LEARNING TO CONTROL ON THE FLY

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes an algorithm which learns to control on the fly. The proposed algorithm has no access to the transition law of the environment, which is actually linear with bounded random noise, and learns to make decisions directly online without training phases or sub-optimal policies as the initial input. Neither estimating the system parameters nor the value functions online, the proposed algorithm adapts the ellipsoid method into the online decision making setting. By adding linear constraints when the feasibility of the decision variable is violated, the volume of the decision variable domain can be collapsed and we upper bound the number of online linear constraints needed for the convergence of the state to be around the desired state under the bounded random state noise. The algorithm is also proved to be of constant bounded online regret given certain range of the bound of the random noise.

1 INTRODUCTION

To regret deeply is to live afresh. - Henry David Thoreau

Differently from the optimal control, which knows the system dynamics, hence has the privilege to stand at the beginning of the time and optimize the cumulative cost up to the terminal time as a function of the control action sequence (Stengel, 1994), (Camacho & Alba, 2013), differently from reinforcement learning (Montague, 1999), which has the training phase to learn the state-action value function, also differently from other online control work which requires a stable policy as the initial input to the algorithm (Dean et al., 2018), (Abbasi-Yadkori et al., 2019), the proposed algorithm in this paper does not know the transition law of the environment, which is actually linear with bounded random noise, and do not have any initial stabilizing policy as input, and can only learn to make decisions directly online. The algorithms neither estimate the system parameters, nor the value functions.

Only knowing the alignment relationship between the agent's sensors and actuators and an upper bound of the noise, with the perfect observer of the state, we would like to design the pure learning to control on the fly algorithms which can learn to steer the state of the agent to converge around the desired state asymptotically with theoretical guarantees.

Related Work:

Online learning (Cesa-Bianchi & Lugosi, 2006) portrays optimization as a process. Such view has been applied in many practical applications, where it is necessary and beneficial to learn and optimize from experience as more aspects of the problem are observed (Hazan, 2019). Recent studies have also considered the theory of online optimization with stochastic or cumulative constraints (Yu et al., 2017), (Yuan & Lamperski, 2018). The ellipsoid method (Bland et al., 1981), which is usually used for black-box represented convex optimizations, has also been adapted for online learning in Yang et al. (2009). However, the above works can all be considered as a special case of our learning to control on the fly problem in stateless systems ($A = 0$ in equation 1) (Agarwal et al., 2019a).

The potential link between online learning to reinforcement learning and control has been studied recently in Cheng et al. (2019) and Wagener et al. (2019). In both works, online learning is managed to be fit into the framework of reinforcement learning or model predictive control, to help design new reinforcement learning, or control algorithms. In contrast, in our work, we would like to adopt the online learning perspective to solve the same decision making problem: to purely learn to make

decisions from online interacting experience as more aspects of the system can be inferred from the online noisy data without any training phase or parameter tuning.

Recently, there has been a renewed interest in learning linear dynamical systems online in the machine learning literature (Arora et al., 2018), (Hazan et al., 2018), (Fazel et al., 2018), (Hazan et al., 2017). Beside, some works also study the online control problem with the guaranteed regret bound, by either assuming known dynamics but adversarially changing loss functions (Agarwal et al., 2019b), or assuming unknown dynamics, but initial sub-optimal policies. The stochastic noise with normal distribution is usually considered in these works.

More specifically, Abbasi-Yadkori & Szepesvári (2011) constructs a high-probability confidence set around the system parameters based on online least-squares estimation, and derives the regret bound around $\tilde{O}(\sqrt{T})$ for the first time for the linear quadratic control problem. However its implementation requires solving a non-convex optimization problem to precision $\tilde{O}(T^{-1/2})$, which can be computationally intractable. Dean et al. (2018) proposes the first polynomial-time algorithm for the adaptive LQR problem that provides high probability guarantees of sub-linear regret. However, in proving the regret upper bound, a stable initial policy is assumed to be given as input.

Instead of the interplay between regret minimization and parameter estimation online, model-free approaches for reinforcement learning (RL) is applied online in Abbasi-Yadkori et al. (2019) to solve the linear quadratic control problem with regret upper bound$O(T^{\frac{2}{3}})$ proved. Least-squares temporal difference learning Tu & Recht (2017) is used to approximate the state-action value functions online. However, these algorithms also require a stable policy as input.

Major Contributions:

1) To the best of our knowledge, the proposed algorithm is the first learning to control on the fly algorithm with a theoretical proof in terms of convergence and online regret that neither uses any information about the transition law of the environment, nor owns a sub-optimal policy as input, and requires neither parameter tuning nor training phase before testing the algorithm.

2) The algorithm is analyzed for the case where the environment is subject to bounded random noise. We propose the algorithm which adapts the idea from the ellipsoid method. By adding linear constraints when the feasibility of the decision variable is violated, we can collapse the volume of the decision variable domain and prove the upper bound of the number of times of activating the separation oracle before convergence of the state to be around the desired state. We also show that the algorithm can be of constant regret given a certain range of the noise bound.

2 PRELIMINARIES

The control theory is built on a strong assumption of the system dynamics model

$$\boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + B\boldsymbol{u}_t + \boldsymbol{w}_{t+1}, \tag{1}$$

where the state $\boldsymbol{x}_t \in \mathbb{R}^n$, the agent's action $\boldsymbol{u}_t \in \mathbb{R}^m$, the state transition matrix $A \in \mathbb{R}^{n\times n}$, the state-action alignment matrix $B \in \mathbb{R}^{n\times m}$. $\boldsymbol{w}_{t+1}$ corresponds to the bounded random state noise, in other words, $||\boldsymbol{w}_{t+1}||_2 \leq \epsilon$. The action $\boldsymbol{u}_t$ is parameterized as

$$\boldsymbol{u}_t = K_t\boldsymbol{x}_t = \Phi_t\boldsymbol{k}_t, \tag{2}$$

where the control gain $K_t \in \mathbb{R}^{m\times n}, \Phi_t \triangleq \boldsymbol{x}_t^{\mathrm{T}} \otimes I_m$, $\otimes$ denotes the kronecker product, $\boldsymbol{k_t} \triangleq \mathrm{vec}(K_t)$, where $\mathrm{vec}(\cdot)$ is the vector operator.

Now if A is unknown to us, B is known to us, and we do not have a training chance to collect any data offline and can only observe the noisy data of $\boldsymbol{x}_t$ online directly, and if we do not know any policy beforehand, can we still find a non-anticipative policy that will still drive the state sequence $\{\boldsymbol{x}_t\}$ to converge around the desired state value? What is the convergence rate? Without loss of generality, we consider the desired state as the equilibrium 0. Hence we would like to trap the state $\{\boldsymbol{x}_t\}$ around the ball $G = \{\boldsymbol{x} : ||\boldsymbol{x}||_2 \leq \epsilon\}$ asymptotically given bounded state noise.

Furthermore, we define the online loss $f_t(\boldsymbol{k}_t)$ that the agent suffers at each step t after taking an action $\boldsymbol{u}_t = \Phi_t\boldsymbol{k}_t$ and observing $\boldsymbol{x}_{t+1}$ as

$$f_t(\boldsymbol{k}_t) = \frac{\eta}{2}\boldsymbol{x}_{t+1}^{\mathrm{T}}\boldsymbol{x}_{t+1} + \frac{\beta}{2}\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_t, \tag{3}$$

where $t = 0, 1, ..., T$, T is the terminal time step, $\eta \in \mathbb{R}^+$ is the state weighting parameter, and $\beta \in \mathbb{R}^+$ is the control weighting parameter.

The cumulative loss (the regret) up to the terminal time step T is defined as

$$\text{Reg}_T = \sum_{t=0}^{T} f_t(\boldsymbol{k}_t) - \sum_{t=0}^{T} f_t(\boldsymbol{k}^*), \tag{4}$$

where $\boldsymbol{k}^* = \text{vec}(K^*)$, and $K^* = \lim_{T\to\infty} \arg\min_K \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[f_t(K)]$, which is the fixed control gain given by the infinite horizon Linear Quadratic Regulator (LQR) knowing the transition matrix A beforehead.

In the online setting, we run the designed algorithm to learn to control on the fly in real time. At each time step t, the agent suffers real loss by taking an action and being at certain state. Hence we are also interested in making as close to the optimal as possible our cumulative loss (regret).

3 The Proposed Algorithm: Learning to Control On the Fly

To learn to control when the environment is subject to bounded noise with theoretical guarantee, we introduce the ellipsoid method to the learning to control on the fly problem and propose the algorithm with analysis in this section.

The following assumptions are made first.

Assumption 1: It is assumed that the state $\boldsymbol{x}_t$, $\forall t = 0, 1, 2, ...$, can be observed exactly.

Assumption 2: There exists K, such that $||A+BK||_2$ can be placed arbitrarily (See Abbasi-Yadkori et al. (2014) and Ibrahimi et al. (2012) for the similar assumptions.)

3.1 The Ellipsoid Method Preliminaries

Vectorized $\boldsymbol{k}_t$: To use the ellipsoid method to keep collapsing the space of the control gain K_t, we need to first rewrite the matrix $K_t \in \mathbb{R}^{n\times m}$ into its vector form $\boldsymbol{k}_t = \text{vec}(K_t) \in \mathbb{R}^{m\cdot n\times 1}$. Similar as equation 2, we can rewrite the fantasy action $\hat{\boldsymbol{u}}_{i-1}(K_t)$ as the following

$$\hat{\boldsymbol{u}}_{i-1}(K_t) = \hat{\boldsymbol{u}}_{i-1}(\boldsymbol{k}_t) = K_t \boldsymbol{x}_{i-1} = \Phi_{i-1}\boldsymbol{k}_t, \tag{5}$$

where $\Phi_{i-1} \triangleq \boldsymbol{x}_{i-1}^{\mathrm{T}} \otimes I_m$.

The ball with Unknown Radius r: Our goal is to achieve the exponential contraction of the state, which means that we would like to trap the online solutions of $\boldsymbol{k}_t$ into a ball which satisfies $||A + BK||_2 \leq \delta < 1$, albeit the radius of the ball r is unknown to us. We define the unknown set as $D_\delta = \{\boldsymbol{k} : ||A + BK||_2 \leq \delta < 1\}$. According to Assumption 2,$\delta \in (0, 1)$ can be chosen arbitrarily by us.

The feasible set Q: We define the feasible set Q of the vectorized control gain variable $\boldsymbol{k_t}$: any $\boldsymbol{k_t}$ which makes $||\boldsymbol{x}_{t+1}||_2 - \epsilon < \delta||\boldsymbol{x}_t||_2$ hold true.

The ball with Radius R: We assume that the feasible set Q is contained in the ball $E_0 = \{\boldsymbol{k} : ||\boldsymbol{k}||_2 \leq R\}$ of a given radius R, which means $Q \subset E_0$. The ball E_0 will be the initial ellipsoid, which yields that

$$\boldsymbol{k}_0 = 0, H_0 = RI, E_0 = \{\boldsymbol{k}_0 + H_0\boldsymbol{z} | \boldsymbol{z}^{\mathrm{T}}\boldsymbol{z} \leq 1\}. \tag{6}$$

The Collapsing Sequence of the Ellipsoids: We define E_t be the ellipsoid at time t, such that $\boldsymbol{k}_t \in E_t$. We also define ρ_t be the radii of the Euclidean ball of the same volume as E_t's. We set $\rho_0 = R$.

Now we introduce the following geometric facts. Let E_{t-1} be an $m \cdot n-$dimensional ellipsoid and

$$\hat{E}_t = \{\boldsymbol{k} \in E_{t-1} | \boldsymbol{a}_t^{\mathrm{T}}\boldsymbol{k} \leq \boldsymbol{a}_t^{\mathrm{T}}\boldsymbol{k}_{t-1}\}, \boldsymbol{a}_t \neq 0$$

be a half of E_{t-1}. If $m \cdot n > 1$, then $\hat{E}_t$ is contained in the ellipsoid E_t of the smallest volume Ben-Tal & Nemirovski (2001):

$$E_t = \{\boldsymbol{k} = \boldsymbol{k}_t + H_t \boldsymbol{z} | \boldsymbol{z}^{\mathrm{T}} \boldsymbol{z} \leq 1\}, \tag{7}$$

$$\boldsymbol{p}_t = \frac{H_{t-1}^{\mathrm{T}} \boldsymbol{a}_t}{\sqrt{\boldsymbol{a}_t^{\mathrm{T}} H_{t-1} H_{t-1}^{\mathrm{T}} \boldsymbol{a}_t}}, \tag{8}$$

$$\boldsymbol{k}_t = \boldsymbol{k}_{t-1} - \frac{1}{m \cdot n + 1} H_{t-1} \boldsymbol{p}_t, \tag{9}$$

$$H_t = \frac{m \cdot n}{\sqrt{(m \cdot n)^2 - 1}} H_{t-1} + (\frac{m \cdot n}{m \cdot n + 1} - \frac{m \cdot n}{\sqrt{(m \cdot n)^2 - 1}})(H_{t-1} \boldsymbol{p}_t) \boldsymbol{p}_t^{\mathrm{T}}, \tag{10}$$

$$\rho_t = |\mathrm{Det} H_t|^{1/n} = (\frac{m \cdot n}{\sqrt{(m \cdot n)^2 - 1}})^{(m \cdot n - 1)/m \cdot n} (\frac{m \cdot n}{m \cdot n + 1})^{1/m \cdot n} \rho_{t-1}. \tag{11}$$

Hence, the $m \cdot n-$dimensional volume $\mathrm{Vol}(E_t)$ of the ellipsoid E_t is less than the volume of $\mathrm{Vol}(E_{t-1})$ in the following relation:

$$\mathrm{Vol}(E_t) = (\frac{m \cdot n}{\sqrt{(m \cdot n)^2 - 1}})^{m \cdot n - 1} \frac{m \cdot n}{m \cdot n + 1} \mathrm{Vol}(E_{t-1}) \leq \exp\{-1/2m \cdot n\} \mathrm{Vol}(E_{t-1}). \tag{12}$$

It also follows from equation 11 that

$$\forall m \cdot n \geq 2, \rho_\tau \leq \exp\{-\tau/2(m \cdot n)^2\} R, \tau = 1, 2, ..., t. \tag{13}$$

In the next part, we will build the algorithm which generates such valid $\boldsymbol{a}_t$ online to collapse the volume of E_t as described above.

3.2 Learning to Control On the Fly

At the initial time $t = 0$, the agent observes the state $\boldsymbol{x}_0$. the action $\boldsymbol{u}_0 = 0$ is taken ($\boldsymbol{k}_0 = 0$), and the environment equation 1 transits the state to $\boldsymbol{x}_1$ at time $t = 1$. The initial ellipsoid E_0 is defined in $equation$ 6.

At time step $t \geq 1$, we do the following to update the control gain $\boldsymbol{k}_t$ and the ellipsoid E_t.

The Separation Oracle: We test if $||\boldsymbol{x}_t||_2 - \epsilon \geq \delta ||\boldsymbol{x}_{t-1}||_2$, where $\delta \in (0, 1)$, and ϵ is the bound of the noise $\boldsymbol{w}_i, i = 1, ..., t$. If it holds true, it means $\boldsymbol{k}_{t-1} \notin Q$. If $\boldsymbol{e}_t^{\mathrm{T}} B \Phi_{e_{t-1}} \neq 0$, we can add a constraint

$$\boldsymbol{e}_t^{\mathrm{T}} B \Phi_{e_{t-1}} (\boldsymbol{k}_{t-1} - \boldsymbol{k}) \geq 0 \tag{14}$$

into the set E_{t-1}, where

$$\boldsymbol{e}_t = \frac{\boldsymbol{x}_t}{||\boldsymbol{x}_t||_2}, \tag{15}$$

$$\Phi_{e_{t-1}} = \boldsymbol{e}_{t-1}^{\mathrm{T}} \otimes I_m. \tag{16}$$

This linear inequality cuts off the ellipsoid E_{t-1} centered at $\boldsymbol{k}_{t-1}$ into a half-ellipsoid which we further cover by the ellipsoid of the smallest volume by applying the standard formulas equation 7, equation 8, equation 9, equation 10. We substitute $\boldsymbol{a}_t = \boldsymbol{e}_t^{\mathrm{T}} B \Phi_{e_{t-1}}$ into equation 7, equation 8, equation 9, equation 10 to update $\boldsymbol{k}_t$ and the ellipsoid E_t. We call the separation oracle is activated if both $||\boldsymbol{x}_t||_2 - \epsilon \geq \delta ||\boldsymbol{x}_{t-1}||_2$ and $\boldsymbol{a}_t \neq 0$ hold true.

The agent updates its action

$$\boldsymbol{u}_t = \Phi_t \boldsymbol{k}_t, \tag{17}$$

where $\Phi_t = \boldsymbol{x}_t^{\mathrm{T}} \otimes I_m$. After taking the action $\boldsymbol{u}_t$, the environment equation 1 transits the state to $\boldsymbol{x}_{t+1}$.

The algorithm is summarized as the following.

Alg: Learning to Control on the Fly $(B, T, R, \delta, \epsilon)$

For $t = 0$ to T :
Alg observes $\boldsymbol{x}_t$
If $t = 0$:
$\boldsymbol{k}_0 = 0$
Else:
If $||\boldsymbol{x}_t||_2 - \epsilon \geq \delta||\boldsymbol{x}_{t-1}||_2$:
$\boldsymbol{a}_t = \boldsymbol{e}_t^{\mathrm{T}} B\Phi_{e_{t-1}}$
If $\boldsymbol{a}_t = 0$:
$\boldsymbol{k}_t = \boldsymbol{k}_{t-1}$
Else:
$\hat{E}_t = \{\boldsymbol{k} \in E_{t-1} | \boldsymbol{a}_t^{\mathrm{T}} \boldsymbol{k} \leq \boldsymbol{a}_t^{\mathrm{T}} \boldsymbol{k}_{t-1}\}$
Update $\boldsymbol{k}_t$ and $E_t = \{\boldsymbol{k} = \boldsymbol{k}_t + H_t \boldsymbol{z} | \boldsymbol{z}^{\mathrm{T}} \boldsymbol{z} \leq 1\}$ according to equation 7, equation 8, equation 9, equation 10
End If
Else:
$\boldsymbol{k}_t = \boldsymbol{k}_{t-1}$
End If
End If
Alg takes action $\boldsymbol{u}_t$ according to equation 17
End For

3.3 The Analysis of the Algorithm: Learning to Control on the Fly

Theorem 3.1. *By adding the constraint equation 14 when the separation oracle is active, we cut off half volume of the ellipsoid E_{t-1} whose $\boldsymbol{k} \notin D_\delta$.*

Proof. It follows from the environment dynamics equation 1 and the parameterized action update equation 2 that

$$\boldsymbol{x}_t - (A + BK_{t-1})\boldsymbol{x}_{t-1} = \boldsymbol{w}_t,$$

multiplying both sides of which by the unit directional vector $\boldsymbol{e}_t \triangleq \frac{\boldsymbol{x}_t}{||\boldsymbol{x}_t||_2}$ yields

$$\boldsymbol{e}_t^{\mathrm{T}}[\boldsymbol{x}_t - (A + BK_{t-1})\boldsymbol{x}_{t-1}] = ||\boldsymbol{x}_t||_2 - \boldsymbol{e}_t^{\mathrm{T}}[A + BK_{t-1}]||\boldsymbol{x}_{t-1}||_2 \boldsymbol{e}_{t-1} = \boldsymbol{e}_t^{\mathrm{T}} \boldsymbol{w}_t.$$

It follows from the fact $||\boldsymbol{w}_t||_2 \leq \epsilon$ that

$$||\boldsymbol{x}_t||_2 - \boldsymbol{e}_t^{\mathrm{T}}[A + BK_{t-1}]||\boldsymbol{x}_{t-1}||_2 \boldsymbol{e}_{t-1} = \boldsymbol{e}_t^{\mathrm{T}} \boldsymbol{w}_t \leq \epsilon.$$

Thus it is also true that

$$||\boldsymbol{x}_t||_2 - \epsilon \leq \boldsymbol{e}_t^{\mathrm{T}}[A + BK_{t-1}]||\boldsymbol{x}_{t-1}||_2 \boldsymbol{e}_{t-1}. \tag{18}$$

When the constraint equation 14 is active, the IF test in the separation oracle holds true, which means $\delta||\boldsymbol{x}_{t-1}||_2 \leq ||\boldsymbol{x}_t||_2 - \epsilon$, substituting which into equation 18 yields

$$\boldsymbol{e}_t^{\mathrm{T}}[A + BK_{t-1}]\boldsymbol{e}_{t-1} \geq \delta. \tag{19}$$

It follows from **Assumption 2** that for any $\delta \in (0, 1)$, there exists K such that $\|A + BK\| \leq \delta < 1$. Such K belongs our desired subset D_δ of K with the unknown radius r, and we denote as K_δ.

It follows that

$$\boldsymbol{e}_t^{\mathrm{T}}[A + BK_\delta]\boldsymbol{e}_{t-1} \leq \|\boldsymbol{e}_t\|_2 \|A + BK_\delta\|_2 \|\boldsymbol{e}_{t-1}\|_2 \leq \delta \tag{20}$$

equation 19 minus equation 20 yields

$$\boldsymbol{e}_t^{\mathrm{T}} B(K_{t-1} - K_\delta)\boldsymbol{e}_{t-1} \geq 0. \tag{21}$$

It follows from equation 2 that $K\boldsymbol{x}_{t-1} = (\boldsymbol{x}_{t-1}^{\mathrm{T}} \otimes I_m)\boldsymbol{k}$. Normalizing both sides by dividing $||\boldsymbol{x}_{t-1}||_2$ yields that $K\boldsymbol{e}_{t-1} = (\boldsymbol{e}_{t-1}^{\mathrm{T}} \otimes I_m)\boldsymbol{k} \triangleq \Phi_{e_{t-1}}\boldsymbol{k}$. We thus plug $K_{t-1}\boldsymbol{e}_{t-1} = \Phi_{e_{t-1}}\boldsymbol{k}_{t-1}$, and $K_\delta \boldsymbol{e}_{t-1} = \Phi_{e_{t-1}}\boldsymbol{k}_\delta$ into equation 21, which yields that

$$\boldsymbol{e}_t^{\mathrm{T}} B \Phi_{e_{t-1}}(\boldsymbol{k}_{t-1} - \boldsymbol{k}_\delta) \geq 0. \tag{22}$$

Thus by adding the above constraints we cut off the other half space of the ellipsoid E_{t-1} which violates equation 22, and those $\boldsymbol{k}$ does not belong to the desired subset $D_\delta = \{\boldsymbol{k} : ||A + BK||_2 \leq \delta < 1\}$. □

Theorem 3.2. *Let $m \cdot n \geq 2$, and under Assumption 1 and 2 such that there exists a desired subset $D_\delta = \{\boldsymbol{k} : ||A + BK||_2 \leq \delta < 1\}$ with unknown radius r. We assume there exists an initial ball $E_0 = \{\boldsymbol{k} : ||\boldsymbol{k}||_2 \leq R\}$ of a given radius R which contains the feasible set Q. No more than $N = 2(m \cdot n)^2 \ln \frac{R}{r}$ times of activation of the separation oracle is needed before the control gain being trapped into the desired subset D_δ.*

Proof. First it can be easily seen that the feasible set Q contains the desired subset D_δ, since if $||A + BK_t||_2 \leq \delta$, then $||\boldsymbol{x}_{t+1}||_2 = ||(A + BK_t)\boldsymbol{x}_t + \boldsymbol{w}_{t+1}|| \leq ||A + BK_t||_2||\boldsymbol{x}_t||_2 + \epsilon \leq \delta||\boldsymbol{x}_t||_2 + \epsilon$.

From Theorem 3.1, we know that by adding the linear constraint equation 14, we cut off the half volume of the previous ellipsoid E_{t-1} whose $\boldsymbol{k} \notin D_\delta$. Then by applying the formulas equation 7, equation 8, equation 9, equation 10, we update the current ellipsoid E_t to be of the smallest volume but covers the remaining half space of the previous ellipsoid E_{t-1}.

Thus every time the separation oracle is activated, and after we do the above procedures to update E_t, equation 13 holds true for $\forall m \cdot n \geq 2$, where the time τ denotes the index of the number of times when the separation oracle is activated, $\tau = 1, 2, ..., N$.

We prove the theorem by contradiction, and we assume that the control gain has not been trapped into D_δ after $N = 2(m \cdot n)^2 \ln \frac{R}{r}$ times of activation of the separation oracle, which means the radii ρ_t of the Euclidean ball of the same volume as E_t's, compared with the radii r of the Euclidean ball of the same volume as D_δ, follows that

$$\frac{\rho_t}{r} > 1. \tag{23}$$

It follows from equation 13 that

$$\rho_t \leq \exp\{-N/2(m \cdot n)^2\}R.$$

Substituting equation 24 into equation 23 yields that

$$N < 2(m \cdot n)^2 \ln \frac{R}{r}, \tag{24}$$

which is the desired contradiction. □

3.4 No Pain, No Gain and No Gain, No Pain

No Pain, No Gain: In Theorem 3.2, it is stated that at most $N = 2(m \cdot n)^2 \ln \frac{R}{r}$ times activation of the separation oracle is needed, before the agent can learn the control gain $\boldsymbol{k}_t \in D_\delta$, which is equivalent of saying that the state $\boldsymbol{x}_t$ can be converged asymptotically.

Each time when the separation oracle is activated, the current state $\boldsymbol{x}_t \neq 0$, and $||\boldsymbol{x}_t||_2 \geq \delta||\boldsymbol{x}_{t-1}||_2 + \epsilon$. Compared to that when the separation oracle is not activated, either $\boldsymbol{x}_t = 0$, or $||\boldsymbol{x}_t||_2 \leq \delta||\boldsymbol{x}_{t-1}||_2 + \epsilon$, we suffer a bigger online loss according to equation 3, and a bigger regret according to equation 4, which can be interpreted as the "pain" that we suffer online.

However, it can also be seen from the learning to control algorithm that: only through activating the separation oracle, we get a chance to update the control gain $\boldsymbol{k}_t$, and collapse the volume of the domain E_t of the control gain $\boldsymbol{k}_t$, by which we learn to control. If our desired goal and gain is to learn a good policy online from scratch such that $\boldsymbol{x}_t$ can be converged asymptotically, we have to take the "pain" to activate the separation oracle no more than N times.

No Gain, No Pain: When the terminal time T is finite, the goal of trapping the control gain $\boldsymbol{k}_t$ into D_δ cannot be guaranteed to be achieved. One counter example is the following: it follows from the learning to control on the fly algorithm design that $\boldsymbol{k}_0 = 0$. Suppose if $||A\boldsymbol{x}_0||_2 \leq \epsilon$, and the random bounded noise, which is adversary, is chosen to be $\boldsymbol{w}_1 = -A\boldsymbol{x}_0$. Then it follows from equation 1 that $\boldsymbol{x}_1 = 0$. If $\boldsymbol{w}_t = 0$, for $t = 2, 3, ..., T-1$, then from $t = 1, ..., T-1$, $\boldsymbol{x}_t = 0$, so we won't have any chance to update $\boldsymbol{k}_t$ and learn to control. From this perspective, if we define our desired gain to be to learn a good policy, for such cases where the separation oracle is not activated, we have no "gain".

In such situations where the separation oracle is not activated, either $\boldsymbol{x}_t = 0$, or $||\boldsymbol{x}_t||_2 \leq \delta||\boldsymbol{x}_{t-1}||_2 + \epsilon$, although we have no "gain" in the knowledge of how to control, our online loss and regret are also small compared with when activated, hence there is also "no pain" along with "no gain".

If we would like to re-define our goal to be to suffer as little regret as possible in the finite running time, we can evaluate the proposed learning to control on the fly algorithm in the following theorem.

Theorem 3.3. *Assume finite running time $T \gg N$. We also assume that there exists an initial ball $E_0 = \{\boldsymbol{k} : ||\boldsymbol{k}||_2 \leq R\}$ of a given radius R which contains the feasible set Q. Under Assumption 1 and 2, with bounded noise $||\boldsymbol{w}_t|| \leq \epsilon, \forall t$, the regret of the proposed learning to control on the fly algorithm is upper bounded by $Reg_T \leq O(T)$. If $\epsilon \leq \frac{c}{\sqrt{T}}$, where $c \geq 0$ is any constant, the proposed algorithm is a no-regret algorithm.*

Proof. Using the proposed learning to control on the fly algorithm, the case in which the agent suffers the maximal regret is when the agent keeps activating the separation oracle for the first $N = 2(m \cdot n)^2 \ln \frac{R}{r}$ steps. Because N is not dependent on T, the cumulative loss for the first N steps $\sum_{t=1}^{N} \frac{\eta}{2}\boldsymbol{x}_{t+1}^{\mathrm{T}}\boldsymbol{x}_{t+1} + \frac{\beta}{2}\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_t$ is bounded by a constant.

It follows from Theorem 3.2 that after the first N steps of activating the oracle, $||A + BK_t||_2 \leq \delta < 1$, and $\boldsymbol{x}_t$ will start to converge.

Next we compute the exact ball that $\boldsymbol{x}_t$ will converge to at the worst case scenario, where $t > N$. Since $\boldsymbol{k}_t \in D_\delta$, the separation oracle will not be activated anymore. Hence, K_t will stay the same. The K_t corresponding to the slowest convergence rate is K_δ such that $||A + BK_\delta||_2 = \delta$. We denote the most adversarial random noise as $\boldsymbol{w}_\delta$, the converged invariant state under $\boldsymbol{w}_\delta$ as $\boldsymbol{x}_\delta$, which satisfies $(A + BK_\delta)\boldsymbol{x}_\delta + \boldsymbol{w}_\delta = \boldsymbol{x}_\delta$. Hence the upper bound for $\boldsymbol{x}_\delta$ yields that

$$||\boldsymbol{x}_\delta||_2^2 \leq ||[I - (A + BK_\delta)]^\dagger \boldsymbol{w}_\delta||_2^2 \leq ||[I - (A + BK_\delta)]^\dagger||_2^2 ||\boldsymbol{w}_\delta||_2^2. \tag{25}$$

We denote $||[I - (A + BK_\delta)]^\dagger||_2 \triangleq \alpha$, which is constant. Hence equation 25 yields that $||\boldsymbol{x}_\delta||_2^2 \leq \alpha^2\epsilon^2$.

We define the exact ball as $G^{'} = \{\boldsymbol{x} : ||\boldsymbol{x}||_2 \leq \alpha\epsilon\}$. The first time when $\boldsymbol{x}_t$ converges around $G^{'}$ is denoted by t_c. The converging speed of $\boldsymbol{x}_t$ when $t > N$ is exponential fast before $\boldsymbol{x}_t$ converges around $G^{'}$. Hence the number of steps $t_c - (N+1)$ is also upper bounded by a constant which does not depend on T. Thus the cumulative loss $\sum_{t=N+1}^{t_c} \frac{\eta}{2}\boldsymbol{x}_{t+1}^{\mathrm{T}}\boldsymbol{x}_{t+1} + \frac{\beta}{2}\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_t$ can also bounded by a constant.

The last part of the cumulative loss is bounded as the following:

$$\begin{aligned}\sum_{t=t_c+1}^{T} \frac{\eta}{2}\boldsymbol{x}_{t+1}^{\mathrm{T}}\boldsymbol{x}_{t+1} + \frac{\beta}{2}\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_t &\leq \sum_{t=t_c+1}^{T} \frac{\eta}{2}||\boldsymbol{x}_\delta||_2^2 + \frac{\beta}{2}||K_\delta||_2^2||\boldsymbol{x}_\delta||_2^2 \\ &\leq \sum_{t=t_c+1}^{T} \frac{\eta}{2}\alpha^2\epsilon^2 + \frac{\beta}{2}r^2\alpha^2\epsilon^2 \leq (\frac{\eta}{2} + \frac{\beta}{2}r^2)\alpha^2\epsilon^2 T = O(T).\end{aligned}$$

If $\epsilon \leq \frac{c}{\sqrt{T}}$, where c is a constant, then $\sum_{t=t_c+1}^{T} \frac{\eta}{2}\boldsymbol{x}_{t+1}^{\mathrm{T}}\boldsymbol{x}_{t+1} + \frac{\beta}{2}\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_t \leq (\frac{\eta}{2} + \frac{\beta}{2}r^2)\alpha^2 c$, which is a constant which does not depend on T.

Since the cumulative loss from time 1 to N and from $N+1$ to t_c are both bounded by constants, and from $t_c + 1$ to T is bounded by $O(T)$, the total cumulative loss $\sum_{t=1}^{T} \frac{\eta}{2}\boldsymbol{x}_{t+1}^{\mathrm{T}}\boldsymbol{x}_{t+1} + \frac{\beta}{2}\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_t$ is bounded by $O(T)$,and specially is bounded by a constant if $\epsilon \leq \frac{c}{\sqrt{T}}$.

The optimal cumulative loss $\sum_{t=1}^{T} f_t(\boldsymbol{k}^*)$ in equation 4 can be lower bounded by a constant, which follows from the property of LQR (Kwakernaak & Sivan, 1972).

Hence we can see that the regret of the proposed learning to control on the fly algorithm is upper bounded by $\text{Reg}_T \leq O(T)$. If $\epsilon \leq \frac{c}{\sqrt{T}}$, the proposed algorithm has constant regret, which is also called to be of no regret. □

REFERENCES

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

Yasin Abbasi-Yadkori, Peter Bartlett, and Varun Kanade. Tracking adversarial targets. In *International Conference on Machine Learning*, pp. 369–377, 2014.

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3108–3117, 2019.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M Kakade, and Karan Singh. Online control with adversarial disturbances. *arXiv preprint arXiv:1902.08721*, 2019a.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pp. 10175–10184, 2019b.

Sanjeev Arora, Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Towards provable control for unknown linear dynamical systems. 2018.

A Ben-Tal and Arkadi Nemirovski. Lectures on modern convex optimization: Analysis. *Algorithms, and Engineering Applications, SIAM, Philadelphia*, 2001.

Robert G Bland, Donald Goldfarb, and Michael J Todd. The ellipsoid method: A survey. *Operations research*, 29(6):1039–1091, 1981.

Eduardo F Camacho and Carlos Bordons Alba. *Model predictive control*. Springer Science & Business Media, 2013.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Ching-An Cheng, Remi Tachet des Combes, Byron Boots, and Geoff Gordon. A reduction from reinforcement learning to no-regret online learning. *arXiv preprint arXiv:1911.05873*, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pp. 4188–4197, 2018.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pp. 6702–6712, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pp. 4634–4643, 2018.

Morteza Ibrahimi, Adel Javanmard, and Benjamin V Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pp. 2636–2644, 2012.

Huibert Kwakernaak and Raphael Sivan. *Linear optimal control systems*, volume 1. Wiley-interscience New York, 1972.

P Read Montague. Reinforcement learning: An introduction, by sutton, rs and barto, ag. *Trends in cognitive sciences*, 3(9):360, 1999.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Robert F Stengel. *Optimal control and estimation*. Courier Corporation, 1994.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. *arXiv preprint arXiv:1712.08642*, 2017.

Nolan Wagener, Ching-An Cheng, Jacob Sacks, and Byron Boots. An online learning approach to model predictive control. *arXiv preprint arXiv:1902.08967*, 2019.

Liu Yang, Rong Jin, and Jieping Ye. Online learning by ellipsoid method. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1153–1160, 2009.

Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pp. 1428–1438, 2017.

Jianjun Yuan and Andrew Lamperski. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*, pp. 6137–6146, 2018.