Online Optimal Control with Affine Constraints

ID 5929, Primary: ML-Online Learning & Bandits, Secondary: ML-Reinforcement Learning

Abstract

This paper considers online optimal control with affine constraints on the states and actions under linear dynamics with random disturbances. We consider convex stage cost functions that change adversarially. Besides, we consider time-invariant and known system dynamics and constraints. To solve this problem, we propose Online Gradient Descent with Buffer Zone (OGD-BZ). Theoretically, we show that OGD-BZ can guarantee the system to satisfy all the constraints despite any realization of the disturbances under proper parameters. Further, we investigate the policy regret of OGD-BZ, which compares OGD-BZ's performance with the performance of the optimal linear policy in hindsight. We show that OGD-BZ can achieve $\tilde{O}(\sqrt{T})$ policy regret under proper parameters, where $\tilde{O}(\cdot)$ absorbs logarithmic terms of T.

1 Introduction

Recent years have witnessed a growing interest on solving control problems by leveraging learning-based techniques, e.g. online learning and/or reinforcement learning (Agarwal et al. 2019; Dean et al. 2018; Ibrahimi, Javanmard, and Roy 2012; Dean et al. 2019a; Fazel et al. 2018; Yang et al. 2019). This is motivated by various applications, such as data center cooling (Lazic et al. 2018), robotics (Fisac et al. 2018), autonomous vehicles (Sallab et al. 2017), etc. However, for real-world implementation, it is crucial to design safe algorithms that guarantee the system to satisfy certain (physical) constraints despite unknown disturbances. For example, the temperatures of a data center should be maintained within certain range to reduce task failures despite possible disturbances from external heat sources, and quadrotors should avoid collision with obstacles even when perturbed by wind, etc. In addition to safety, many applications involve time-varying environments, such as varying electricity prices and moving targets. Therefore, the safe algorithm design should not be over-conservative and should adapt to time-varying environments for desirable online performance.

In this paper, we design safe algorithms for time-varying environments by considering the following constrained online optimal control problem. Specifically, we consider a linear system with random disturbances

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t \geq 0, \tag{1}$$

where disturbance w_t is random and satisfies $\|w_t\|_\infty \leq \bar{w}$. Consider affine constraints on the state x_t and the action u_t:

$$D_x x_t \leq d_x, \quad D_u u_t \leq d_u, \quad \forall\, t \geq 0. \tag{2}$$

For simplicity, we assume the system parameters $A, B, \bar{w}$ and the constraints are known. At stage $0 \leq t \leq T$, a convex cost function $c_t(x_t, u_t)$ is adversarially generated and the decision maker selects a feasible action u_t before $c_t(x_t, u_t)$ is revealed. We aim to achieve two goals simultaneously: (i) to minimize the sum of the adversarially varying costs, (ii) to satisfy the constraints (2) for all t despite the disturbances. There is a rich body of work addressing each goal separately but lack results on both goals together as discussed below.

Firstly, there is recent progress on online optimal control to address the goal (i). A commonly adopted performance metric is policy regret, which compares the online cost with the cost of the optimal linear policy in hindsight (Agarwal et al. 2019). Sublinear policy regrets have been achieved for linear systems with either stochastic disturbances (Cohen et al. 2018; Agarwal, Hazan, and Singh 2019) or adversarial disturbances (Agarwal et al. 2019; Foster and Simchowitz 2020; Goel and Hassibi 2020). However, most literature only considers the unconstrained control problem. Recently, (Nonhoff and Müller 2020) studies constrained online optimal control but assumes no disturbances.

Secondly, there are many papers from the control community to address goal (ii): constraints satisfaction. Perhaps the most famous algorithms are Model Predictive Control (MPC) (Rawlings and Mayne 2009) and its variants, such as robust MPC which guarantees constraints satisfaction in the presence of disturbances (Bemporad and Morari 1999; Kouvaritakis, Rossiter, and Schuurmans 2000; Mayne, Seron, and Raković 2005; Limon et al. 2010; Zafiriou 1990). However, robust MPC tends to sacrifice optimality for safety. Further, there lacks regret/optimality analysis for robust MPC under adversarially varying costs.

Therefore, an important question remains to be addressed: *Q: how to design online algorithms to both satisfy the constraints despite disturbances and yield $o(T)$ policy regrets?*

Our contributions. In this paper, we answer the question above by proposing an online control algorithm–Online Gradient Descent with Buffer Zones (OGD-BZ). To develop OGD-BZ, we first convert the constrained online optimal control problem as an OCO problem with temporal-coupled

stage costs and temporal-coupled stage constraints, and then convert the temporal-coupled OCO problem to a classical OCO problem. The problem conversion leverages the techniques from recent unconstrained online control literature and robust optimization literature. Since the conversion is not exact/equivalent, we tighten the constraint set by adding buffer zones to account for approximation errors caused by the problem conversion. We then apply classical OCO algorithms such as OGD to solve the problem and call the resulting algorithm as OGD-BZ.

Theoretically, we show that, with proper parameters, OGD-BZ can ensure all the states and actions to satisfy the constraints (2) for any disturbances bounded by $\bar{w}$. In addition, we show that OGD-BZ's policy regret can be bounded by $\tilde{O}(\sqrt{T})$ for general convex cost functions $c_t(x_t, u_t)$ under proper assumptions and parameters. As far as we know, OGD-BZ is the first algorithm with theoretical guarantees on both sublinear policy regret and robust constraints satisfaction. Further, our theoretical results explicitly characterizes a trade-off between the constraints satisfaction and the low regret when deciding the size of the buffer zone of OGD-BZ: a larger buffer zone, which indicates a more conservative search space, is preferred for constraints satisfaction; while a smaller buffer zone is preferred for low regret.

Related work. We provide more literature review below.

Safe reinforcement learning for control systems. There is a rich body of literature on safe RL and safe learning-based control that studies how to learn optimal policies without violating constraints and without knowing the system (Fisac et al. 2018; Aswani et al. 2013; Wabersich and Zeilinger 2018; Garcıa and Fernández 2015; Cheng et al. 2019; Zanon and Gros 2019; Fulton and Platzer 2018). Perhaps the most relevant paper is (Dean et al. 2019b), which proposes algorithms to learn optimal linear policies for a constrained linear quadratic regulator problem. However, most theoretical guarantees in the safe RL literature require time-invariant environment and there lacks policy regret analysis when facing time-varying objectives. This paper addresses the time-varying objectives but considers known system dynamics. It is our ongoing work to combine both safe RL and our approach to design safe learning algorithms with policy regret guarantees in time-varying problems.

Another important notion of safety is the system stability, which is also studied in the safe RL/learning-based control literature (Dean et al. 2018, 2019a; Chow et al. 2018).

Online convex optimization (OCO). (Hazan 2019) provides a review on classical (decoupled) OCO. OCO with memory considers coupled costs and decoupled constraints (Anava, Hazan, and Mannor 2015). There are also papers on OCO with coupled constraints (Yuan and Lamperski 2018; Cao, Zhang, and Poor 2018; Kveton et al. 2008), where constraint violation is usually allowed. Further, OCO does not consider dynamical systems, let alone system disturbances.

Constrained optimal control. Constrained optimal control enjoys a long history of research. Without disturbances, it is known that the optimal controller for linearly constrained LQR is piecewise linear (Bemporad et al. 2002). With disturbances (as considered in this paper), the problem is much more challenging. Existing approaches, such as robust tube-based MPC (Limon et al. 2008; Rawlings and Mayne 2009; Limon et al. 2010), usually sacrifices optimality for feasibility. Linear policies are also deployed in the literature (Limon et al. 2008; Dean et al. 2019b; Schildbach, Goulart, and Morari 2015), although linear policies may not be optimal for constrained optimal control with disturbances.

Notations and conventions We let $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ denote the L_1, L_2, L_∞ norms respectively for vectors and matrices. We also use $\|\cdot\|$ to denote L_2 norm. For two vectors $a, b \in \mathbb{R}^n$, we write $a \leq b$ if $a_i \leq b_i$ for any entry i. For better exposition, some bounds use $\Theta(\cdot)$ to omit constants that do not depend on T or the problem dimensions explicitly.

2 Problem Formulation

In this paper, we consider an online optimal control problem with linear dynamics and affine constraints. Specifically, at each stage $t = 0, 1, \ldots, T$, an agent observes the current state x_t and implements an action u_t, which incurs a cost $c_t(x_t, u_t)$. The stage cost function $c_t(\cdot, \cdot)$ is generated adversarially and revealed to the agent after the action u_t is taken. The system evolves to the next state according to (1), where x_0 is fixed, w_t is a random disturbance bounded by $w_t \in \mathcal{W} = \{w \in \mathbb{R}^n : \|w\|_\infty \leq \bar{w}\}$, and states and actions should satisfy the affine constraints (2). We denote the corresponding constraint sets as $\mathcal{X} = \{x \in \mathbb{R}^n : D_x x \leq d_x\}$, $\mathcal{U} = \{u \in \mathbb{R}^m : D_u u \leq d_u\}$, where $d_x \in \mathbb{R}^{k_x}$ and $d_u \in \mathbb{R}^{k_u}$. Define $k_c = k_x + k_u$ as the total number of the constraints.

For simplicity, we consider that the parameters $A, B, \bar{w}, D_x, d_x, D_u, d_u$ are known a priori, and leave the unknown case as future work.

Definition 1 (Feasible controller)**.** Consider a controller (or an algorithm) $\mathcal{A}$ that chooses action $u_t^{\mathcal{A}}$ based on history states $\{x_k^{\mathcal{A}}\}_{k=0}^t$ and cost functions $\{c_k(\cdot, \cdot)\}_{k=0}^{t-1}$. The controller $\mathcal{A}$ is called *feasible* if $x_t^{\mathcal{A}} \in \mathcal{X}$ and $u_t^{\mathcal{A}} \in \mathcal{U}$ for all $0 \leq t \leq T$ and all disturbances $\{w_k \in \mathcal{W}\}_{k=0}^T$.

For a feasible algorithm/controller $\mathcal{A}$, the total cost is defined as

$$J_T(\mathcal{A}) = \mathop{\mathbb{E}}_{\{w_k\}} \left[\sum_{t=0}^{T} c_t(x_t^{\mathcal{A}}, u_t^{\mathcal{A}}) \right]. \tag{3}$$

Benchmark policy and policy regret. In this paper, we consider linear policy of the form $u_t = -Kx_t$ as our benchmark policy for simplicity, though the optimal policy for the constrained control of noisy systems may be nonlinear (Rawlings and Mayne 2009). We leave the discussion on nonlinear policies as future work.

Based on (Cohen et al. 2018), we introduce the definition of strong stability, which is a quantitative version of the stability property that permits non-asymptotic regret analysis.

Definition 2 (Strong Stability)**.** A linear controller $u_t = -Kx_t$ is (κ, γ)-strongly stable for $\kappa \geq 1$ and $\gamma \in (0, 1]$ if there exists a matrix L and an invertible matrix H such that $A - BK = H^{-1}LH$, with $\|L\|_2 \leq 1 - \gamma$ and $\max(\|H\|_2, \|H^{-1}\|_2, \|K\|_2) \leq \kappa$.

Our benchmark policy class includes any linear controller $u_t = -Kx_t$ satisfying the conditions below:

$$\mathcal{K} = \{K : K \text{ is feasible and } (\kappa, \gamma)\text{-strongly stable.}\}$$

where K is called feasible if the controller $u_t = -Kx_t$ is feasible according to Definition 1.

The policy regret of online algorithm $\mathcal{A}$ is defined as

$$\text{Reg}(\mathcal{A}) = J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K). \tag{4}$$

Assumptions and definitions. For the rest of the paper, we assume $x_0 = 0$ for simplicity and define $\kappa_B = \max(\|B\|_2, 1)$. In addition, we introduce the following assumptions on the disturbances and the cost functions, which are standard in literature (Agarwal, Hazan, and Singh 2019).

Assumption 1. $\{w_t\}$ are i.i.d. with $\mathbb{E}[w_t] = 0$, covariance matrix Σ_w, and bounded range $\|w_t\|_\infty \le \bar{w}$, where $\bar{w} > 0$.

Assumption 2. For any $t \ge 0$, cost function $c_t(x_t, u_t)$ is convex and differentiable with respect to x_t and u_t. Further, there exists $G > 0$, such that for any $\|x\|_2 \le b$, $\|u\|_2 \le b$, we have $\|\nabla_x c_t(x, u)\|_2 \le Gb$ and $\|\nabla_u c_t(x, u)\|_2 \le Gb$.

Next, we define strictly and loosely feasible controllers.

Definition 3 (Strict and loose feasibility)**.** A feasible controller $\mathcal{A}$ is called *ϵ-strictly feasible* for some $\epsilon > 0$ if $D_x x_t^{\mathcal{A}} \le d_x - \epsilon \mathbb{1}_{k_x}$ and $D_u u_t^{\mathcal{A}} \le d_u - \epsilon \mathbb{1}_{k_u}$ for all $0 \le t \le T$ under any disturbance sequence $\{w_k \in \mathcal{W}\}_{k=0}^T$.

A controller $\mathcal{A}$ is called *ϵ-loosely feasible* for some $\epsilon > 0$ if $D_x x_t^{\mathcal{A}} \le d_x + \epsilon \mathbb{1}_{k_x}$ and $D_u u_t^{\mathcal{A}} \le d_u + \epsilon \mathbb{1}_{k_u}$ for all $0 \le t \le T$ under any disturbance sequence $\{w_k \in \mathcal{W}\}_{k=0}^T$.

Besides, we assume that there exists a strictly feasible policy in $\mathcal{K}$. We note that the existence of a strictly feasible solution instead of just a feasible solution is usually required in constrained optimization and control theory (Boyd and Vandenberghe 2004; Limon et al. 2010).

Assumption 3. There exists $K_* \in \mathcal{K}$ such that the policy $u_t = -K_* x_t$ is ϵ_*-strictly feasible for some $\epsilon_* > 0$.

Intuitively, Assumption 3 requires the sets $\mathcal{X}$ and $\mathcal{U}$ to have non-empty interiors; and that the disturbance set $\mathcal{W}$ is small enough so that a disturbed linear system $x_{t+1} = (A - BK_*)x_t + w_t$ stays in the interiors of $\mathcal{X}$ and $\mathcal{U}$ for any $\{w_k \in \mathcal{W}\}_{k=0}^T$. Since $x_0 = 0$, Assumption 3 also implicitly assumes that 0 belongs to the interiors of $\mathcal{X}$ and $\mathcal{U}$. Finally, although it is challenging to verify Assumption 3 directly, (Limon et al. 2010) provides a sufficient condition to verify Assumption 3, which is by solving a convex optimization involving linear matrix inequalities (LMI) (Boyd et al. 1994).

3 Preliminaries

This section briefly reviews the unconstrained online optimal control and robust constrained optimization literature; techniques from which motivates our algorithm design.

3.1 Unconstrained online optimal control.

In our setting if one considers $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{U} = \mathbb{R}^m$, then the problem reduces to an unconstrained online optimal control. For such unconstrained online control problems, (Agarwal, Hazan, and Singh 2019; Agarwal et al. 2019) propose a disturbance-action policy class to design an online policy.

Definition 4 (Disturbance-Action Policy (Agarwal, Hazan, and Singh 2019))**.** Fix an arbitrary (κ, γ)-strongly stable matrix $\mathbb{K}$ a priori. Given an $H \in \{1, 2, \ldots, T\}$, a disturbance-action policy defines the control policy as:

$$u_t = -\mathbb{K}x_t + \sum_{i=1}^{H} M^{[i]} w_{t-i}, \quad \forall\, t \ge 0, \tag{5}$$

where, $M^{[i]} \in \mathbb{R}^{m \times n}$ and $w_t = 0$ for $t \le 0$. Let $\boldsymbol{M} = \{M^{[i]}\}_{i=1}^H$ denote the list of parameter matrices for the disturbance-action policy.

In (5), $\mathbb{K}$ can be computed efficiently by SDP formulation (Cohen et al. 2018). Further, (Agarwal, Hazan, and Singh 2019) introduces a bounded convex constraint set on policy $\boldsymbol{M}$ for technical simplicity and without loss of generality:[1]

$$\mathcal{M}_2 = \{\boldsymbol{M} = \{M^{[i]}\}_{i=1}^H : \|M^{[i]}\|_2 \le \kappa^3 \kappa_B (1-\gamma)^i, \forall i\} \tag{6}$$

The following Proposition 1 derives the approximations of the states and actions when implementing disturbance-action policies.

Proposition 1 ((Agarwal et al. 2019))**.** *When implementing a disturbance-action policy* (5) *with time-varying $\boldsymbol{M}_t = \{M_t^{[i]}\}_{i=1}^H$ at each stage $t \ge 0$, the states and actions satisfy:*

$$x_t = A_{\mathbb{K}}^H x_{t-H} + \tilde{x}_t \;\; \textit{and} \;\; u_t = -\mathbb{K}A_{\mathbb{K}}^H x_{t-H} + \tilde{u}_t, \tag{7}$$

where $A_{\mathbb{K}} = A - B\mathbb{K}$. The approximate/surrogate state and action, $\tilde{x}_t$ and $\tilde{u}_t$, are defined as:

$$\tilde{x}_t = \sum_{k=1}^{2H} \Phi_k^x(\boldsymbol{M}_{t-H:t-1}) w_{t-k},$$

$$\tilde{u}_t = -\mathbb{K}\tilde{x}_t + \sum_{i=1}^{H} M_t^{[i]} w_{t-i} = \sum_{k=1}^{2H} \Phi_k^u(\boldsymbol{M}_{t-H:t}) w_{t-k},$$

$$\Phi_k^x(\boldsymbol{M}_{t-H:t-1}) = A_{\mathbb{K}}^{k-1} \mathbb{1}_{(k \le H)} + \sum_{i=1}^{H} A_{\mathbb{K}}^{i-1} B M_{t-i}^{[k-i]} \mathbb{1}_{(1 \le k-i \le H)}$$

$$\Phi_k^u(\boldsymbol{M}_{t-H:t}) = M_t^{[k]} \mathbb{1}_{(k \le H)} - \mathbb{K}\Phi_k^x(\boldsymbol{M}_{t-H:t-1}),$$

where $\boldsymbol{M}_{t-H:t} \coloneqq \{\boldsymbol{M}_{t-H}, \ldots, \boldsymbol{M}_t\}$, the superscript k in $A_{\mathbb{K}}^k$ denotes the kth power of $A_{\mathbb{K}}$, and $M_t^{[k]}$ with superscript $[k]$ denotes the kth matrix in list $\boldsymbol{M}_t$. Further, define $\mathring{\Phi}_k^x(\boldsymbol{M}) = \Phi_k^x(\boldsymbol{M}, \ldots, \boldsymbol{M})$, $\mathring{\Phi}_k^u(\boldsymbol{M}) = \Phi_k^u(\boldsymbol{M}, \ldots, \boldsymbol{M})$.

Notice that $\tilde{x}_t$ and $\tilde{u}_t$ are affine functions of $\boldsymbol{M}_{t-H:t}$. Based on $\tilde{x}_t$ and $\tilde{u}_t$, (Agarwal, Hazan, and Singh 2019) introduces an approximate cost function:

$$f_t(\boldsymbol{M}_{t-H:t}) = \mathbb{E}[c_t(\tilde{x}_t, \tilde{u}_t)],$$

which is convex with respect to $\boldsymbol{M}_{t-H:t}$ since $\tilde{x}_t$ and $\tilde{u}_t$ are affine functions of $\boldsymbol{M}_{t-H:t}$ and $c_t(\cdot, \cdot)$ is convex.

OCO with memory. In (Agarwal, Hazan, and Singh 2019), the unconstrained online optimal control problem is converted to *OCO with memory*, i.e., at each stage t, the

[1] This is without loss of generality because (Agarwal et al. 2019) shows that any (κ, γ)-strongly stable linear policy can be approximated by a disturbance-action policy in $\mathcal{M}_2$.

agent selects a policy $\boldsymbol{M}_t \in \mathcal{M}_2$ and then incurs a cost $f_t(\boldsymbol{M}_{t-H:t})$. Notice that the cost function at stage t couples the current policy $\boldsymbol{M}_t$ with the H-stage historical policies $\boldsymbol{M}_{t-H:t-1}$, but the constraint set $\mathcal{M}_2$ is decoupled and only depends on the current $\boldsymbol{M}_t$.

To solve this OCO with memory problem, (Agarwal, Hazan, and Singh 2019) defines decoupled cost functions

$$\mathring{f}_t(\boldsymbol{M}_t) \coloneqq f_t(\boldsymbol{M}_t, \dots, \boldsymbol{M}_t), \tag{8}$$

by letting the H-stage historical policies be identical to the current policy. Notice that $\mathring{f}_t(\boldsymbol{M}_t)$ is still convex. Accordingly, the OCO with memory is reformulated as a classical OCO problem with stage cost $\mathring{f}_t(\boldsymbol{M}_t)$, which is solved by classical OCO algorithms such as online gradient descent (OGD) in (Agarwal, Hazan, and Singh 2019). The stepsizes of OGD are chosen to be sufficiently small so that the variation between the current policy $\boldsymbol{M}_t$ and the H-stage historical policies $\boldsymbol{M}_{t-H}, \dots, \boldsymbol{M}_{t-1}$ are sufficiently small, which guarantees small approximation error between $\mathring{f}_t(\boldsymbol{M}_t)$ and $f_t(\boldsymbol{M}_{t-H:t})$, and thus low regrets. For more details, we refer the reader to (Agarwal, Hazan, and Singh 2019).

3.2 Robust optimization with constraints.

Consider a robust optimization problem with linear constraints (Ben-Tal, El Ghaoui, and Nemirovski 2009):

$$\min_x f(x) \quad \text{s.t. } a_i^\top x \le b_i,\ \forall\, a_i \in \mathcal{C}_i,\ \forall\, 1 \le i \le k, \tag{9}$$

where the (box) uncertainty sets are defined as $\mathcal{C}_i = \{a_i = \tilde{a}_i + P_i z : \|z\|_\infty \le \bar{z}\}$ for any i. Notice that the robust constraint $\{a_i^\top x \le b_i,\ \forall\, a_i \in \mathcal{C}_i\}$ is equivalent to the standard constraint $\{\sup_{a_i \in \mathcal{C}_i}[a_i^\top x] \le b_i\}$. Further, one can derive

$$\begin{aligned}\sup_{a_i \in \mathcal{C}_i} a_i^\top x &= \sup_{\|z\|_\infty \le \bar{z}} (\tilde{a}_i + P_i z)^\top x \\ &= \tilde{a}_i^\top x + \sup_{\|z\|_\infty \le \bar{z}} z^\top (P_i^\top x) = \tilde{a}_i^\top x + \|P_i^\top x\|_1 \bar{z}\end{aligned} \tag{10}$$

Therefore, the robust optimization (9) can be equivalently reformulated as the linearly constrained optimization below:

$$\min_x f(x) \quad \text{s.t. } \tilde{a}_i^\top x + \|P_i^\top x\|_1 \bar{z} \le b_i,\ \forall\, 1 \le i \le k.$$

4 Online Algorithm Design

This section introduces our online algorithm design for online disturbance-action policies (Definition 4).

Roughly speaking, to develop our online algorithm, we first convert the constrained online optimal control to OCO with memory and coupled constraints, which is later converted to classical OCO and solved by OCO algorithms. The conversion leverages the approximation and the reformulation techniques reviewed in Section 3. During the conversion, we ensure that the outputs of the OCO algorithms are feasible for the original control problem. This is achieved by tightening the original constraints (adding buffer zones) to allow for approximation errors. Besides, we ensure small buffer zones and small approximation errors so that the optimality/regret is not sacrificed significantly for feasibility. The details of algorithm design are discussed below.

Step 1: Constraints on approximate states and actions. When applying the disturbance-action policies (5), we can use (7) to rewrite the state constraint $x_{t+1} \in \mathcal{X}$ as

$$D_x A_{\mathbb{K}}^H x_{t-H+1} + D_x \tilde{x}_{t+1} \le d_x,\ \ \forall \{w_k \in \mathcal{W}\}_{k=0}^T, \tag{11}$$

where $\tilde{x}_{t+1}$ is the approximate state. Note that the term $D_x A_{\mathbb{K}}^H x_{t-H+1}$ decays exponentially with H. If there exists H such that $D_x A_{\mathbb{K}}^H x_{t-H+1} \le \epsilon_1 \mathbb{1}_{k_x}$, $\forall \{w_k \in \mathcal{W}\}_{k=0}^T$, then a tightened constraint on the approximate state, i.e.

$$D_x \tilde{x}_{t+1} \le d_x - \epsilon_1 \mathbb{1}_{k_x}, \quad \forall \{w_k \in \mathcal{W}\}_{k=0}^T, \tag{12}$$

can guarantee the original constraint on the true state (11).

The action constraint $u_t \in \mathcal{U}$ can similarly be converted to a tightened constraint on the approximate action $\tilde{u}_t$, i.e.

$$D_u \tilde{u}_t \le d_u - \epsilon_1 \mathbb{1}_{k_u}, \quad \forall \{w_k \in \mathcal{W}\}_{k=0}^T, \tag{13}$$

if $D_u(-\mathbb{K} A_{\mathbb{K}}^H x_{t-H}) \le \epsilon_1 \mathbb{1}_{k_u}$ for any disturbances.

Step 2: Constraints on the policy parameters. Next, we reformulate the robust constraints (12) and (13) on $\tilde{x}_{t+1}$ and $\tilde{u}_t$ as polytopic constraints on policy parameters $\boldsymbol{M}_{t-H:t}$ based on the robust optimization techniques in Section 3.2.

Firstly, we consider the ith row of the constraint (12), i.e. $D_{x,i}^\top \tilde{x}_{t+1} \le d_{x,i} - \epsilon_1\ \forall \{w_k \in \mathcal{W}\}_{k=0}^T$, where $D_{x,i}^\top$ denotes the ith row of the matrix D_x. Note that this constraint is equivalent to $\sup_{\{w_k \in \mathcal{W}\}_{k=0}^T} (D_{x,i}^\top \tilde{x}_{t+1}) \le d_{x,i} - \epsilon_1$. Further, by the definitions of $\tilde{x}_{t+1}$ and $\mathcal{W}$, and (10), we obtain

$$\begin{aligned}\sup_{\{w_k \in \mathcal{W}\}} D_{x,i}^\top \tilde{x}_{t+1} &= \sup_{\{w_k \in \mathcal{W}\}} D_{x,i}^\top \sum_{s=1}^{2H} \Phi_s^x(\boldsymbol{M}_{t-H+1:t}) w_{t+1-s} \\ &= \sum_{s=1}^{2H} \sup_{w_{t+1-s} \in \mathcal{W}} D_{x,i}^\top \Phi_s^x(\boldsymbol{M}_{t-H+1:t}) w_{t+1-s} \\ &= \sum_{s=1}^{2H} \|D_{x,i}^\top \Phi_s^x(\boldsymbol{M}_{t-H+1:t})\|_1 \bar{w}\end{aligned}$$

Define $g_i^x(\boldsymbol{M}_{t-H+1:t}) = \sum_{s=1}^{2H} \|D_{x,i}^\top \Phi_s^x(\boldsymbol{M}_{t-H+1:t})\|_1 \bar{w}$. Hence, the robust constraint (12) on $\tilde{x}_{t+1}$ is equivalent to the following polytopic constraints on $\boldsymbol{M}_{t-H+1:t}$:

$$g_i^x(\boldsymbol{M}_{t-H+1:t}) \le d_{x,i} - \epsilon_1, \quad \forall\, 1 \le i \le k_x. \tag{14}$$

Similarly, the constraint (13) on $\tilde{u}_t$ is equivalent to:

$$g_j^u(\boldsymbol{M}_{t-H:t}) \le d_{u,j} - \epsilon_1, \quad \forall\, 1 \le j \le k_u, \tag{15}$$

where $g_j^u(\boldsymbol{M}_{t-H:t}) = \sum_{s=1}^{2H} \|D_{u,j}^\top \Phi_s^u(\boldsymbol{M}_{t-H:t})\|_1 \bar{w}$.

Step 3: OCO with memory and temporal-coupled constraints. Based on Step 2 and Section 3.1, we can convert the constrained online optimal control problem to OCO with memory and temporal-coupled constraints. That is, at each stage t, the decision maker selects a policy $\boldsymbol{M}_t$ satisfying constraints (14) and (15), and then incurs a cost $f_t(\boldsymbol{M}_{t-H:t})$. In our framework, the constraints (14), (15) and the cost function $f_t(\boldsymbol{M}_{t-H:t})$ couple the current policy with the historical policies. This makes the problem far more challenging than OCO with memory which only considers coupled costs (Anava, Hazan, and Mannor 2015).

Step 4: Benefits of the slow variation of online policies. We approximate the coupled constraint functions $g_i^x(\boldsymbol{M}_{t-H+1:t})$ and $g_j^u(\boldsymbol{M}_{t-H:t})$ as decoupled ones below:

$$\mathring{g}_i^x(\boldsymbol{M}_t)=g_i^x(\boldsymbol{M}_t,\dots,\boldsymbol{M}_t), \mathring{g}_i^u(\boldsymbol{M}_t)=g_i^u(\boldsymbol{M}_t,\dots,\boldsymbol{M}_t),$$

by letting the historical policies $\boldsymbol{M}_{t-H:t-1}$ be identical to the current $\boldsymbol{M}_t$. If the online policy $\boldsymbol{M}_t$ varies slowly with t, which is satisfied by most OCO algorithms (e.g. OGD with a diminishing stepsize (Hazan 2019)), one may be able to bound the approximation errors by $g_i^x(\boldsymbol{M}_{t-H+1:t}) - \mathring{g}_i^x(\boldsymbol{M}_t) \le \epsilon_2$ and $g_j^u(\boldsymbol{M}_{t-H:t}) - \mathring{g}_j^u(\boldsymbol{M}_t) \le \epsilon_2$ for a small $\epsilon_2 > 0$. Thus, the constraints (14) and (15) are ensured by the polytopic constraints on $\boldsymbol{M}_t$:

$$\mathring{g}_i^x(\boldsymbol{M}_t) \le d_{x,i} - \epsilon_1 - \epsilon_2,\ \mathring{g}_j^u(\boldsymbol{M}_t) \le d_{u,j} - \epsilon_1 - \epsilon_2, \quad (16)$$

where the buffer zone ϵ_2 allows for the approximation error caused by neglecting the variation of online policies.

Step 5: Conversion to OCO. By Step 4, we define a decoupled search space/constraint set on each policy below.

$$\Omega_\epsilon = \{\boldsymbol{M} \in \mathcal{M} : \mathring{g}_i^x(\boldsymbol{M}) \le d_{x,i} - \epsilon, \forall 1 \le i \le k_x, \\ \mathring{g}_j^u(\boldsymbol{M}) \le d_{u,j} - \epsilon, \forall\, 1 \le j \le k_u\}. \quad (17)$$

where $\mathcal{M}$ is a bounded convex constraint set defined as

$$\mathcal{M} = \{\boldsymbol{M} :\ \|M^{[i]}\|_\infty \le 2\sqrt{n}\kappa^3(1-\gamma)^{i-1},\ \forall\, 1 \le i \le H\}.$$

Our set $\mathcal{M}$ is slightly different from $\mathcal{M}_2$ in (6) to ensure that Ω_ϵ is a polytope.[2] Notice that Ω_ϵ provides buffer zones with size ϵ to account for the approximation errors ϵ_1 and ϵ_2.

Based on Ω_ϵ and Section 3.1, we can further convert the OCO with memory and coupling constraints in Step 3 to a classical OCO problem below. That is, at each stage t, the agent selects a policy $\boldsymbol{M}_t \in \Omega_\epsilon$, and then suffers a convex stage cost $\mathring{f}_t(\boldsymbol{M}_t)$ defined in (8). We apply online gradient descent to solve this OCO problem, as described in Algorithm 1. We select the stepsizes of OGD to be small enough to ensure small approximation errors from the problem conversion and small buffer zones, but also to be large enough to allow online policies to adapt to time-varying environments. Conditions for suitable stepsizes are discussed in Section 5.

In Algorithm 1, the most computationally demanding step at each stage is the projection onto the polytope Ω_ϵ, which requires solving a quadratic program. Nevertheless, one can reduce the online computational burden via offline computation by leveraging the solution structure of quadratic programs (see (Alessio and Bemporad 2009) for more details).

Lastly, we note that other OCO algorithms can be applied to solve this problem too, e.g. online natural gradient.

Remark 1. To ensure safety, safe RL literature usually constructs a safe set for the state (Fisac et al. 2018), while this paper constructs a safe search space Ω_ϵ for the policies directly. Further, safe RL literature may employ unsafe policies occassionally, for example, (Fisac et al. 2018) allows unsafe exploration policies within the safe set and changes to a safe policy on the boundary of the safe set. However, our search space Ω_ϵ only contains safe/feasible policies. Despite a smaller policy search space, our OGD-BZ still achieves desirable performance in Section 5. Nevertheless, when the system is unknown, larger sets of exploration policies may benefit the performance, which is left as future work.

[2]Compared with $\mathcal{M}_2$, our $\mathcal{M}$ uses the L_∞ norm; the $\sqrt{n}$ factor accounts for the change of norms; and κ_B disappears because we can prove that κ_B is not necessary to ensure no loss of generality.

Algorithm 1: OGD-BZ

Input: A (κ,γ)-strongly stable matrix $\mathbb{K}$, parameter $H > 0$, buffer size ϵ, stepsize η_t.

Determine the polytopic constraint set Ω_ϵ by (17) with buffer size ϵ and initialize $\boldsymbol{M}_0 \in \Omega_\epsilon$.

for $t = 0, 1, 2, \dots, T$ **do**

  Implement action $u_t = -\mathbb{K}x_t + \sum_{i=1}^H M_t^{[i]} w_{t-i}$.

  Observe the next state x_{t+1} and record $w_t = x_{t+1} - Ax_t - Bu_t$.

  Run projected OGD

$$\boldsymbol{M}_{t+1} = \Pi_{\Omega_\epsilon}\left[\boldsymbol{M}_t - \eta_t \nabla \mathring{f}_t(\boldsymbol{M}_t)\right]$$

  where $\mathring{f}_t(\boldsymbol{M})$ is defined in (8).

5 Theoretical Results

In this section, we show that OGD-BZ guarantees both feasibility and $\tilde{O}(\sqrt{T})$ policy regret under proper parameters.

Preparation. To establish the conditions on the parameters for our theoretical results, we introduce three quantities $\epsilon_1(H), \epsilon_2(\eta, H), \epsilon_3(H)$ below. We note that $\epsilon_1(H)$ and $\epsilon_2(\eta, H)$ bound the approximation errors in Step 1 and Step 4 of Section 4 respectively (see Lemma 1, Lemma 2 in Section 5.1 for more details). $\epsilon_3(H)$ bounds the constraint violation of the disturbance-action policy $\boldsymbol{M}(K)$, where $\boldsymbol{M}(K)$ approximates the standard linear controller $u_t = -Kx$ for any $K \in \mathcal{K}$ (see Lemma 3 in Section 5.1 for more details).

Definition 5. We define

$$\epsilon_1(H) = c_1 n\sqrt{m}H(1-\gamma)^H, \epsilon_2(\eta, H) = c_2\eta \cdot n^2\sqrt{m}H^2$$
$$\epsilon_3(H) = c_3\sqrt{n}(1-\gamma)^H$$

where c_1, c_2, c_3 are constant factors that depend polynomially on $\|D_x\|_2, \|D_u\|_2, \kappa, \kappa_B, \gamma^{-1}, \bar{w}, G$.

Feasibility of OGD-BZ

Theorem 1 (Feasibility)**.** *Consider constant stepsize $\eta_t = \eta$, $\epsilon \ge 0, H \ge \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})}$. If the buffer size ϵ and H satisfy*

$$\epsilon \le \epsilon_* - \epsilon_1(H) - \epsilon_3(H),$$

the set Ω_ϵ is non-empty. Further, if η, ϵ and H also satisfy

$$\epsilon \ge \epsilon_1(H) + \epsilon_2(\eta, H),$$

our OGD-BZ is feasible, i.e. $x_t^{\text{OGD-BZ}} \in \mathcal{X}$ and $u_t^{\text{OGD-BZ}} \in \mathcal{U}$ for all t and for any disturbances $\{w_k \in \mathcal{W}\}_{k=0}^T$.

Discussions: Firstly, Theorem 1 shows that ϵ should be small enough to ensure a nonempty Ω_ϵ and thus valid outputs of OGD-BZ. This is intuitive since the constraints

become more conservative as ϵ increases. Since $\epsilon_1(H) + \epsilon_3(H) = \Theta(H(1-\gamma)^H)$ decays with H by Definition 5, the first condition also implicitly requires a large enough H.

Secondly, Theorem 1 shows that, to ensure feasibility, the buffer size ϵ should also be large enough to allow for the total approximation errors $\epsilon_1(H) + \epsilon_2(\eta, H)$, which is consistent with our discussion in Section 4. To ensure the compatibility of the two conditions on ϵ, the approximation errors $\epsilon_1(H) + \epsilon_2(\eta, H)$ should be small enough, which requires a large enough H and a small enough η by Definition 5.

In conclusion, the feasibility requires a large enough H, a small enough η, and an ϵ which is not too large or too small, for example, we can select $\eta \leq \frac{\epsilon_*}{8c_2 n^2 \sqrt{m} H^2}$, $\epsilon_*/4 \leq \epsilon \leq 3\epsilon_*/4$, and $H \geq \max(\frac{\log(\frac{4(2c_1+c_2)n\sqrt{m}}{\epsilon_*}T)}{\log((1-\gamma)^{-1}}, \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})})$.

Policy Regret Bound for OGD-BZ.

Theorem 2 (Regret Bound)**.** *Under the conditions in Theorem 1, OGD-BZ enjoys the regret bound below:*

$$\begin{aligned}\text{Reg(OGD-BZ)} \leq O\Bigg(& n^3 m H^3 \eta T + \frac{mn}{\eta} \\ &+ (1-\gamma)^H H^{2.5} T\left(\frac{n^{4.5}m^2}{\epsilon_*} + \sqrt{k_c} m n^{2.5}\right) \\ &+ \epsilon T H^{1.5}\left(\frac{n^{3.5}m^{1.5}}{\epsilon_*} + \sqrt{k_c m n^3}\right)\Bigg)\end{aligned}$$

where the hidden constant depends polynomially on $\kappa, \kappa_B, \gamma^{-1}, \|D_x\|_2, \|D_u\|_2, \|d_x\|_2, \|d_u\|_2, \bar{w}, G$.

Theorem 2 provides a regret bound for OGD-BZ as long as OGD-BZ is feasible. Notice that as the buffer size ϵ increases, the regret bound becomes worse. This is intuitive since our OGD-BZ will have to search for policies in a smaller set Ω_ϵ if ϵ increases. Consequently, the buffer size ϵ can serve as a tuning parameter for the trade-off between safety and regrets, i.e., a small ϵ is preferred for low regrets while a large ϵ is preferred for feasibility (as long as $\Omega_\epsilon \neq \emptyset$). In addition, although a small stepsize η is preferred for feasibility in Theorem 1, Theorem 2 suggests that the stepsize should not be too small for low regrets since the regret bound contains a $\Theta(\eta^{-1})$ term. This is intuitive since the stepsize η should be large enough to allow OGD-BZ to adapt to the varying objectives for better online performance.

Next, we provide a regret bound with specific parameters.

Corollary 1. *For sufficiently large* T*, when* $H \geq \frac{\log(4(2c_1+c_2)n\sqrt{m}T/\epsilon_*)}{\log((1-\gamma)^{-1})}$, $\eta = \Theta(\frac{1}{n^2\sqrt{m}H\sqrt{T}})$, $\epsilon = \epsilon_1(H) + \epsilon_2(\eta, H) = \Theta(\frac{\log(n\sqrt{m}T)}{\sqrt{T}})$*, OGD-BZ is feasible and* $\text{Reg(OGD-BZ)} \leq \tilde{O}\left((n^{3.5}m^{1.5}{k_c}^{0.5})\sqrt{T}\right)$.

Corollary 1 shows that OGD-BZ achieves $\tilde{O}(\sqrt{T})$ regrets when $H \geq \Theta(\log T)$, $\eta = \tilde{\Theta}(1/\sqrt{T})$ and $\epsilon = \tilde{\Theta}(1/\sqrt{T})$. This demonstrates that OGD-BZ can ensure both constraint satisfaction and sublinear regrets under proper parameters of the algorithm. We remark that although a larger H is preferred for better performance, the computational complexity of OGD-BZ increases with H. Besides, though the choices of H, η and ϵ above require the prior knowledge of T, one can apply doubling tricks (Hazan 2019) to avoid this requirement. Lastly, we note that our $\tilde{O}(\sqrt{T})$ regret bound is consistent with the unconstrained online optimal control literature for convex cost functions (Agarwal et al. 2019). For strongly convex costs, the regret for the unconstrained case is logarithmic in T (Agarwal, Hazan, and Singh 2019), and we conjecture that logarithmic regret can also be achieved for the constrained case, which is our ongoing work.

5.1 Proof of Theorem 1

To prove Theorem 1, we first establish three lemmas that bound errors by $\epsilon_1(H), \epsilon_2(\eta, H)$ and $\epsilon_3(H)$ respectively. The proofs of these lemmas are in the supplementary file.

Firstly, we show that the approximation error in Step 1 of Section 4 can be bounded by $\epsilon_1(H)$.

Lemma 1. *When* $\boldsymbol{M}_t \in \mathcal{M}$ *and* $H \geq \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})}$*, we have*

$$\max_{\|w_k\|_\infty \leq \bar{w}} \|D_x A_{\mathbb{K}}^H x_{t-H}\|_\infty \leq \epsilon_1(H)$$
$$\max_{\|w_k\|_\infty \leq \bar{w}} \|D_{u,j}^\top \mathbb{K} A_{\mathbb{K}}^H x_{t-H}\|_\infty \leq \epsilon_1(H)$$

Secondly, we show that the error incurred by the Step 3 of Section 4 can be bounded by $\epsilon_2(\eta, H)$.

Lemma 2. *When* $H \geq \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})}$*, the policies* $\{\boldsymbol{M}_t\}_{t=0}^T$ *generated by OGD-BZ with a constant stepsize* η *satisfy*

$$\max_{1\leq i\leq k_x} |\mathring{g}_i^x(\boldsymbol{M}_t) - g_i^x(\boldsymbol{M}_{t-H+1:t})| \leq \epsilon_2(\eta, H)$$
$$\max_{1\leq j\leq k_u} |\mathring{g}_j^u(\boldsymbol{M}_t) - g_j^u(\boldsymbol{M}_{t-H:t})| \leq \epsilon_2(\eta, H)$$

Thirdly, we show that for any $K \in \mathcal{K}$, there exists a disturbance-action policy $\boldsymbol{M}(K) \in \mathcal{M}$ to approximate the policy $u_t = -Kx_t$. However, $\boldsymbol{M}(K)$ may not be feasible and is only $\epsilon_3(H)$-loosely feasible.

Lemma 3. *For any* $K \in \mathcal{K}$*, there exists a disturbance-action policy* $\boldsymbol{M}(K) = \{M^{[i]}(K)\}_{i=1}^H \in \mathcal{M}$ *defined as* $M^{[i]}(K) = (\mathbb{K} - K)(A - BK)^{i-1}$ *such that*

$$\max(\|x_t^K - x_t^{\boldsymbol{M}(K)}\|_2, \|u_t^K - u_t^{\boldsymbol{M}(K)}\|_2) \leq \epsilon_3(H),$$

where (x_t^K, u_t^K) *and* $(x_t^{\boldsymbol{M}(K)}, u_t^{\boldsymbol{M}(K)})$ *are produced by controller* $u_t = -Kx_t$ *and disturbance-action policy* $\boldsymbol{M}(K)$ *respectively. Hence,* $\boldsymbol{M}(K)$ *is* $\epsilon_3(H)$*-loosely feasible.*

Based on Lemma 3, we can further show that $\boldsymbol{M}(K)$ belongs to a polytopic constraint set in the following corollary. For the rest of the paper, we will omit the arguments in $\epsilon_1(H), \epsilon_2(\eta, H), \epsilon_3(H)$ for notational simplicity.

Corollary 2. *Consider* $K \in \mathcal{K}$*, if* K *is* ϵ_0*-strictly feasible for* $\epsilon_0 \geq 0$*, then* $\boldsymbol{M}(K) \in \Omega_{\epsilon_0 - \epsilon_1 - \epsilon_3}$.

Next, we prove that Ω_ϵ is non-empty by showing that $\boldsymbol{M}(K_*) \in \Omega_\epsilon$. Specifically, since K_* defined in Assumption 3 is ϵ_*-strictly feasible, by Corollary 2, there exists $\boldsymbol{M}(K_*) \in \Omega_{\epsilon_* - \epsilon_1 - \epsilon_3}$. Since the set Ω_ϵ becomes smaller as ϵ increases, when $\epsilon_* - \epsilon_1 - \epsilon_3 \geq \epsilon$, we have $\boldsymbol{M}(K_*) \in \Omega_{\epsilon_* - \epsilon_1 - \epsilon_3} \subseteq \Omega_\epsilon$, which proves that Ω_ϵ is non-empty.

Finally, we prove the feasibility by Lemma 1 and Lemma 2 based on the discussions in Section 4. Specifically, OGD-BZ guarantees that $\boldsymbol{M}_t \in \Omega_\epsilon$ for all t. Thus, by Lemma 2, we have $g_i^x(\boldsymbol{M}_{t-H:t-1}) = g_i^x(\boldsymbol{M}_{t-H:t-1}) - \mathring{g}_i^x(\boldsymbol{M}_{t-1}) + \mathring{g}_i^x(\boldsymbol{M}_{t-1}) \le \epsilon_2 + d_{x,i} - \epsilon$ for any i. Further, by Step 2 of Section 4 and Lemma 1, we have $D_{x,i}^\top x_t = D_{x,i}^\top A_{\mathbb{K}}^H x_{t-H} + D_{x,i}^\top \tilde{x}_t \le \|D_x A_{\mathbb{K}}^H x_{t-H}\|_\infty + g_i^x(\boldsymbol{M}_{t-H:t-1}) \le \epsilon_1 + \epsilon_2 + d_{x,i} - \epsilon \le d_{x,i}$ if $\epsilon \ge \epsilon_1 + \epsilon_2$ for any $\{w_k \in \mathcal{W}\}_{k=0}^T$. Therefore, $x_t \in \mathcal{X}$ for all $w_k \in \mathcal{W}$. Similarly, we can show $u_t \in \mathcal{U}$ for any $w_k \in \mathcal{W}$. Thus, OGD-BZ is feasible.

5.2 Proof of Theorem 2

We divide the regret into three parts and bound each part.

$$\begin{aligned}\text{Reg}(OGD-BZ) =& \ J_T(\mathcal{A}) - \min_{K\in\mathcal{K}} J_T(K)\\ =& \underbrace{J_T(\mathcal{A}) - \sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}_t)}_{\text{Part i}} + \underbrace{\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}_t) - \min_{\boldsymbol{M}\in\Omega_\epsilon}\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M})}_{\text{Part ii}}\\ &+ \underbrace{\min_{\boldsymbol{M}\in\Omega_\epsilon}\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}) - \min_{K\in\mathcal{K}} J_T(K)}_{\text{Part iii}}\end{aligned}$$

Bound on Part ii. Firstly, we bound Part ii based on the regret bound of OGD in the literature (Hazan 2019).

Lemma 4. *With a constant stepsize η, we have* Part ii $\le \delta^2/2\eta + \eta G_f^2 T/2$, *where* $\delta = \sup_{\boldsymbol{M},\tilde{\boldsymbol{M}}\in\Omega_\epsilon} \|\boldsymbol{M} - \tilde{\boldsymbol{M}}\|_F \le 4\sqrt{mn}\kappa^3/\gamma$ *and* $G_f = \max_t \sup_{\boldsymbol{M}\in\Omega_\epsilon} \|\nabla \mathring{f}_t(\boldsymbol{M})\|_F \le \Theta(\sqrt{n^3H^3m})$.

The bounds on δ, G_f are proved in the supplementary file.

Bound on Part iii. For notational simplicity, we denote $\boldsymbol{M}^* = \arg\min_{\Omega_\epsilon}\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}), K^* = \arg\min_{\mathcal{K}} J_T(K)$. By Lemma 3, we can construct a loosely feasible $\boldsymbol{M}_{\text{ap}} = \boldsymbol{M}(K^*)$ to approximate K^*. By Corollary 2, we have

$$\boldsymbol{M}_{\text{ap}} \in \Omega_{-\epsilon_1-\epsilon_3} \tag{18}$$

We will bound Part iii by leveraging $\boldsymbol{M}_{\text{ap}}$ as middle-ground and bounding the Part iii-A and Part iii-B defined below.

$$\text{Part iii} = \underbrace{\sum_{t=0}^T (\mathring{f}_t(\boldsymbol{M}^*) - \mathring{f}_t(\boldsymbol{M}_{\text{ap}}))}_{\text{Part iii-A}} + \underbrace{\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}_{\text{ap}}) - J_T(K^*)}_{\text{Part iii-B}}$$

Lemma 5. *Consider $K^* \in \mathcal{K}$ and $\boldsymbol{M}_{ap} = \boldsymbol{M}(K^*)$, then* Part iii-B $\le \Theta(Tn^2mH^2(1-\gamma)^H)$.

Lemma 6. *Under the conditions in Theorem 2, we have*

$$\text{Part iii-A} \le \Theta\left((\epsilon_1+\epsilon_3+\epsilon)TH^{\frac{3}{2}}\left(\frac{n^{3.5}m^{1.5}}{\epsilon_*} + \sqrt{k_c mn^3}\right)\right)$$

We highlight that $\boldsymbol{M}_{\text{ap}}$ may not belong to Ω_ϵ by (18). Therefore, even though $\boldsymbol{M}^*$ is optimal in Ω_ϵ, Part iii-A can still be non-negative and has to be bounded to yield a regret bound. This is different from the unconstrained online control literature (Agarwal, Hazan, and Singh 2019), where Part iii-A is non-positive because $\boldsymbol{M}_{\text{ap}} \in \mathcal{M}$ and $\boldsymbol{M}^*$ is optimal in the same set $\mathcal{M}$ when there are no constraints (see (Agarwal, Hazan, and Singh 2019) for more details).

Bound on Part i. Finally, we provide a bound on Part i.

Lemma 7. *Apply Algorithm 1 with constant stepsize η, then* Part i $\le O(Tn^2mH^2(1-\gamma)^H + n^3mH^3\eta T)$.

The proof is similar to (Agarwal, Hazan, and Singh 2019).

Finally, Theorem 2 can be proved by summing up the bounds on Part i, Part ii, Part iii-A, and Part iii-B in Lemma 4-7 and only explicitly showing the highest order terms.

5.3 Proof of Lemma 6

We define $\boldsymbol{M}^\dagger = \arg\min_{\Omega_{-\epsilon_1-\epsilon_3}}\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M})$. By (18), we have $\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}_{\text{ap}}) \ge \sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}^\dagger)$. Therefore, it suffices to bound $\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}^*) - \sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}^\dagger)$, which can be viewed as the difference in the optimal values when perturbing the feasible set from Ω_ϵ to $\Omega_{-\epsilon_1-\epsilon_3}$. To bound Part iii-A, we establish a perturbation result by leveraging the polytopic structure of Ω_ϵ and $\Omega_{-\epsilon_1-\epsilon_3}$.

Proposition 2. *Consider two polytopes $\Omega_1 = \{x : Cx \le h\}$, $\Omega_2 = \{x : Cx \le h - \Delta\}$, where $\Delta_i \ge 0$ for all i. Consider a convex function $f(x)$ that is L-Lipschitz continuous on Ω_1. If Ω_1 is bounded, i.e. $\sup_{x_1,x_1'\in\Omega_1}\|x_1 - x_1'\|_2 \le \delta_1$ and if Ω_2 is non-empty, i.e. there exists $\mathring{x} \in \Omega_2$, then*

$$|\min_{\Omega_1} f(x) - \min_{\Omega_2} f(x)| \le L\frac{\delta_1\|\Delta\|_\infty}{\min_{i:\Delta_i>0}(h - C\mathring{x})_i}. \tag{19}$$

The proof of Proposition 2 is in the supplementary file.

The rest of the proof is by applying Proposition 2. Firstly, we provide bounds on the variables in (19) for our problem.

Lemma 8 (A sketch version)**.** *There exists an enlarged polytope $\Gamma_\epsilon = \{\vec{W} : C\vec{W} \le h_\epsilon\}$ that is equivalent to Ω_ϵ for any $\epsilon \in \mathbb{R}$, where $\vec{W}$ contains elements of $\boldsymbol{M}$ and auxiliary variables (to handle the constraints with absolute values).*

Further, (i) $\Gamma_{-\epsilon_1-\epsilon_3}$ is bounded by $\delta_1 = \Theta(\sqrt{m}n + \epsilon_*\sqrt{k_c})$; (ii) $\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M})$ is Lipschitz continuous with $L = \Theta(T(nH)^{1.5}\sqrt{m})$; (iii) The difference Δ between Γ_ϵ and $\Gamma_{-\epsilon_1-\epsilon_3}$ satisfies $\|\Delta\|_\infty = \epsilon + \epsilon_1 + \epsilon_3$; (iv) There exists $\vec{W}^\circ \in \Gamma_\epsilon$ s.t. $\min_{i:\Delta_i>0}(h_{-\epsilon_1-\epsilon_3} - C\vec{W}^\circ)_i \ge \epsilon_*$.

The full Lemma 8 and its proof are in the supplementary file. The proof is completed by Lemma 8 and Proposition 2.

6 Conclusion and Future Work

This paper studies online optimal control with linear constraints and linear dynamics with random disturbances. We propose OGD-BZ and show that OGD-BZ can satisfy all the constraints despite disturbances and ensure $\tilde{O}(\sqrt{T})$ policy regret. The paper focuses on the theoretical results and defer the numerical results to the supplementary file. There are many interesting future directions, e.g. (i) consider adversarial disturbances , (ii) consider soft constraints, (iii) consider more general disturbances, (iv) consider bandit feedback, (v) reduce the regret bound's dependence on dimensions, (vi) consider unknown systems, (vii) consider more general policies than linear policies, (viii) prove logarithmic regrets for strongly convex costs, etc.

Ethics Statement

The primary motivation for this paper is to develop an online control algorithm under linear constraints on the states and actions, and under noisy linear dynamics. Some practical physical systems can be approximated by noisy linear dynamics and most practical systems have to satisfy certain constraints on the states and actions, such as data center cooling and robotics, etc. Our proposed approach ensures to generate control policies that satisfies the constraints even under the uncertainty of unknown noises. Thus our algorithm can potentially be very beneficial for safety critical applications. However, note that our approach relies on a set of technical assumptions, as mentioned in the paper, which may not directly hold for all practical applications. Hence, when applying our algorithm, particular care are needed when modeling the system and the constraints.

References

Agarwal, N.; Bullins, B.; Hazan, E.; Kakade, S. M.; and Singh, K. 2019. Online control with adversarial disturbances. In *36th International Conference on Machine Learning, ICML 2019*, 154–165. International Machine Learning Society (IMLS).

Agarwal, N.; Hazan, E.; and Singh, K. 2019. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, 10175–10184.

Alessio, A.; and Bemporad, A. 2009. A survey on explicit model predictive control. In *Nonlinear model predictive control*, 345–369. Springer.

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

Aswani, A.; Gonzalez, H.; Sastry, S. S.; and Tomlin, C. 2013. Provably safe and robust learning-based model predictive control. *Automatica* 49(5): 1216–1226.

Bemporad, A.; and Morari, M. 1999. Robust model predictive control: A survey. In *Robustness in identification and control*, 207–226. Springer.

Bemporad, A.; Morari, M.; Dua, V.; and Pistikopoulos, E. N. 2002. The explicit linear quadratic regulator for constrained systems. *Automatica* 38(1): 3–20.

Ben-Tal, A.; El Ghaoui, L.; and Nemirovski, A. 2009. *Robust optimization*, volume 28. Princeton University Press.

Boyd, S.; El Ghaoui, L.; Feron, E.; and Balakrishnan, V. 1994. *Linear matrix inequalities in system and control theory*. SIAM.

Boyd, S.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Cao, X.; Zhang, J.; and Poor, H. V. 2018. A virtual-queue-based algorithm for constrained online convex optimization with applications to data center resource allocation. *IEEE Journal of Selected Topics in Signal Processing* 12(4): 703–716.

Cheng, R.; Orosz, G.; Murray, R. M.; and Burdick, J. W. 2019. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3387–3395.

Chow, Y.; Nachum, O.; Duenez-Guzman, E.; and Ghavamzadeh, M. 2018. A lyapunov-based approach to safe reinforcement learning. In *Advances in neural information processing systems*, 8092–8101.

Cohen, A.; Hasidim, A.; Koren, T.; Lazic, N.; Mansour, Y.; and Talwar, K. 2018. Online Linear Quadratic Control. In *International Conference on Machine Learning*, 1029–1038.

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

Dean, S.; Mania, H.; Matni, N.; Recht, B.; and Tu, S. 2019a. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics* 1–47.

Dean, S.; Tu, S.; Matni, N.; and Recht, B. 2019b. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, 5582–5588. IEEE.

Fazel, M.; Ge, R.; Kakade, S.; and Mesbahi, M. 2018. Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator. In *International Conference on Machine Learning*, 1467–1476.

Fisac, J. F.; Akametalu, A. K.; Zeilinger, M. N.; Kaynama, S.; Gillula, J.; and Tomlin, C. J. 2018. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control* 64(7): 2737–2752.

Foster, D. J.; and Simchowitz, M. 2020. Logarithmic regret for adversarial online control. *arXiv preprint arXiv:2003.00189* .

Fulton, N.; and Platzer, A. 2018. Safe reinforcement learning via formal methods. In *AAAI Conference on Artificial Intelligence*.

Garcıa, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16(1): 1437–1480.

Goel, G.; and Hassibi, B. 2020. The Power of Linear Controllers in LQR Control. *arXiv preprint arXiv:2002.02574* .

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

Ibrahimi, M.; Javanmard, A.; and Roy, B. V. 2012. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, 2636–2644.

Kouvaritakis, B.; Rossiter, J. A.; and Schuurmans, J. 2000. Efficient robust predictive control. *IEEE Transactions on automatic control* 45(8): 1545–1549.

Kveton, B.; Yu, J. Y.; Theocharous, G.; and Mannor, S. 2008. Online Learning with Expert Advice and Finite-Horizon Constraints. In *AAAI*, 331–336.

Lazic, N.; Boutilier, C.; Lu, T.; Wong, E.; Roy, B.; Ryu, M.; and Imwalle, G. 2018. Data center cooling using model-predictive control. In *Advances in Neural Information Processing Systems*, 3814–3823.

Limon, D.; Alvarado, I.; Alamo, T.; and Camacho, E. 2008. On the design of Robust tube-based MPC for tracking. *IFAC Proceedings Volumes* 41(2): 15333–15338.

Limon, D.; Alvarado, I.; Alamo, T.; and Camacho, E. 2010. Robust tube-based MPC for tracking of constrained linear systems with additive disturbances. *Journal of Process Control* 20(3): 248–260.

Mayne, D. Q.; Seron, M. M.; and Raković, S. 2005. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica* 41(2): 219–224.

Nonhoff, M.; and Müller, M. A. 2020. An online convex optimization algorithm for controlling linear systems with state and input constraints. *arXiv* arXiv–2005.

Rawlings, J. B.; and Mayne, D. Q. 2009. *Model predictive control: Theory and design*. Nob Hill Pub.

Sallab, A. E.; Abdou, M.; Perot, E.; and Yogamani, S. 2017. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging* 2017(19): 70–76.

Schildbach, G.; Goulart, P.; and Morari, M. 2015. Linear controller design for chance constrained systems. *Automatica* 51: 278–284.

Wabersich, K. P.; and Zeilinger, M. N. 2018. Linear model predictive safety certification for learning-based control. In *2018 IEEE Conference on Decision and Control (CDC)*, 7130–7135. IEEE.

Yang, Z.; Chen, Y.; Hong, M.; and Wang, Z. 2019. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Advances in Neural Information Processing Systems*, 8353–8365.

Yuan, J.; and Lamperski, A. 2018. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*, 6137–6146.

Zafiriou, E. 1990. Robust model predictive control of processes with hard constraints. *Computers & Chemical Engineering* 14(4-5): 359–371.

Zanon, M.; and Gros, S. 2019. Safe reinforcement learning using robust MPC. *arXiv preprint arXiv:1906.04005* .