Online Optimal Control with Affine Constraints

Yingying Li,[1] **Subhro Das,**[2] **Na Li**[1]
[1] John A. Paulson School of Engineering and Applied Sciences, Harvard University
[2] MIT-IBM Watson AI Lab, IBM Research
yingyingli@g.harvard.edu, subhro.das@ibm.com, nali@seas.harvard.edu

Abstract

This paper considers online optimal control with affine constraints on the states and actions under linear dynamics with bounded random disturbances. The system dynamics and constraints are assumed to be known and time invariant but the convex stage cost functions change adversarially. To solve this problem, we propose Online Gradient Descent with Buffer Zones (OGD-BZ). Theoretically, we show that OGD-BZ with proper parameters can guarantee the system to satisfy all the constraints despite any admissible disturbances. 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 a policy regret upper bound that is square root of the horizon length multiplied by some logarithmic terms of the horizon length under proper algorithm parameters.

Introduction

Recently, there is a lot of interest in solving control problems by learning-based techniques, e.g. online learning and reinforcement learning (Agarwal et al. 2019; Li, Chen, and Li 2019; Ibrahimi, Javanmard, and Roy 2012; Dean et al. 2018; Fazel et al. 2018; Yang et al. 2019; Li et al. 2019a). This is motivated by applications such as data centers (Lazic et al. 2018; Li et al. 2019b), robotics (Fisac et al. 2018), autonomous vehicles (Sallab et al. 2017), power systems (Chen et al. 2021), etc. For real-world implementation, it is crucial to design safe algorithms that ensure the system to satisfy certain (physical) constraints despite unknown disturbances. For example, temperatures in data centers should be within certain ranges to reduce task failures despite disturbances from unmodeled heat sources, quadrotors should avoid collisions even when perturbed by wind, etc. In addition to safety, many applications involve time-varying environments, e.g. varying electricity prices, moving targets, etc. Hence, safe algorithms should not be over-conservative and should adapt to varying environments for desirable 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 are many studies to address each goal separately but lack results on both goals together as discussed below.

Firstly, there is recent progress on online optimal control to address 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 2020a,b). 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 (hard) constraint 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) as well as stochastic MPC which considers soft constraints and allows constraints violation (Oldewurtel, Jones, and Morari 2008; Mesbah 2016). However, there lack algorithms with both regret/optimality guarantees and constraint satisfaction guarantees.

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 into an online convex optimization (OCO) problem with temporal-coupled stage costs and temporal-coupled stage constraints, and then convert the temporal-coupled OCO problem into 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 method 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 constraint satisfaction. Further, our theoretical results explicitly characterize a trade-off between the constraint satisfaction and the low regret when deciding the size of the buffer zone of OGD-BZ. That is, 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. 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; Li, Qu, and Li 2020; Li and Li 2020). The papers on OCO with coupled constraints usually allow constraint violation (Yuan and Lamperski 2018; Cao, Zhang, and Poor 2018; Kveton et al. 2008). Besides, OCO does not consider system dynamics or 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 linear quadratic regulator is piecewise linear (Bemporad et al. 2002). With disturbances (as considered in this paper), the problem is much more challenging. Current methods such as robust MPC (Limon et al. 2008, 2010; Rawlings and Mayne 2009) and stochastic MPC (Mesbah 2016; Oldewurtel, Jones, and Morari 2008) usually deploy linear policies for fast computation even though linear policies are suboptimal. Besides, most theoretical analysis of robust/stochastic MPC focus on stability, recursive feasibility, and constraints satisfaction, instead of policy regrets.

Notations and Conventions We let $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ denote the L_1, L_2, L_∞ norms respectively for vectors and matrices. Let $\mathbb{1}_n$ denote an all-one vector in $\mathbb{R}^n$. For two vectors $a, b \in \mathbb{R}^n$, we write $a \leq b$ if $a_i \leq b_i$ for any entry i. Let $\text{vec}(A)$ denote the vectorization of matrix A. For better exposition, some bounds use $\Theta(\cdot)$ to omit constants that do not depend on T or the problem dimensions explicitly.

Problem Formulation

In this paper, we consider an online optimal control problem with linear dynamics and affine constraints. Specifically, at each stage $t \in \{0, 1, \ldots, T\}$, 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

$$\mathcal{X} = \{x \in \mathbb{R}^n : D_x x \leq d_x\}, \quad \mathcal{U} = \{u \in \mathbb{R}^m : D_u u \leq d_u\},$$

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 that the initial value satisfies $x_0 = 0$. We leave the study of unknown parameters and general x_0 for the future.

Definition 1 (Safe controller)**.** Consider a controller (or an algorithm) $\mathcal{A}$ that chooses action $u_t^{\mathcal{A}} \in \mathcal{U}$ 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 *safe* 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 $\{w_k \in \mathcal{W}\}_{k=0}^T$.

Define the total cost of a safe algorithm/controller $\mathcal{A}$ 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 policies 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 define strong stability, which is a quantitative version of stability and is commonly introduced to ease non-asymptotic regret analysis in the online control literature (Agarwal, Hazan, and Singh 2019; Agarwal et al. 2019).

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 Q such that $A - BK = Q^{-1}LQ$, with $\|L\|_2 \leq 1 - \gamma$ and $\max(\|Q\|_2, \|Q^{-1}\|_2, \|K\|_2) \leq \kappa$.

As shown in Cohen et al. (2018), strongly stable controllers can be computed efficiently by SDP formulation.

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

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

where K is called safe if the controller $u_t = -Kx_t$ is safe 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 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. and bounded by $\|w_t\|_\infty \leq \bar{w}$, where $\bar{w} > 0$.[1]

Assumption 2. For any $t \geq 0$, 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 $\|x\|_2 \leq b$, $\|u\|_2 \leq b$, we have $\|\nabla_x c_t(x, u)\|_2 \leq Gb$ and $\|\nabla_u c_t(x, u)\|_2 \leq Gb$.

Next, we define strictly and loosely safe controllers.

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

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

In the following, we assume the existence of a *strictly* safe linear policy. The existence of a safe linear policy is necessary since otherwise our policy regret is not well-defined. The existence of a strictly safe policy provides some flexibility for the approximation steps in our algorithm design and is a common assumption in constrained optimization and control (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 safe 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 $\{w_k \in \mathcal{W}\}_{k=0}^T$. In addition, Assumption 3 implicitly assumes that 0 belongs to the interiors of $\mathcal{X}$ and $\mathcal{U}$ since we let $x_0 = 0$. Finally, though it is challenging to verify Assumption 3 directly, there are numerical verification methods, e.g. by solving linear matrix inequalities (LMI) programs (Limon et al. 2010).[2]

[1]The results of this paper can be extended to adversarial noises.

[2](Limon et al. 2010) provides an LMI program to compute a near-optimal linear controller for a time-invariant constrained control problem, which can be used to verify the existence of a safe solution. To verify Assumption 3, one could run the LMI program with the constraints tightened by ϵ and continue to reduce ϵ if no solution is found until ϵ is smaller than a certain threshold.

Preliminaries

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

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:

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

where, $M^{[i]} \in \mathbb{R}^{m \times n}$ and $w_t = 0$ for $t \leq 0$. Let $\boldsymbol{M} = \{M^{[i]}\}_{i=1}^H$ denote the list of parameter matrices for the disturbance-action policy.[3]

For the rest of the paper, we will fix $\mathbb{K}$ and discuss how to choose parameter $\boldsymbol{M}$. In Agarwal, Hazan, and Singh (2019), a bounded convex constraint set on policy $\boldsymbol{M}$ is introduced for technical simplicity and without loss of generality:[4]

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

The next proposition introduces state and action approximations 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 $t \geq 0$, the states and actions satisfy:*

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

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

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

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

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

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

[3]The disturbance-action policy is mainly useful for non-zero disturbances. Nevertheless, our theoretical results do not require $w_t \neq 0$ because for no-disturbance systems, any strongly stable controller $u_t = -\mathbb{K}x_t$ will only result in a constant $O(1)$ regret.

[4]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 $\mathcal{M}_2$.

where $\boldsymbol{M}_{t-H:t} \coloneqq \{\boldsymbol{M}_{t-H}, \dots, \boldsymbol{M}_t\}$, 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 $\mathring{\Phi}_k^x(\boldsymbol{M}) = \Phi_k^x(\boldsymbol{M}, \dots, \boldsymbol{M})$, $\mathring{\Phi}_k^u(\boldsymbol{M}) = \Phi_k^u(\boldsymbol{M}, \dots, \boldsymbol{M})$.

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.

Remark 1. The disturbance-action policy is related to *affine disturbance feedback* in stochastic MPC (Oldewurtel, Jones, and Morari 2008; Mesbah 2016), which also considers policies that are linear with disturbances to convexify the control problem in MPC's lookahead horizon.

OCO with memory. In Agarwal, Hazan, and Singh (2019), the unconstrained online optimal control problem is converted into *OCO with memory*, i.e. at each stage t, the 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 $\boldsymbol{M}_{t-H}, \dots, \boldsymbol{M}_{t-1}$ is sufficiently small, which guarantees a 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).

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:

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

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 into *OCO with memory and coupled constraints*, which is later converted into classical OCO and solved by OCO algorithms. The conversions leverage the approximation and the reformulation techniques in the **Preliminaries**. During the conversions, we ensure that the outputs of the OCO algorithms are safe for the original control problem. This is achieved by tightening the original constraints (adding buffer zones) to allow for approximation errors. Besides, our method ensures small approximation errors and thus small buffer zones so that the optimality/regret is not sacrificed significantly for safety. 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 into 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 $\{w_k \in \mathcal{W}\}_{k=0}^T$.

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 reviewed in **Robust Optimization with Constraints**.

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. 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 (10) and the definitions of $\tilde{x}_{t+1}$ and $\mathcal{W}$, 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 By Step 2 and our review of robust optimization, we can convert the constrained online optimal control problem into *OCO with memory and coupled constraints*. That is, at each 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, both 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 $\boldsymbol{M}_t$.[5] 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 $\epsilon_2 > 0$. Consequently, the constraints (14) and (15) are ensured by the polytopic constraints that only depend on $\boldsymbol{M}_t$:

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

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

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

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

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.[6] 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 technique (8), we can further convert the "OCO with memory and coupled constraints" in Step 3 into a classical OCO problem below. That is, at each 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 Step 4 and thus 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 **Theoretical Results**.

[5] Though we consider $\boldsymbol{M}_t = \cdots = \boldsymbol{M}_{t-H}$ here, the component $M_t^{[i]}$ of $\boldsymbol{M}_t = \{M_t^{[i]}\}_{i=1}^H$ can be different for different i.

[6] 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 here (see (Li, Das, and Li 2020)).

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).

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, online mirror descent, etc. One can also apply projection-free methods, e.g. (Yuan and Lamperski 2018), to reduce the computational burden at the expense of $o(T)$ constraint violation.

Remark 2. 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. Besides, safe RL literature may employ unsafe policies occasionally, 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 policies. Despite a smaller policy search space, our OGD-BZ still achieves desirable (theoretical) performance. Nevertheless, when the system is unknown, larger sets of exploration policies may benefit the performance, which is left as future work.

Remark 3. It is worth comparing our method with a well-known robust MPC method: tube-based robust MPC (see e.g. Rawlings and Mayne (2009)). Tube-based robust MPC also tightens the constraints to allow for model inaccuracy and/or disturbances. However, tube-based robust MPC considers constraints on the states, while our method converts the state (and action) constraints into the constraints on the policy parameters by leveraging the properties of disturbance-action policies.

Theoretical Results

In this section, we show that OGD-BZ guarantees both safety 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 the previous section respectively (see Lemma 1 and Lemma 3 in the proof of Theorem 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 linear controller $u_t = -Kx_t$ for any $K \in \mathcal{K}$ (see Lemma 4 in the proof of Theorem 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, and c_3 are polynomials of $\|D_x\|_\infty, \|D_u\|_\infty$, $\kappa, \kappa_B, \gamma^{-1}, \bar{w}, G$.

Safety of OGD-BZ

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

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

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

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

our OGD-BZ is safe, 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/feasible outputs of OGD-BZ. This is intuitive since the constraints are 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 implies a large enough H.

Secondly, Theorem 1 shows that, to ensure safety, 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 the previous section. 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 safety requires a large enough H, a small enough η, and an ϵ which is neither too large nor 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 $H \geq \max(\frac{\log(\frac{8(c_1+c_3)n\sqrt{m}}{\epsilon_*}T)}{\log((1-\gamma)^{-1}}, \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})})$.

Remark 4. It can be shown that it is safe to implement any $\boldsymbol{M} \in \Omega_\epsilon$ for an infinite horizon ($0 \leq t \leq +\infty$) under the conditions of Theorem 1 based on the proof of Theorem 1. For more details, please refer to the end of the proof of Theorem 1.

Policy Regret Bound for OGD-BZ

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

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

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

Theorem 2 provides a regret bound for OGD-BZ as long as OGD-BZ is safe. 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 safety (as long as $\Omega_\epsilon \neq \emptyset$). In addition, although a small stepsize η is preferred for safety 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 $H \geq \frac{\log(8(c_1+c_2)n\sqrt{m}T/\epsilon_*)}{\log((1-\gamma)^{-1})}$, $\eta = \Theta(\frac{1}{n^2\sqrt{m}H\sqrt{T}})$, $\epsilon = \epsilon_1(H) + \epsilon_2(\eta, H) = \Theta(\frac{\log(n\sqrt{m}T)}{\sqrt{T}})$, OGD-BZ is safe and* $\text{Reg(OGD-BZ)} \leq \tilde{O}\left((n^3 m^{1.5} {k_c}^{0.5})\sqrt{T}\right)$.

Corollary 1 shows that OGD-BZ achieves $\tilde{O}(\sqrt{T})$ regrets when $H \geq \Theta(\log T)$, $\eta^{-1} = \tilde{\Theta}(\sqrt{T})$, and $\epsilon = \tilde{\Theta}(1/\sqrt{T})$. This demonstrates that OGD-BZ can ensure both constraint satisfaction and sublinear regrets under the proper parameters of the algorithm. We remark that a larger H is preferred for better performance due to smaller approximation errors and a potentially larger policy search space Ω_ϵ, but 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). We leave the study on the constrained control with strongly convex costs for the future.

Proof of Theorem 1

To prove Theorem 1, we first provide lemmas to bound errors by $\epsilon_1(H), \epsilon_2(\eta, H)$, and $\epsilon_3(H)$, respectively. The proofs of Lemmas 1-4 and Corollary 2 in this subsection are provided in the arXiv version (Li, Das, and Li 2020).

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

Lemma 1 (Error bound $\epsilon_1(H)$)**.** *When $\boldsymbol{M}_k \in \mathcal{M}$ for all k and $H \geq \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})}$, we have*

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

The proof of Lemma 1 relies on the boundedness of x_t when implementing $\boldsymbol{M}_t \in \mathcal{M}$ as stated below.

Lemma 2 (Bound on x_t)**.** *With $\boldsymbol{M}_k \in \mathcal{M}$ for all k and $\kappa^2(1-\gamma)^H < 1$, we have*

$$\|x_t\|_2 \le b,$$

where $b = \frac{\kappa\sqrt{n}\bar{w}(\kappa^2+2\kappa^5\kappa_B\sqrt{mn}H)}{(1-\kappa^2(1-\gamma)^H)\gamma} + 2\sqrt{mn}\kappa^3\bar{w}/\gamma$. Hence, when $H \ge \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})}$, we have $b \le 8\sqrt{mn^2}H\bar{w}\kappa^6\kappa_B/\gamma$.

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

Lemma 3 (Error bound $\epsilon_2(\eta, H)$)**.** *When $H \ge \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})}$, the policies $\{\boldsymbol{M}_t\}_{t=0}^T$ generated by OGD-BZ with a constant stepsize η satisfy*

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

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 safe and is only $\epsilon_3(H)$-loosely safe.

Lemma 4 (Error bound $\epsilon_3(H)$)**.** *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*

$$\max(\|D_x[x_t^K - x_t^{\boldsymbol{M}(K)}]\|_\infty, \|D_u[u_t^K - u_t^{\boldsymbol{M}(K)}]\|_\infty) \le \epsilon_3(H)$$

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 safe.

Based on Lemma 4, 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 safe for $\epsilon_0 \ge 0$, then $\boldsymbol{M}(K) \in \Omega_{\epsilon_0-\epsilon_1-\epsilon_3}$.*

Proof of Theorem 1. For notational simplicity, we denote the states and actions generated by OGD-BZ as x_t and u_t in this proof. First, we show $\boldsymbol{M}(K_*) \in \Omega_\epsilon$ below. Since K_* defined in Assumption 3 is ϵ_*-strictly safe, by Corollary 2, there exists $\boldsymbol{M}(K_*) \in \Omega_{\epsilon_*-\epsilon_1-\epsilon_3}$. Since the set Ω_ϵ is smaller as ϵ increases, when $\epsilon_* - \epsilon_1 - \epsilon_3 \ge \epsilon$, we have $\boldsymbol{M}(K_*) \in \Omega_{\epsilon_*-\epsilon_1-\epsilon_3} \subseteq \Omega_\epsilon$, so Ω_ϵ is non-empty.

Next, we prove the safety by Lemma 1 and Lemma 3 based on the discussions in the previous section. Specifically, OGD-BZ guarantees that $\boldsymbol{M}_t \in \Omega_\epsilon$ for all t. Thus, by Lemma 3, 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 the previous section and Lemma 1, we have $D_{x,i}^\top x_t = D_{x,i}^\top A_{\mathbb{K}}^H x_{t-H} + D_{x,i}^\top \tilde{x}_t \le \|D_x A_{\mathbb{K}}^H x_{t-H}\|_\infty + g_i^x(\boldsymbol{M}_{t-H:t-1}) \le \epsilon_1 + \epsilon_2 + d_{x,i} - \epsilon \le d_{x,i}$ for any $\{w_k \in \mathcal{W}\}_{k=0}^T$ if $\epsilon \ge \epsilon_1 + \epsilon_2$. 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 safe. □

Proof of Remark 4. When implementing $\boldsymbol{M} \in \Omega_\epsilon$ for an infinite horizon, we have $\mathring{g}_i^x(\boldsymbol{M}) = g_i^x(\boldsymbol{M}, \dots, \boldsymbol{M}) \le d_{x,i} - \epsilon$. Since Proposition 1 holds for any $t \ge 0$, we still have $D_{x,i}^\top x_t = D_{x,i}^\top A_{\mathbb{K}}^H x_{t-H} + D_{x,i}^\top \tilde{x}_t \le \|D_x A_{\mathbb{K}}^H x_{t-H}\|_\infty + g_i^x(\boldsymbol{M}, \dots, \boldsymbol{M}) \le \epsilon_1 + d_{x,i} - \epsilon \le d_{x,i}$ for any $\{w_k \in \mathcal{W}\}_{k=0}^T$ if $\epsilon \ge \epsilon_1 + \epsilon_2$. Thus, $x_t \in \mathcal{X}$ for all $w_k \in \mathcal{W}$ for $t \ge 0$. Constraint satisfaction of u_t can be proved similarly.

Proof of Theorem 2

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

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

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

Lemma 5. *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 $G_f = \max_t \sup_{\boldsymbol{M}\in\Omega_\epsilon} \|\nabla \mathring{f}_t(\boldsymbol{M})\|_F \le Gb(1+\kappa)\sqrt{n}\bar{w}\kappa^2\kappa_B\sqrt{H}\frac{1+\gamma}{\gamma}$. Consequently, when $H \ge \frac{\log(2\kappa^2)}{\log((1-\gamma)^{-1})}$, we have $G_f \le \Theta(\sqrt{n^3H^3m})$ and the hidden factor is quadratic on $\bar{w}$.*

The proof details are provided in (Li, Das, and Li 2020).

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 4, we can construct a loosely safe $\boldsymbol{M}_{\text{ap}} = \boldsymbol{M}(K^*)$ 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 6. *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 7. *Under the conditions in Theorem 2, we have*

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

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 positive 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 8. *Apply Algorithm 1 with constant stepsize η, then* Part i $\leq O(Tn^2mH^2(1-\gamma)^H + n^3mH^3\eta T)$.

The proofs of Lemma 6 and Lemma 8 are similar to those in 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 Lemmas 5-8 and only explicitly showing the highest order terms.

Proof of Lemma 7

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 $\sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}_{\text{ap}}) \geq \sum_{t=0}^T \mathring{f}_t(\boldsymbol{M}^\dagger)$. 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/safe 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 $\Omega_1 = \{x : Cx \leq h\}$, $\Omega_2 = \{x : Cx \leq h - \Delta\}$, where $\Delta_i \geq 0$ for all i. Consider a convex function $f(x)$ that is L-Lipschitz continuous on Ω_1. If Ω_1 is bounded, i.e. $\sup_{x_1, x_1' \in \Omega_1} \|x_1 - x_1'\|_2 \leq \delta_1$ and if Ω_2 is non-empty, i.e. there exists $\mathring{x} \in \Omega_2$, then*

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

To prove Lemma 7, it suffices to bound the quantities in (19) for our problem and then plug them in (19).

Lemma 9. *There exists an enlarged polytope $\Gamma_\epsilon = \{\vec{W} : C\vec{W} \leq h_\epsilon\}$ that is equivalent to Ω_ϵ for any $\epsilon \in \mathbb{R}$, where $\vec{W}$ contains elements of $\boldsymbol{M}$ and auxiliary variables.*

Further, under the conditions of Theorem 1, (i) $\Gamma_{-\epsilon_1-\epsilon_3}$ is bounded by $\delta_1 = \Theta(\sqrt{mn} + \sqrt{k_c})$; (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. $\min_{\{i:\Delta_i>0\}}(h_{(-\epsilon_1-\epsilon_3)} - C\vec{W}^\circ)_i \geq \epsilon_*$.

Numerical Experiments

In this section, we numerically test our OGD-BZ on a thermal control problem with a Heating Ventilation and Air Conditioning (HVAC) system. Specifically, we consider the linear thermal dynamics studied in Zhang et al. (2016) with additional random disturbances, that is, $\dot{x}(t) = \frac{1}{\upsilon\zeta}(\theta^{\text{o}}(t) - x(t)) - \frac{1}{\upsilon}u(t) + \frac{1}{\upsilon}\pi + \frac{1}{\upsilon}w(t)$, where $x(t)$ denotes the room temperature at time t, $u(t)$ denotes the control input that is related with the air flow rate of the HVAC system, $\theta^{\text{o}}(t)$ denotes the outdoor temperature, $w(t)$ represents random disturbances, π represents external heat sources' impact, υ and ζ are physical constants. We discretize the thermal dynamics with $\Delta_t = 60$s. For human comfort and/or safe operation of device, we impose constraints on the room temperature, $x(t) \in [x_{\min}, x_{\max}]$, and the control inputs, $u(t) \in [u_{\min}, u_{\max}]$. Consider a desirable temperature θ^{set} set by the user and a control setpoint u^{set}. Consider the cost function $c(t) = q_t(x(t) - \theta^{set})^2 + r_t(u(t) - u^{set})^2$.

In our experiments, we consider $\upsilon = 100$, $\zeta = 6$, $\theta^o = 30°$C, $\pi = 1.5$, and let w_t be i.i.d. generated from Unif$(-2, 2)$. Besides, we consider $\theta^{set} = 24°$C, $x_{\min} = 22°$C, $x_{\max} = 26°$C, $u_{\min} = 0$, $u_{\max} = 5$. We consider $q_t = 2$ for all t and time-varying r_t generated i.i.d. from Unif$(0.1, 4)$. When applying OGD-BZ, we select $H = 7$ and a diminishing stepsize $\eta_t = \Theta(t^{-0.5})$, i.e. we let $\eta_t = 0.5(40)^{-0.5}$ for $t < 40$ and $\eta_t = 0.5(t+1)^{-0.5}$ for $t \geq 40$.

Figure 1 plots the comparison of OGD-BZ with different buffer sizes. Specifically, $\epsilon = 0.04$ is a properly chosen buffer size and $\epsilon = 0.4$ offers larger buffer zones. From Figure 1-(a), we can observe that the averaged regret with a properly chosen buffer size $\epsilon = 0.04$ quickly diminishes to 0, which is consistent with Theorem 2. In addition, Figure 1-(b) and Figure 1-(c) plot the range of $x(t)$ and $u(t)$ under random disturbances in 1000 trials to demonstrate the safety of OGD-BZ. With a larger buffer zone, i.e. $\epsilon = 0.4$, the range of x_t is smaller and further from the boundaries, thus being safer. Interestingly, the range of $u(t)$ becomes slightly larger, which still satisfies the control constraints because the control constraints are not binding/active in this experiment and which indicates more control power is used here to ensure a smaller range of $x(t)$ under disturbances. Finally, the regret with $\epsilon = 0.4$ is worse than that with $\epsilon = 0.04$, which demonstrates the trade-off between safety and performance and how the choices of the buffer size affect this trade-off.

Supplementary Proofs for Lemma 7

Proof of Proposition 2

Since $\Omega_2 \subseteq \Omega_1$, we have $\min_{\Omega_2} f(x) - \min_{\Omega_1} f(x) \geq 0$. Let $x_1^* = \arg\min_{\Omega_1} f(x)$. We will show that there exists $x_2^\dagger \in \Omega_2$ such that $\|x_1^* - x_2^\dagger\|_2 \leq \frac{\delta_1\|\Delta\|_\infty}{\min_{i\in S}(h - C\mathring{x})_i}$, where $S = \{i : \Delta_i > 0\}$. Then, by the Lipschitz continuity, we can prove the bound: $\min_{\Omega_2} f(x) - \min_{\Omega_1} f(x) \leq f(x_2^\dagger) - f(x_1^*) \leq \frac{L\delta_1\|\Delta\|_\infty}{\min_{i\in S}(h - C\mathring{x})_i}$.

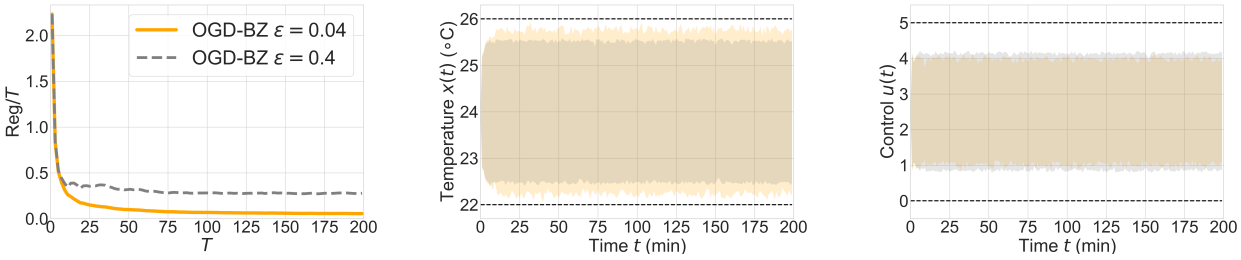
In the following, we will show, more generally, that there exists $x_2 \in \Omega_2$ that is close to x_1 for any $x_1 \in \Omega_1$. For ease of notation, we define $y = x - \mathring{x}$, $\Omega_1^y = \{y : Cy \leq h - C\mathring{x}\}$, and $\Omega_2^y = \{y : Cy \leq h - C\mathring{x} - \Delta\}$. Notice that $0 \in \Omega_2^y$ and $(h - C\mathring{x} - \Delta)_i \geq 0$. Besides, we have $y_1 = x_1 - \mathring{x} \in \Omega_1^y$. Further, by the convexity of Ω_1^y, we have $\lambda y_1 \in \Omega_1^y$ for $0 \leq \lambda \leq 1$.

If $(Cy_1)_i \leq (h - C\mathring{x} - \Delta)_i$ for all i, then $y_1 \in \Omega_2^y$ and $x_1 \in \Omega_2$. So we can let $x_2 = x_1$ and $\|x_2 - x_1\|_2 = 0$.

If, instead, there exists a set S' such that for any $i \in S'$, $(Cy_1)_i > (h - C\mathring{x} - \Delta)_i$. Then, define

$$\lambda = \min_{i\in S'} \frac{(h - C\mathring{x} - \Delta)_i}{(Cy_1)_i}.$$

Notice that $\lambda \in [0, 1)$. We can show that $\lambda y_1 \in \Omega_2^y$ below. When $i \in S'$, $(\lambda Cy_1)_i \leq (Cy_1)_i \frac{(h - C\mathring{x} - \Delta)_i}{(Cy_1)_i} = (h - C\mathring{x} - \Delta)_i$. When $i \notin S'$, we have $(\lambda Cy_1)_i \leq \lambda(h - C\mathring{x} - \Delta)_i \leq$



(a) Averaged regret (b) $x(t)$'s range (c) $u(t)$'s range

Figure 1: Comparison of OGD-BZ with buffer sizes $\epsilon = 0.04$ and $\epsilon = 0.4$. In Figure (b) and (c), the yellow shade represents the range of $x(t)$ generated by OGD-BZ with $\epsilon = 0.04$, while the grey shade is generated by OGD-BZ with $\epsilon = 0.4$.

$(h - C\mathring{x} - \Delta)_i$. Therefore, $\lambda y_1 \in \Omega_2^y$. Define $x_2 = \lambda y_1 + \mathring{x}$, then $x_2 \in \Omega_2$. Notice that $\|x_1 - x_2\|_2 = \|y_1 - y_2\|_2 = (1-\lambda)\|y_1\|_2 \le (1-\lambda)\delta_1$. Since $y_1 \in \Omega_1^y$, when $i \in S'$, we have $0 \le (h - C\mathring{x} - \Delta)_i < (Cy_1)_i \le (h - C\mathring{x})_i$. Therefore, $\frac{(h-C\mathring{x}-\Delta)_i}{(Cy_1)_i} \ge \frac{(h-C\mathring{x}-\Delta)_i}{(h-C\mathring{x})_i} = 1 - \frac{\Delta_i}{(h-C\mathring{x})_i}$. Consequently, by $S' \subseteq S$, we have $1 - \lambda \le \max_{i\in S'} \frac{\Delta_i}{(h-C\mathring{x})_i} \le \frac{\|\Delta\|_\infty}{\min_{i\in S'}(h-C\mathring{x})_i} \le \frac{\|\Delta\|_\infty}{\min_{i\in S}(h-C\mathring{x})_i}$.

Proof of Lemma 9

We first provide an explicit expression for Γ_ϵ and then prove the bounds on Γ_ϵ based on the explicit expression.

Lemma 10. *For any $\epsilon \in \mathbb{R}$, $\boldsymbol{M} \in \Omega_\epsilon$ if and only if there exist $\{Y^x_{i,k,l}\}_{(1\le i\le k_x, 1\le k\le 2H, 1\le l\le n)}$, $\{Y^u_{j,k,l}\}_{(1\le j\le k_u, 1\le k\le 2H, 1\le l\le n)}$, $\{Z^{[i]}_{k,j}\}_{(1\le i\le H, 1\le k\le m, 1\le j\le n)}$ such that*

$$\begin{cases} \sum_{k=1}^{2H}\sum_{l=1}^{n} Y^x_{i,k,l}\bar{w} \le d_{x,i} - \epsilon, \ \forall 1 \le i \le k_x \\ \sum_{k=1}^{2H}\sum_{l=1}^{n} Y^u_{j,k,l}\bar{w} \le d_{u,j} - \epsilon, \ \forall 1 \le j \le k_u \\ \sum_{j=1}^{n} Z^{[i]}_{k,j} \le 2\sqrt{n}\kappa^3(1-\gamma)^{i-1}, \forall 1 \le i \le H, 1 \le k \le m \\ -Y^x_{i,k,l} \le (D^\top_{x,i}\mathring{\Phi}^x_k(\boldsymbol{M}))_l \le Y^x_{i,k,l}, \quad \forall i,k,l \\ -Y^u_{j,k,l} \le (D^\top_{u,j}\mathring{\Phi}^u_k(\boldsymbol{M}))_l \le Y^u_{j,k,l} \quad \forall i,k,l \\ -Z^{[i]}_{k,j} \le M^{[i]}_{k,j} \le Z^{[i]}_{k,j}, \quad \forall i,k,j. \end{cases}$$

Let $\vec{W}$ denote the vector containing the elements of $\boldsymbol{M}, \boldsymbol{Y}^x = \{Y^x_{i,k,l}\}, \boldsymbol{Y}^u = \{Y^u_{j,k,l}\}, \boldsymbol{Z} = \{Z^{[i]}_{k,j}\}$. Thus, the constraints above can be written as $\Gamma_\epsilon = \{\vec{W} : C\vec{W} \le h_\epsilon\}$.

Since Lemma 10 holds for any $\epsilon \in \mathbb{R}$, we can similarly define $\Gamma_{-\epsilon_1-\epsilon_3} = \{\vec{W} : C\vec{W} \le h_{-\epsilon_1-\epsilon_3}\}$ which is equivalent to $\Omega_{-\epsilon_1-\epsilon_3}$. Lemma 10 is based on a standard reformulation method in constrained optimization to handle inequalities involving absolute values so the proof is omitted.

Proof of (i). Firstly, notice that $\sum_{j=1}^n (M^{[i]}_{k,j})^2 \le \sum_{j=1}^n (Z^{[i]}_{k,j})^2 \le (\sum_{j=1}^n Z^{[i]}_{k,j})^2 \le 4n\kappa^6(1-\gamma)^{2i-2}$. Then,

$$\sum_{k=1}^{m}\sum_{i=1}^{H}\sum_{j=1}^{n}(M^{[i]}_{k,j})^2 \le \sum_{k=1}^{m}\sum_{i=1}^{H}\sum_{j=1}^{n}(Z^{[i]}_{k,j})^2 \le 4nm\kappa^6\frac{1}{2\gamma-\gamma^2}$$

Similarly, by the first two constraints in Lemma 10 and by the definition of $\Gamma_{-\epsilon_1-\epsilon_3}$, we have $\sum_{k=1}^{2H}\sum_{l=1}^{n}(Y^x_{i,k,l})^2 \le (d_{x,i} + \epsilon_1 + \epsilon_3)^2/\bar{w}^2 \le (d_{x,i} + \epsilon_*)^2/\bar{w}^2$ and $\sum_{k=1}^{2H}\sum_{l=1}^{n}(Y^u_{j,k,l})^2 \le (d_{u,j} + \epsilon_1 + \epsilon_3)^2/\bar{w}^2 \le (d_{u,j} + \epsilon_*)^2/\bar{w}^2$. Therefore, $\sum_{i=1}^{k_x}\sum_{k=1}^{2H}\sum_{l=1}^{n}(Y^x_{i,k,l})^2 \le \sum_{i=1}^{k_x}(d_{x,i} + \epsilon_*)^2/\bar{w}^2$, and $\sum_{j=1}^{k_u}\sum_{k=1}^{2H}\sum_{l=1}^{n}(Y^u_{j,k,l})^2 \le \sum_{j=1}^{k_u}(d_{u,j} + \epsilon_*)^2/\bar{w}^2$. Consequently,

$$\begin{aligned} &\|\boldsymbol{M}\|_F^2 + \|\boldsymbol{Y}^x\|_F^2 + \|\boldsymbol{Y}^u\|_F^2 + \|\boldsymbol{Z}\|_F^2 \\ &\le \frac{8nm\kappa^6}{2\gamma-\gamma^2} + \frac{\sum_{i=1}^{k_x}(d_{x,i}+\epsilon_*)^2 + \sum_{j=1}^{k_u}(d_{u,j}+\epsilon_*)^2}{\bar{w}^2} = \delta_1^2 \end{aligned}$$

where $\delta_1 = \Theta(\sqrt{mn} + \sqrt{k_c})$ by the boundedness of ϵ_*, d_x, d_u. (Although δ_1 depends linearly on $1/\bar{w}$, we will show $L = TG_f$ and G_f is quadratic on $\bar{w}$ by Lemma 5, hence, $L\delta_1$ is still linear with $\bar{w}$.)

Proof of (ii). Since the gradient of $\mathring{f}_t(\boldsymbol{M})$ is bounded by $G_f = \Theta(\sqrt{n^3mH^3})$, the gradient of $\sum_{t=0}^{T}\mathring{f}_t(\boldsymbol{M})$ is bounded by $LG_f = \Theta(T\sqrt{n^3mH^3})$.

Proof of (iii). Notice that the differences between Γ_ϵ and $\Gamma_{-\epsilon_1-\epsilon_3}$ come from the first two lines of the right-hand-side of inequalities in Lemma 10, which is $\epsilon + \epsilon_1 + \epsilon_3$ in total.

Proof of (iv). From the proof of Theorem 1, we know that $\boldsymbol{M}(K_*) \in \Omega_{\epsilon_*-\epsilon_1-\epsilon_3} \subseteq \Omega_\epsilon$. Therefore, there exist corresponding $\boldsymbol{Y}^x(K_*), \boldsymbol{Y}^u(K_*), \boldsymbol{Z}(K_*)$ such that $\vec{W}^\circ = \mathrm{vec}(\boldsymbol{M}(K_*), \boldsymbol{Y}^x(K_*), \boldsymbol{Y}^u(K_*), \boldsymbol{Z}(K_*)) \in \Gamma_{\epsilon_*-\epsilon_1-\epsilon_3} \subseteq \Gamma_\epsilon$. Therefore, $\min_{\{i:\Delta_i>0\}}(h_{-\epsilon_1-\epsilon_3} - C\vec{W}^\circ)_i \ge \epsilon_1 + \epsilon_3 - (-\epsilon_* + \epsilon_1 + \epsilon_3) = \epsilon_*$.

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. There are many interesting future directions, e.g. (i) consider adversarial disturbances and robust stability, (ii) consider soft constraints and unbounded noises, (iii) consider bandit feedback, (iv) reduce the regret bound's dependence on dimensions, (v) consider unknown systems, (vi) consider more general policies than linear policies, (vii) prove logarithmic regrets for strongly convex costs, etc.

Acknowledgements

This work was conducted while the first author was doing internship at the MIT-IBM Watson AI Lab. We thank the helpful suggestions from Jeff Shamma, Andrea Simonetto, Yang Zheng, Runyu Zhang, and the reviewers.

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 satisfy 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 cares 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.; 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.

Chen, X.; Qu, G.; Tang, Y.; Low, S.; and Li, N. 2021. Reinforcement Learning for Decision-Making and Control in Power Systems: Tutorial, Review, and Vision. *arXiv preprint arXiv:2102.01168* .

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. 2020a. The Power of Linear Controllers in LQR Control. *arXiv preprint arXiv:2002.02574* .

Goel, G.; and Hassibi, B. 2020b. Regret-optimal control in dynamic environments. *arXiv preprint arXiv:2010.10473* .

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.

Li, Y.; Chen, X.; and Li, N. 2019. Online optimal control with linear dynamics and predictions: Algorithms and regret analysis. In *Advances in Neural Information Processing Systems*, 14887–14899.

Li, Y.; Das, S.; and Li, N. 2020. Online Optimal Control with Affine Constraints. *arXiv preprint arXiv:2010.04891* .

Li, Y.; and Li, N. 2020. Leveraging Predictions in Smoothed Online Convex Optimization via Gradient-based Algorithms. *Advances in Neural Information Processing Systems* 33.

Li, Y.; Qu, G.; and Li, N. 2020. Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit. *IEEE Transactions on Automatic Control* .

Li, Y.; Tang, Y.; Zhang, R.; and Li, N. 2019a. Distributed Reinforcement Learning for Decentralized Linear Quadratic Control: A Derivative-Free Policy Optimization Approach. *arXiv preprint arXiv:1912.09135* .

Li, Y.; Zhong, A.; Qu, G.; and Li, N. 2019b. Online markov decision processes with time-varying transition probabilities and rewards. In *ICML workshop on Real-world Sequential Decision Making*.

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.

Mesbah, A. 2016. Stochastic model predictive control: An overview and perspectives for future research. *IEEE Control Systems Magazine* 36(6): 30–44.

Nonhoff, M.; and Müller, M. A. 2020. An online convex optimization algorithm for controlling linear systems with state and input constraints. *arXiv preprint arXiv:2005.11308* .

Oldewurtel, F.; Jones, C. N.; and Morari, M. 2008. A tractable approximation of chance constrained stochastic MPC based on affine disturbance feedback. In *2008 47th IEEE conference on decision and control*, 4731–4736. IEEE.

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.

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* .

Zhang, X.; Shi, W.; Li, X.; Yan, B.; Malkawi, A.; and Li, N. 2016. Decentralized temperature control via HVAC systems in energy efficient buildings: An approximate solution procedure. In *Proceedings of 2016 IEEE Global Conference on Signal and Information Processing*, 936–940.