Tracking Adversarial Targets

Yasin Abbasi-Yadkori YASIN.ABBASIYADKORI@QUT.EDU.AU
Queensland University of Technology

Peter Bartlett BARTLETT@EECS.BERKELEY.EDU
University of California, Berkeley and QUT

Varun Kanade VKANADE@EECS.BERKELEY.EDU
University of California, Berkeley

Abstract

We study linear control problems with quadratic losses and adversarially chosen tracking targets. We present an efficient algorithm for this problem and show that, under standard conditions on the linear system, its regret with respect to an optimal linear policy grows as $O(\log^2 T)$, where T is the number of rounds of the game. We also study a problem with adversarially chosen transition dynamics; we present an exponentially-weighted average algorithm for this problem, and we give regret bounds that grow as $O(\sqrt{T})$.

1. Introduction

Consider a robot that controls an electron microscope to track a microorganism. Given the entire trajectory of the microorganism and the dynamics of the system, the optimal control can be computed. The trajectory, however, is not known in advance and the target might behave in an arbitrary fashion. In such situations, designing a controller based on some prior knowledge about the target location might be sub-optimal. It is important to take the behavior of the target into account.

We consider problems with linear transition functions and quadratic tracking losses. When the target trajectory is known in advance, the problem is called the linear quadratic (LQ) problem in the control community. The LQ problem is one of the most studied problems in the control literature and is widely applied in practice (Lai and Wei, 1982; 1987; Chen and Guo, 1987; Chen and Zhang, 1990; Fiechter, 1997; Lai and Ying, 2006; Campi and Kumar, 1998; Bittanti and Campi, 2006; Abbasi-Yadkori and Szepesvári, 2011). By principles of dynamic programming the optimal controller can be computed analytically. The solution is obtained by computing value functions, starting from the last round and recursing backward. This method needs to know the entire target sequence from the beginning and its computational complexity scales linearly with the length of the trajectory. It turns out that the optimal controller is linear in the state vector, and the value functions are quadratic in state and action.

As we discussed earlier, the assumption that the entire trajectory is known in advance is not always realistic. But what would tracking mean without a reference trajectory? To make the problem well-defined, we fix a class of mappings from states to actions (also known as policies) as our competitors. Our objective is to track the trajectory nearly as well as the best policy in the comparison class in hindsight. The standard dynamic programming procedures are not applicable when the entire sequence is not known in advance. We show that we can still have an effective tracking algorithm even if the sequence is not known in advance. The proposed algorithm is perhaps the first tracking method that can deal with infinite and unknown sequences.

We study the adversarial version of the LQ problem where an adversary designs the trajectory of the target and reveals the target location only at the end of each round. Formally, we study problems with transition dynamics and loss functions

$$x_{t+1} = Ax_t + Ba_t \,,$$
$$\ell_t(x_t, a_t) = (x_t - g_t)^\top Q(x_t - g_t) + a_t^\top a_t \,,$$

where $x_t \in \mathbb{R}^n$ is the state at round t; $a_t \in \mathbb{R}^d$ is the action; g_t is the target location; $\ell_t : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is the loss function; and A, B, Q are known matrices.[1] The matrix Q is symmetric and positive definite. The learner observes the

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

[1] We can also use this formulation for loss functions of the

sequence of states $(x_t)_t$. We make no assumptions on the trajectory sequence $(g_t)_t$, apart from its boundedness. The sequence might even be generated by an adversary. An LQ problem is defined by a 4-tuple (A, B, Q, G), where G is an upper bound on the norm of vectors g_t.

Let T be a time horizon, x_t^π be the state of the system if we run policy π for t rounds, and Π be a class of policies. Let

$$\rho_T(g_1, \dots, g_T) = \sum_{t=1}^{T} \ell_t(x_t, a_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(x_t^\pi, \pi(x_t^\pi)) \ .$$

The objective of the learner is to suffer low loss. The performance is measured by the regret defined by

$$R_T(A, B, Q, G, x_1, \Pi) = \sup_{g_1,\dots,g_T : \|g_t\| \le G} \rho_T(g_1, \dots, g_T) \ .$$

where $\|\cdot\|$ denotes the 2-norm. In what follows, we use $R_T \overset{\text{def}}{=} R_T(A, B, Q, G, x_1, \Pi)$. For simplicity, we assume that $x_1 = 0$. As the optimal policy for the classical problem with a constant target is linear (strictly speaking, affine), a reasonable choice for the class of competitors is the set of linear (affine) policies.

The problem that we describe is an instance of Markov Decision Process (MDP) problems with fixed and known dynamics, and changing loss functions. Note that the loss function depends on the target location, which can change in an arbitrary fashion. Such MDP problems were previously studied by Even-Dar et al. (2009) and Neu et al. (2010b;a). However, these papers provide results for finite MDP problems and are not applicable to problems with large state/action spaces. Our key finding is that the algorithm of Even-Dar et al. (2009) can be modified to be applicable in adversarial LQ problems. Interestingly, the resultant algorithm is identical with a Policy Iteration method (see, for example, (Howard, 1960)) with changing loss functions. Another interesting observation is that the *gain matrix* is independent of target vectors (see Lemma 4). This simplifies the design and analysis of our algorithm. We prove that the regret of the algorithm is logarithmic in the number of rounds of the game.

Finally, we also study a more adversarial problem:

$$\begin{aligned} x_{t+1} &= A_t x_t + B_t a_t, \\ \ell_t(x_t) &= (x_t - g_t)^\top Q (x_t - g_t) + a_t^\top a_t \ . \end{aligned}$$

Here, the time-varying transition matrices A_t and B_t and the target vector g_t are chosen by an adversary. In this problem, we show that under a uniform stability assumption, an exponentially-weighted average algorithm recently proposed by Abbasi-Yadkori et al. (2013) enjoys an $O(\sqrt{T})$ regret bound with respect to the class of linear policies.

2. Notation

We use $\sigma_{\min}(M)$ and $\sigma_{\max}(M)$ to denote the minimum and maximum eigenvalues of the positive semidefinite matrix M, respectively. We use $\|\cdot\|$ to denote the 2-norm of matrices and vectors, where the 2-norm of a matrix M is defined by $\|M\| = \sqrt{\sigma_{\max}(M^\top M)}$. We use $M \succ 0$ to denote that M is positive definite, while we use $M \succeq 0$ to denote that it is positive semidefinite. We use M_{ij} to denote a block of matrix M. The indices and the dimensionality of the block will be understood from the context. Similarly, v_i denotes a block of vector v.

3. Tracking Adversarial Targets

Even-Dar et al. (2009) study finite state MDP problems with fixed and known transition functions and adversarial loss functions. Their algorithm (MDP-E) in its present form is not applicable to our problem with a continuous state space. Somewhat surprisingly, we can design a variant of the MDP-E algorithm that is applicable to our tracking problem with continuous state and action spaces.

The MDP-E algorithm, shown in Figure 1, maintains an expert algorithm in each state; that is, it treats each state as a separate problem of prediction with expert advice, where each action corresponds to an expert. At round t, the product of expert recommendations over the state space defines the policy, denoted by π_t. The algorithm takes action $a_t \sim \pi_t(x_t)$ and observes the loss function ℓ_t. It computes the value function V_{π_t, ℓ_t} defined by the *Bellman Optimality Equation*

$$\begin{aligned} \exists \lambda, \forall x, a, \quad \lambda + V_{\pi_t, \ell_t}(x, a) &= \ell_t(x, a) \\ &+ \mathbb{E}_{x' \sim m(.|x,a)} \left[V_{\pi_t, \ell_t}(x', \pi_t(x')) \right] \ , \end{aligned}$$

where m defines the state transition probabilities.[2] Then, the algorithm feeds the expert algorithm in state x with $V_t(x, .) = V_{\pi_t, \ell_t}(x, .)$ as the loss function at time t. Thus, the computational cost of the MDP-E algorithm per round is $O(W + |\mathcal{X}|)$, where W is the cost of obtaining the value function and $\mathcal{X}$ is the finite state space.

Applied to the LQ problem, the value functions are defined

form $\ell_t(x_t, a_t) = (x_t - g_t)^\top Q (x_t - g_t) + a_t^\top R a_t$ for a positive definite matrix R, by writing $x_{t+1} = A x_t + (B R^{-1/2}) R^{1/2} a_t = A x_t + \widetilde{B} \widetilde{a}_t$ and $\ell_t(x_t, a_t) = (x_t - g_t)^\top Q (x_t - g_t) + a_t^\top R^{1/2} R^{1/2} a_t = (x_t - g_t)^\top Q (x_t - g_t) + \widetilde{a}_t^\top \widetilde{a}_t$.

[2] In the Bellman Optimality Equation, scalar λ is the average loss of policy π_t, and $V_{\pi_t, \ell_t}(x, a)$ is the relative goodness of state-action pair (x, a) under policy π_t. Note that under certain assumptions, the average loss is independent of the initial state, however, some states are more favorable as the policy incurs lower losses during the transient phase, starting from those states.

```
Initialize an expert algorithm in each state
for $t := 1, 2, \dots$ do
    Let $\pi_t(x_t)$ be the prediction of the expert algorithm in state $x_t$
    Take action $a_t \sim \pi_t(x_t)$
    Observe loss function $\ell_t$
    Compute $V_t = V_{\pi_t,\ell_t}$
    For all $x$, feed the expert algorithm in state $x$ with loss $V_t(x, .)$
end for
```

Figure 1. The MDP-E Algorithm

by

$$\exists\lambda, \forall x, a, \quad \lambda + V_t(x,a) = \ell_t(x,a) + V_t(Ax + Ba, \pi_t(Ax + Ba))\,, \tag{1}$$

where we use the notation $V_t = V_{\pi_t,\ell_t}$. The linear structure allows us to compute V_t implicitly for all states, thus overcoming the difficulty of the infinite state space. As we will show, a suitable expert algorithm for our problem is the Follow The Leader (FTL) algorithm that we define next.

Consider an online quadratic optimization problem where at round t the adversary chooses a quadratic loss function f_t that is defined over a convex set $D \in \mathbb{R}^k$. Simultaneously, the learner makes a prediction $p_t \in D$, suffers loss $f_t(p_t)$ and observes the loss function f_t. The regret of the learner is defined by $B_T = \sum_{t=1}^T f_t(p_t) - \min_{p\in D}\sum_{t=1}^T f_t(p)$. The FTL algorithm makes the prediction

$$p_t = \operatorname*{argmin}_{p\in D} \sum_{s=1}^{t-1} f_s(p)\,.$$

The FTL algorithm enjoys the following regret bound for quadratic losses (Cesa-Bianchi and Lugosi, 2006, Theorem 3.1):

Theorem 1 (FTL Regret bound)**.** *Assume f_t is convex, maps to $[0, C_1]$, is Lipschitz with constant C_2, and is twice differentiable everywhere with Hessian $H \succ C_3 I$. Then the regret of the Follow The Leader algorithm is bounded by $B_T \le \frac{4C_1C_2^2}{C_3}(1 + \log T)$.*

Figure 2 shows the FTL-MDP algorithm for the linear quadratic tracking problem. It corresponds to the MDP-E algorithm, with FTL as the expert algorithm for each state. The algorithm starts at state $x_1 = 0$ and the first policy is chosen to be $\pi_1(x) = -K_* x$, where K_* is a gain matrix that will be defined later.[3] The algorithm computes the total loss in each state, shown by $V'(x, .) = \sum_{s=1}^{t-1} V_s(x, .)$.

```
$x_1 = 0$
$\forall x, \pi_1(x) = -K_* x$ (6)
for $t := 1, 2, \dots$ do
    Take action $a_t = \pi_t(x_t)$ and suffer the loss $\ell_t(x_t, a_t)$
    Move to the state $x_{t+1} = Ax_t + Ba_t$
    Compute the value function $V_t = V_{\pi_t,\ell_t}$ (2)
    Let $V' = \sum_{s=1}^t V_s$
    Obtain the policy by solving $\nabla_a V'(x, a) = 0$: $\pi_{t+1}(x) = -K_{t+1}x + c_{t+1}$
end for
```

Figure 2. FTL-MDP: The Follow the Leader Algorithm for Markov Decision Processes

The FTL strategy chooses the greedy action in each state, which is obtained by minimizing $V'(x, .)$. As the next lemma shows, value functions computed in the FTL-MDP algorithm are always quadratic and thus, the function V' is always quadratic in state and action. This implies that policies are linear in state.

Lemma 2. *Consider the* MDP-E *algorithm applied to the adversarial LQ problem (A, B, Q, G). Let the expert algorithm be the FTL strategy. Assume the first policy is chosen to be an arbitrary linear policy, $\pi_1(x) = -K_1 x + c_1$. Then, for appropriate matrices P_t and L_t, the value function at time t has the form of*

$$V_t(x,a) = \begin{pmatrix} x^\top & a^\top \end{pmatrix} P_t \begin{pmatrix} x \\ a \end{pmatrix} + L_t^\top \begin{pmatrix} x \\ a \end{pmatrix}\,,$$

and the policy chosen by the algorithm at time t is $\pi_t(x) = -K_t x + c_t$, where $K_t = (\sum_{s=1}^{t-1} P_{s,22})^{-1} \sum_{s=1}^{t-1} P_{s,21}$ and $c_t = -(\sum_{s=1}^{t-1} P_{s,22})^{-1} \sum_{s=1}^{t-1} L_{s,2}/2$ and $P_{s,ij}$ and $L_{s,i}$ are the ijth and ith blocks of matrix P_s and vector L_s, respectively. (Here the block structure naturally appears as components corresponding to the state and action.)

The proof uses the following lemma that shows that the value of a linear policy is quadratic.

Lemma 3. *Consider the LQ problem (A, B, Q, G) with fixed target g_*. Let K be a matrix such that $\|A - BK\| < 1$. The value function of policy $\pi(x) = -Kx + c$ has a quadratic form*

$$V_{\pi,\ell}(x,a) = \begin{pmatrix} x^\top & a^\top \end{pmatrix} P \begin{pmatrix} x \\ a \end{pmatrix} + L^\top \begin{pmatrix} x \\ a \end{pmatrix}\,, \tag{2}$$

where $P = P(K)$ and L are solutions to equations

$$P = \begin{pmatrix} A^\top \\ B^\top \end{pmatrix} \begin{pmatrix} I & -K^\top \end{pmatrix} P \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} + \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \tag{3}$$

[3] Adopting a convention from feedback control, we represent linear policies with a negative sign.

and

$$L^\top = \left(L^\top + 2\begin{pmatrix}0 & c^\top\end{pmatrix} P\right)\begin{pmatrix} I \\ -K\end{pmatrix}\begin{pmatrix}A & B\end{pmatrix} - \begin{pmatrix}2g_*^\top Q & 0\end{pmatrix} . \tag{4}$$

Further, the loss in the limiting state x_∞^π is

$$\lambda = g_*^\top Q g_* + c^\top P_{22} c + L_2^\top c \, . \tag{5}$$

The proof can be found in Appendix A.

Proof of Lemma 2. We prove the lemma by induction. By Lemma 3, the value of the first policy, $\pi_1(x) = -K_1 x + c_1$, has the form of

$$V_1 = V_{\pi_1,\ell_1}(x,a) = \begin{pmatrix}x^\top & a^\top\end{pmatrix} P_1 \begin{pmatrix}x \\ a\end{pmatrix} + L_1^\top \begin{pmatrix}x \\ a\end{pmatrix} ,$$

for some matrices P_1 and L_1. This establishes the base case.

Next assume the value functions up to time $t-1$ are all quadratic. Because we use a FTL strategy as our expert algorithm in states, the policy in state x in round t can be computed by

$$\pi_t(x) = \operatorname*{argmin}_a \sum_{s=1}^{t-1} V_s(x,.) \, .$$

We obtain the policy by setting $\nabla_a \sum_{s=1}^{t-1} V_s(x,.) = 0$. Letting

$$V_s(x,a) = \begin{pmatrix}x^\top & a^\top\end{pmatrix} P_s \begin{pmatrix}x \\ a\end{pmatrix} + L_s^\top \begin{pmatrix}x \\ a\end{pmatrix}$$

for $s = 1 \dots (t-1)$, we get that

$$V'(x,a) = \sum_{s=1}^{t-1} V_s(x,a) = \begin{pmatrix}x^\top & a^\top\end{pmatrix} \widetilde{P}_t \begin{pmatrix}x \\ a\end{pmatrix} + \widetilde{L}_t^\top \begin{pmatrix}x \\ a\end{pmatrix} ,$$

where $\widetilde{P}_t = \sum_{s=1}^{t-1} P_s$ and $\widetilde{L}_t = \sum_{s=1}^{t-1} L_s$. Taking the gradient with respect to the second argument and setting to zero, we get that,

$$\nabla_a V'(x,a) = 2\widetilde{P}_{t,21} x + 2\widetilde{P}_{t,22} a + \widetilde{L}_{t,2} = 0$$

and thus,

$$a = -\widetilde{P}_{t,22}^{-1} \widetilde{P}_{t,21} x - \frac{1}{2} \widetilde{P}_{t,22}^{-1} \widetilde{L}_{t,2} \, .$$

Thus, the policy can be compactly written as

$$\forall x, \quad \pi_t(x) = -K_t x + c_t \, ,$$

where $K_t = \widetilde{P}_{t,22}^{-1} \widetilde{P}_{t,21}$ and $c_t = -\widetilde{P}_{t,22}^{-1} \widetilde{L}_{t,2}/2$. Given this linear policy, we get the quadratic value function V_t from Equation (2).

□

Lemma 2 implies that the MDP-E algorithm can be efficiently implemented in the adversarial LQ problem.

Before stating the main result of this paper, we describe certain assumptions and definitions from the control literature. (See, for example, (Bertsekas, 2001)).

Definition 1. *A pair (A, B), where A is an $n \times n$ matrix and B is an $n \times d$ matrix, is said to be controllable if the $n \times nd$ matrix $[B \; AB \; \dots \; A^{n-1}B]$ has full rank. A pair (A, C), where A is an $n \times n$ matrix and C is an $d \times n$ matrix, is said to be observable if the pair $(A^\top, C^\top)$ is controllable.*

Roughly speaking, controllability implies that the state can be moved arbitrarily by changing the actions, while observability implies that the state can be externally measured. We assume that the system is controllable and observable. These assumptions are standard in the literature, and will allow a closed form expression for the optimal control law.

Assumption A1. ***(Controllability and observability)*** The pair (A, B) is controllable and the pair $(A, Q^{1/2})$ is observable.

Under this assumption, the *gain matrix* is stable, i.e. there exists $\rho \in (0, 1)$ such that $\|A - BK_*\| \le \rho < 1$, where

$$K_* = (I + B^\top S B)^{-1} B^\top S A \tag{6}$$

is the gain matrix (Bertsekas, 2001), and S is the solution of the *Riccati equation*

$$S = Q + A^\top S A - A^\top S B (I + B^\top S B)^{-1} B^\top S A \, .$$

Interestingly, as the next lemma shows, all gain matrices are equal. The proof can be found in Appendix A.

Lemma 4. *Consider the* FTL-MDP *algorithm. Let $P_* = P(K_*)$ as defined by (3). If we choose $K_1 = K_*$, then all gain matrices are equal, $K_* = K_1 = K_2 = K_3 = \dots$, and hence $P_* = P_1 = P_2 = P_3 = \dots$.*

Lemma 4 shows that gain matrices are independent of target vectors and can be computed by assuming that all target vectors are zero. Given the fixed gain matrix, the system is driven to a desired target position by changing the bias term of the linear policy.

We represent the linear policy $\pi(x) = -Kx + c$ by the pair $\pi = (K, c)$. The class of (K', C')-bounded, ρ-stable linear policies is defined by $\Pi = \{\pi = (K, c) \, : \, \|A - BK\| \le \rho, \, \|K\| \le K', \, \|c\| \le C'\}$.

Theorem 5. *For a controllable, observable LQ problem (A, B, Q, G), the regret of the* FTL-MDP *algorithm with respect to the class of (K', C')-bounded, ρ-stable linear policies is $O(\log^2 T)$, where the hidden constants are polynomials in*

$\|A\|, \|B\|, \|Q\|, G, S, \|P_*\|, 1/\lambda_{\min}(P_*), \|K_*\|, K', C'$, *and* $1/(1-\rho)$.

The hidden constants are all small order polynomials. As we will show in the next section, a careful asymptotic analysis gives us an asymptotic $O(\log T)$ bound. It is an open problem to show a finite-time $O(\log T)$ regret bound with polynomial constants.

4. Analysis

Let $x_\infty^\pi = \lim_{t\to\infty} x_t^\pi$ be the limiting state under (K', C')-bounded and ρ-stable linear policy $\pi = (K, c)$. In Lemma 6 we show that this limit exists. Let $\lambda_t(\pi) = \ell_t(x_\infty^\pi, \pi(x_\infty^\pi))$ be the loss of policy π in state x_∞^π. We decompose the regret

$$\begin{aligned} R_T &= \sum_{t=1}^T \ell_t(x_t, a_t) - \sum_{t=1}^T \ell_t(x_t^\pi, \pi) \\ &= \sum_{t=1}^T \ell_t(x_t, a_t) - \sum_{t=1}^T \lambda_t(\pi_t) + \sum_{t=1}^T \lambda_t(\pi_t) \\ &\quad - \sum_{t=1}^T \lambda_t(\pi) + \sum_{t=1}^T \lambda_t(\pi) - \sum_{t=1}^T \ell_t(x_t^\pi, \pi) \end{aligned}$$

Let

$$\begin{aligned} \alpha_T &= \sum_{t=1}^T \ell_t(x_t, a_t) - \sum_{t=1}^T \lambda_t(\pi_t)\,, \\ \beta_T &= \sum_{t=1}^T \lambda_t(\pi_t) - \sum_{t=1}^T \lambda_t(\pi)\,, \\ \gamma_T &= \sum_{t=1}^T \lambda_t(\pi) - \sum_{t=1}^T \ell_t(x_t^\pi, \pi)\,. \end{aligned}$$

The terms α_T and γ_T correspond to the difference between the losses of the policies between their stationary and transient states. The term β_T measures the regret with respect to the optimal policy. The rest of this section is devoted to providing bounds on these terms.

4.1. Bounding α_T

To bound α_T, we need to show that sum of terms $\ell_t(x_t, a_t) - \lambda_t(\pi_t)$ is small. Let $x_\infty^{\pi_t} = \lim_{s\to\infty} x_s^{\pi_t}$. We will show that this limit exists. Because $\lambda_t(\pi_t) = \ell_t(x_\infty^{\pi_t}, \pi_t(x_\infty^{\pi_t}))$ and $\ell_t(x_t, a_t) = \ell_t(x_t, \pi_t(x_t))$, we need to show that $x_\infty^{\pi_t}$ is close to x_t. This is done in a number of steps. First, we obtain the limiting state $x_\infty^{\pi_t}$ (Lemma 6 and the discussion after that). Then, we show that the chosen policy is slowly changing. Given these two results, we bound $\|x_t - x_\infty^{\pi_t}\|$, which is then used to bound α_T.

First, we study the behavior of the state vector under any bounded and stable policy. We show that the policy converges to its stationary state exponentially fast.

Lemma 6. *The limiting state $x_\infty^\pi = \lim_{t\to\infty} x_t^\pi$ under a (K', C')-bounded and ρ-stable linear policy $\pi = (K, c)$ exists and is equal to $x_\infty^\pi = (I - A + BK)^{-1}Bc$. Further, we have that $\|x_t^\pi\| \le \|B\| C'/(1-\rho)$ and*

$$\left\|x_{t+1}^\pi - x_\infty^\pi\right\| \le \rho^t \left\|(I - A + BK)^{-1}Bc\right\| .$$

Proof. We have that

$$\begin{aligned} x_{t+1}^\pi &= (A - BK)x_t^\pi + Bc \\ &= (A - BK)^2 x_{t-1}^\pi + (A - BK)Bc + Bc \\ &= \dots \\ &= (A - BK)^t x_1 + \sum_{s=1}^t (A - BK)^{t-s}Bc \\ &= 0 + \sum_{s=0}^{t-1} (A - BK)^s Bc\,, \end{aligned}$$

where we used $x_1 = 0$ in the last equality. Thus, as t goes to infinity, $x_t^\pi \to (I - A + BK)^{-1}Bc$. This also implies that $\|x_t^\pi\| \le \|B\| C'/(1-\rho)$. It is also easy to see that

$$\begin{aligned} \left\|x_{t+1}^\pi - x_\infty^\pi\right\| &= \left\|\sum_{s=0}^\infty (A - BK)^s Bc - \sum_{s=0}^{t-1} (A - BK)^s Bc\right\| \\ &= \left\|(A - BK)^t (I - A + BK)^{-1}Bc\right\| \\ &\le \|A - BK\|^t \left\|(I - A + BK)^{-1}Bc\right\| \\ &\le \rho^t \left\|(I - A + BK)^{-1}Bc\right\| . \end{aligned}$$

□

Note that even if $\pi = (K, c) \notin \Pi$, but $(A - BK)$ is stable, the above argument is still valid and we get a similar result. In particular,

$$x_\infty^{\pi_t} = (I - A + BK_*)^{-1}Bc_t\,. \tag{7}$$

Letting C be an upper bound on $\|c_t\|$ for $t \le T$, with a similar argument we can also show that

$$\|x_t\| \le \frac{\|B\| C}{1-\rho}\,. \tag{8}$$

In what follows, we use X to denote the upper bound on the norm of state vector, $X \overset{\text{def}}{=} \frac{\|B\|C}{1-\rho}$. Next, we prove that the chosen policy is slowly changing and the bias term in policies is bounded by C, where

$$\begin{aligned} C &= \|D\| G H'\,, \\ D &= P_{*,22}^{-1} B^\top (I - A + BK_*)^{-\top} Q\,, \\ H'' &= \sqrt{1 + \|K_*\|^2 \|B\|^2/(1-\rho)^2} + \frac{\|K_*\| \|B\|}{1-\rho}\,, \\ H' &= H'' \sqrt{1 + \|K_*\|^2 \|B\|^2/(1-\rho)^2}\,, \end{aligned}$$

where P_* denotes the solution to Equation (3), corresponding to gain matrix K_*. The proof can be found in Appendix A.

Lemma 7. *We have that*

$$(i).\ \|c_t\| \le C\,, \qquad (ii).\ \|c_t - c_{t-1}\| \le \frac{\|D\|\,G + 2C}{t-1}\ .$$

Next, we show that the limiting state of policy π_t (i.e. $\lim_{s\to\infty} x_s^{\pi_t}$) is close to the state at time t. The proof can be found in Appendix A.

Lemma 8. *If* $t > \lceil \log(T-1)/\log(1/\rho)\rceil$*, then*

$$\begin{aligned}\|x_t - x_\infty^{\pi_t}\| \le \frac{\|B\|\,(\|D\|\,G + 2C)}{1-\rho}\Bigg(&\frac{1+\log(t-1)}{t-1}\\ &+ \frac{\log(t-1)}{\log(1/\rho)}\left(\frac{1}{t - \log(t-1)/\log(1/\rho)}\right)\Bigg)\\ &+ \rho^{t-1}\frac{\|B\|\,C}{1-\rho}\ .\end{aligned}$$

Also, we have that

$$\begin{aligned}\sum_{t=1}^{T}\|x_t - x_\infty^{\pi_t}\| \le &\frac{1}{1-\rho}\Bigg(4\,\|B\|\,C\left\lceil\frac{\log T}{\log(1/\rho)}\right\rceil + \frac{\|B\|\,C}{1-\rho}\\ &+ \|B\|\,(\|D\|\,G + 2C)(1+\log T)\\ &\times\left(1 + \log T + \frac{\log T}{\log(1/\rho)}\right)\Bigg)\ .\end{aligned}$$

Remark 9. *This lemma shows an* $O(\log^2 T)$ *bound on* $\sum_{t=1}^{T}\|x_t - x_\infty^{\pi_t}\|$*. As we will see, this leads to an* $O(\log^2 T)$ *regret bound. Let* $\epsilon_t = \|x_t - x_\infty^{\pi_t}\|$*. To get an* $O(\log T)$ *regret bound, we need to show that* $\epsilon_t = H_1/t$ *for a constant* H_1*. A careful examination of proof of Lemma 8 reveals that* $\epsilon_t = H_2\sum_{s=1}^{t-1}\rho^s/(t-s) + H_3$ *for constants* H_2 *and* H_3*. Let* $f(t) = \sum_{s=1}^{t-1}\rho^s/(t-s)$*. We want that* $f(t) = H_4/t$ *for a constant* H_4 *so that we get an* $O(\log T)$ *regret bound.*

To show this, we argue as follows: first establish the simple recurrence $f(t+1) = \rho f(t) + \rho/t$*. This implies that* $f(t) \to 0$ *as* $t \to \infty$*. Now, define* $g(t+1) = t f(t+1)$*. Thus* $g(t+1) = \rho g(t) + \rho f(t) + \rho$*. Since* $\rho < 1$ *and* $\lim_{t\to\infty} f(t) = 0$*,* $\lim_{t\to\infty} g(t)$ *exists and a simple calculation shows that this is* $\rho/(1-\rho)$*. This in fact implies that* $f(t) \to \rho/(t(1-\rho))$ *asymptotically, which in turn implies that regret is* $O(\log T)$ *asymptotically.*

Now we are ready to bound α_T.

Lemma 10. *Let*

$$\begin{aligned}Z_1 &= 2(C\,\|K_*\| + G\,\|Q\|) + 2(\|Q\| + \left\|K_*^2\right\|)\frac{\|B\|\,C}{1-\rho}\,,\\ Z_2 &= 4\,\|B\|\,C\left\lceil\frac{\log T}{\log(1/\rho)}\right\rceil + \frac{\|B\|\,C}{1-\rho}\\ Z_3 &= \|B\|\,(\|D\|\,G + 2C)(1+\log T)\\ &\qquad\times\left(1 + \log T + \frac{\log T}{\log(1/\rho)}\right)\ .\end{aligned}$$

Then we have that

$$\alpha_T \le Z_1(Z_2 + Z_3)/(1-\rho)\ .$$

Proof. For policy $\pi = (K, c)$, we have

$$\begin{aligned}\ell_t(x,\pi) = &\, x^\top(Q + K^\top K)x\\ &- 2(c^\top K + g_t^\top Q)x + c^\top c + g_t^\top Q g_t\ .\end{aligned}$$

For policy $\pi_t = (K_t, c_t) = (K_*, c_t)$, define $S_t = Q + K_*^\top K_*$ and $d_t = 2(c_t^\top K_* + g_t^\top Q)$. We write

$$\begin{aligned}\alpha_T &= \sum_{t=1}^{T}\left(x_t^\top S_t x_t - d_t x_t\right) - \sum_{t=1}^{T}\left(x_\infty^{\pi_t\top} S_t x_\infty^{\pi_t} - d_t x_\infty^{\pi_t}\right)\\ &= \sum_{t=1}^{T} d_t(x_\infty^{\pi_t} - x_t) + \sum_{t=1}^{T}\left(\left\|S_t^{1/2}x_t\right\| - \left\|S_t^{1/2}x_\infty^{\pi_t}\right\|\right)\\ &\qquad\times\left(\left\|S_t^{1/2}x_t\right\| + \left\|S_t^{1/2}x_\infty^{\pi_t}\right\|\right)\\ &\le \sum_{t=1}^{T} d_t(x_\infty^{\pi_t} - x_t) + \sum_{t=1}^{T}\left\|S_t^{1/2}(x_t - x_\infty^{\pi_t})\right\|\\ &\qquad\times\left(\left\|S_t^{1/2}x_t\right\| + \left\|S_t^{1/2}x_\infty^{\pi_t}\right\|\right)\\ &\le \sum_{t=1}^{T}\left(\|d_t\| + \left\|S_t^{1/2}\right\|\left(\left\|S_t^{1/2}x_t\right\| + \left\|S_t^{1/2}x_\infty^{\pi_t}\right\|\right)\right)\\ &\qquad\times\|x_\infty^{\pi_t} - x_t\|\\ &\le Z_1\sum_{t=1}^{T}\|x_\infty^{\pi_t} - x_t\|\\ &\le Z_1(Z_2 + Z_3)/(1-\rho)\ .\end{aligned}$$

where we used Lemma 8 in the last step. □

4.2. Bounding β_T

The term β_T is bounded by showing a reduction to regret minimization algorithms (in this case, the FTL algorithm). To use the regret bound of the FTL algorithm (Theorem 1), we need to show boundedness of the value functions.

Lemma 11. *Let* $X' = \|B\|\,C'/(1-\rho)$*,* $U = \max\{\|K_*\|, K'\}\max\{X, X'\} + \max\{C, C'\}$*,* $V = \|P_*\|\,(X' + U)^2 + \frac{2}{1-\rho}\,(G\,\|Q\| + \rho C\,\|P_*\|)\,(X' + U)$*, and*

$F = 2\|P_{*,22}\| U + \|P_{*,21}\| X' + \frac{2}{1-\rho}(G\|Q\| + \rho C\|P_*\|)$. *For any t, and any (K', C')-bounded, ρ-stable linear policy $\pi = (K, c)$,*

$$(i).\ \|a_t\| \le U\,,$$
$$(ii).\ \|-Kx_\infty^\pi + c\| \le U\,,$$
$$(iii).\ \|-K_* x_\infty^\pi + c_t\| \le U\;.$$

For any action such that $\|a\| \le U$,

$$(iv).\ V_t(x_\infty^\pi, a) \le V\;.$$

Further, $V_t(x_\infty^\pi, .)$ is Lipschitz in its second argument with constant F. Finally, the Hessians of the value functions are positive definite and $H(V_t(x_\infty^\pi, .)) \succ 2I$.

The proof can be found in Appendix A.

Lemma 12. *We have*

$$\beta_T \le 2VF^2(1 + \log T)\;.$$

Proof. To bound β_T, first note that

$$\begin{aligned} V_{\pi,\ell}(x_\infty^{\pi'}, \pi') &= \ell(x_\infty^{\pi'}, \pi') - \lambda_{\pi,\ell} + V_{\pi,\ell}(x_\infty^{\pi'}, \pi) \\ &= \lambda_{\pi',\ell} - \lambda_{\pi,\ell} + V_{\pi,\ell}(x_\infty^{\pi'}, \pi)\;. \end{aligned}$$

Thus,

$$\lambda_{\pi,\ell} - \lambda_{\pi',\ell} = V_{\pi,\ell}(x_\infty^{\pi'}, \pi) - V_{\pi,\ell}(x_\infty^{\pi'}, \pi')\;.$$

Thus,

$$\beta_T \le \sum_{t=1}^T (V_{\pi_t,\ell_t}(x_\infty^\pi, \pi_t) - V_{\pi_t,\ell_t}(x_\infty^\pi, \pi))\;.$$

Now notice that $V_{\pi_t,\ell_t}(x_\infty^\pi, .)$ is the loss function that is fed to the FTL strategy in state x_∞^π. Lemma 11 shows that conditions of Theorem 1 are satisfied. Thus, we get the result from the regret bound for the FTL algorithm (Theorem 1):

$$\beta_T \le 2VF^2(1 + \log T)\;.$$

□

4.3. Bounding γ_T

Finally, we bound γ_T. The proof is similar to the proof of Lemma 10 and can be found in Appendix A.

Lemma 13. *Let $Z_1' = (CK' + G\|Q\|) + (\|Q\| + K'^2)\|B\| C'/(1-\rho)$. Then we have that*

$$\gamma_T \le \frac{2Z_1' \|B\| C'}{(1-\rho)^2}\;.$$

4.4. Putting Everything Together

Proof of Theorem 5. The regret bound follows from Lemmas 10, 12, and 13.

□

5. Adversarially Chosen Transition Matrices

Our results can be extended to LQ problems with adversarial transition matrices,

$$\begin{aligned} x_{t+1} &= A_t x_t + B_t a_t, \qquad (9) \\ \ell_t(x_t, a_t) &= (x_t - g_t)^\top Q (x_t - g_t) + a_t^\top a_t\;. \end{aligned}$$

Here, transition matrices A_t and B_t and the target vector g_t are chosen by an adversary. Once again, we measure the regret with respect to the class of linear policies. Thus, any policy π is identified by some pair (K, c) such that $\pi(x) = -Kx + c$.

The only no-regret algorithm for this setting is the result of Abbasi-Yadkori et al. (2013) who propose an exponentially-weighted average algorithm and analyze it under a mixing assumption. Similarly, we make the following assumption:

Assumption A2. ***(Uniform Stability)*** The choices of the learner and the adversary are restricted to sets $\mathcal{K} \times \mathcal{C} \subset \mathbb{R}^{d\times n} \times \mathbb{R}^d$ and $\mathcal{A} \times \mathcal{B} \subset \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d}$, respectively. There exists $0 < \rho < 1$ such that for any $A \in \mathcal{A}$ and $B \in \mathcal{B}$, and any $K \in \mathcal{K}$,

$$\|A - BK\| < \rho\;.$$

Further, there exits $K', C' > 0$ such that for any $K \in \mathcal{K}$ and $c \in \mathcal{C}$, $\|K\| \le K'$ and $\|c\| \le C'$.

The proposed algorithm for the LQ problem (9) is shown in Figure 3. The algorithm maintains a distribution over policies. The distribution has the form of

$$q_t(\pi) \propto e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_s^\pi, \pi)}\,, \quad \eta > 0\;. \qquad (10)$$

The following theorem bounds the regret of this algorithm.

Theorem 14. *Consider a uniformly stable system. The regret of the algorithm in Figure 3 with respect to a class of policies $|\Pi|$ is bounded by $O(\sqrt{T \log |\Pi|} + \log |\Pi|)$.*

Proof. We prove the theorem by showing that conditions of Theorem 1 in (Abbasi-Yadkori et al., 2013) are satisfied.

Uniform mixing assumption: Let $P(\pi, A, B)$ be the transition probability matrix of policy $\pi = (K, c) \in \mathcal{K} \times \mathcal{C}$ under transition dynamics $(A, B) \in \mathcal{A} \times \mathcal{B}$. Let p_1 and p_1' be two distributions over the state space and $p_2 = p_1 P(\pi, A, B)$

```
$q_0$: the uniform distribution over $\Pi$, $\eta = 1/\sqrt{T}$
for $t := 1, 2, \ldots$ do
    With probability $\beta_t = w_{\pi_{t-1},t-1}/w_{\pi_{t-1},t-2}$ choose the previous policy, $\pi_t = \pi_{t-1}$, while with probability $1 - \beta_t$, choose $\pi_t$ based on the distribution $q_t$.
    Learner takes the action $a_t = -K_t x_t + c_t$. Simultaneously, adversary chooses transition matrices $A_t$ and $B_t$ and target vector $g_t$.
    Learner suffers loss $\ell_t(x_t, a_t)$ and observes $A_t, B_t, g_t$.
    Update state: $x_{t+1} = A_t x_t + B_t a_t$.
    Update the distribution $q_t(\pi) \propto w_{\pi,t}$, where $w_{\pi,t} = e^{-\eta \sum_{s=1}^{t} \mathbb{E}[\ell_s(x_s^\pi, \pi)]}$
end for
```

Figure 3. The Exponentially Weighted Algorithm for Linear Quadratic Problems

and $p'_2 = p'_1 P(\pi, A, B)$. Let $1 \le k \le n$ be rank of $M = (A - BK)$. Let M' be a $k \times k$ matrix whose eigenvalues are the nonzero eigenvalues of M. For $x \in \mathbb{R}^n$, let $x_r(x) \in \mathbb{R}^k$ be a parameterization of the component of x that is on the row space of M. Similarly, define $x_n(x) \in \mathbb{R}^{n-k}$ that corresponds to the orthogonal component on the null space of M. For $u \in \mathbb{R}^k$ and $v \in \mathbb{R}^{n-k}$, let $x(u, v)$ be a vector in $\mathbb{R}^n$ such that $x_r(x(u, v)) = u$ and $x_n(x(u, v)) = v$. Finally, let $p_r(u) = \int_{\mathbb{R}^{n-k}} p_1(x(u, v))dv$. Using integration by substitution, we get that

$$\begin{aligned}
\|p_2 - p'_2\|_1 &= \int_{\mathbb{R}^k} |p_2(y) - p'_2(y)| \, dy \\
&= \int_{\mathbb{R}^k} |p_r(u) - p'_r(u)| \, |\det(M')| \, du \\
&\le \rho^k \int_{\mathbb{R}^k} |p_r(u) - p'_r(u)| \, du \\
&\le \rho^k \int_{\mathbb{R}^k} \left| \int_{\mathbb{R}^{n-k}} p_1(x(u, v))dv \right. \\
&\qquad\qquad \left. - \int_{\mathbb{R}^{n-k}} p'_1(x(u, v))dv \right| du \\
&\le \rho \int_{\mathbb{R}^k} \int_{\mathbb{R}^{n-k}} |p_1(x(u, v)) - p'_1(x(u, v))| \, dv du \\
&= \rho \int_{\mathbb{R}^n} |p_1(x) - p'_1(x)| \, dx = \rho \, \|p_1 - p'_1\|_1 \ .
\end{aligned}$$

This shows that the uniform mixing assumption of Abbasi-Yadkori et al. (2013) is satisfied with the choice of mixing time $\tau = 1/\log(1/\rho)$.

Bounded losses: With an argument similar to the proof of Lemma 6, we can show that the state is bounded. This, together with the boundedness of sets $\mathcal{K}$ and $\mathcal{C}$, give that the action is bounded. Thus, all losses are bounded. □

Results of Abbasi-Yadkori et al. (2013) also apply to the simpler setting of Section 3. However, sampling from the distribution (10) can be computationally expensive, whereas the FTL-MDP algorithm is computationally efficient.

6. Conclusions and Future Work

We studied the problem of controlling linear systems with adversarial quadratic tracking losses, competing with the family of policies that compute actions as linear functions of the state. We presented an algorithm whose regret, with respect to such linear policies, is logarithmic in the number of rounds of the game. An interesting direction for future work is to consider more complex families of policies, such as the class of linear policies with a limited number of switches.

Existing tracking algorithms require the target sequence to be known in advance. Also their computational complexity scales linearly with the length of the trajectory. The main difficulty in the setting studied here, is the adversarial nature of target vectors, which is very different from the classical setting. The key advance is to show how the idea of Even-Dar et al. (2009) (instantiating expert algorithms in all states) can be applied to the LQ problem, which has a continuous and unbounded state space. This is done by showing that the sequence of value functions and policies will be quadratic and linear respectively, if we choose the right expert algorithm (FTL). The compact representation of value functions and policies allows an efficient implementation of the FTL algorithm.

We showed how a related approach can be applied to adversarially chosen changing linear dynamics. Unfortunately, this algorithms is computationally expensive. A more challenging problem is to design efficient algorithms for the case of adversarially chosen changing transition matrices. An interesting open problem is whether there is an efficient no-regret algorithm, or whether a *computational lower* bound can be established.

It might be possible to extend our results to LQ problems with fixed, but unknown transition matrices of the form:

$$\begin{aligned}
x_{t+1} &= Ax_t + Ba_t + w_{t+1}, \\
\ell_t(x_t, a_t) &= (x_t - g_t)^\top Q(x_t - g_t) + a_t^\top a_t \ ,
\end{aligned}$$

where w_{t+1} is a sub-Gaussian noise and matrices A and B are unknown. We expect this extension to be fairly straighfoward using techniques from (Neu et al., 2012) and (Abbasi-Yadkori and Szepesvári, 2011). Our approach is similar to Neu et al. (2012), with the difference that we use FTL instead of FPL.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, 2011.

Yasin Abbasi-Yadkori, Peter Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *NIPS*, 2013.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2001.

S. Bittanti and M. C. Campi. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information and Systems*, 6(4):299–320, 2006.

M. C. Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6): 1890–1907, 1998.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

H. Chen and L. Guo. Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost. *SIAM Journal on Control and Optimization*, 25(4): 845–867, 1987.

H. Chen and J. Zhang. Identification and adaptive control for systems with unknown orders, delay, and coefficients. *Automatic Control, IEEE Transactions on*, 35(8): 866 –877, August 1990.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

C. Fiechter. Pac adaptive control of linear systems. In *in Proceedings of the 10th Annual Conference on Computational Learning Theory, ACM*, pages 72–80. Press, 1997.

R. A. Howard. *Dynamic Programming and Markov Processes*. MIT, 1960.

T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):pp. 154–166, 1982.

T. L. Lai and C. Z. Wei. Asymptotically efficient self-tuning regulators. *SIAM Journal on Control and Optimization*, 25:466–481, March 1987.

T. L. Lai and Z. Ying. Efficient recursive estimation and adaptive control in stochastic regression and armax models. *Statistica Sinica*, 16:741–772, 2006.

Gergely Neu, András György, and András Antos Csaba Szepesvári. Online Markov decision processes under bandit feedback. In *NIPS*, 2010a.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest path problem. In *COLT*, 2010b.

Gergely Neu, András György, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *AISTATS*, 2012.

A. Proofs

Proof of Lemma 3. Consider the Bellman equation

$$\lambda + V_{\pi,\ell}(x,a) = \ell(x,a) + V_{\pi,\ell}(Ax+Ba, \pi(Ax+Ba)) \ .$$

We prove the lemma by showing that the given quadratic form is the unique solution of the Bellman equation.

Let $z = (x \quad a)$ and

$$z' = \begin{pmatrix} Ax+Ba \\ -K(Ax+Ba)+c \end{pmatrix} = \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} x \\ a \end{pmatrix} + \begin{pmatrix} 0 \\ c \end{pmatrix} \ .$$

We guess a quadratic form for the value functions and write

$$\lambda + z^\top P z + L^\top z = (x-g_*)^\top Q(x-g_*) + a^\top a + z'^\top P z' + L^\top z' \ .$$

The above equation has a solution if

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} A^\top \\ B^\top \end{pmatrix} \begin{pmatrix} I & -K^\top \end{pmatrix} P \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} + \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} , \tag{11}$$

and

$$L^\top = \begin{pmatrix} L_1^\top & L_2^\top \end{pmatrix} = \left(L^\top + 2\begin{pmatrix} 0 & c^\top \end{pmatrix} P\right) \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} - \begin{pmatrix} 2g_*^\top Q & 0 \end{pmatrix} , \tag{12}$$

and

$$\lambda = g_*^\top Q g_* + c^\top P_{22} c + L_2^\top c \ .$$

We have that

$$\left\| \begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} I \\ -K \end{pmatrix} \right\| = \|A - BK\| < 1 \ .$$

This implies that iterative equations (11) and (12) have a unique solution. Thus, the quadratic form is the solution of the Bellman equation. □

Proof of Lemma 4. From Lemma 3, we have that

$$P_t = \begin{pmatrix} A^\top \\ B^\top \end{pmatrix} \begin{pmatrix} I & -K_t^\top \end{pmatrix} P_t \begin{pmatrix} I \\ -K_t \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} + \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix}$$

and

$$L_t^\top = \left(L_t^\top + 2\begin{pmatrix} 0 & c_t^\top \end{pmatrix} P_t\right) \begin{pmatrix} I \\ -K_t \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} - \begin{pmatrix} 2g_t^\top Q & 0 \end{pmatrix} \ .$$

Notice that the value of P_t depends only on the values of A, B, and K_t, which in turn, by Lemma 2, depend only on $\{K_1, P_1, \dots, P_{t-1}\}$. Thus, matrix P_t is determined by K_1 independently of the adversarial choices $\{g_1, \dots, g_t\}$.

In the absence of adversarial vectors, the optimal policy has the form of $\pi(x) = -K_* x$, where $K_* = (I + B^\top S B)^{-1} B^\top S A$ and S is the solution of the Riccati equation. Consider a problem where $g_1 = g_2 = \cdots = 0$, $c_1 = c_2 = \cdots = 0$, and $K_1 = K_*$ is the gain matrix of the optimal policy. Then, V_1 is the value function of the optimal policy. Because π_2 is the greedy policy with respect to V_1, it is the optimal policy and thus K_2 is also the gain matrix of the optimal policy, and so $K_2 = K_1$. Repeating the same argument shows that all gain matrices are the same. Thus, if we choose K_1 to be the optimal gain matrix in the non-adversarial problem, we will get $K_1 = \cdots = K_t$ and hence $P_1 = P_2 = \cdots = P_t$. □

Proof of Lemma 7. First we prove (i). Under policy $\pi_t(x) = -K_* x + c_t$, we have that

$$\left(x_\infty^{\pi_t},\, \pi_t(x_\infty^{\pi_t})\right) = \left(Ax_\infty^{\pi_t} + B\pi_t(x_\infty^{\pi_t}),\, \pi_t(Ax_\infty^{\pi_t} + B\pi_t(x_\infty^{\pi_t}))\right) \ .$$

Thus, by (1) and (7),

$$\begin{aligned}\lambda &= (x_\infty^{\pi_t} - g_t)^\top Q(x_\infty^{\pi_t} - g_t) + (-K_* x_\infty^{\pi_t} + c_t)^\top(-K_* x_\infty^{\pi_t} + c_t)\\ &= g_t^\top Q g_t + c_t^\top (I + B^\top (I - A + BK_*)^{-\top}(Q + K_*^\top K_*)(I - A + BK_*)^{-1}B)c_t\\ &\quad + 2(-g_t^\top Q - c_t^\top K_*)(I - A + BK_*)^{-1}Bc_t\ .\end{aligned}$$

Then (5) implies that

$$\begin{aligned}L_{t,2}^\top &= 2(-g_t^\top Q - c_t^\top K_*)(I - A + BK_*)^{-1}B\,,\\ P_{*,22} &= I + B^\top (I - A + BK_*)^{-\top}(Q + K_*^\top K_*)(I - A + BK_*)^{-1}B\ .\end{aligned}$$

By Lemmas 2 and 4, $c_t = -\frac{1}{2}P_{*,22}^{-1}\left(\frac{1}{t-1}\sum_{s=1}^{t-1} L_{s,2}\right)$. Thus,

$$\begin{aligned}c_t &= -P_{*,22}^{-1}\left(\frac{1}{t-1}\sum_{s=1}^{t-1} L_{s,2}\right)\\ &= -\frac{P_{*,22}^{-1}B^\top}{t-1}(I - A + BK_*)^{-\top}\sum_{s=1}^{t-1}(-Qg_s - K_*^\top c_s)\\ &= \frac{1}{t-1}\left(D\sum_{s=1}^{t-1} g_s + H\sum_{s=1}^{t-1} c_s\right)\ ,\end{aligned} \tag{13}$$

where $H = P_{*,22}^{-1}B^\top(I - A + BK_*)^{-\top}K_*^\top$. To obtain a bound on $\max_t \|c_t\|$ from the above equation, we need to show that $\|H\|$ is sufficiently smaller than one. Let $N = (I - A + BK_*)^{-1}$, $M = K_*NB$, and $L = (I + M^\top M)^{-1}M^\top$. We have that

$$\begin{aligned}H &= (I + B^\top N^\top (Q + K_*^\top K_*)NB)^{-1}M^\top\\ &\prec (I + B^\top N^\top K_*^\top K_* NB)^{-1}M^\top\\ &= (I + M^\top M)^{-1}M^\top\\ &= L\,,\end{aligned} \tag{14}$$

and

$$\begin{aligned}LL^\top &= (I + M^\top M)^{-1}M^\top M(I + M^\top M)^{-1}\\ &= (I + M^\top M)^{-1}(M^\top M + I - I)(I + M^\top M)^{-1}\\ &= (I + M^\top M)^{-1}\left(I - (I + M^\top M)^{-1}\right)\ .\end{aligned}$$

Because $\left\|M^\top M\right\| = \lambda_{\max}(M^\top M)$, $\|N\| \le 1/(1-\rho)$, and $\left\|M^\top M\right\| \le \|K_*\|^2 \|B\|^2 / (1-\rho)^2$, we get that

$$\begin{aligned}\left\|LL^\top\right\| &\le \left\|(I + M^\top M)^{-1}\right\| \left\|I - (I + M^\top M)^{-1}\right\|\\ &\le 1 - \frac{1}{1 + \|M^\top M\|}\\ &\le 1 - \frac{1}{1 + \|K_*\|^2 \|B\|^2/(1-\rho)^2}\\ &= \frac{\|K_*\|^2\|B\|^2/(1-\rho)^2}{1 + \|K_*\|^2\|B\|^2/(1-\rho)^2}\ .\end{aligned}$$

By (14) and the above inequality, we get that

$$\begin{aligned}\|H\| \le \|L\| = \left\|L^\top\right\| &= \sqrt{\lambda_{\max}(LL^\top)} = \sqrt{\|LL^\top\|}\\ &\le \frac{\|K_*\|\,\|B\|/(1-\rho)}{\sqrt{1 + \|K_*\|^2\|B\|^2/(1-\rho)^2}}\ .\end{aligned}$$

Let $v = 1/(1 - \|H\|)$. We get that

$$\begin{aligned} v &\le \frac{1}{1 - \frac{\|K_*\|\|B\|/(1-\rho)}{\sqrt{1+\|K_*\|^2\|B\|^2/(1-\rho)^2}}} \\ &= \frac{\sqrt{1 + \|K_*\|^2 \|B\|^2/(1-\rho)^2}}{\sqrt{1 + \|K_*\|^2 \|B\|^2/(1-\rho)^2} - \|K_*\| \|B\|/(1-\rho)} \\ &= \sqrt{1 + \|K_*\|^2 \|B\|^2/(1-\rho)^2} \left(\sqrt{1 + \|K_*\|^2 \|B\|^2/(1-\rho)^2} + \frac{\|K_*\| \|B\|}{1-\rho} \right) \\ &= H' \; . \end{aligned}$$

Now we are ready to bound $\|c_t\|$. By (13), we get that for any $t \ge 1$,

$$\|c_t\| \le \|D\| G + \frac{1}{t-1} \sum_{s=1}^{t-1} \|c_s\| \le \|D\| G + \|H\| \max_{s \ge 1} \|c_s\| \; .$$

Thus, $\max_{t\ge1} \|c_t\| \le \|D\| G + \|H\| \max_{t\ge1} \|c_t\|$ and thus, $\max_{t\ge1} \|c_t\| \le \frac{\|D\|G}{1-\|H\|} \le \|D\| GH' = C$.

Proof of (ii). First we write c_t in terms of c_{t-1}:

$$\begin{aligned} c_t &= \frac{1}{t-1} \left(D \sum_{s=1}^{t-1} g_s + H \sum_{s=1}^{t-1} c_s \right) \\ &= \frac{Dg_{t-1}}{t-1} + \frac{Hc_{t-1}}{t-1} + \frac{t-2}{t-1} \left(\frac{D}{t-2} \sum_{s=1}^{t-2} g_s + \frac{H}{t-2} \sum_{s=1}^{t-2} c_s \right) \\ &= \frac{Dg_{t-1}}{t-1} + \frac{Hc_{t-1}}{t-1} + \frac{t-2}{t-1} c_{t-1} \\ &= \frac{1}{t-1} (Dg_{t-1} + ((t-2)I + H)c_{t-1}) \; . \end{aligned}$$

This implies that $c_t - c_{t-1} = \frac{1}{t-1}(Dg_{t-1} - (I-H)c_{t-1})$. Then we use the facts that $\|c_t\| \le C$ and $\|H\| < 1$ to obtain

$$\|c_t - c_{t-1}\| \le \frac{\|D\| G + 2C}{t-1} \; .$$

□

Proof of Lemma 8. Let $f^\pi : \mathcal{X} \to \mathcal{X}$ be the transition function under policy $\pi = (K, c)$, i.e. $f^\pi(x) = (A - BK)x + Bc$. Let $\epsilon_{k,t} = \|x_k - x_\infty^{\pi_t}\|$ and $\epsilon_t = \|x_t - x_\infty^{\pi_t}\|$ denote the difference between the state variable and the limiting state under the chosen policy. We write[4]

$$\begin{aligned} \epsilon_{k,t} &= \|f^{\pi_k}(x_{k-1}) - f^{\pi_t}(x_{k-1}) + f^{\pi_t}(x_{k-1}) - x_\infty^{\pi_t}\| \\ &\le \|f^{\pi_k}(x_{k-1}) - f^{\pi_t}(x_{k-1})\| + \|f^{\pi_t}(x_{k-1}) - f^{\pi_t}(x_\infty^{\pi_t})\| \; . \end{aligned}$$

From this decomposition, we get that

$$\begin{aligned} \epsilon_{k,t} &\le \|B\| \|c_k - c_t\| + \|f^{\pi_t}(x_{k-1}) - f^{\pi_t}(x_\infty^{\pi_t})\| \\ &\le \|B\| \|c_k - c_t\| + \rho \|x_{k-1} - x_\infty^{\pi_t}\| \\ &\le \|B\| (\|D\| G + 2C) \sum_{s=k}^{t-1} \frac{1}{s} + \rho \|x_{k-1} - x_\infty^{\pi_t}\| \; . \end{aligned}$$

[4] A similar decomposition, but with a different norm, was used in (Even-Dar et al., 2009, proof of Lemma 5.2.) to bound the difference between the stationary distribution of the chosen policy and the distribution of the state variable in a finite MDP problem.

Thus,

$$\begin{aligned}
\epsilon_t &\leq \|B\| \left(\|D\| G + 2C\right) \sum_{k=1}^{t} \rho^{t-k} \sum_{s=k}^{t-1} \frac{1}{s} + \rho^{t-1} \left\|x_1 - x_\infty^{\pi_t}\right\| \\
&= \|B\| \left(\|D\| G + 2C\right) \sum_{s=1}^{t-1} \frac{1}{t-s} \sum_{k=s}^{t-1} \rho^k + \rho^{t-1} \frac{\|B\| C}{1-\rho} \\
&\leq \frac{\|B\| \left(\|D\| G + 2C\right)}{1-\rho} \sum_{s=1}^{t-1} \frac{\rho^s}{t-s} + \rho^{t-1} \frac{\|B\| C}{1-\rho} ,
\end{aligned}$$

where the second step follows from Equation (7), Lemma 7, and the fact that $x_1 = 0$. If $t > \lceil \log(T-1)/\log(1/\rho) \rceil$, we get that

$$\begin{aligned}
\sum_{s=1}^{t-1} \frac{\rho^s}{t-s} &= \sum_{s:\rho^s \leq 1/(t-1)} \frac{\rho^s}{t-s} + \sum_{s:1>\rho^s>1/(t-1)} \frac{\rho^s}{t-s} \\
&\leq \frac{1}{t-1} \sum_{s=1}^{t-1} \frac{1}{t-s} + \frac{\log(t-1)}{\log(1/\rho)} \left(\frac{1}{t - \log(t-1)/\log(1/\rho)} \right) \\
&\leq \frac{1+\log(t-1)}{t-1} + \frac{\log(t-1)}{\log(1/\rho)} \left(\frac{1}{t - \log(t-1)/\log(1/\rho)} \right) .
\end{aligned}$$

Thus,

$$\begin{aligned}
\epsilon_t \leq \frac{\|B\| \left(\|D\| G + 2C\right)}{1-\rho} &\left(\frac{1+\log(t-1)}{t-1} + \frac{\log(t-1)}{\log(1/\rho)} \left(\frac{1}{t - \log(t-1)/\log(1/\rho)} \right) \right) \\
&+ \rho^{t-1} \frac{\|B\| C}{1-\rho} .
\end{aligned}$$

To prove the second part of lemma, let $u_T = \lceil \log(T-1)/\log(1/\rho) \rceil$. We have that

$$\sum_{t>u_T} \frac{1}{t - \log(T-1)/\log(1/\rho)} \leq \sum_{t>u_T} \frac{1}{t-u_T} \leq \sum_{t=1}^{T-u_T} \frac{1}{t} \leq \sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log(T) . \tag{15}$$

Thus, by (8) and (15),

$$\begin{aligned}
\sum_{t=1}^{T} \epsilon_t &\leq \sum_{t\leq u_T} \epsilon_t + \sum_{t>u_T} \epsilon_t \\
&\leq \frac{1}{1-\rho} \Bigg(4\|B\| C \left\lceil \frac{\log T}{\log(1/\rho)} \right\rceil + \frac{\|B\| C}{1-\rho} \\
&\qquad + \|B\| \left(\|D\| G + 2C\right)(1+\log T) \left(1 + \log T + \frac{\log T}{\log(1/\rho)} \right) \Bigg) .
\end{aligned}$$

□

The fact that all gain matrices are identical greatly simplifies the boundedness proof.

Proof of Lemma 11. First, it is easy to verify that $P_{*,22} \succ I$ and thus, $H(V_t) = P_{*,22} \succ 2I$. The gradient of the value function can be written as

$$\nabla_a V_t(x_\infty^\pi, a) = 2P_{*,22} a + P_{*,21} x_\infty^\pi + L_{t,2}^\top .$$

Thus, $\|\nabla_a V_t(x_\infty^\pi, a)\| \leq F$ for any $\|a\| \leq U$.

Proof of (i). By (8), $\|x_t\| \leq X$, and by Lemma 7, $\|c_t\| \leq C$. Thus, all actions are bounded by

$$\|a_t\| = \|-K_* x_t + c_t\| \leq \|K_*\| X + C \leq U \ .$$

Proof of (ii) and (iii). By Lemma 6,

$$\|-K x_\infty^\pi + c\| \leq K' X' + C' \leq U \ .$$

Similarly,

$$\|-K_* x_\infty^\pi + c_t\| \leq \|K_*\| X' + C \leq U \ .$$

Proof of (iv). By (4) and the fact that $K_t = K_*$ and $P_t = P_*$, we get that

$$\|L_t\| \leq \frac{2}{1-\rho} \left(G \|Q\| + \rho C \|P_*\| \right) \ .$$

Further, by (2), for any policy $\pi \in \Pi$ and any action satisfying $\|a\| \leq U$, the value functions are bounded by

$$\begin{aligned} V_t(x_\infty^\pi, a) &= \begin{pmatrix} x_\infty^{\pi\top} & a^\top \end{pmatrix} P_* \begin{pmatrix} x_\infty^\pi \\ a \end{pmatrix} + L_t^\top \begin{pmatrix} x_\infty^\pi \\ a \end{pmatrix} \\ &\leq \|P_*\| (X' + U)^2 + \frac{2}{1-\rho} \left(G \|Q\| + \rho C \|P_*\| \right) (X' + U) \\ &= V \ . \end{aligned}$$

□

Proof of Lemma 13. For policy $\pi = (K, c)$, we have $\ell_t(x, \pi) = x^\top (Q + K^\top K) x - 2(c^\top K + g_t^\top Q) x + c^\top c + g_t^\top Q g_t$. Define $S = Q + K^\top K$ and $d_t = 2(c^\top K + g_t^\top Q)$. We write

$$\begin{aligned} \gamma_T &= \sum_{t=1}^T \left(x_\infty^{\pi\top} S x_\infty^\pi - d_t x_\infty^\pi \right) - \sum_{t=1}^T \left(x_t^{\pi\top} S x_t^\pi - d_t x_t^\pi \right) \\ &= \sum_{t=1}^T d_t (x_t^\pi - x_\infty^\pi) + \sum_{t=1}^T \left(\left\| S^{1/2} x_\infty^\pi \right\| - \left\| S^{1/2} x_t^\pi \right\| \right) \left(\left\| S^{1/2} x_t^\pi \right\| + \left\| S^{1/2} x_\infty^\pi \right\| \right) \ . \end{aligned}$$

Thus,

$$\begin{aligned} \gamma_T &\leq \sum_{t=1}^T d_t (x_t^\pi - x_\infty^\pi) + \sum_{t=1}^T \left\| S^{1/2} (x_t^\pi - x_\infty^\pi) \right\| \left(\left\| S^{1/2} x_t^\pi \right\| + \left\| S^{1/2} x_\infty^\pi \right\| \right) \\ &\leq \sum_{t=1}^T \left(\|d_t\| + \left\| S^{1/2} \right\| \left(\left\| S^{1/2} x_t^\pi \right\| + \left\| S^{1/2} x_\infty^\pi \right\| \right) \right) \|x_t^\pi - x_\infty^\pi\| \\ &\leq Z_1' \sum_{t=1}^T \|x_t^\pi - x_\infty^\pi\| \ . \end{aligned}$$

We get the desired result by Lemma 6. □