(X_t, u_t^*) | X_k \} \) has a closed-form expression given by (21).

3. Approximate Policy Optimization in Finite-Horizon Learning by Doing
As pointed out in Section 2.1, the rollout algorithm for approximate policy optimization in infinite-horizon Markov decision problems is based on stationarity of the optimal policy and the method of policy iteration to solve the DP equation (5). For finite-horizon problems, the optimal policy is nonstationary and DP is carried out by backward induction, to which the policy iteration approach in Section 2 is not applicable. On the other hand, by allowing discount factors to vary with the time to maturity, we can still use the rollout algorithm to approximate the optimal policy. We show in particular how this can be carried out for the finite-horizon analog of (3) in the simple linear regression model (2). Letting $T$ denote the horizon (or maturity period) and defining $X_t$ and $g(X_t, u)$ by (9) and (10), note that

$$
\int E_{\alpha, \beta} \left\{ \sum_{t=1}^{T} \ell(y_t, u_t) \right\} d\pi(\alpha, \beta) = E \left\{ \sum_{t=1}^{T} g(X_t, u_t) | X_1 \right\}.
$$

(22)

Let $u_t^{opt}$ denote the optimal control at current time $t$. For $t$ near $T$, the myopic rule $u_t^{mvp}$ given by (12) is expected to perform like $u_t^{opt}$ since there are not many stages in the future to benefit from active experimentation. Therefore we can use a rollout algorithm which uses the myopic policy as the base policy and which sets the discount factor to be 1 from $t$ to $T$ and to be 0 after $T$. When the time to maturity $T - t$ is not too large, this rollout algorithm should still be able to achieve a good balance between the current-period loss and the information content of the experimental design on $\alpha$ and $\beta$. For large $T - t$, however, the rollout may over-perturb $u_t^{mvp}$ as its base policy is the myopic policy that does not incorporate active experimentation after current time $t$. Recall in this connection that the rollout chooses at time $t$ the control $u$ to minimize

$$
g(X_t, u) + E \left\{ \sum_{i=t+1}^{T} g(X_i, u_i^{mvp}) | X_t, u_t = u \right\}.
$$

(23)

For large $T - t$, the sum over $i + 1 \leq i \leq T$ in (23) can outweigh the term $g(X_t, u)$, leading to over-experimentation at time $t$. To avoid this pitfall, we restrict the number of summands in (23), whose sum is to be replaced by a truncated sum over $t + 1 \leq i \leq \min(T, t + m)$.

Note that whereas the number of summands in (23) relates to the future $T - t$ periods, (23) also incorporates the past and present in the state $X_t$. As pointed out in Section 2C, and by Lai and Robbins (1979, 1982) for the finite-horizon case, the myopic policy is optimal when $\beta$ is known, and we expect $u_t^{mvp}$ to differ not much from $u_t^{opt}$ if $\Var_{t-1}(\beta)$, which will be denoted by $V_t$, is small. On the other hand, if $V_t$ is large, active experimentation can be used to improve $u_t^{mvp}$ in (23). We shall modify $u_t^{mvp}$ by $\mu_{i,t}(u)$ that depends on a parameter $u$ to be chosen by minimizing

$$
g(X_t, u) + E \left\{ \sum_{i=t+1}^{t+m} g(X_i, \mu_{i,t}(u)) | X_t, u_t = u \right\}
$$

(24)
when $T - t > m$. Thus, we replace the criterion (23), which we use to choose $u_i$ when $T - t \leq m$, by (24) when $T - t > m$.

We next describe the control $\mu_{i,s}(u)$ in (24) for $t + 1 \leq i \leq t + m$. Let

$$\Delta_{i,t}(u) = |u - u_i^{\text{mvp}}| \min\{1, (V_{i-1}/V_{t-1})^{1/2}\}$$

(25)

denote the distance between $u$ and $u_i^{\text{mvp}}$ scaled by the ratio of the standard deviations of the posterior distributions of $\beta$ at times $i - 1$ and $t - 1$. Note that the posterior variance of $\beta$ at time $k$ is closely related to the variance $\sigma_\epsilon^2/\Sigma_{j=1}^k (u_j - \bar{u}_k)^2$ of the least squares estimate of $\beta$, where $\bar{u}_k = k^{-1}\sum_{j=1}^k u_j$. Moreover, as shown by Lai and Robbins (1979),

$$\sum_{j=1}^k (u_j - \bar{u}_k)^2 = \sum_{j=1}^{k-1} (u_j - \bar{u}_{j-1})^2 + k^{-1}(k-1)(u_k - \bar{u}_{k-1})^2.$$  

(26)

Therefore, to increase the information content of the experimental design at time $k$ for reduction of future posterior variances of $\beta$, one should move $u_k$ as far away from $\bar{u}_{k-1}$ as possible. This suggests the following perturbation of the myopic rule at time $i (> t)$ when active experimentation at time $t$ results in the control $u$ instead of $u_i^{\text{mvp}}$:

$$\mu_{i,t}(u) = u_i^{\text{mvp}} + \Delta_{i,t}(u) \text{sgn}(u_i^{\text{mvp}} - \bar{u}_{i-1}).$$

(27)

For the discounted infinite-horizon problem in Section 2 where we have used the myopic policy as the base policy of a rollout, we can replace the myopic policy by $\mu_{i,1}$ when the discount factor $\delta$ is very near 1, for which the discounted problem is asymptotically equivalent to a finite-horizon problem with horizon $T \approx (1 - \delta)^{-1}$. This is tantamount to setting $u_i^{(k)}$ in (8) to be $\mu_{i,1}(u)$. The preceding policy can also be modified for the frequentist setting that does not assume a prior distribution on $(\alpha, \beta)$, as in Anderson and Taylor (1976) and Lai and Robbins (1979, 1982). Details are given in Section 4.3.

4. Implementation and Numerical Examples

Throughout this section we assume the quadratic loss function $l(y_t, u_t) = (y_t - y^*)(2) + \omega u_t^2$ in adaptive control of the linear regression model (2). To implement the approximate policy optimization schemes in Sections 2 and 3, we make use of the Kalman filter for recursive updates of the quantities in (13)-(15). Let $\theta = (\alpha, \beta)'$, $\theta_t = E\theta$, $P_t = \sigma_\epsilon^{-2} \text{Cov}t(\theta)$, $\varphi_t = (1, u_t)'$. Initializing at the prior mean vector and covariance matrix, the Kalman filter provides the following recursion:

$$\theta_t = \theta_{t-1} + (1 + \varphi_t'P_{t-1}\varphi_t)^{-1}\varphi_t(y_t - \theta_{t-1}'\varphi_t),$$

(28a)

$$P_t = P_{t-1} - (1 + \varphi_t'P_{t-1}\varphi_t)^{-1}\varphi_t\varphi_t'P_{t-1}.$$  

(28b)

9
Note that $E_t((\alpha - y^*)^2 ) = \text{Var}_t(\alpha) + (E_t(\alpha - y^*))^2$ and $E_t[|\alpha - y^*|\beta] = \text{Cov}_t(\alpha, \beta) + (E_t(\alpha - y^*))E_t\beta$.

As pointed out in Section 2.1, the infinite series in (8) is replaced by a finite sum over $2 \leq t \leq n_\delta$, whose expectation is computed by Monte Carlo. We now describe briefly the Monte Carlo procedure. From the prior normal distribution $\pi$ (with mean $\theta_0$ and covariance matrix $\sigma_2^2 P_0$ given by $X_1$), we can sample $(\alpha, \beta)'$ and combine with the control $u_t = u$ to generate $y_1 \sim N(\alpha + \beta u, \sigma_e^2)$. From $(u_1, y_1)$, we obtain the mean $\theta_1$ and covariance matrix $\sigma_e^2 P_1$ of the posterior distribution which then yields $X_2$. We can then compute $g(X_2, u_2^{mup})$ by using the explicit formula (15) and generate $y_2 \sim N(\alpha + \beta u_2^{mup}, \sigma_e^2)$. Proceeding inductively in this way yields one realization of the cumulative discounted cost $C = \sum_{t=2}^{n_\delta} \delta^{t-1} g(X_t, u_t^{mup})$. Repeating the procedure $N$ times yields $(\alpha^{(i)}(t), \beta^{(i)}(t), C_i)$, $i = 1, \cdots, N$, and the average $N^{-1} \sum_{i=1}^N C_i$ gives a Monte Carlo estimate $\hat{C}(u)$ of the cost-to-go $C(u) := E\{\sum_{t=2}^{n_\delta} \delta^{t-1} g(X_t, u_t^{mup})|X_1, u_1 = u\}$. By making use of the inequality

$$\left| u_t^{opt} - u_t^{mup} \right| \leq \frac{\sqrt{\delta}}{1 - \delta} \left\{ \frac{[E_{t-1}(\alpha - y^*)^2)(\omega + E_{t-1}\beta^2) - [E_{t-1}(\alpha - y^*)\beta]^2]}{\omega + E_{t-1}\beta^2} \right\}^{\frac{1}{2}},$$

we can restrict the search for the minimizer of $g(X_1, u) + \hat{C}(u)$ to an interval centered at $u_1^{mup}$ and having half-width given by the right-hand side of (29) (with $t = 1 = 0$). The proof of (29) is given in the Appendix.

For the finite-horizon case considered in Section 3, the proof in the Appendix shows that (29) still holds with $\delta/(1 - \delta)$ replaced by $T - t$. The preceding Monte Carlo procedure can also be used to evaluate the conditional expectation in (24) that uses $\mu_t(u)$ (instead of $\mu_t^{mup}$) defined by (27) as the base policy for the rollout algorithm. Choosing $m$ in (24) between 10 and 30 for $T$ around 100 not only saves computation time but also avoids the potential pitfall of over-weighting the discounted cost $g(X_t, u)$ by too many terms in the cost-to-go based on sub-optimal future controls. This remark also applies to the choice of $n_\delta$ in truncating the infinite series in (8) for the discounted problem when $\delta$ is very near 1, for which the discounted problem behaves like a finite-horizon problem with $T \approx (1 - \delta)^{-1}$.

4.1. A SIMULATION STUDY OF INFINITE-HORIZON ADAPTIVE CONTROL

We implement the rollout policy by using the procedure described above in a simulation study of its performance for the infinite-horizon adaptive control problem of minimizing the expected total cost (3) with discount factor $\delta = 0.9$ and $l(y_t, u_t) = y_t^2$. We choose the truncation horizon $n_\delta$ to be 80, noting that $(0.9)^{80} < 0.00022$, and use $N = 100$ simulation runs for the Monte Carlo evaluation of the cost-to-go. We compare the expected cost of the rollout policy with that of the myopic policy for three choices of the prior normal distribution $\pi$, which all have $E(\alpha) = 5$ and $E(\beta) = 1$ but which have different covariance matrices for $(\alpha, \beta)$. In addition, we also consider the expected cost of the super-optimal policy SP(3) to
provide some benchmark for comparison. To see the learning rate in these procedures, we consider the regret \( R_n \) of a policy as it changes over time, where the regret is defined by

\[
R_n = E_{\pi} \left\{ \sum_{t=1}^{n} \delta^{t-1} (\alpha + \beta u_t)^2 \right\}, \quad 1 \leq n \leq n_\delta.
\]

(30)

Note that the expected total cost up to time \( n \) can be expressed as

\[
E_{\pi} \left( \sum_{t=1}^{n} \delta^{t-1} y_t^2 \right) = \sigma_\epsilon^2 \sum_{t=1}^{n} \delta^{t-1} + R_n,
\]

where the first summand is the same for all policies and is due to the random errors \( \epsilon_t \). Besides \( I(y_t, u_t) = y_t^2 \), we also assume that \( \sigma_\epsilon = 1 \) throughout our numerical studies. Figure 1 plots \( R_n \), which is evaluated by Monte Carlo using 1000 simulations, for the three policies. As shown in the figure, the rollout policy has performance near that of the super-optimal policy and considerably outperforms the myopic policy. Moreover, its learning rate is highest in the first 3 observations, where regret rises sharply before flattening out for \( n \geq 4 \).

INSERT FIGURE 1 ABOUT HERE

4.2. A SIMULATION STUDY OF FINITE-HORIZON ADAPTIVE CONTROL

Figure 2 summarizes the results of a simulation study of the modified rollout policy in Section 3 (with \( \mu_{\text{opt}} \) in lieu of \( \mu^{\text{opt}} \) to form the base policy). Its performance is again measured by the regret, which we now define as

\[
R^*_n = E_{\pi} \left\{ \sum_{t=1}^{n} (\alpha + \beta u_t)^2 \right\}, \quad 1 \leq n \leq T;
\]

(31)

noting that \( E_{\pi}(\sum_{t=1}^{n} y_t^2) = n\sigma_\epsilon^2 + R^*_n \). Here \( T = 80 \) and we choose \( m = 30 \) to implement the modified rollout policy, for which we again use \( N = 100 \) simulations to evaluate the cost-to-go. Figure 2 also gives for comparison the \( R^*_n \) values for the myopic and SP(4) policies, and each \( R^*_n \) value is based on 1000 simulations. It shows that the modified rollout policy has performance comparable to the super-optimal policy and is considerably better than the myopic policy.

INSERT FIGURE 2 ABOUT HERE

4.3. MODIFICATION TO FREQUENTIST SETTING AND A COMPARATIVE STUDY

Without assuming a prior distribution on the unknown parameters, Anderson and Taylor (1976) used the least squares estimate \( \hat{\theta}_t = (\hat{\alpha}_t, \hat{\beta}_t) \) to estimate \((\alpha, \beta)\) from \(\{(y_i, u_t), i \leq t\}\) and proposed the certainty equivalence rule \( u_t = \max(-K, \min\{y^* - \hat{\alpha}_{t-1} / \hat{\beta}_{t-1}, K\}) \) for \( t \geq 3 \), where \( K \) is some \textit{a priori} bound on \((y^* - \alpha) / \beta\) and is one of three tuning parameters
of the control rule, the other two being \( u_1 \) and \( u_2 \). They performed a simulation study using different values of the three tuning parameters and 100 Monte Carlo simulations to evaluate \( E_{\alpha, \beta}(\alpha + \beta u_t)^2 \) at \( \alpha = 0 \) and \( \beta = 1 \), assuming that \( \alpha_t = 1 \) and \( y_t = 0 \). They also used their simulation results to study the statistical properties of \((\hat{\alpha}_t, \hat{\beta}_t)\). Anderson and Taylor (1976, p.1293) conjectured \((\hat{\alpha}_t, \hat{\beta}_t)\) to be consistent but mentioned at the end of the paper that the evidence pertaining to the consistency of \( \hat{\beta}_t \) from the simulation study was inconclusive. Lai and Robbins (1982) subsequently disproved the conjecture by constructing an event, with positive probability, on which \( u_t \) eventually gets stuck at one of the endpoints. Our simulation studies, one of which is reported in Figure 3 below, have shown that this probability may be substantial for certain values of \((\alpha, \beta)\).

In the Bayesian context, Prescott (1972) has shown that the myopic rule outperforms the certain equivalence rule. Although there is no prior distribution in the Anderson-Taylor framework and we do not want to impose one as the frequentist performance of the Bayesian myopic rule will depend on how much weight the prior distribution assigns to the vicinity of the actual \((\alpha, \beta)\), we can modify the Bayesian myopic rule of Section 3 for \( t \geq 3 \) by using \((u_1, y_1)\) and \((u_2, y_2)\) to generate a "synthetic" prior distribution. Letting \( \varphi_t = (1, u_t)' \) as before, we use \( N(\hat{\beta}_2, \sigma^2_t(\sum_{i=1}^{T} \varphi_i \varphi_i')^{-1}) \) as the prior distribution to define the myopic rule \( u_t^{my} \) for \( t \geq 3 \).

With \( N(\hat{\beta}_2, \sigma^2_t(\sum_{i=1}^{T} \varphi_i \varphi_i')^{-1}) \) as the prior distribution, we can also apply the modified rollout rule of Section 3 to the present context when \( t \geq 3 \). Since \( u_1 \) and \( u_2 \) are predetermined control parameters in the Anderson-Taylor framework and do not depend on which control policy is used for \( t \geq 3 \), we define the regret as

\[
R_n = E_{\alpha, \beta} \left\{ \sum_{t=3}^{n} (\alpha + \beta u_t)^2 \right\}, \quad 3 \leq n \leq T. \tag{32}
\]

Figure 3 gives the results of a simulation study of \( R_n \) for the modified rollout rule, the myopic rule and the certainty equivalence rule. Following the simulation study of Anderson and Taylor (1976), we use the \textit{a priori} bound \( K = 10 \) for the certainty equivalence rule.

We can also modify the super-optimal policy \( SP(k) \) for the present context as follows. For \( 3 \leq t \leq k \), use the myopic rule that assumes \( N(\hat{\beta}_2, \sigma^2_t(\sum_{i=1}^{T} \varphi_i \varphi_i')^{-1}) \) as the prior distribution. For \( t > k \), the value of \( \beta \) is completely revealed and one can use the maximum likelihood estimate \( \hat{\alpha}_t = t^{-1} \sum_{i=1}^{t} (y_i - \beta u_i) \) to estimate \( \alpha \). Thus \( SP(k) \) uses the control \( u_t = (y_t - \hat{\alpha}_{t-1})/\beta \) for \( t > k \). Figure 3 also gives \( R_n \) of \( SP(6) \) for benchmark comparison. It shows that the modified rollout rule has considerable improvement over the certainty equivalence rule and also over the myopic rule except when the regret of the myopic rule is close to that of the super-optimal rule.

\textbf{INSERT FIGURE 3 ABOUT HERE}
4.4. THE CASE OF UNKNOWN $\sigma_\varepsilon$

We first consider the frequentist setting of Section 4.3. The Anderson-Taylor certainty equivalence rule does not involve $\sigma_\varepsilon^2$. The myopic and rollout rules involve $\sigma_\varepsilon^2$ in (13)-(15), noting that $\text{Cov}(\hat{\theta}_t) = \sigma_\varepsilon^2 (\sum_{i=1}^{t} \varphi_i \varphi_i')^{-1}$ in the absence of a prior distribution on $\theta$. One can estimate $\sigma_\varepsilon^2$ by dividing the residual sum of squares by the degrees of freedom, i.e.,

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{s} (y_i - \hat{\theta}_s \varphi_i)^2 / (s - 2).$$

Note that this requires $s \geq 3$; in particular, both the numerator and denominator of $\hat{\sigma}_\varepsilon^2$ are 0. Replacing $\hat{\sigma}_\varepsilon^2$ by $\hat{\sigma}_{t-1}^2$ for $t \geq 4$, we can evaluate the state $X_t$ in (9) to implement the modified rollout or the myopic policy of Section 4.3. With $u_1 = -3$, $u_2 = -4$ and $u_3$ defined by the Anderson-Taylor rule for all policies considered, Table 1 gives the regret (32), at $\alpha = 0$ and $\beta = 1$, of this modified myopic policy when $\sigma_\varepsilon$ is unknown and of SP(6) that assumes $\sigma_\varepsilon$ and $\beta$ to be known for $t \geq 6$. It shows that the modified myopic and SP(6) policies still outperform the certainty equivalence policy, whose regret is also given in the table, although the improvement is substantially less than that in the top panel of Figure 3 which assume $\sigma_\varepsilon$ to be known. Each result in Table 1 is based on 1000 simulations. Because the estimate $\hat{\sigma}_{t-1}^2$ of $\sigma_\varepsilon^2$ is unreliable for $t \leq 30$, we recommend to use the modified rollout policy of Section 4.3 only for $30 < t \leq T - m$, and to use $u_t^{\text{myr}}$ for other values of $t$. Note that $V_{t-1}/V_{t-1}$ in (25) does not involve $\sigma_\varepsilon^2$ because of its cancellation in the numerator and the denominator.

In the present example with $T = 80$ and $m = 30$, and with most of the active learning in the top panel of Figure 3 occurring in the first few stages, this modified rollout policy differs little from the myopic policy and is therefore omitted from the table.

**INSERT TABLE 1 ABOUT HERE**

We can also extend the rollout and modified rollout policies in Sections 2 and 3 to the Bayesian setting in which $\sigma_\varepsilon^2$ is unknown and has an inverse gamma prior distribution. Specifically, we assume that $(2\sigma_\varepsilon^2)^{-1} \sim \text{Gamma}(g, \lambda)$ and that conditional on $\sigma_\varepsilon^2$, $\theta = (\alpha, \beta)'$ has the normal distribution with mean $\theta_0$ and covariance matrix $\sigma_\varepsilon^2 P_0$. Then the posterior distribution $L_t$ of $(\sigma_\varepsilon^2, \theta)$ given $\{(y_i, u_i), i \leq t\}$ can be described as follows: $L((2\sigma_\varepsilon^2)^{-1})$ is $\text{Gamma}(g + \frac{1}{2}t, 1/\lambda_t)$ and $L_t(\theta|\sigma_\varepsilon^2)$ is $N(\theta_t, \sigma_\varepsilon^2 P_t)$, where $\theta_t$ and $P_t$ are given by the recursion (28a, b) and the scale parameter $\lambda_t$ of the posterior gamma distribution of $(2\sigma_\varepsilon^2)^{-1}$ is given by

$$\lambda_t = \lambda^{-1} + \theta_0 P_0^{-1} \theta_0 + \sum_{i=1}^{t} y_i^2 - \theta_i' P_t^{-1} \theta_t;$$

see Box and Tiao (1973). By augmenting the state $X_t$ in (9) to

$$X_t = (E_{t-1} \alpha, E_{t-1} \beta, E_{t-1} \alpha^2, E_{t-1} \beta^2, E_{t-1} \alpha \beta, \lambda_{t-1}),$$

13
we can carry out the rollout and modified rollout procedures for the augmented Bayesian model. Note in this connection that

\[ \text{Cov}_{t-1}(\theta) = E_{t-1}\{\text{Cov}_{t-1}(\theta|\sigma^2_t)\} = E_{t-1}(\sigma^2_t P_{t-1}) = \frac{\lambda_{t-1}}{2g + t - 3} P_{t-1}, \]  

(34)

since if \((2W)^{-1}\) has a Gamma\((\tilde{g}, \tilde{\lambda})\) distribution with \(\tilde{g} > 1\), then \(W\) has the inverse gamma IG\((\tilde{g}, 2\tilde{\lambda})\) distribution and \(E(W) = 1/(2\tilde{\lambda}(\tilde{g} - 1))\). We can use (33) and (34) to compute \(v_\alpha\), \(v_\beta\) and \(v_{\alpha\beta}\) in (13), and replace \(\sigma^2_t\) in (13) and (15) by \(E_{t-1}(\sigma^2_t) = \lambda_{t-1}/(2g + t - 3)\).

5. Concluding Remarks

This paper shows how the rollout algorithm with a suitably chosen base policy can be used to compute nearly optimal policies for adaptive control in regression models. The manageable computational complexity of the algorithm enables one to carry out extensive simulation studies of the value and extent of active learning in these adaptive control problems, as we have demonstrated in Section 4. While the rollout algorithm has its roots in the policy iteration approach to DP, another direction of recent research in computationally tractable approximations for DP is related to the value iteration approach. This involves (i) approximation of the value function by basis functions and regression methods and (ii) Monte Carlo simulations to evaluate the cost-to-go function; see Bertsekas and Tsitsiklis (1996). Approximate value iteration schemes have recently been developed and applied to the valuation of path-dependent financial derivatives; see Longstaff and Schwartz (2001), Tsitsiklis and Van Roy (2001), Lai and Wong (2004) and the references therein.

Appendix

To prove (29), first note that

\[ E_{t-1}\left\{ \sum_{i=t}^{\infty} \delta^{i-1} g(X_i, u_i^{mp}) \right\} \leq E_{t-1}\left\{ \sum_{i=t}^{\infty} \delta^{i-1} g(X_i, u_i^{mp}) \right\}. \]  

(35)

From (12) and (13), it follows that

\[
E_{t-1}g(X_{i+1}, u_{i+1}^{mp}) \leq E_{t-1}\left\{ \min_u g(X_{i+1}, u) \right\} \leq \min_u E_{t-1}g(X_{i+1}, u) \\
\leq \min_u \left\{ \sigma^2 + E_{t-1}[E_i(\alpha - y)^2] + u^2[\omega + E_{t-1}(E_i\beta^2)] + 2ue_{t-1}[E_i(\alpha - y)\beta] \right\} \\
= \min_u \left\{ \sigma^2 + E_{t-1}(\alpha - y)^2 + u^2(\omega + E_{t-1}\beta^2) + 2ue_{t-1}[(\alpha - y)\beta] \right\} \\
= \min_u g(X_i, u) = g(X_i, u_{i}^{mp}),
\]

noting that \(E_{t-1}(E_iZ) = E_{t-1}Z\) for any integrable random variable \(Z\) (the so-called “tower property” of conditional expectations) and recalling the definition (9) of \(X_i\). Applying \(E_{t-2}\)
to the above inequality and then repeating with \( E_{t-3} \), etc., we obtain

\[
E_{t-1} \left\{ \sum_{i=t}^{\infty} \delta^{i-1} g(X_i, u_i^{\text{mvp}}) \right\} \leq \sum_{i=t}^{\infty} \delta^{i-1} g(X_i, u_i^{\text{mvp}}) \leq \frac{\delta^{t-1}}{1 - \delta} g(X_t, u_t^{\text{mvp}}).
\]

Putting this in (35) yields

\[
\delta^{t-1} g(X_t, u_t^{\text{opt}}) \leq E_{t-1} \left\{ \sum_{i=t}^{\infty} \delta^{i-1} g(X_i, u_i^{\text{opt}}) \right\} \leq \frac{\delta^{t-1}}{1 - \delta} g(X_t, u_t^{\text{mvp}}),
\]

implying that \( u = u_t^{\text{opt}} \) satisfies the inequality

\[
g(X_t, u) \leq g(X_t, u_t^{\text{mvp}})/(1 - \delta). \tag{36}
\]

Since \( g(X_t, u) \) is a quadratic function of \( u \) given by (13), (29) follows from (36) by solving a quadratic equation in \( u \) and using the expressions (14) and (15) for \( u_t^{\text{mvp}} \) and \( g(X_t, u_t^{\text{mvp}}) \).

The preceding argument also applies to the finite-horizon case, with \( \sum_{i=t}^{\infty} \) replaced by \( \sum_{i=t}^{T} \) and \( \delta \) set equal to 1 in (35), and with \( 1/(1 - \delta) \) replaced by \( T - t + 1 \) in (36).

**Acknowledgment**

This research is supported by the National Science Foundation.

**References**


Figure 1. Discounted regret $R_n$ of rollout (solid line), myopic (dot), and SP(3) (dash-dot), for different choices of prior covariance matrices. Top panel: $\text{Var}(\alpha) = 4$, $\text{Var}(\beta) = 9$, $\text{Cov}(\alpha, \beta) = 0$. Middle panel: $\text{Var}(\alpha) = 3$, $\text{Var}(\beta) = 6$, $\text{Cov}(\alpha, \beta) = -1$. Bottom panel: $\text{Var}(\alpha) = 2$, $\text{Var}(\beta) = 2$, $\text{Cov}(\alpha, \beta) = -0.5$. 
Figure 2. Regret $R_n$ of rollout (solid line), myopic (dot), and SP(4) (dash-dot), for different choices of prior covariance matrices. Top panel: $\text{Var}(\alpha) = 4$, $\text{Var}(\beta) = 9$, $\text{Cov}(\alpha, \beta) = 0$. Middle panel: $\text{Var}(\alpha) = 3$, $\text{Var}(\beta) = 6$, $\text{Cov}(\alpha, \beta) = -1$. Bottom panel: $\text{Var}(\alpha) = 2$, $\text{Var}(\beta) = 2$, $\text{Cov}(\alpha, \beta) = -0.5$. 
Figure 3. Regret $\bar{R}_n$ of rollout (solid line), myopic (dot), SP(6) (dash-dot), and certainty equivalence (dash) for different choices of prior covariance matrices. Top panel: $\alpha = 0$, $\beta = 1$, $u_1 = -3$, $u_2 = -4$. Middle panel: $\alpha = 5$, $\beta = 1$, $u_1 = -3$, $u_2 = -4$. Bottom panel: $\alpha = 0$, $\beta = 1$, $u_1 = 6$, $u_2 = 4$. 
Table 1. Regret $\bar{R}_n$ of certainty equivalence (CE), myopic and super-optimal SP(6) policies when $\sigma_\epsilon = 1$ is unknown, $\alpha = 0$, $\beta = 1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>25.03</td>
<td>33.95</td>
<td>34.89</td>
<td>35.46</td>
<td>35.98</td>
<td>36.31</td>
<td>37.28</td>
<td>38.23</td>
<td>39.47</td>
<td>40.54</td>
</tr>
<tr>
<td>Myopic</td>
<td>25.03</td>
<td>30.06</td>
<td>31.43</td>
<td>32.45</td>
<td>33.64</td>
<td>34.38</td>
<td>35.84</td>
<td>36.83</td>
<td>38.02</td>
<td>39.41</td>
</tr>
<tr>
<td>SP(6)</td>
<td>25.03</td>
<td>30.06</td>
<td>31.43</td>
<td>32.45</td>
<td>32.75</td>
<td>32.98</td>
<td>33.70</td>
<td>34.50</td>
<td>35.51</td>
<td>36.44</td>
</tr>
</tbody>
</table>