Online Control with Adversarial Disturbances

Naman Agarwal [1] **Brian Bullins** [2 1] **Elad Hazan** [2 1] **Sham M. Kakade** [3 4 1] **Karan Singh** [2 1]

Abstract

We study the control of a linear dynamical system with adversarial disturbances (as opposed to statistical noise). The objective we consider is one of regret: we desire an online control procedure that can do nearly as well as that of a procedure that has full knowledge of the disturbances in hindsight. Our main result is an efficient algorithm that provides nearly tight regret bounds for this problem. From a technical standpoint, this work generalizes upon previous work in two main aspects: our model allows for adversarial noise in the dynamics, and allows for general convex costs.

1. Introduction

This paper studies the robust control of linear dynamical systems. minimization in online learning. A linear dynamical system is governed by the dynamics equation

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1.1}$$

where x_t is the state, u_t is the control and w_t is a disturbance to the system. At every time step t, the controller suffers a cost $c_t(x_t, u_t)$ to enforce the control. In this paper, we consider the setting of online control with *arbitrary* disturbances. Formally, the setting involves, at every time step t, an adversary selecting a convex cost function $c_t(x, u)$ and a disturbance w_t, and the goal of the controller is to generate a sequence of controls u_t such that a sequence of convex costs $c_t(u_t, x_t)$ is minimized. This generalization offers several challenges that have remain unaddressed by the literature on Linear Quadratic Regulators.

[1]Google AI Princeton [2]Department of Computer Science, Princeton University [3]Allen School of Computer Science and Engineering, University of Washington [4]Department of Statistics, University of Washington. Correspondence to: Naman Agarwal <namanagarwal@google.com>, Brian Bullins <bbullins@cs.princeton.edu>, Elad Hazan <ehazan@google.com>, Sham Kakade <sham@cs.washington.edu>, Karan Singh <karans@princeton.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Challenge 1. Perhaps the most important challenge we address is in dealing with arbitrary disturbances w_t *in the dynamics*. This is a difficult problem, and so standard approaches almost exclusively assume i.i.d. Gaussian noise. Worst-case approaches in the control literature, also known as H_∞-control and its variants, are overly pessimistic. Instead, we take an online (adaptive) approach to dealing with adversarial disturbances.

Challenge 2. Another limitation for efficient methods is the classical assumption that the costs $c_t(x_t, u_t)$ are quadratic, as is the case for the *linear quadratic regulator*. Part of the focus in the literature on the quadratic costs is due to special properties that allow for efficient computation of the best linear controller in hindsight. One of our main goals is to introduce a more general technique that allows for efficient algorithms even when faced with arbitrary convex costs.

Our contributions. In this paper, we tackle both challenges outlined above: coping with adversarial noise, and general loss functions in an online setting. To define the performance metric, denote for any control algorithm $\mathcal{A}$,

$$J_T(\mathcal{A}) = \sum_{t=1}^{T} c_t(x_t, u_t).$$

The standard comparator in control is a linear controller, which generates a control signal as a linear function of the state, i.e. $u_t = -Kx_t$. Let $J(K)$ denote the cost of a linear controller from a certain class $K \in \mathcal{K}$. For an algorithm $\mathcal{A}$, we define the regret as the sub-optimality of its cost with respect to the best linear controller from a certain set

$$\texttt{Regret} = J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K).$$

Our main result is an efficient algorithm for control which achieves regret $O(\sqrt{T})$ in the setting described above. Our results above are obtained using a host of techniques from online convex optimization, notably online learning for loss functions with memory and improper learning using convex relaxation.

2. Related Work

Non-stochastic MDPs: The setting we consider, control in systems with linear transition dynamics (Bertsekas, 2005)

in presence of adversarial disturbances, can be cast as that of planning in an adversarially changing MDP (Arora et al., 2012b; Dekel & Hazan, 2013). The results obtained via this reduction are unsatisfactory because these regret bounds scale with the size of the state space, which is usually exponential in the dimension of the system. In addition, the regret in these scale as $\Omega(T^{\frac{2}{3}})$. In comparison, (Yu et al., 2009; Even-Dar et al., 2009) solve the online planning problem for MDPs with fixed dynamics and changing costs. The satisfying aspect of their result is that the regret bound does not explicitly depend on the size of the state space, and scales as $O(\sqrt{T})$. However, the dynamics are fixed and without (adversarial) noise.

Robust Control: The most notable attempts to handle adversarial perturbations in the dynamics are called H_∞ control (Zhou et al., 1996; Stengel, 1994). In this setting, the controller solves for the best linear controller assuming worst case future disturbances, i.e.

$$\min_{K_1} \max_{\varepsilon_{1:T}} \min_{K_2} \ldots \min_{K_t} \max_{\varepsilon_T} \sum_t c_t(x_t, u_t),$$

assuming similar linear dynamics as in equation (1.1). In comparison, we do not solve for the entire noise trajectory in advance, but adjust for it iteratively. metric of regret (for us) vs. competitive ration (H_∞ control). Another difference is computational: the above mathematical program may be hard to compute for general cost functions, as compared to our efficient gradient-based algorithm.

LQR with changing costs: For the Linear Quadratic Regulator problem, (Cohen et al., 2018) consider changing quadratic costs with stochastic noise to get a $O(\sqrt{T})$ regret bound. This work is well aligned with results, and the present paper employs some notions developed therein (eg. strong stability). However, the techniques used in (Cohen et al., 2018) (eg. the SDP formulation for a linear controller) are strongly reliant on the quadratic nature of the cost functions and stochasticity of the disturbances. In particular, even for the offline problem, to the best of our knowledge, there does not exist a SDP formulation to determine the best linear controller for convex losses. In an earlier work, (Abbasi-Yadkori & Szepesvári, 2011) considers a more restricted setting with fixed, deterministic dynamics (hence, noiseless) and changing quadratic costs.

3. Problem Setting

3.1. Interaction Model

The Linear Dynamical System is a Markov decision process on continuous state and action spaces, with linear transition dynamics. In each round t, the learner outputs an action u_t on observing the state x_t and incurs a cost of $c_t(x_t, u_t)$, where $c_t(\cdot, \cdot)$ is convex. The system then transitions to a new state x_{t+1} according to

$$x_{t+1} = Ax_t + Bu_t + w_t.$$

In the above definition, w_t is the disturbance sequence the system suffers at each time step. In this paper, we make no distributional assumptions on w_t. The sequence w_t is not made known to the learner in advance.

For any algorithm $\mathcal{A}$, the cost we attribute to it is

$$J_T(\mathcal{A}) = \sum_{t=1}^{T} c_t(x_t, u_t)$$

where $x_{t+1} = Ax_t + Bu_t + w_t$ and $u_t = \mathcal{A}(x_1, \ldots x_t)$. With some abuse of notation, we shall use $J(K)$ to denote the cost of a linear controller $\pi(K)$ which chooses the action as $u_t = -Kx_t$.

3.2. Assumptions

We make the following assumptions throughout the paper. We remark that they are less restrictive, and hence, allow for more general systems than those considered by the previous works. In particular, we allow for adversarial (rather than i.i.d. stochastic) noise, and convex cost functions. Also, the non-stochastic nature of the disturbances permits, without loss of generality, the assumption that $x_0 = 0$.

Assumption 3.1. *The matrices that govern the dynamics are bounded, ie.,* $\|A\| \le \kappa_A, \|B\| \le \kappa_B$. *The perturbation introduced per time step is bounded, ie.,* $\|w_t\| \le W$.

Assumption 3.2. *The costs* $c_t(x, u)$ *are convex. Further, as long as it is guaranteed that* $\|x\|, \|u\| \le D$, *it holds that*

$$|c_t(x, u)| \le \beta D^2, \|\nabla_x c_t(x, u)\|, \|\nabla_u c_t(x, u)\| \le G_c D.$$

Following the definitions in (Cohen et al., 2018), we work on the following class of linear controllers.

Definition 3.3. *A linear policy* K *is* (κ, γ)*-strongly stable if there exist matrices* L, H *satisfying* $A - BK = HLH^{-1}$, *such that following two conditions are met:*

1. *The spectral norm of* L *is strictly smaller than unity, ie.,* $\|L\| \le 1 - \gamma$.
2. *The controller and the transforming matrices are bounded, ie.,* $\|K\| \le \kappa$ *and* $\|H\|, \|H^{-1}\| \le \kappa$.

3.3. Regret Formulation

Let $\mathcal{K} = \{K : K \text{ is } (\kappa, \gamma)\text{-strongly stable}\}$. For an algorithm $\mathcal{A}$, the regret is the sub-optimality of its cost with respect to a best linear controller.

$$\texttt{Regret} = J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K).$$

4. Preliminaries

4.1. A Disturbance-Action Policy Class

We put forth the notion of a *disturbance-action controller* which chooses the action as a linear map of the past disturbances. Any disturbance-action controller ensures that the state of a system executing such a policy may be expressed as a linear function of the parameters of the policy. This property is convenient in that it permits efficient optimization over the parameters of such a policy.

Definition 4.1 (Disturbance-Action Policy)**.** *A disturbance-action policy $\pi(M, K)$ is specified by parameters $M = (M^{[1]}, \ldots, M^{[H]})$ and a fixed matrix K. At every time t, such a policy $\pi(M, K)$ chooses the recommended action u_t at a state x_t[1], defined as*

$$u_t = -Kx_t + \sum_{i=1}^{H} M^i w_{t-i}.$$

For notational convenience, here it may be considered that $w_i = 0$ for all $i < 0$.

We refer to the policy played at time t as $M_t = \{M_t^{[i]}\}$ where the subscript t refers to the time index and the superscript $[i]$ refers to the action of M_t on w_{t-i}. Note that such a policy can be executed because w_{t-1} is perfectly determined on the specification of x_t as $w_{t-1} = x_t - Ax_{t-1} - Bu_{t-1}$. It shall be established in later sections that such a policy class can approximate any linear policy with a strongly stable matrix in terms of the total cost suffered.

4.2. Evolution of State

In this section, we reason about the evolution of the state of a linear dynamical system under a non-stationary policy $\pi = (\pi_0, \ldots, \pi_{T-1})$ composed of T policies, where each π_t is specified by $\pi_t(M_t = (M_t^{[1]}, \ldots, M_t^{[H]}), K)$. Again, with some abuse of notation, we shall use $\pi((M_0, \ldots, M_{T-1}), K)$ to denote such a non-stationary policy. The following definitions serve to ease the burden of notation.

1. Define $\tilde{A}_K = A - BK$. $\tilde{A}_K$ shall be helpful in describing the evolution of state starting from a non-zero state in the absence of disturbances.

2. $x_t^K(M_0, \ldots, M_{t-1})$ is the state attained by the system upon execution of a non-stationary policy $\pi((M_0, \ldots, M_{t-1}), K)$. We drop the arguments M_i and the K from the definition of x_t when it is clear from the context. If the same policy M is used across all time steps, we compress the notation to $x_t^K(M)$. Note that $x_t^K(0)$ refers to running the linear policy K in the standard way.

3. $\Psi_{t,i}^K(M_0 \ldots, M_{t-1})$ is a transfer matrix that describes the effect of w_{t-i} on the state x_{t+1}, formally defined below. When the arguments to $\Psi_{t,i}^K$ are clear from the context, we drop the arguments. When M is the same across all arguments we suppress the notation to $\Psi_{t,i}^K(M)$.

[1] x_t is completely determined given $w_0 \ldots w_{t-1}$. Hence, the use of x_t only serves to ease presentation.

Definition 4.2. *Define the disturbance-state transfer matrix $\Psi_{t,i}^K$ to be*

$$\Psi_{t,i}^K(M_{t-H}, \ldots, M_{t-1}) = \tilde{A}_K^i \mathbf{1}_{i \le H} + \sum_{j=1}^{H} \tilde{A}_K^j B M_{t-j}^{[i-j]} \mathbf{1}_{i-j \in [1,H]}.$$

Lemma 4.3. *If u_t is chosen as a non-stationary policy $\pi((M_1, \ldots, M_T), K)$ recommends, then the state sequence is governed as follows:*

$$x_{t+1} = \sum_{i=0}^{t} \Psi_{t,i} w_{t-i}, \tag{4.1}$$

which can equivalently be written as

$$x_{t+1} = \tilde{A}_K^{H+1} x_{t-H} + \sum_{i=0}^{2H} \Psi_{t,i} w_{t-i}. \tag{4.2}$$

4.3. Idealized Setting

Note that the counter-factual nature of regret in the control setting implies in the loss at a time step t, depends on all the choices made in the past. To efficiently deal with this we propose that our optimization problem only consider the effect of the past H steps while planning, forgetting about the state, the system was at time $t - H$. We will show later that the above scheme tracks the true cost suffered upto a small additional loss. To formally define this idea, we need the following definition on *ideal* state.

Definition 4.4 (Ideal State & Action)**.** *Define an* ideal *state y_{t+1}^K which is the state the system would have reached if it played the non-stationary policy $(M_{t-H}, \ldots, M_t)$ at all time steps from $t - H$ to t, assuming the state at $t - H$ is 0. Similarly, define $v_t^K(M_{t-H}, \ldots, M_t)$ to be an idealized action that would have been executed at time t if the state observed at time t is $y_t^K(M_{t-H}, \ldots, M_{t-1})$. Formally,*

$$y_{t+1}^K(M_{t-H}, \ldots, M_t) = \sum_{i=0}^{2H} \Psi_{t,i} w_{t-i},$$

$$v_t^K(M_{t-H}, \ldots, M_t) = -K y_t^K + \sum_{i=1}^{H} M_t^{[i]} w_{t-i}.$$

We can now consider the loss of the *ideal* state and the *ideal* action.

Definition 4.5 (Ideal Cost)**.** *Define the idealized cost function f_t to be the cost associated with the idealized state and idealized action, i.e.,*

$$f_t(M_{t-H}, \dots, M_t) = c_t(y_t^K(M_{t-H}, \dots, M_{t-1}), v_t^K(M_{t-H}, \dots, M_t)).$$

The linearity of y_t^K in past controllers and the linearity of v_t^K in its immediate state implies that f_t is a convex function of a linear transformation of $M_{t-H}, \dots, M_t$ and hence convex in $M_{t-H}, \dots, M_t$. This renders it amenable to algorithms for online convex optimization.

In Theorem 5.3 we show that f_t and c_t on a sequence are close by and this reduction allows us to only consider the truncated f_t while planning allowing for efficiency. The precise notion of minimizing regret such truncated f_t was considered in online learning literature (Anava et al., 2015) before as online convex optimization(OCO) with memory. We present an overview of this framework next.

4.4. OCO with Memory

We now present an overview of the online convex optimization (OCO) with memory framework, as established by (Anava et al., 2015). In particular, we consider the setting where, for every t, an online player chooses some point $x_t \in \mathcal{K} \subset \mathbb{R}^d$, a loss function $f_t : \mathcal{K}^{H+1} \mapsto \mathbb{R}$ is revealed, and the learner suffers a loss of $f_t(x_{t-H}, \dots, x_t)$. We assume a certain coordinate-wise Lipschitz regularity on f_t of the form such that, for any $j \in \{1, \dots, H\}$, for any $x_1, \dots, x_H, \tilde{x}_j \in \mathcal{K}$,

$$|f_t(\dots, x_j, \dots) - f_t(\dots, \tilde{x}_j, \dots)| \le L\|x_j - \tilde{x}_j\|. \tag{4.3}$$

In addition, we define $\tilde{f}_t(x) = f_t(x, \dots, x)$, and we let

$$G_f = \sup_{t \in \{1,\dots,T\}, x \in \mathcal{K}} \|\nabla \tilde{f}_t(x)\|, \quad D = \sup_{x,y \in \mathcal{K}} \|x - y\|. \tag{4.4}$$

The resulting goal is to minimize the *policy regret* (Arora et al., 2012b), which is defined as

$$\texttt{Regret} = \sum_{t=H}^{T} f_t(x_{t-H}, \dots, x_t) - \min_{x \in \mathcal{K}} \sum_{t=H}^{T} f_t(x, \dots, x).$$

As shown by (Anava et al., 2015), by running a memory-based OGD, we may bound the policy regret by the following theorem.

Theorem 4.6. *Let $\{f_t\}_{t=1}^T$ be Lipschitz continuous loss functions with memory such that $\tilde{f}_t$ are convex, and let L, D, and G_f be as defined in (4.3) and (4.4). Then, Algorithm 2 generates a sequence $\{x_t\}_{t=1}^T$ such that*

$$\sum_{t=H}^{T} f_t(x_{t-H}, \dots, x_t) - \min_{x \in \mathcal{K}} \sum_{t=H}^{T} f_t(x, \dots, x) \le \frac{D^2}{\eta} + TG_f^2\eta + LH^2\eta G_f T.$$

Furthermore, setting $\eta = \frac{D}{\sqrt{G_f(G_f + LH^2)T}}$ implies that

$$\texttt{Regret} \le O\left(D\sqrt{G_f(G_f + LH^2)T}\right).$$

Algorithm 1 Online Control Algorithm

Input: Step size η, Control Matrix K, Parameters $\kappa_B, \kappa, \gamma, T$.
Define $H = 2\kappa_B\kappa^3\gamma^{-1}\log(T)$
Define $\mathcal{M} = \{M = \{M^{[1]} \dots M^{[H]}\} : \|M^{[i]}\| \le \kappa^3\kappa_B(1-\gamma)^i\}$.
Initialize $M_0 \in \mathcal{M}$ arbitrarily.
for $t = 0, \dots, T-1$ **do**
  Choose the action $u_t = c_t - Kx_t + \sum_{i=1}^{H} M^{[i]} w_{t-i}$.
  Observe the new state x_{t+1} and record $w_t = x_{t+1} - Ax_t - Bu_t$.
  Define the function $g_t(M)$ as $g_t(M) = f_t(M, \dots M)$ (refer Definition 4.5)
  Set $M_{t+1} = \Pi_{\mathcal{M}}(M_t - \eta\nabla g_t(M))$
end for

5. Algorithm & Main Result

Algorithm 1 describes our proposed algorithm for controlling linear dynamical systems with adversarial disturbances which at all times maintains a disturbance-action controller. The algorithm implements the memory based OGD on the loss $f_t(\cdot)$ as described in the previous section. The algorithm requires the specification of a (κ, γ)-strongly stable matrix K once before the online game. Such a matrix can be obtained offline using an SDP relaxation as described in (Cohen et al., 2018). The following theorem states the regret bound Algorithm 1 guarantees.

Theorem 5.1 (Main Theorem)**.** *Suppose Algorithm 1 is executed with $\eta = \Theta\left(G_c W\sqrt{T}\right)^{-1}$, on an LDS satisfying Assumption 3.1 with control costs satisfying Assumption 3.2. Then, it holds true that*

$$J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K) \le O\left(G_c W^2\sqrt{T}\log(T)\right),$$

Furthermore, the algorithm maintains at most $O(1)$ parameters can be implemented in time $O(1)$ per time step. Here $O(\cdot)$, $\Theta(\cdot)$ contain polynomial factors in $\gamma^{-1}, \kappa_B, \kappa, d$.

5.1. Sufficiency of Disturbance-Action Policies

The class of policies described in Definition 4.1 is powerful enough in its representational capacity to capture any fixed linear policy. Lemma 5.2 establishes this equivalence in terms of the state and action sequence each policy produces.

Lemma 5.2 (Sufficiency)**.** *For any two (κ, γ)-strongly stable matrices K^*, K, there exists a policy $\pi(M_*, K)$, with $M_* = (M_*^{[1]}, \dots, M_*^{[H]})$ defined as*

$$M_*^{[i]} = (K^* - K)(A - BK^*)^{i-1}$$

such that

$$\sum_{t=0}^{T} \left(c_t(x_t^K(M_*), u_t^K(M_*)) - c_t(x_t^{K^*}(0), u_t^{K^*}(0)) \right) \tag{5.1}$$

$$\leq T \cdot \frac{2G_c DWH\kappa_B^2\kappa^5(1-\gamma)^{H+1}}{\gamma} \tag{5.2}$$

5.2. Approximation Theorems

The following theorem relates the cost of $f_t(M_{t-H}, \dots M_t)$ with the actual cost $c_t(x_t, u_t)$.

Theorem 5.3. *For any (κ, γ)-strongly stable K, any number a and any sequence of policies $M_1 \dots M_T$ satisfying $\|M_t^{[i]}\| \leq a(1-\gamma)^i$, if the perturbations are bounded by W, we have that*

$$\sum_{t=1}^{T} f_t(M_{t-H}, \dots M_t) - \sum_{t=1}^{T} c_t(x_t^K, u_t^K) \tag{5.3}$$

$$\leq 2TG_c D^2\kappa^3(1-\gamma)^{H+1} \tag{5.4}$$

where

$$D \triangleq \frac{W(\kappa^2 + H\kappa_B\kappa^2 a)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{aW}{\gamma}$$

Before giving the proof of the above theorem, we will need a few lemmas which will be useful.

Lemma 5.4. *Let K be a (κ, γ)-strongly stable matrix, a be any number and M_t be a sequence such that for all i, t, we have $\|M_t^{[i]}\| \leq a(1-\gamma)^i$, then we have that for all i, t*

$$\|\Psi_{t,i}^K\| \leq \kappa^2(1-\gamma)^i \cdot \mathbf{1}_{i \leq H} + H\kappa_B\kappa^2 a(1-\gamma)^i$$

Proof of Lemma 5.4. The proof follows by noticing that

$$\begin{aligned}
\|\Psi_{t,i}^K\| &\leq \|\tilde{A}_K^i\| \mathbf{1}_{i \leq H} + \sum_{j=1}^{H} \|\tilde{A}_K^j\| \|B\| \|M_{t-j}^{[i-j]}\| \mathbf{1}_{i-j \in [1,H]} \\
&\leq \kappa^2(1-\gamma)^i \cdot \mathbf{1}_{i \leq H} + \sum_{j=1}^{H} \kappa_B\kappa^2 a(1-\gamma)^i \\
&\leq \kappa^2(1-\gamma)^i \cdot \mathbf{1}_{i \leq H} + H\kappa_B\kappa^2 a(1-\gamma)^i,
\end{aligned}$$

where the second and the third inequalities follow by using the fact that K is a (κ, γ)-strongly stable matrix and the conditions on the spectral norm of M. □

We now derive a bound on the norm of each of the states.

Lemma 5.5. *Suppose the system satisfies Assumption 3.1 and let M_t be a sequence such that for all i, t, we have that $\|M_t^{[i]}\| \leq a(1-\gamma)^i$ for a number a. Define*

$$D \triangleq \frac{W(\kappa^2 + H\kappa_B\kappa^2 a)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{aW}{\gamma}$$

Further suppose K^* is a (κ, γ)-strongly stable matrix. We have that for all t

$$\max(\|x_t^K\|, \|y_t^K(M_{t-H-1} \dots M_{t-1})\|, \|x_t(K^*)\|) \leq D$$

$$\max(\|u_t^K\|, \|v_t^K(M_{t-H} \dots M_t)\|) \leq D$$

$$\|x_t^K - y_t^K(M_{t-H-1} \dots M_{t-1})\| \leq \kappa^2(1-\gamma)^{H+1} D$$

$$\|u_t^K - v_t^K(M_{t-H} \dots M_t)\| \leq \kappa^3(1-\gamma)^{H+1} D$$

5.3. Bounding the properties of the OCO game with Memory

5.3.1. Bounding the Lipschitz Constant

Lemma 5.6. *Consider two policy sequences $\{M_{t-H} \dots M_{t-k} \dots M_t\}$ and $\{M_{t-H} \dots \tilde{M}_{t-k} \dots M_t\}$ which differ in exactly one policy played at a time step $t-k$ for $k \in \{0, \dots, H\}$. Then we have that*

$$\begin{aligned}
&|f_t(M_{t-H} \dots M_{t-k} \dots M_t) - f_t(M_{t-H} \dots \tilde{M}_{t-k} \dots M_t)| \\
&\leq 2G_c DW\kappa_B\kappa^3(1-\gamma)^k \sum_{i=0}^{H} \left(\|M_{t-k}^{[i]} - \tilde{M}_{t-k}^{[i]}\| \right).
\end{aligned}$$

Therefore using assumption 3.2 and Lemma 5.5, we immediately get that

$$\begin{aligned}
&f_t(M_{t-H} \dots M_{t-k} \dots M_t) - f_t(M_{t-H} \dots \tilde{M}_{t-k} \dots M_t) \\
&\leq 2G_c DW\kappa_B\kappa^3(1-\gamma)^k \sum_{i=0}^{H} \left(\|M_{t-k}^{[i]} - \tilde{M}_{t-k}^{[i]}\| \right)
\end{aligned}$$

5.3.2. Bounding the Gradient

Lemma 5.7. *For all M such that $\|M^{[j]}\| \leq a(1-\gamma)^j$ for all $j \in [1, H]$, we have that*

$$\|\nabla_M f_t(M \dots M)\|_F \leq G_c DWHd \left(\frac{2\kappa_B\kappa^3}{\gamma} + H \right)$$

Note that since M is a matrix, the ℓ_2 norm of the gradient $\nabla_M f_t$ corresponds to to the Frobenius norm of the $\nabla_M f_t$ matrix. Due to space constraints, we provide the proof in the appendix.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011, Budapest, Hungary*, pp. 1–26, 2011. URL http://www.jmlr.org/proceedings/papers/v19/abbasi-yadkori11a/abbasi-yadkori11a.pdf.

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

Abbasi-Yadkori, Y., Bartlett, P., and Kanade, V. Tracking adversarial targets. In *International Conference on Machine Learning*, pp. 369–377, 2014.

Abbasi-Yadkori, Y., Lazic, N., and Szepesvári, C. Regret bounds for model-free linear quadratic control. *CoRR*, abs/1804.06021, 2018. URL http://arxiv.org/abs/1804.06021.

Anava, O., Hazan, E., and Mannor, S. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pp. 784–792, 2015.

Arora, R., Dekel, O., and Tewari, A. Deterministic mdps with adversarial rewards and bandit feedback. *arXiv preprint arXiv:1210.4843*, 2012a.

Arora, R., Dekel, O., and Tewari, A. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012b.

Arora, S., Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Towards provable control for unknown linear dynamical systems. 2018.

Audenaert, K. M. A generalisation of mirsky's singular value inequalities. *arXiv preprint arXiv:1410.4941*, 2014.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Beckermann, B. and Townsend, A. On the singular values of matrices with displacement structure. *arXiv preprint arXiv:1609.09494*, 2016.

Bertsekas, D. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 2005.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Choi, M.-D. Tricks or treats with the hilbert matrix. *The American Mathematical Monthly*, 90(5):301–312, 1983.

Cohen, A., Hassidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. Online linear quadratic control. *arXiv preprint arXiv:1806.07104*, 2018.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. *CoRR*, abs/1805.09388, 2018. URL http://arxiv.org/abs/1805.09388.

Dekel, O. and Hazan, E. Better rates for any adversarial deterministic mdp. In *International Conference on Machine Learning*, pp. 675–683, 2013.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Fazel, M., Ge, R., Kakade, S. M., and Mesbahi, M. Global convergence of policy gradient methods for linearized control problems. *arXiv preprint arXiv:1801.05039*, 2018.

Ghahramani, Z. and Hinton, G. E. Parameter estimation for linear dynamical systems. Technical report, Technical Report CRG-TR-96-2, University of Toronto, Deptartment of Computer Science, 1996.

Grünbaum, F. A. A remark on hilbert's matrix. *Linear Algebra and its Applications*, 43:119–124, 1982.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL http://dx.doi.org/10.1561/2400000013.

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

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Spectral filtering for general linear dynamical systems. *arXiv preprint arXiv:1802.03981*, 2018.

Hilbert, D. Ein beitrag zur theorie des legendre'schen polynoms. *Acta mathematica*, 18(1):155–159, 1894.

Huang, W., Sun, F., Cao, L., Zhao, D., Liu, H., and Harandi, M. Sparse coding and dictionary learning with linear dynamical systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3938–3947, 2016.

Kalai, A. and Vempala, S. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Kalman, R. E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82.1: 35–45, 1960.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. ISSN 0890-5401. doi: http://dx.doi.org/10.1006/inco.1994.1009.

Ljung, L. *System identification: Theory for the User*. Prentice Hall, Upper Saddle Riiver, NJ, 2 edition, 1998.

Ljung, L. Prediction error estimation methods. *Circuits, Systems and Signal Processing*, 21(1):11–21, 2002.

Martens, J. Learning the linear dynamical system with asos. In Fürnkranz, J. and Joachims, T. (eds.), *Proceedings of the 27th International Conference on Machine Learning*, pp. 743–750. Omnipress, 2010a. URL http://www.icml2010.org/papers/532.pdf.

Martens, J. Learning the linear dynamical system with asos. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 743–750, 2010b.

Neu, G. and Gómez, V. Fast rates for online learning in linearly solvable markov decision processes. *arXiv preprint arXiv:1702.06341*, 2017.

Roweis, S. and Ghahramani, Z. A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345, 1999.

Schur, J. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und Angewandte Mathematik*, 140:1–28, 1911.

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

Shumway, R. H. and Stoffer, D. S. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Slepian, D. Prolate spheroidal wave functions, fourier analysis, and uncertainty: The discrete case. *Bell Labs Technical Journal*, 57(5):1371–1430, 1978.

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

Van Overschee, P. and De Moor, B. *Subspace Identification for Linear Systems*. Springer Science & Business Media, 2012.

Wan, E. A. and Van Der Merwe, R. The unscented kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pp. 153–158. IEEE, 2000.

Wang, Y.-S., Matni, N., and Doyle, J. C. A system level approach to controller synthesis. *IEEE Transactions on Automatic Control*, 2019.

Yu, J. Y. and Mannor, S. Online learning in markov decision processes with arbitrarily changing rewards and transitions. In *Game Theory for Networks, 2009. GameNets' 09. International Conference on*, pp. 314–322. IEEE, 2009.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

Zhou, K., Doyle, J. C., Glover, K., et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 928–936, 2003a.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 928–936, 2003b.

A. Appendix

B. Proofs

Proof of Theorem 5.1. Note that by the definition of the algorithm we have that all $M_t \in \mathcal{M}$, where

$$\mathcal{M} = \{M = \{M^{[1]} \dots M^{[H]}\} : \|M^{[i]}\| \leq \kappa^3 \kappa_B (1-\gamma)^i\}.$$

Let D be defined as

$$D \triangleq \frac{W(\kappa^2 + H\kappa_B \kappa^2 a)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{\kappa_B \kappa^3 W}{\gamma}.$$

Let K^* be the optimal linear policy in hindsight. By definition K^* is a (κ, γ)-strongly stable matrix. Using Lemma 5.2 and Theorem 5.3, we have that

$$\min_{M_* \in \mathcal{M}} \left(\sum_{t=0}^{T} f_t(M_*, \dots, M_*) \right) - \sum_{t=0}^{T} c_t(x_t^{K^*}(0), u_t^{K^*}(0)) \tag{B.1}$$

$$\leq \min_{M_* \in \mathcal{M}} \left(\sum_{t=0}^{T} c_t(x_t^{K}(M_*), u_t^{K}(M_*)) \right) - \sum_{t=0}^{T} c_t(x_t^{K^*}(0), u_t^{K^*}(0)) + 2TG_c D^2 \kappa^3 (1-\gamma)^{H+1}$$

$$\leq 2TG_c D(1-\gamma)^{H+1} \left(\frac{WH\kappa_B^2 \kappa^5}{\gamma} + D\kappa^3 \right). \tag{B.2}$$

Let $M_1 \dots M_T$ be the sequence of policies played by the algorithm. Note that by definition of the constraint set S, we have that

$$\forall t \in [T], \forall i \in [H] \quad \|M_t^{[i]}\| \leq \kappa_B \kappa^3 (1-\gamma)^i.$$

Using Theorem 5.3 we have that

$$\sum_{t=0}^{T} c_t(x_t^K, u_t^K) - \sum_{t=0}^{T} f_t(M_{t-H} \dots M_t) \leq 2TG_c D^2 \kappa^3 (1-\gamma)^{H+1}. \tag{B.3}$$

Finally using Theorem 4.6 and using Lemmas 5.6, 5.7 to bound the constants G_f and L associated with the function f_t and by noting that

$$\max_{M_1, M_2 \in \mathcal{M}} \|M_1 - M_2\| \leq \frac{\kappa_B \kappa^3 \sqrt{d}}{\gamma},$$

we have that

$$\sum_{t=0}^{T} f_t(M_{t-H} \dots M_t) - \min_{M_* \in \mathcal{M}} \sum_{t=0}^{T} f_t(M_*, \dots, M_*) \leq 8G_c W D d^{3/2} \kappa_B^2 \kappa^6 H^{2.5} \gamma^{-1} \sqrt{T}. \tag{B.4}$$

Summing up (B.1), (B.3) and (B.4), and using the condition that $H = \frac{\kappa^2}{\gamma} \log(T)$, we get the result. □

Proof of Lemma 5.2. By definition we have that

$$x_{t+1}(K^*) = \sum_{i=0}^{t} \tilde{A}_K^i w_{t-i}$$

Consider the following calculation for M_* with $M_*^{[i]} \triangleq (K^* - K)(A - BK^*)^{i-1}$ and for any $i \in \{0 \dots H\}$. We have that

$$\begin{aligned}
\Psi_{t,i}^K(M_*) &= \tilde{A}_K^i + \sum_{j=1}^{i} \tilde{A}_K^{i-j} B M^{[j]} \\
&= \tilde{A}_K^i + \sum_{j=1}^{i} \tilde{A}_K^{i-j} B(K^* - K)\tilde{A}_{K^*}^{j-1} \\
&= \tilde{A}_K^i + \sum_{j=1}^{i} \tilde{A}_K^{i-j} (\tilde{A}_{K^*} - \tilde{A}_K)\tilde{A}_{K^*}^{j-1} \\
&= \tilde{A}_K^i + \sum_{j=1}^{i} \left(\tilde{A}_K^{i-j}\tilde{A}_{K^*}^{j} - \tilde{A}_K^{i-j+1}\tilde{A}_{K^*}^{j-1} \right) \\
&= \tilde{A}_{K^*}^i
\end{aligned}$$

The final equality follows as the sum telescopes. Therefore, we have that

$$x_{t+1}^K(M_*) = \sum_{i=0}^{H} \tilde{A}_{K^*}^i w_{t-i} + \sum_{i=H+1}^{t} \Psi_{t,i}^K(M_*) w_{t-i}.$$

From the above we get that

$$\|x_t^{K^*}(0) - x_t^K(M_*)\| \le W \sum_{i=H+1}^{t} \|\Psi_{t,i}^K(M_*)\| \le \frac{WH\kappa_B^2\kappa^5(1-\gamma)^{H+1}}{\gamma}, \tag{B.5}$$

where the last inequality follows from using Lemma 5.4 and using the fact that $\|M_*^{[i]}\| \le \kappa_B\kappa^3(1-\gamma)^i$.

Further comparing the actions taken by the two policies we get that

$$\begin{aligned}
\|u_t^{K^*} - u_t^K(M_*)\| &= \left\| -K^* x_t^{K^*} + K x_t^K(M_*) - \sum_{i=0}^{t} (K^* - K)\tilde{A}_{K^*}^i w_{t-i} \right\| \\
&\le \left\| \sum_{i=H+1}^{t} K \left(\tilde{A}_{K^*}^i + \Psi_{t,i}^K(M_*) \right) w_{t-i} \right\| \\
&\le \frac{2WH\kappa_B^2\kappa^5(1-\gamma)^{H+1}}{\gamma}.
\end{aligned}$$

Using the above, Assumption 3.2 and Lemma 5.5, we get that

$$\sum_{t=0}^{T} \left(c_t(x_t^K(M_*), u_t^K(M_*)) - c_t(x_t^{K^*}, u_t^{K^*}) \right) \le T \cdot \frac{2G_c DWH\kappa_B^2\kappa^5(1-\gamma)^{H+1}}{\gamma}$$ □

Proof of Lemma 5.5. Using the definition of x_t we have that

$$\begin{aligned}
\|x_t^K\| &\le \kappa^2(1-\gamma)^{H+1}\|x_{t-H}\| + W \cdot \left(\sum_{i=0}^{2H} \|\Psi_{t,i}\| \right) \\
&\le \kappa^2(1-\gamma)^{H+1}\|x_{t-H}\| + W \cdot \left(\frac{\kappa^2 + H\kappa_B\kappa^2 a}{\gamma} \right)
\end{aligned}$$

The above recurrence can be seen to easily satisfy the following upper bound.

$$\|x_t^K\| \le \frac{W(\kappa^2 + H\kappa_B\kappa^2 a)}{\gamma(1 - \kappa^2(1-\gamma)^{H+1})} \le D \tag{B.6}$$

A similar bound can easily be established for

$$\|y_t^K(M_{t-H-1}\dots M_{t-1})\| \le W\cdot\left(\frac{\kappa^2+H\kappa_B\kappa^2 a}{\gamma}\right)\le D \tag{B.7}$$

It is also easy to see via the definitions that

$$\|x_t^K - y_t^K(M_{t-H-1}\dots M_{t-1})\| \le \|\tilde{A}_K^i\|\|x_{t-H}\| \le \kappa^2(1-\gamma)^{H+1}D \tag{B.8}$$

We can finally bound

$$\|x_t^{K^*}(0)\| \le \frac{W\kappa^2}{\gamma} \le D$$

For the actions we can use the definitions to bound the actions as follows using (B.6) and (B.7)

$$\|u_t^K\| \le \|Kx_t\| + \sum_{i=1}^H \|M_t^{[i]}w_{t-i}\| \le \kappa\|x_t^K\| + \frac{aW}{\gamma} \le D$$

$$\|v_t^K(M_{t-H}\dots M_t)\| \le \|Ky_t^K(M_{t-H-1}\dots M_{t-1})\| + \sum_{i=1}^H \|M_t^{[i]}w_{t-i}\| \le D.$$

We also have that using (B.8)

$$\begin{aligned}
&\|u_t^K - v_t^K(M_{t-H}\dots M)\| \\
&= K(x_t^K - y_t^K(M_{t-H-1}\dots M_{t-1})) \\
&\le \kappa^3(1-\gamma)^{H+1}D.
\end{aligned}$$

□

Finally, we prove Theroem 5.3.

Proof of Theorem 5.3. Using the above lemmas we can now bound the approximation error between f_t and c_t using Assumption 3.2

$$\begin{aligned}
&|c_t(x_t,u_t) - f_t(M_{t-H}\dots M_t)| \\
&= |c_t(x_t,u_t) - c_t(y_t^K(M_{t-H-1},\dots M_{t-1}), v_t^K(M_{t-H},\dots M_t))| \\
&\le G_cD\|x_t - y_t^K(M_{t-H-1},\dots M_{t-1})\| + G_cD\|u_t - v_t^K(M_{t-H},\dots M_t))\| \\
&\le 2G_cD^2\kappa^3(1-\gamma)^{H+1}.
\end{aligned}$$

This finishes the proof of Theorem 5.3. □

Proof of Lemma 5.6. For the rest of the proof, we will denote $y_{t+1}^K(\{M_{t-H}\dots M_{t-k}\dots M_t\})$ as y_{t+1}^K and $y_{t+1}^K(\{M_{t-H}\dots \tilde{M}_{t-k}\dots M_t\})$ as $\tilde{y}_{t+1}^K$. Similarly define v_t^K and $\tilde{v}_t^K$. It follows immediately from the definitions that

$$\begin{aligned}
\|y_t^K - \tilde{y}_t^K\| &= \|\tilde{A}_K^k B\sum_{i=0}^{2H}\left(M_{t-k}^{[i-k]} - \tilde{M}_{t-k}^{[i-k]}\right)w_{t-i}\mathbf{1}_{i-k\in[1,H]}\| \\
&\le \kappa_B\kappa^2(1-\gamma)^k W\sum_{i=0}^{H}\left(\|M_{t-k}^{[i]} - \tilde{M}_{t-k}^{[i]}\|\right).
\end{aligned}$$

Furthermore, we have that

$$\begin{aligned}
\|v_t^K - \tilde{v}_t^K\| &= \| - K(y_t - \tilde{y}_t) + \mathbf{1}_{k=0}\sum_{i=0}^{H}(M_t^{[i]} - \tilde{M}_t^{[i]})w_{t-i}\| \\
&\le 2\kappa_B\kappa^3(1-\gamma)^k W\sum_{i=0}^{H}\left(\|M_{t-k}^{[i]} - \tilde{M}_{t-k}^{[i]}\|\right).
\end{aligned}$$

□

Proof of Lemma 5.7. To derive a crude bound on the quantity in question, it will be sufficient to derive an absolute value bound on $\nabla_{M_{p,q}^{[r]}} f_t(M,\dots,M)$ for all r,p,q. To this end, we consider the following calculation. Using Lemma 5.5, we get that $y_t^K(M\dots M), v_t^K(M\dots M) \le D$. Therefore, using assumption 3.2, we have that

$$|\nabla_{M_{p,q}^{[r]}} f_t(M\dots M)| \le G_c D \left(\left\| \frac{\partial y_t^K(M)}{\partial M_{p,q}^{[r]}} + \frac{\partial v_t^K(M\dots M)}{\partial M_{p,q}^{[r]}} \right\| \right).$$

We now bound the quantities on the right-hand side:

$$\begin{aligned} \left\| \frac{\delta y_t^K(M\dots M)}{\delta M_{p,q}^{[r]}} \right\| &= \left\| \sum_{i=0}^{2H} \sum_{j=1}^{H} \left[\frac{\partial \tilde{A}_K^j B M^{[i-j]}}{\partial M_{p,q}^{[r]}} \right] w_{t-i} \mathbf{1}_{i-j\in[1,H]} \right\| \\ &\le \sum_{i=r}^{r+H} \left\| \left[\frac{\partial \tilde{A}_K^{i-r} B M^{[r]}}{\partial M_{p,q}^{[r]}} \right] w_{t-i} \right\| \le \frac{W\kappa_B\kappa^2}{\gamma}. \end{aligned}$$

Similarly,

$$\begin{aligned} \left\| \frac{\partial v_t^K(M\dots M)}{\partial M_{p,q}^{[r]}} \right\| &\le \kappa \left\| \frac{\delta y_t^K(M\dots M)}{\delta M_{p,q}^{[r]}} \right\| + \left\| \sum_{i=0}^{H} \frac{\partial M^{[i]}}{\partial M_{p,q}^{[r]}} w_{t-i} \right\| \\ &\le W \left(\frac{\kappa_B \kappa^3}{\gamma} + H \right). \end{aligned}$$

Combining the above inequalities gives the bound in the lemma. □

C. Proof of Theorem 4.6

Proof. By the standard OGD analysis, we know that

$$\sum_{t=H}^{T} \tilde{f}_t(x_t) - \min_{x\in\mathcal{K}} \sum_{t=H}^{T} \tilde{f}_t(x) \le \frac{D^2}{\eta} + TG^2\eta.$$

In addition, we know by (4.3) that, for any $t \ge H$,

$$\begin{aligned} |f_t(x_{t-H},\dots,x_t) - f_t(x_t,\dots,x_t)| &\le L\sum_{j=1}^{H} \|x_t - x_{t-j}\| \le L\sum_{j=1}^{H}\sum_{l=1}^{j} \|x_{t-l+1} - x_{t-l}\| \\ &\le L\sum_{j=1}^{H}\sum_{l=1}^{j} \eta \|\nabla \tilde{f}_{t-l}(x_{t-l})\| \le LH^2\eta G, \end{aligned}$$

and so we have that

$$\left| \sum_{t=H}^{T} f_t(x_{t-H},\dots,x_t) - \sum_{t=H}^{T} f_t(x_t,\dots,x_t) \right| \le TLH^2\eta G.$$

It follows that

$$\sum_{t=H}^{T} f_t(x_{t-H},\dots,x_t) - \min_{x\in\mathcal{K}} \sum_{t=H}^{T} f_t(x,\dots,x) \le \frac{D^2}{\eta} + TG_f^2\eta + LH^2\eta G_f T.$$

□

Algorithm 2 OGD with Memory (OGD-M).

Input: Step size η, functions $\{f_t\}_{t=m}^{T}$
Initialize $x_0, \ldots, x_{H-1} \in \mathcal{K}$ arbitrarily.
for $t = H, \ldots, T$ **do**
    Play x_t, suffer loss $f_t(x_{t-H}, \ldots, x_t)$
    Set $x_{t+1} = \Pi_{\mathcal{K}} \left(x_t - \eta \nabla \tilde{f}_t(x) \right)$
end for