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 linear dynamical systems with adversarial disturbances, as opposed to statistical noise. We present an efficient algorithm that achieves nearly-tight regret bounds in this setting. Our result generalizes upon previous work in two main aspects: the algorithm can accommodate adversarial noise in the dynamics, and can handle general convex costs.

1. Introduction

This paper studies the robust control of linear dynamical systems. A linear dynamical system (LDS) 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, under known transition dynamics specified by the matrices A and B. 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 $c_t(x_t, u_t)$ is minimized.

The above setting generalizes a fundamental problem in control theory, such as the Linear Quadratic Regulator, which has been studied over several decades. However, despite the significant amount of literature on the problem, several challenges remained.

[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, the standard approaches often assume i.i.d. Gaussian noise. Worst-case approaches in the control literature, also known as H_∞-control and its variants, are considered 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 $c_t(x, u)$ 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 via dynamic programming. 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 of the challenges outlined above: coping with adversarial noise, and general loss functions in an online setting. For this we turn to the algorithmic methodology of regret minimization in online learning. To define the performance metric, we denote the cost for a control algorithm $\mathcal{A}$ as

$$J_T(\mathcal{A}) = \sum_{t=0}^{T} c_t(x_t, u_t).$$

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, i.e.,

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

Our main result is an efficient algorithm for control which achieves $O(\sqrt{T})$ regret in the setting described above. While a similar setting has been considered in the literature before (Cohen et al., 2018), our work generalizes previous work in the following ways:

1. Our algorithm achieves regret $O(\sqrt{T})$ even in the presence of bounded adversarial disturbances. Previous

regret bounds needed to assume that the disturbances w_t are drawn from a distribution with zero mean and bounded variance.

2. Our regret bounds apply to any sequence of adversarially chosen convex loss functions. Previous efficient algorithms applied to convex quadratic costs only.

Our results above are obtained using several techniques from online convex optimization, notably online learning for loss functions with memory, and improper learning using convex relaxations.

2. Related Work

Online Learning: This work advocates for worst-case regret as a robust performance metric in the presence of adversarial noise. A special case of our setting is that of regret minimization in stateless systems (where $A = 0$), which is a well-studied problem in machine learning. We refer the reader to various books and surveys on the topic (Cesa-Bianchi & Lugosi, 2006; Shalev-Shwartz et al., 2012; Hazan, 2016). Of particular interest to our study is the setting of online learning with memory (Anava et al., 2015).

Learning and Control in Linear Dynamical Systems: The modern setting for linear dynamical systems arose in the seminal work of Kalman (1960), who introduced the Kalman filter as a recursive least-squares solution for maximum likelihood estimation (MLE) of Gaussian perturbations to the system in latent-state systems. The framework and filtering algorithm have proven to be a mainstay in control theory and time-series analysis. We refer the reader to the classic survey (Ljung, 1998), and the extensive overview of recent literature in (Hardt et al., 2018). Most of this literature, as well as much of classical control theory, deals with zero-mean random noise, typically normally distributed.

Recently, there has been a renewed interest in learning both fully-observable and latent-state linear dynamical systems. For fully-observable systems, sample complexity and regret bounds for control (under Gaussian noise) were obtained in (Abbasi-Yadkori & Szepesvári, 2011; Dean et al., 2018; Abbasi-Yadkori et al., 2019). The technique of spectral filtering for learning and open-loop control of non-observable systems was introduced and studied in (Hazan et al., 2017; Arora et al., 2018; Hazan et al., 2018). Provable control in the Gaussian noise setting was also studied in (Fazel et al., 2018).

Robust Control: A notable attempt to handle adversarial perturbations in the dynamics takes place in H_∞ control (Stengel, 1994; Zhou et al., 1996). In this setting, the controller solves for the best linear controller assuming worst case noise to come, i.e.,

$$\min_{K_1} \max_{\varepsilon_1} \min_{K_2} \ldots \min_{K_T} \max_{\varepsilon_T} \sum_t c_t(x_t, u_t),$$

assuming similar linear dynamics as in equation (1.1). This approach is overly pessimistic, and leads to sub-optimal performance in various cases of interest. In comparison, we do not solve for the entire noise trajectory in advance, but adjust for it iteratively, and, in this manner, offer an instance-dependent guarantee. 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.

Non-stochastic MDPs: The setting we consider, namely control in systems with linear transition dynamics (Bertsekas, 2005) in the presence of adversarial disturbances, can be cast as that of planning in an adversarially changing MDP (Arora et al., 2012; 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 exponential in the dimension of the system. In addition, the regret in these scales 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, these assume that the dynamics are fixed and without (adversarial) noise.

LQR with Changing Costs: For the Linear Quadratic Regulator problem, Cohen et al. (2018) consider changing quadratic costs with stochastic noise to achieve a $O(\sqrt{T})$ regret bound. This work is well aligned with our results, and the present paper employs some notions developed therein (e.g., strong stability). However, the techniques used in (Cohen et al., 2018) (such as the SDP formulation for a linear controller) heavily rely 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 an SDP formulation to determine the best linear controller for convex losses. In an earlier work, Abbasi-Yadkori & Szepesvári (2011) consider a more restricted setting with fixed, deterministic dynamics (hence, noiseless) and changing quadratic costs. Ng & Kim (2005) consider the adaptation of online learning to control in a manner that ensures stability. While the work does not provide any regret guarantee, the proposed notion of stability continues to be useful in the literature (e.g. (Cohen et al., 2018)) and this paper.

Iterative Learning Control: We also note the similarity of our proposed algorithm to methods involving iterative learning control (Ahn et al., 2007; Owens & Hätönen, 2005)

and feedback error learning (Nakanishi & Schaal, 2004). These procedures, often operating over deterministic systems in the episodic setting, attempt to learn a control law via iteratively adjusting the controller as a function of the tracking error.

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 upon observing the state x_t and incurs a cost of $c_t(x_t, u_t)$, where c_t 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 the system suffers at each time step. In this paper, we make no distributional assumptions on w_t, and the sequence w_t is not made known to the learner in advance.

For any algorithm $\mathcal{A}$, we attribute a cost of

$$J_T(\mathcal{A}) = \sum_{t=0}^{T} c_t(x_t, u_t),$$

where $x_{t+1} = Ax_t + Bu_t + w_t$ and $u_t = \mathcal{A}(x_1, \dots x_t)$. Overloading notation, we shall use $J(K)$ to denote the cost of a linear controller 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 than those considered by previous works, and hence allow for more general systems. In particular, we allow for adversarial (rather than i.i.d. stochastic) noise, and we also handle general convex cost functions. In addition, 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, i.e., $\|A\| \le \kappa_A, \|B\| \le \kappa_B$. The perturbation introduced per time step is bounded, i.e., $\|w_t\| \le W$.*

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*

$$|c_t(x, u)| \le \beta D^2, \text{ and}$$

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

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, Q satisfying $A - BK = QLQ^{-1}$, such that following two conditions are met:*

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

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 the best linear controller, i.e.,

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

3.4. Proof Techniques and Overview

Choice of Policy Class: We begin by parameterizing the policy we execute at every step as a linear function of the disturbances in the past in Definition 4.1. Similar parameterization has been considered in the system level synthesis framework (see (Wang et al., 2019)). This leads to a convex relaxation of the problem. Optimization on alternative parameterizations, including an SDP based framework (Cohen et al., 2018) or a direct parameterization (Fazel et al., 2018), has been studied in the literature, but they seem unable to capture general convex functions as well as adversarial disturbance or lead to a non-convex loss. To avoid a linear dependence on time for the number of parameters in our policy, we additionally include a stable linear controller in our policy allowing us to effectively consider only $O(\gamma^{-1} \log(T))$ previous perturbations. Lemma 5.2 makes this notion of approximation precise.

Reduction to OCO with Memory: The choice of policy class with an appropriately chosen horizon H allows us to reduce the problem to compete with functions with truncated memory. This naturally falls under the class of online convex optimization with memory (see Section 4.5). Theorem 5.3 makes this reduction precise. Finally to bound the regret on truncated functions we use the Online Gradient Descent based approach specified in (Anava et al., 2015), which requires a bound on Lipschitz constants which we provide in Section 5.3.1. This reduction is inspired by the ideas introduced in (Even-Dar et al., 2009).

3.5. Roadmap

The following section provides the notation required to define our algorithm and regret bounds. Section 5 describes our primary method (Algorithm 1) and establishes the central regret guarantees (Theorem 5.1), along with the requisite lemmas and their proofs.

4. Preliminaries

In this section, we establish some important definitions that will prove to be useful throughout the paper.

4.1. Notation

We reserve the letters x, y for states and u, v for control actions. We denote by $d = \max(\dim(x), \dim(u))$, i.e., a bound on the dimensionality of the problem. We reserve capital letters A, B, K, M for matrices associated with the system and the policy. Other capital letters are reserved for universal constants in the paper. We use the shorthand $M_{i:j}$ to denote a subsequence $\{M_i, \ldots, M_j\}$.

4.2. A Disturbance-Action Policy Class

We establish 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. The situation may be contrasted with that of a linear controller. While the action recommended by a linear controller is also linear in past disturbances (a consequence of being linear in the current state), the state sequence produced on the execution of a linear policy is not a linear function of its parameters.

Definition 4.1 (Disturbance-Action Policy)**.** *A disturbance-action policy $\pi(M, K)$ is specified by parameters $M = (M^{[0]}, \ldots, M^{[H-1]})$ and a fixed matrix K, for horizon $H \geq 1$. At every time t, such a policy $\pi(M, K)$ chooses the recommended action u_t at a state x_t[1], defined as*

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

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.

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

4.3. Evolution of State

This section describes 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 $\pi_t(M_t = (M_t^{[0]}, \ldots, M_t^{[H-1]}), K)$. 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 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. $x_t^K(M_{0:t-1})$ is the state attained by the system upon execution of a non-stationary policy $\pi(M_{0:t-1}, K)$. We similarly define $u_t^K(M_{0:t-1})$ to be the action executed at time t. If the same policy M is used across all time steps, we compress the notation to $x_t^K(M), u_t^K(M)$. Note that $x_t^K(0), u_t^K(0)$ refers to running the linear policy K.

3. $\Psi_{t,i}^{K,h}(M_{t-h:t})$ is a transfer matrix that describes the effect of w_{t-i} with respect to the past $h+1$ policies on the state x_{t+1}, formally defined below. When M is the same across all arguments we compress the notation to $\Psi_{t,i}^{K,h}(M)$.

Definition 4.2. *For any $t, h \leq t, i \leq H + h$, define the disturbance-state transfer matrix $\Psi_{t,i}^{K,h}$ to be a function with $h+1$ inputs defined as*

$$\Psi_{t,i}^{K,h}(M_{t-h:t}) = \tilde{A}_K^i \mathbf{1}_{i \leq h} + \sum_{j=0}^{h} \tilde{A}_K^j B M_{t-j}^{[i-j-1]} \mathbf{1}_{i-j \in [1,H]}.$$

It will be worthwhile to note that $\Psi_{t,i}^{K,h}$ is linear in its arguments $M_{t-h:t}$. For the rest of the paper we will set h to be H unless specified otherwise.

Lemma 4.3. *If u_t is chosen as what the non-stationary policy $\pi((M_0, \ldots, M_T), K)$ recommends, then the state sequence satisfies the following recurrence for any time t and $h \geq 0$:*

$$x_{t+1} = \tilde{A}_K^{h+1} x_{t-h} + \sum_{i=0}^{H+h} \Psi_{t,i}^{K,h}(M_{t-h:t}) w_{t-i}. \quad (4.1)$$

Proof. For any time t, we will prove the claim inductively

on h. For $h = 0$, we have that

$$x_{t+1} = \tilde{A}_K x_t + \sum_{i=1}^{H} B M_t^{[i-1]} w_{t-i} + w_t$$
$$= \tilde{A}_K x_t + \sum_{i=0}^{H} \Psi_{t,i}^{K,0}(M_t) w_{t-i}$$

Further suppose the lemma holds for some $h \geq 0$, we prove it for $h+1$. We have that

$$x_{t+1} = \tilde{A}_K^{h+1} x_{t-h} + \sum_{i=0}^{H+h} \Psi_{t,i}^{K,h}(M_{t-h:t}) w_{t-i}$$
$$= \tilde{A}_K^{h+1}(\tilde{A}_K x_{t-h-1} + \sum_{i=1}^{H} B M_{t-h-1}^{[i-1]} w_{t-h-i-1} +$$
$$w_{t-h-1}) + \sum_{i=0}^{H+h} \Psi_{t,i}^{K,h}(M_{t-h:t}) w_{t-i}$$
$$= \tilde{A}_K^{h+2} x_{t-h-1} + \sum_{i=0}^{H+h} \Big(\Psi_{t,i}^{K,h}(M_{t-h:t}) +$$
$$\tilde{A}_K^{i} \mathbf{1}_{i=h+1} + \tilde{A}_K^{h+1} B M_{t-h-1}^{[i-h-2]} \mathbf{1}_{i-h-1 \in [1,H]} \Big) w_{t-i}$$
$$= \tilde{A}_K^{h+2} x_{t-h-1} + \sum_{i=0}^{H+h+1} \Psi_{t,i}^{K,h+1}(M_{t-h-1:t}) w_{t-i}.$$

The proof now follows by induction. □

4.4. 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 notion of an *ideal* state.

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

$$y_{t+1}^K(M_{t-H:t}) = \sum_{i=0}^{2H} \Psi_{t,i}^{K,H}(M_{t-H:t}) w_{t-i},$$
$$v_{t+1}^K(M_{t-H:t+1}) = -K y_{t+1}^K(M_{t-H:t})$$
$$+ \sum_{i=1}^{H} M_{t+1}^{[i-1]} w_{t+1-i}.$$

When M is the same across all arguments we compress the notation to $y_{t+1}^K(M), v_{t+1}^K(M)$.

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.,*

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

When M is the same across all arguments we compress the notation to $f_t(M)$.

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 $M_{t-H-1:t}$ and hence convex in $M_{t-H-1:t}$. This renders it amenable to algorithms for online convex optimization.

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

4.5. 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 $f_t(x_{t-H:t})$. We assume a certain coordinate-wise Lipschitz regularity on f_t of the form such that, for any $j \in \{0, \ldots, H\}$, for any $x_{0:H}, \tilde{x}_j \in \mathcal{K}$,

$$|f_t(x_{0:j-1}, x_j, x_{j+1:H}) - f_t(x_{0:j-1}, \tilde{x}_j, x_{j+1:H})|$$
$$\leq L\|x_j - \tilde{x}_j\|. \quad (4.2)$$

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

$$G_f = \sup_{t \in \{0,\ldots,T\}, x \in \mathcal{K}} \|\nabla \tilde{f}_t(x)\|, \quad D = \sup_{x,y \in \mathcal{K}} \|x - y\|. \quad (4.3)$$

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

$$\texttt{PolicyRegret} = \sum_{t=H}^{T} f_t(x_{t-H:t}) - \min_{x \in \mathcal{K}} \sum_{t=0}^{T} \tilde{f}_t(x).$$

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

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 $\mathcal{M} = \{M = \{M^{[0]} \dots M^{[H-1]}\} : \|M^{[i-1]}\| \leq \kappa^3 \kappa_B (1-\gamma)^i\}$.
Initialize $M_0 \in \mathcal{M}$ arbitrarily.
for $t = 0, \dots, T-1$ **do**
  Choose the action $u_t = c_t - Kx_t + \sum_{i=1}^{H} M^{[i-1]} w_{t-i}$.
  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 to Definition 4.5)
  Set $M_{t+1} = \Pi_{\mathcal{M}}(M_t - \eta \nabla g_t(M_t))$
end for

Theorem 4.6. *Let $\{f_t\}_{t=H}^{T}$ 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 (4.2) and (4.3). Then, there exists an algorithm which generates a sequence $\{x_t\}_{t=0}^{T}$ such that*

$$\sum_{t=H}^{T} f_t(x_{t-H:t}) - \min_{x \in \mathcal{K}} \sum_{t=H}^{T} \tilde{f}_t(x) \leq 3D\sqrt{G_f(G_f + LH^2)T}.$$

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*

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

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

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

$$\mathcal{M} = \{M = \{M^{[0]} \dots M^{[H-1]}\} : \|M^{[i-1]}\| \leq \kappa^3 \kappa_B (1-\gamma)^i\}.$$

Let D be defined as

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

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

$$\min_{M_* \in \mathcal{M}} \left(\sum_{t=0}^{T} f_t(M_*) \right) - \sum_{t=0}^{T} c_t(x_t^{K^*}(0), u_t^{K^*}(0)) \quad (5.1)$$

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

$$\leq 2TG_c D (1-\gamma)^{H+1} \left(\frac{WH\kappa_B^2 \kappa^5}{\gamma} + D\kappa^3 \right). \quad (5.2)$$

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

$$\sum_{t=0}^{T} c_t(x_t^K, u_t^K) - \sum_{t=0}^{T} f_t(M_{t-H-1:t}) \leq 2TG_c D^2 \kappa^3 (1-\gamma)^{H+1}. \quad (5.3)$$

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

$$\sum_{t=0}^{T} f_t(M_{t-H} \dots M_t) - \min_{M_* \in \mathcal{M}} \sum_{t=0}^{T} f_t(M_*, \dots, M_*) \leq 8G_c W D d^{3/2} \kappa_B^2 \kappa^6 H^{2.5} \gamma^{-1} \sqrt{T}. \quad (5.4)$$

Summing up (5.1), (5.3) and (5.4), and using the condition that $H = \frac{\kappa^2}{\gamma} \log(T)$, we get the result. □

5.1. Sufficiency of Disturbance-Action Policies

The class of policies described in Definition 4.1 is powerful enough to capture any fixed linear policy. Lemma 5.2 establishes this equivalence in terms of the state and action sequence that 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 $M_* = (M_*^{[0]}, \dots, M_*^{[H-1]})$ defined as*

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

such that

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

Proof of Lemma 5.2. By definition we have that

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

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

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

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

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

From the above we get that

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

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

Further comparing the actions taken by the two policies, we may see that

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

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

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

5.2. Approximation Theorems

The following theorem relates the cost of $f_t(M_{t-H-1:t})$ with the actual cost $c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t}))$.

Theorem 5.3. *For any (κ, γ)-strongly stable K, any $\tau > 0$, and any sequence of policies $M_1 \dots M_T$ satisfying $\|M_t^{[i]}\| \leq \tau(1-\gamma)^i$, if the perturbations are bounded by W, we have that*

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

where

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

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, $\tau > 0$, and M_t be a sequence such that for all i, t, we have $\|M_t^{[i]}\| \leq \tau(1-\gamma)^i$. Then we have that for all i, t, h,*

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

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 $\|M_t^{[i]}\| \leq \tau(1-\gamma)^i$ for a number τ. Define*

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

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

$$\|x_t^K(M_{0:t-1})\|, \|y_t^K(M_{t-H-1:t-1})\|, \|x_t^{K^*}(0)\| \leq D$$

$$\|u_t^K(M_{0:t-1})\|, \|v_t^K(M_{t-H-1:t})\| \leq D$$

$$\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1})\| \leq \kappa^2(1-\gamma)^{H+1}D$$

$$\|u_t^K(M_{0:t}) - v_t^K(M_{t-H-1:t})\| \leq \kappa^3(1-\gamma)^{H+1}D.$$

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

$$\begin{aligned}
&|c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - f_t(M_{t-H-1:t})| \\
&= |c_t(x_t^K(M_{0:t}), u_t^K(M_{0:t-1})) \\
&\quad - c_t(y_t^K(M_{t-H-1:t-1}), v_t^K(M_{t-H-1:t}))| \\
&\leq G_c D\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1})\| + \\
&\quad G_c D\|u_t^K(M_{0:t}) - v_t^K(M_{t-H-1:t}))\| \\
&\leq 2G_c D^2\kappa^3(1-\gamma)^{H+1}.
\end{aligned}$$

This finishes the proof of Theorem 5.3. □

5.3. Bounding the Properties of OCO with Memory

Thus far, Lemma 5.2 establishes that it is sufficient to compare against the class of disturbance-action policies. For such policies, Theorem 5.3 reduces the counter-factual regret minimization problem to online learning on loss functions with memory. It remains to quantify the constants with which Theorem 4.6 may be invoked to obtain a regret bound. Note that Line 3 of Algorithm 1 places on upper bound on the diameter D.

5.3.1. Bounding the Lipschitz Constant

We begin by establishing an upper bound on the lipschitz constant L as defined in Equation 4.2.

Lemma 5.6. *Consider two policy sequences $\{M_{t-H-1}\dots M_{t-k}\dots M_t\}$ and $\{M_{t-H-1}\dots \tilde{M}_{t-k}\dots M_t\}$ which differ in exactly one policy played at a time step $t-k$ for $k \in \{0,\dots,H\}$. Then we have that*

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

5.3.2. Bounding the Gradient

A bound on the norm of the gradient follows similarly.

Lemma 5.7. *For all M such that $\|M^{[j]}\| \leq \tau(1-\gamma)^j$ for all $j \in [0, H-1]$, we have that*

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

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.

6. Conclusion

In this paper, we demonstrate an algorithmic methodology to control linear dynamical systems with adversarial disturbances through regret minimization, as well as how to handle general convex costs. Our notion of a robust controller is able to learn and adapt according to the noise encountered during the process. This deviates from the study of robust control in the framework of H_∞ control, that attempts to find a control with the anticipation of worst-case instances for all future perturbations.

Acknowledgements

Elad Hazan acknowledges funding from the NSF, award number 1704860. Sham Kakade acknowledges funding from the Washington Research Foundation for Innovation in Data-intensive Discovery, the DARPA award FA8650-18-2-7836, and the ONR award N00014-18-1-2247.

References

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

Ahn, H.-S., Chen, Y., and Moore, K. L. Iterative learning control: Brief survey and categorization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1099–1121, 2007.

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. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1503–1510, 2012.

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

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.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. Online linear quadratic control. In *International Conference on Machine Learning*, pp. 1028–1037, 2018.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pp. 4188–4197, 2018.

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

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 the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 1466–1475, 2018.

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. In *Advances in Neural Information Processing Systems*, pp. 4634–4643, 2018.

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

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

Nakanishi, J. and Schaal, S. Feedback error learning and nonlinear adaptive control. *Neural Networks*, 17(10): 1453–1465, 2004.

Ng, A. Y. and Kim, H. Stable adaptive control with online learning. In *Advances in Neural Information Processing Systems*, pp. 977–984, 2005.

Owens, D. H. and Hätönen, J. Iterative learning controlan optimization paradigm. *Annual reviews in control*, 29(1): 57–70, 2005.

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

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

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., 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.

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

Appendix

A. Proof of Theorem 4.6

Proof. By the standard OGD analysis, we know that

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

In addition, we know by (4.2) that, for any $t \geq H$,

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

and so we have that

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

It follows that

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

The theorem follows after setting $\eta = \frac{D}{\sqrt{G_f(G_f + LH^2)T}}$. □

B. Proof of Lemma 5.4

Proof of Lemma 5.4. The proof follows by noticing that

$$\begin{aligned} \|\Psi_{t,i}^{K,h}\| &\leq \|\tilde{A}_K^i\| \mathbf{1}_{i \leq H} + \sum_{j=0}^{h} \|\tilde{A}_K^j\| \|B\| \|M_{t-j}^{[i-j-1]}\| \mathbf{1}_{i-j \in [1,H]} \\ &\leq \kappa^2 (1-\gamma)^i \cdot \mathbf{1}_{i \leq H} + \sum_{j=1}^{H} \kappa_B \kappa^2 \tau (1-\gamma)^{i-1} \\ &\leq \kappa^2 (1-\gamma)^i \cdot \mathbf{1}_{i \leq H} + H\kappa_B \kappa^2 \tau (1-\gamma)^{i-1}, \end{aligned}$$

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. □

C. Proof of Lemma 5.5

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

$$\|x_t^K(M_{0:t-1})\| \le \kappa^2(1-\gamma)^{H+1}\|x_{t-H-1}^K(M_{0:t-H-2})\| + W \cdot \left(\sum_{i=0}^{2H}\|\Psi_{t-1,i}^{K,H}(M_{t-H,t-1})\|\right)$$
$$\le \kappa^2(1-\gamma)^{H+1}\|x_{t-H-1}(M_{0:t-H-2})\| + W \cdot \left(\frac{\kappa^2 + H\kappa_B\kappa^2\tau}{\gamma}\right).$$

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

$$\|x_t^K(M_{0:t-1})\| \le \frac{W(\kappa^2 + H\kappa_B\kappa^2 a)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} \le D. \tag{C.1}$$

A similar bound can be established for

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

It is also simple to see via the definitions that

$$\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1})\| \le \|\tilde{A}_K^i\|\|x_{t-H-1}\| \le \kappa^2(1-\gamma)^{H+1}D. \tag{C.3}$$

We can now 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 (C.1) and (C.2)

$$\|u_t^K(M_{0:t})\| \le \|Kx_t^K(M_{0:t-1})\| + \sum_{i=1}^{H}\|M_t^{[i-1]}w_{t-i}\| \le \kappa\|x_t^K(M_{0:t-1})\| + \frac{\tau W}{\gamma} \le D,$$

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

We also have that using (C.3)

$$\|u_t^K(M_{t-H-1:t}) - v_t^K(M_{t-H-1:t})\| = K(x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1})) \le \kappa^3(1-\gamma)^{H+1}D.$$

□

D. Proof of Lemma 5.6

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

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

Furthermore, we have that

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

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

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

□

E. Proof of Lemma 5.7

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 $\nabla_{M_{p,q}^{[r]}} f_t(M)$ for all r, p, q. To this end, we consider the following calculation. Using Lemma 5.5, we get that $y_t^K(M), v_t^K(M) \leq D$. Therefore, using Assumption 3.2, we have that

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

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

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

Similarly,

$$\left\| \frac{\partial v_t^K(M)}{\partial M_{p,q}^{[r]}} \right\| \leq \kappa \left\| \frac{\delta y_t^K(M)}{\delta M_{p,q}^{[r]}} \right\| + \left\| \sum_{i=1}^{H} \frac{\partial M^{[i-1]}}{\partial M_{p,q}^{[r]}} w_{t-i} \right\| \leq W \left(\frac{\kappa_B \kappa^3}{\gamma} + H \right).$$

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