Online Bandit Control with Dynamic Batch Length and Adaptive Learning Rate

Jihun Kim 1  Javad Lavaei 1

Abstract

This paper is concerned with the online bandit control problem, where the goal is to learn the best stabilizing controller from a pool of stabilizing and destabilizing controllers of unknown types for a given dynamical system. We develop an algorithm, named Dynamic Batch length and Adaptive learning Rate (DBAR), and study its stability and regret. Dynamic batch length in DBAR enables the system to attain asymptotic stability, where the algorithm behaves as if there were no destabilizing controllers in the candidate pool. Adaptive learning rate based on the state norm enhances the system to alleviate the exponentially increasing regret in the number of destabilizing controllers. Unlike the existing Exp3 algorithm requiring an exponential stability assumption and leading to a regret bound with an exponential term, DBAR uses a much broader notion of stability and mitigates the exponential term in the regret.

1. Introduction

The multi-armed bandit (MAB) problem aims to minimize the total cost of pulling a series of arms while receiving immediate cost feedback for each arm pulled. Given a finite number of arms, the problem balances between exploration and exploitation of arms without knowing the exact cost structure of each arm. On the other hand, the online optimal control problem considers a transition dynamic $x_{t+1} = f(x_t, u_t, w_t)$ and a set of cost functions $c_t(x_t, u_t)$, $t = 0, \ldots, T$, where the goal is to minimize the sum of costs over time, while both $f$ and $c_t$ are fully or partially unknown. Basically, MAB is a special type of the online optimal control problem in the sense that MAB is stateless and simply selects an action each time, while the online control problem has a countable or an uncountable number of states and selects a controller, acting as a function from states into actions, each time without knowing the cost functions. Bandit algorithms can thus be leveraged for online control, wherein the average cost incurred with a controller can be interpreted as the bandit feedback of pulling the controller-arm (Lin et al., 2023; Li et al., 2023).

In this paper, we address the online nonstochastic control problem where both a transition dynamic $f$ and cost functions $c_t$ can be unbounded and nonlinear. We only have knowledge about $x_t$ and the bandit feedback $c_t(x_t, u_t)$ at time $t$, with adversarial disturbances $w_t$ injected at each time step as in Gradu et al. (2020) and Cassel & Koren (2020). We operate the system with a single trajectory, meaning that the system state cannot be reset. Meanwhile, to overcome the difficulties of an unknown nonlinear system, we are given a finite set of $N$ controllers in advance, where we are not aware of whether each controller can stabilize the system but are allowed to alternate between these controllers within a single trajectory according to a specific logic.

To deal with this online bandit control with logic-based switching, Li et al. (2023) adopted their Exp3-ISS algorithm, which uses the well-known Exp3 algorithm (Auer et al., 2002) with a mini-batch approach (Arora et al., 2012), while successively removing $|\mathcal{U}|$ existing destabilizing controllers when detected in terms of input-to-state stability (ISS). Their Exp3-ISS algorithm turns out to achieve both finite-gain stability and $O(T^{2/3}N^{1/3} + \exp(O(|\mathcal{U}|)))$ regret guarantees, even with unbounded cost functions. In this case, we aim to significantly relax the notion of stability and yet sharpen the regret bound by designing our algorithm DBAR (Dynamic Batch length and Adaptive learning Rate).

Related work. Optimal control problems have been widely leveraged in a variety of fields with the influential dynamic programming approach (Bellman, 1957). Recent successes of reinforcement learning (RL) in safety-critical systems, such as aircraft (Razaghi et al., 2022), robotics (Ibarz et al., 2021), and autonomous driving (Kiran et al., 2021), are also deeply rooted in optimal control methods (Bertsekas, 2019).

Online nonstochastic control considers a dynamical system with adversarial disturbances, which is more challenging than having statistical noise. Early papers assumed full access to cost functions, enabling us to leverage optimal policy structure with cost function gradients (Agarwal et al., 2019; Foster & Simchowitz, 2020; Hazan et al., 2020; Hazan & Singh, 2022). Later, studies were generalized to address the
problem without cost gradients information (Gradu et al., 2020; Cassel & Koren, 2020; Ghai et al., 2023; Sun et al., 2023); instead, they estimated the cost gradients, using the history of scalar cost (bandit feedback) along the trajectory. However, the above research restricts the system to linear transition dynamics. Instead, our work considers the candidate controller pool to handle unknown nonlinear systems.

*Logic-based switching control* (Morse, 1995; Narendra & Balakrishnan, 1997; Narendra et al., 2003) has been demonstrated to gain stability of optimal control problems in various contexts, including feedforward systems (Yu, 2020), multi-agent systems (Lv et al., 2022), and fuzzy control (Balta et al., 2023). The common idea to ensure system stability is to falsify the detected destabilizing controller, meaning that one can completely remove those controllers failing to satisfy certain stability criteria from the controller pool (Baldi et al., 2010; Battistelli et al., 2010; 2014; 2018; Stefanovic & Safonov, 2011; Li et al., 2023).

*Multi-armed bandits* with adversarial disturbances were first addressed in the pioneering work by Auer et al. (2002) under bounded costs in their notable Exp3 algorithm. Arora et al. (2012) later improved the algorithm using the same controller within a mini-batch, attaining a regret bound equivalent to the lower bound presented in Dekel et al. (2014). As we have access to the candidate controller pool in our problem setting, we adopt a bandit-related approach. Meanwhile, as the cost deeply relies on the system state as a context, adopting contextual bandit (Luo et al., 2018; Shen et al., 2019; Gur et al., 2022) seems appealing. However, in our scenario of observing a single trajectory, we are not allowed to interact with the oracle of the contextual bandit.

*Dynamic batching* gained considerable attention for training deep neural networks by increasing the batch size over time and adaptively increasing the learning rate to maintain the ratio between the two (Devarakonda et al., 2017; Bollapragada et al., 2018; Shallue et al., 2019; Ma et al., 2023). Although this has been widely used in the machine learning literature, we adopt this idea to our control problem, progressively increasing the batch length within a single trajectory to achieve asymptotic stability.

*Adaptive learning rate* in machine learning is generally determined by a set of gradients observed so far (Ruder, 2016). As we do not have access to the gradients in our problem, we focus on the learning rate for bandit algorithms. Recently, it was shown in Aubert et al. (2023) that two different constant learning rates for bandits cannot be distinguished in the learning process, thus they emphasized the necessity of using a polynomially decreasing learning rate. Several works (van Erven et al., 2011; de Rooij et al., 2014) also suggested using decreasing learning rate as the batch length increases. Building on this idea, Li et al. (2023) proposed to use a non-increasing learning rate over time, while no theoretical guarantee was presented. To the best of our knowledge, this paper is the first work to provide theoretical guarantees for the adaptive learning rate scheme based on the stability of state norm, where the rate is not necessarily non-increasing.

**Contribution.** The idea of our algorithm is three-fold:

1. We adopt a dynamic batch length instead of a fixed length to achieve both asymptotic and finite-gain stability. The batch length is scheduled to be non-decreasing and growing unboundedly over time.

2. As the batch length becomes eventually unbounded, it is practical to define the concept of a destabilizing controller as one that violates (asymptotic) ISS rather than exponential ISS. Thus, we allow more controllers to qualify as a stabilizing controller within a *priori* controller pool.

3. Instead of using a fixed learning rate, we adopt an adaptive learning rate that relies on the system state norm. While the conventional way to apply the Exp3 Algorithm is to use a non-increasing learning rate, we decrease the learning rate if the state is unstable and subsequently increase the learning rate if the state returns to a stable region. By implementing this approach, we can alleviate the exponential term $\exp(O(|\mathcal{U}|))$ and attain a regret bound $O(T^{2/3}N^{1/3}(|\mathcal{U}| + 1)^{1/3})$ if $|\mathcal{U}|$ is known and $O(T^{2/3}N^{1/3}(|\mathcal{U}| + 1)^{1/2})$ if $|\mathcal{U}|$ is unknown.

We additionally apply our algorithm DBAR to the switched system (Tousi et al., 2008; Zhao et al., 2022), where the transition dynamic and controller pool can change infrequently according to either the detection of a destabilizing controller or pre-determined time instants (Battistelli et al., 2011).

**Outline.** The paper is organized as follows. In Section 2, we formulate the problem with definitions and assumptions. In Section 3, we propose our DBAR algorithm. In Section 4, we study the stability of the algorithm, the regret bound, and its applications in switched systems. In Section 5, we present numerical experiments on the DBAR algorithm with an ablation study on batch length and learning rate. Finally, concluding remarks are provided in Section 6.

**Notation.** For a vector $z$, $\|z\|$ denotes the Euclidean norm of a vector. We use $O(\cdot)$ for the big-O notation, $o(\cdot)$ for the small-o notation, and $\tilde{O}(\cdot)$ for the big-O notation hiding logarithmic factors. For an event $A$, $\mathcal{I}_A$ denotes an indicator function, where $\mathcal{I}_A = 1$ if an event $A$ occurs and $\mathcal{I}_A = 0$ otherwise. $Pr(A)$ denotes the probability of an event $A$. Let $E[A]$ denote the expectation operator. For a set $Z$, we use $|Z|$ for the cardinality and $Z^c$ for the complement of a set $Z$. For a real number $e$, we use $\lfloor e \rfloor$ for the floor and $\lceil e \rceil$ for the ceiling of $e$. Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{Z}_+$ denote the set of nonnegative integers. For $e_1, e_2 \in \mathbb{Z}_+$ where $e_2 \leq e_1$, let $\{i_{e_1:e_2}\}$ denote the set $\{i_e : e_2 \leq e \leq e_1, e \in \mathbb{Z}_+\}$. 


2. Problem Formulation

Consider the general discrete-time dynamical system $x_{t+1} = f(x_t, u_t, w_t)$, $t = 0, \ldots, T - 1$, where $x_t \in \mathbb{R}^n$ is the system state at time $t$, $u_t \in \mathbb{R}^m$ is the control input at time $t$ to be designed via an algorithm, and $w_t \in \mathbb{R}^W$ is the adversarial noise at time $t$. The control input $u_t$ is determined by selecting a controller from a $\pi$-priori finite number of controller pool consisting of $\pi_i : \mathbb{R}^n \to \mathbb{R}^m$, $i = 1, \ldots, N$. Each time instance $t$ is associated with a cost function $c_t : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. The state transition is governed by the dynamic $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^W \to \mathbb{R}$. We have the following assumptions on the transition dynamic $f$.

**Assumption 2.1** (Transition dynamics). The transition dynamic $f$ is $L_f$-Lipschitz continuous with $L_f \geq 1$; i.e., $|f(x, u, w) - f(\tilde{x}, \tilde{u}, \tilde{w})| \leq L_f (|x - \tilde{x}| + |u - \tilde{u}| + |w - \tilde{w}|)$ for all $x, \tilde{x} \in \mathbb{R}^n$, $u, \tilde{u} \in \mathbb{R}^m$, $w, \tilde{w} \in \mathbb{R}^W$, where $W = \{ w \in \mathbb{R}^W : ||w|| \leq w_{\text{max}} \}$ and the bounding constant $w_{\text{max}} > 0$ is assumed to be known. We let $f(0, 0, 0) = f_0$.

We also adopt the notion of locally Lipschitz continuous cost functions $c_t$ given in Li et al. (2023), which contains quadratic tracking costs along an arbitrary bounded state trajectory and the corresponding action sequence.

**Assumption 2.2** (Cost functions). There exist $L_{c1}, L_{c2} > 0$ such that $|c_t(x, u) - c_t(\tilde{x}, \tilde{u})| \leq (L_{c1} \max\{||x||, ||\tilde{x}||\} + \max\{||u||, ||\tilde{u}||\}) + L_{c2} (||x - \tilde{x}|| + ||u - \tilde{u}||)$ for all $x, \tilde{x} \in \mathbb{R}^n$, $u, \tilde{u} \in \mathbb{R}^m$, $t \in \mathbb{Z}_+$. There exists $c_{0,\text{max}} \geq 0$ such that $|c_t(0, 0)| \leq c_{0,\text{max}}$ for all $t \in \mathbb{Z}_+$.

Input-to-state (asymptotic) stability (ISS) is a classical notion of stability implying that the controller successfully stabilizes the system under any bounded noises (Sontag, 2008; Khalil, 2015). Incremental (asymptotic) stability extends the input-to-state stability to describe the asymptotic behavior of some trajectory towards a different trajectory (Tran et al., 2016). It is worth noting that Li et al. (2023) also adopted these concepts under an exponential stability assumption; i.e., they require some controllers to satisfy exponential ISS and exponential incremental stability. However, it is far more practical to include general asymptotic concepts within stabilizing controllers. We will address this issue below.

**Definition 2.3** (Input-to-state stable controller). A controller $\pi$ is input-to-state stable (ISS) if there exists a non-increasing function $\beta(\cdot) : \mathbb{Z}_+ \to \mathbb{R}$ that satisfies $\beta(0) = 1$ with $\lim_{t \to \infty} \beta(t) = 0$ and $\beta > 0$ such that for any $x_0 \in \mathbb{R}^n$ and $||w_t|| \leq w_{\text{max}}$ for all $t \geq 0$, the sequence $\{x_t\}_{t \geq 0}$ determined by $x_{t+1} = f(x_t, \pi(x_t), w_t)$ satisfies $||x_t|| \leq \beta(t)||x_0|| + \gamma w_{\text{max}}$.

**Definition 2.4** (Incrementally stable controller). A controller $\pi$ is incrementally stable if there exists a non-increasing function $\beta(\cdot) : \mathcal{N} \cup \{0\} \to \mathbb{R}$ that satisfies $\beta(t) \to 0$ as $t \to \infty$ such that for any $x_0, \tilde{x}_0 \in \mathbb{R}^n$ and $||w_t|| \leq w_{\text{max}}$ for all $t \geq 0$, it holds that $||x_t - \tilde{x}_t|| \leq \beta(t)||x_0 - \tilde{x}_0||$ for any two sequences determined by $x_{t+1} = f(x_t, \pi(x_t), w_t)$ and $\tilde{x}_{t+1} = f(\tilde{x}_t, \pi(\tilde{x}_t), w_t)$.

**Definition 2.5** (Stabilizing and destabilizing controller). Let $\mathcal{S}$ denote an index set of stabilizing controllers that satisfy both of Definitions 2.3 and 2.4. We also let $\mathcal{U}$ denote an index set of destabilizing controllers that do not satisfy Definition 2.3. Thus, we have $|\mathcal{S}| \geq 1$ and $\mathcal{U} \subseteq \mathcal{N}$.

**Assumption 2.6** (Controller pool). Consider the candidate controller index set $\mathcal{P}_0 = \{ 1, \ldots, N \}$. There exists at least one controller in $\mathcal{P}_0$ that satisfies Definitions 2.3 and 2.4. All candidate controllers are $L_\pi$-Lipschitz continuous. There exists $\pi_{0,\text{max}} \geq 0$ such that $||\pi_i(0)|| \leq \pi_{0,\text{max}}$ for all $i \in \mathcal{P}_0$.

**Remark 2.7**. Definition 2.4 is a stronger notion than Definition 2.3, by triangle inequality. However, for a system with additive noise; i.e., $f(x_t, u_t, w_t) = (F(x_t, u_t) + H(w_t))$, a controller satisfying Definition 2.3 implies Definition 2.4. In such a case, Assumption 2.6 boils down to requiring only one ISS controller in the controller pool.

Now, we define different notions of system stability with bounded adversarial disturbances $w_t$, where $||w_t|| \leq w_{\text{max}}$ holds. Asymmetric stability and finite-gain stability both shed light on the connection between the disturbance input and the corresponding state output, where none of them implies the other (Hill & Moylan, 1980). Hence, it is desirable to achieve both stability notions.

**Definition 2.8** (Asymptotic stability). A system is asymptotically stable if the sum of state norms satisfies $\lim_{t \to \infty} \frac{1}{t} \sum_{t=0}^{T} ||x_t|| \leq \gamma w_{\text{max}}$.

**Definition 2.9** (Finite-gain stability). A system is finite-gain $L_1$ stable if there exist constants $A_1, A_2 > 0$ such that $\sum_{t=0}^{T} ||x_t|| \leq A_1 \cdot w_{\text{max}} T + A_2$.

Recall that $x_t$ and $u_t$ denote the state and action sequence for the system according to the algorithm. We also let $x_t^* \ast u_t^*$ denote the optimal state and action sequence generated by the best stabilizing controller $i^*$ that satisfies both of Definitions 2.3 and 2.4; i.e., $i^* = \arg \min_{i \in \mathcal{S}} \sum_{t=0}^{T} c_t(x_t, u_t) - c_t(x_t^*, u_t^*)$.

3. Algorithm Description

Denote the number of batches in the algorithm by $B$. Denote by $t_b$ the start time for each batch $b = 0, 1, \ldots, B - 1$, and let $t_B = T + 1$. We implement the same policy within the mini-batch. Our idea of dynamic batch length aims to attain $\lim_{t \to \infty} \frac{B}{t} = 0$ (see Lemma A.9 in the Appendix).
Adversarial Noise

Remark

denote the number of times that the Break statement in
is found to violate Definition 2.3; generate the state trajectory based on the selected controller
below the red line after a certain time.

> has higher values than the red line, which is our asymptotic
> than the orange line. Moreover, the blue line occasionally
> adversarial noise, the blue line shows a larger state norm
> both relatively easier statistical noise and more challenging
> orange lines represent the state norms generated by a fixed
> after the first two batches.
> and therefore one can adjust
> of dynamic batch length only requires
> exponential notions in Li et al. (2023). Moreover, our design
> unboundedly over time, excessively strict stability criteria
> type of batch length formulation as polynomial batches with
> sampled from Uniform [−0.2, 0.5], and (b) 0.15 + 0.35 \sin(\frac{t}{\pi}).

Assumption 3.1 (Dynamic batch length). We design our
batch length \((\tau_b)_{b \geq 0}\) as follows:

1. \(\tau_b\) is non-decreasing in \(b\) and \(\lim_{b \to \infty} \tau_b = +\infty\).
2. \(\arg \max_{b \geq 0} \frac{\tau_{b+1}}{\tau_b} = 0\) and \(\lim_{b \to \infty} \frac{\tau_{b+1}}{\tau_b} = 1\).

For example, \(\tau_b = [z_1(z_2)^b] > 0\) and \(\tau_b = [z_1(\nu b + z_2)]^b\) for every \(b \geq 1\) with the constants \(z_1, z_2, \nu > 0\)

satisfy Assumption 3.1. For future use, we refer to this
type of batch length formulation as polynomial batches with
\((z_1, z_2, \nu, \nu)\).

Remark 3.2. As our dynamic batch length eventually grows
unboundedly over time, excessively strict stability criteria
may result in most of the candidate controllers violating
these criteria. Thus, it is crucial to adopt (asymptotic) ISS
and incremental stability as our criteria, instead of
exponential notions in Li et al. (2023). Moreover, our design
of dynamic batch length only requires \(\max_{b \geq 0} \frac{\tau_{b+1}}{\tau_b} = \tau_0\),
and therefore one can adjust \(\frac{\tau_{b+1}}{\tau_b}\) in a more flexible manner
after the first two batches.

Figure 1 demonstrates the necessity of a dynamic batch
length regardless of the noise assumption. The blue and
orange lines represent the state norms generated by a fixed
batch length and a dynamic batch length, respectively. With
both relatively easier statistical noise and more challenging
adversarial noise, the blue line shows a larger state norm
than the orange line. Moreover, the blue line occasionally
has higher values than the red line, which is our asymptotic
stability bound \(\gamma w_{\max} = 1.5\), while the orange line remains
below the red line after a certain time.

We propose our DBAR algorithm in Algorithm 1. Lines 3-9
generate the state trajectory based on the selected controller
\(\pi_{K_b}\) for the current batch \(b\), and falsify the controller if it
is found to violate Definition 2.3; i.e., \(K_b \in \mathcal{U}\). Here, let
\(U\) denote the number of times that the Break statement in
Line 7 is activated. In the rest of the paper, when we say
the Break statement is activated, it means that Line 7 of
Algorithm 1 has been activated. As the controllers in \(\mathcal{U}\) do
not suffer from the Break statement, they always remain in
the controller pool. Accordingly, we have \(U \leq |\mathcal{U}|\).

Lines 11-20 determine \(\alpha_{b+1}\) and \(s_{b+1}\) indicating the mag-
nitude of the next batch’s initial state norm compared to
\(\|x_0\|\). It is worth noting that Line 14 avoids \(s_{b+1} > s_b + 1\), and
simultaneously, Lipschitz constants in Assumptions 2.1
and 2.6 ensure that \(\alpha(.)\) does not tend towards infinity (see
Lemma A.5 in the Appendix). It is later discussed formally
in Lemma 4.5 that these observations cause \(s_b \neq 0\) to occur
at most \(O(U)\) times throughout the algorithm.

Lines 21-26 determine the weight \(W_{b+1}(k)\) for each con-
troller \(k\). In Line 21, we use the sum of costs at the current
batch \(b\) to add up to the weight in Line 25. In Lines 22-26,
we reset the weight if $s_{b+1} \neq s_b$. This resetting weight idea to forget the costs in the past is also proposed in van Erven et al. (2011). In the scenario that the Lipschitz constant $L_f$ is very large, it may help to forget the time-varying costs $c_0, \ldots, c_{t-1}$ and restart gathering the information from the outset. Line 22 is exactly the case where the next batch’s state norm significantly deviates from the current state norm.

Lines 27-29 calculate the adaptive learning rate $\eta_{b+1}$ for the next batch $b+1$ used to apply the Exp3 algorithm to our problem. It indicates that the learning rate decreases in unstable states and increases back to the initial value when the state norm returns to a stable region. Thus, the learning rate fluctuates depending on the state norm. However, it is essential to note that the effective learning rate, determined by the ratio $\frac{s_{b+1}}{s_b}$, indeed decreases as the batch length increases even if $s_{b+1} = s_b$. The only plausible situation in which the effective rate may increase is $s_{b+1} < s_b$ with $(\alpha_{b+1})^2 > \frac{s_{b+1}}{s_b}$. Apart from this scenario, the effective learning rate experiences a polynomial decay with polynomial batches defined in Assumption 3.1, which does not cause any contradiction with the polynomially decreasing learning rate concept proposed in Aubert et al. (2023).

Our adaptive learning rate basically intends to stabilize the cost of current batch, which can possibly be unbounded with a dynamic batch length. This helps alleviate the exponential term $\exp(O(\|U\|))$ that potentially appears in the regret bound. Moreover, our selection of the policy for the next batch is based on the state norm, which can be interpreted as context-driven. This approach enables us to harness a form of contextual bandit without requiring strict assumptions.

4. Main Results

4.1. Stability

In Algorithm 1, we define $H(t) := \sum_{i=0}^{t-1} \beta(i)$, which determines the scope of stabilizing controllers throughout the entire horizon. Before presenting the main theorems on stability, we provide the following useful lemma.

**Lemma 4.1.** Define $H(t) := \sum_{i=0}^{t-1} \beta(i)$. Then, we have

$$\lim_{{t \to \infty}} \frac{H(t)}{t} = 0.$$  

**Proof.** Recall that we designed $\beta(\cdot)$ to be non-increasing and nonnegative. Then, we have $\beta(i) \leq \int_{i-1}^{t-1} \beta(x)dx$ for every $i \geq 1$. Using the inequality, one can write

$$0 \leq H(t) = \beta(0) + \sum_{i=1}^{t-1} \beta(i) \leq \beta(0) + \int_{0}^{t-1} \beta(x)dx.$$  

(1)

If $\lim_{{t \to \infty}} H(t) < \infty$, clearly $\lim_{{t \to \infty}} \frac{H(t)}{t} = 0$ holds. If $\lim_{{t \to \infty}} H(t) = \infty$, we leverage L’Hôpital’s rule with $\beta(t) \to 0$ as $t \to \infty$ to derive

$$\lim_{{t \to \infty}} \frac{H(t)}{t} \leq \lim_{{t \to \infty}} \frac{\beta(0) + \int_{0}^{t-1} \beta(x)dx}{t} = \lim_{{t \to \infty}} \frac{\beta(t-1)}{1} = 0,$$

where the first inequality follows from (1). $\square$

Now, we present the stability results of Algorithm 1.

**Theorem 4.2** (Asymptotic stability). In Algorithm 1, suppose that $\frac{\tau_x}{\tau_y}(\gamma) < 1$. Then, it holds that

$$\lim_{{T \to \infty}} \frac{1}{T} \sum_{{t=0}}^{T} \|x_t\| \leq \gamma w_{\text{max}}.$$  

**Theorem 4.3** (Finite-gain stability). In Algorithm 1, suppose that $\frac{\tau_x}{\tau_y}(\gamma) < 1$. Assume that $\lim_{{t \to \infty}} H(t) < \infty$. Then, Algorithm 1 achieves finite-gain $L_1$ stability; i.e., there exist constants $A_1, A_2 > 0$ such that for all $T \in Z_+$,

$$\sum_{t=0}^{T} \|x_t\| \leq A_1 \cdot w_{\text{max}} T + A_2.$$  

**Proof sketch:** We have $\lim_{{t \to \infty}} H(t) = 0$ by Lemma 4.1. Using this result with the non-decreasing property of both $\tau_x$ and $H(\tau_y)$, we obtain that $\sum_{{b=0}}^{B-1} H(\gamma_0) = o(T)$ according to Assumption 3.1 for the dynamic batch length. This assumption further indicates that falsifying destabilizing controllers in Lines 5-8 results in the existence of a constant $M > 0$ such that the following inequality holds for all $T$:

$$\sum_{t=0}^{T} \|x_t\| \leq M + \gamma w_{\text{max}} \cdot (O(\sum_{{b=0}}^{B-1} H(\gamma_0)) + T).$$  

(2)

Thus, $\sum_{{b=0}}^{B-1} H(\gamma_0) = o(T)$ along with (2) proves both Theorems 4.2 and 4.3. More details about the proof are provided in Appendix A. $\square$

**Remark 4.4.** With a fixed batch length $\tau$ presented in Li et al. (2023), one cannot achieve asymptotic stability since $B \neq o(T)$. Instead, one obtains that $\lim_{{T \to \infty}} \frac{1}{T} \sum_{{t=0}}^{T} \|x_t\| = \gamma w_{\text{max}}$$ (1 + O(\frac{1}{T})) > \gamma w_{\text{max}}$. This necessitates our dynamic batch length strategy in Algorithm 1. It is also crucial to note that we have achieved asymptotic stability even when $\lim_{{t \to \infty}} H(t) = \infty$. Additionally, finite-gain stability can be achieved for every $\beta(\cdot)$ that satisfies $H(\cdot) < \infty$, which incorporates exponentially stable controllers.

4.2. Regret

In Algorithm 1, define $L := \{0 \leq b \leq B - 1, b \in Z_+: s_{b+1} \neq s_b\}$ and let $b^0, \ldots, b^{|L|}$ denote the batch where Line 22 is satisfied; i.e., $s_{b^l+1} \neq s_b$ for $l = 1, \ldots, |L|$. For convenience, we let $b^0 = 0, b^{|L|+1} = B - 1,$ and $s_B = s_{B-1}$. Also, define $V := \{0 \leq b \leq B - 1, b \in Z_+: s_b \neq s_0\}$. Before presenting the main theorems on regret, we provide the following useful lemma.
Lemma 4.5. In Algorithm 1, suppose that \( \beta(\tau_0) < 1 \) and let \( U \) denote the number of times that the Break statement is activated. Then, it holds that \( |L| = O(U) \) and \( |V| = O(U) \).

Proof. For every batch \( b = 0, \ldots, B - 1 \), we have

\[
\|x_{t+b}\| < (\alpha_b)^{s_b+1}\|x_0\| + \delta
\]

by Lines 11-20. If the Break statement is not activated, since we designed \( \delta \geq \frac{\tau_{\max}}{1-\beta(\tau_0)} \), it yields that

\[
\|x_{t+b}\| \leq \beta(\tau_b)\|x_t\| + \gamma U
\]

\[
\leq \beta(\tau_0)(\alpha_b)^{s_b+1}\|x_0\| + \beta(\tau_0)\delta + \gamma U
\]

\[
\leq \beta(\tau_0)(\alpha_b)^{s_b+1}\|x_0\| + \delta < (\alpha_b)^{s_b+1}\|x_0\| + \delta,
\]

which implies that \( s_{b+1} > s_b \) cannot occur when the Break statement is not activated. As a result, starting from \( s_0 = 0 \), the event \( s_{b+1} = s_b + 1 \) can occur at most \( U \) times. Also, since Line 14 avoids \( s_{b+1} > s_b + 1 \), the event \( s_{b+1} < s_b \) can occur at most \( U \) times as well, leading to \( |L| \leq 2U \).

Now, we observe the number of batches \( \tilde{b} > 0 \) needed to stabilize the state norm; i.e., \( \min\{\tilde{b} > 0 : s_{b+\tilde{b}} < s_b\} \).

When the Break statement is not activated, one can write

\[
\|x_{t+b}\| \leq \beta(\tau_b)\|x_t\| + \gamma U
\]

\[
\leq \beta(\tau_0)^\tilde{b}\|x_0\| + \gamma U\max\frac{1}{1-\beta(\tau_0)}\|x_0\| + \delta
\]

\[
\leq \beta(\tau_0)^\tilde{b}\|x_0\| + \delta
\]

\[
< (\alpha_b)^{s_b+1}\|x_0\| + \delta,
\]

where the last two inequalities are by the design of \( \delta \) and (3). It is desirable to find the minimum value of \( \tilde{b} > 0 \) that makes the right-hand side of (4) smaller than \( (\alpha_b)^{s_b}\|x_0\| + \delta \):

\[
(\beta(\tau_0)^\tilde{b})(\alpha_b)^{s_b+1}\|x_0\| + \delta < (\alpha_b)^{s_b}\|x_0\| + \delta
\]

\[
\implies \beta(\tau_0)^\tilde{b} \geq (\alpha_b)^{s_b}\|x_0\| + \delta.
\]

(5)

where the right-hand side of (5) can be upper-bounded by \( \alpha_b + \frac{\delta}{\|x_0\|} \) since \( \alpha_b > 1 \). Thus, if \( s_b \neq 0 \), we have

\[
\min\{\tilde{b} > 0 : s_{b+\tilde{b}} < s_b\} \leq \log(\frac{\alpha_b + \frac{\delta}{\|x_0\|}}{-\log(\beta(\tau_0))}),
\]

(6)

when the Break statement is not activated. In other words, starting from an arbitrary batch \( b \) where \( s_b > 0 \), within the number of batches on the right-hand side of (6), either the Break statement is activated or the value of \( s_b \) decreases. Thus, considering that \( |L| \leq 2U \), we have

\[
|V| \leq (2U - 1) \log(\frac{\alpha_b + \frac{\delta}{\|x_0\|}}{-\log(\beta(\tau_0))})
\]

which completes the proof. More proof details can be found in the Appendix (see Lemma B.3).
to factor in every potential exponential term to be multiplied with the initial learning rate \( \eta_0 \), which is on the order of \( T^{-2/3} \), thus inherently being a mitigating factor (see the term \( \frac{\alpha_0}{\eta_0} \sum_{t=0}^{B-1} E_{K_{b-1}}(u_i^0(k))^2 \) in Lemma B.5). Alleviating the exponential term guarantees safety even if the base of the term is large, for example, by drastic variation of transition dynamics and the policy along the state.

Remark 4.9. Our dynamic (non-decreasing) batch length design enables the reduction of the number of batches needed to recover the stable state after the Break statement is activated, unlike the fixed batch length case that requires a constant number of batches to do so. Thus, an adaptive learning rate with a dynamic batch length is effective when the number of destabilizing controllers \(|\mathcal{U}|\) is large.

Meanwhile, in Line 14 in Algorithm 1, we update the value of \( \alpha_b \) to an arbitrary number satisfying the range condition. In practice, we recommend setting \( \alpha_b \) to be small since one can update this value on the next batch whenever needed. If \( \alpha_b \) is too large, the term

\[
\mathbb{E}_{K_{B-1}} \sum_{b=0}^{B-1} \sum_{t=t_b}^{t_{b+1}-1} c_i(x^b_t(u^b_t), u^b_t) - c_i(x^b_t, u^b_t)
\]

may increase, leading to a deterioration in the overall performance of the algorithm.

A question arises as what happens if \(|\mathcal{U}|\) is not known in advance. With Algorithm 1, one can leverage \(|\mathcal{U}|+1 \leq N\) to upper-bound the regret in Theorem 4.7 and achieve \( \tilde{O}(T^{2/3}N^{2/3}) \) at best with determining \( \eta_0 \) and \( (\tau_b)_{b>0} \) as if there were only one stabilizing controller. It turns out that we can reduce the bound to \( \tilde{O}(T^{2/3}N^{1/3})(|\mathcal{U}|+1)^{1/2} \) by adaptively changing the value of \( \eta_0 \) as in Algorithm 2, where the first two statements imply that we increase the value of \( \mu_b \) if the Break statement in Algorithm 1 is activated and keep it unchanged otherwise.

**Theorem 4.10** (Regret bound with unknown \(|\mathcal{U}|\)). Consider Algorithm 2 with polynomial batches defined in Assumption 3.1 with \( \frac{1}{N^{1/2}}, z, \frac{1}{2}, \nu \), where the constants \( z, \nu \geq 0 \) satisfy \( \tau_0 > 0 \) and \( \frac{\beta(\tau_0)}{\tau_0} 2 < \frac{1}{2\nu^2} \). Also, let \( \eta_0 = O((\frac{1}{T^{2/3}N^{1/3}})) \) and \( y = \frac{1}{2} \). When \( \lim_{t \to \infty} H(t) < \infty \) and \( T \geq \max\{\frac{\exp(\Omega(|\mathcal{U}|))}{|\mathcal{U}|^{1/2}}, \frac{N^{1/2}}{|\mathcal{U}|^{1/2+1/3}}, N\} \), we have

\[
\text{Regret}_T = \tilde{O}(T^{2/3}N^{1/3}(|\mathcal{U}|+1)^{1/2}).
\]

**Proof sketch:** Define \( \eta_{0,r} := \eta_0 \sqrt{r} + 1 \). It turns out that for every \( r = 0, \ldots, |\mathcal{U}|, \tilde{O}(\frac{1}{\eta_{0,r}}) \) appears in the regret instead of the integrated term \( \tilde{O}(\frac{1}{\eta_0}) \) in Theorem 4.6. The constant \(|\mathcal{U}|+1\) is distributed among each \( \tilde{O}(\frac{1}{\eta_{0,r}}) \) term. Under the constraints given by the disintegration rule using Lemma 4.5 for each \( r \), one can establish an upper bound of \( \tilde{O}(\frac{|\mathcal{U}|+1}{\eta_0}) \) on the sum of \( \tilde{O}(\frac{1}{\eta_{0,r}}) \) terms over \( r = 0, \ldots, |\mathcal{U}| \) by attaining the coefficients of these terms with complementary slackness in Karush-Kuhn-Tucker (KKT) conditions. The details are available in Appendix C.

### 4.3. Applications: Switched systems

So far, we have used the best stabilizing controller \( i^* \in S \) for all time steps \( t = 0, \ldots, T \) as the baseline of regret. However, the proofs of the theorems stated above imply one can even use any set of controllers \( \{i^0, i^1, \ldots\} \subseteq S \) as a baseline, where the controller is switched from \( i^t \) to \( i^{t+1} \) whenever the cumulative weight \( W(.) \) resets. This motivates the application of our DBAR algorithm to scenarios such as the switched systems (Toussi et al., 2008; Zhao et al., 2022) for which the transition dynamics and the associated controller pool may undergo changes, as well as the balloning problem (Ghalme et al., 2021) where the controller pool may expand up to some finite set. We propose Algorithm 3, the switching version of DBAR, which resets the weight whenever the system is faced with a finite number of \( O(U) \) switches. Here, we consider the regret with switching costs where the unit cost \( d \geq 1 \) is additionally incurred when the controller is switched; i.e., \( d \sum_{t=1}^{T} I_{\{i_t \neq i_t-1\}} \) done in Altschuler & Talwar (2018) and Arora et al. (2019). Let \( x_{t}^i \) and \( u_{t}^i \) denote the state and action sequence generated by the set of best stabilizing controllers \( \{i^0, i^1, \ldots\} \subseteq S \).

The regret bound provided in the following theorem is similar to the lower bound presented in Dekel et al. (2014), except that there is an extra term \( (|\mathcal{U}|+1)^{1/3} \), reflecting the unbounded costs for the bandits.
Theorem 4.11 (Regret with switching costs bound with known $|\mathcal{U}|$). In Algorithm 3, consider $(\tau_b)_{b \geq 0}$ and $\eta_b$ as defined in Theorem 4.7. When $\lim_{t \to \infty} H(t) < \infty$ and $T \geq \max\{\frac{\exp(\Theta(|\mathcal{U}|))}{\eta N(|\mathcal{U}|+1)d}, \frac{1}{\eta N(|\mathcal{U}|+1)d} + N(|\mathcal{U}|+1)\}$, we have
\[
\text{Regret}_T + \mathbb{E}_{K_{b-1},0} \left[ d \sum_{b=1}^{B-1} \mathcal{I}(K_b \neq K_{b-1}) - d \sum_{t=1}^{T} \mathcal{I}(\hat{c}_t \neq c_{t-1}) \right] = \tilde{O}(T^{2/3}N^{1/3}|\mathcal{U}|+1)^{1/3}d^{1/3}).
\]

Proof. The proof is provided in Appendix D. $\square$

5. Numerical Experiments

To demonstrate the main results of this paper, we provide illustrative examples on both linear and nonlinear dynamics. While the simulations are on low-dimensional systems for illustration purposes, similar observations can be made for high-dimensional systems.

Example 1: Consider the following linear dynamical system for $t = 0, 1, \ldots$ with $x_t \in \mathbb{R}^2$ and $u_t \in \mathbb{R}^2$:
\[
x_{t+1} = \begin{bmatrix} 2 & 1.2 \\ 1 & 2.5 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0.3 \\ 0.4 & 0.9 \end{bmatrix} u_t + w_t, \tag{7}
\]
where $x_0 = [100, 200]'$ and $w_t = [\sin(\frac{t}{10}), \sin(\frac{t}{11})]'$.

We restrain the policy class to $u_t = K x_t = \begin{bmatrix} k_1 & k_2 \\ k_3 & k_4 \end{bmatrix} x_t$, where $K \in \mathbb{R}^{2 \times 2}$. We consider a controller pool $K' = \{K : k_1, k_3, k_4 \in \{-3, -2, -1\}, k_2 \in \{-1, 0, 1\}\}$ that has $|\mathcal{U}| = 53$ out of 81 candidate controllers. The goal is to keep the state near the origin, where the cost function is quadratic at each time, namely $c_t(x_t, u_t) = \|x_t\|^2$.

In Figure 2(a), falsifying destabilizing controllers effectively stabilizes the state norm as in Li et al. (2023). Figures 2(a) and 2(b) show that both dynamic batch length and adaptive learning rate, integral components of our algorithm DBAR, successfully lowers the regret and stabilizes the system, where approximately 2/3 of controllers in $K'$ are destabilizing the system. In this case, Figures 2(c) and 2(d) demonstrate that the two components of our algorithm mutually reinforce each other, as discussed in Remark 4.9.

Example 2: Consider the following nonlinear noise-injected ball-beam system (Hauser et al., 1992) with $B = 0.7143$:
\[
\ddot{x} = B(x^2 \dot{\theta}^2 - 9.81 \sin \theta) + 3w, \quad \dot{\theta} = u, \tag{8}
\]
where $x$ is the ball position, $\theta$ is the beam angle, $u$ is the action, and $w = \sin(\frac{t}{10})$. We can now adopt broader notion of stabilizing controllers and choose the policy class to be the nested saturating control (Teel, 1992), without making exponential assumptions. In Figure 3(b), we observe that DBAR effectively stabilizes the explosion of the nonlinear system, even when there exist few stabilizing controllers.

Figure 2. The stability and the regret in the linear system under sinusoidal noise. Ablation study of the algorithm is presented.

In an online bandit control problem, an agent makes decisions with the bandit feedback information, while suffering from adversarial disturbances. To address such challenges, this paper develops a novel Exp3-type algorithm with theoretical guarantees. The proposed algorithm uses a dynamic batch length to achieve asymptotic stability of the system without requiring the exponential assumptions on stabilizing controllers in the pool. Our adaptive learning rate scheme observes the stability of state norm to overcome the inherent exponential term in the regret, thereby greatly improving overall regret. Future directions include extending these results to problems with explicit safety constraints.
Acknowledgements

This work was supported by grants from ARO, ONR, AFOSR, NSF, and the UC Noyce Initiative.

References


A. Stability Proof

Let \( b_1, \ldots, b_U \) denote the next batch after the Break statement is activated; i.e., \( \| x_{t_{bu}} \| > \beta(t_{bu} - t_{bu-1}) \| x_{t_{bu-1}} \| + \gamma w_{\text{max}} \) for every \( u = 1, \ldots, U \). For future use, let \( b_0 = 0 \) and \( b_{U+1} = B \). Accordingly, \( t_{bu} = t_0 = 0 \) and \( t_{b_{U+1}} = t_B = T + 1 \).

**Lemma A.1** (Restatement of Lemma 4.1). Define \( H(t) := \sum_{i=0}^{t-1} \beta(i) \). Then, we have

\[
\lim_{t \to \infty} \frac{H(t)}{t} = 0.
\]

**Proof.** The proof is given in Lemma 4.1. \( \square \)

**Lemma A.2.** For \( 0 \leq j \leq k \), we have

\[
\frac{H(\tau_k)}{H(\tau_j)} \leq \frac{\tau_k}{\tau_j}.
\]

**Proof.** For \( 0 \leq j \leq k \),

\[
\frac{H(\tau_k)}{H(\tau_j)} \leq \frac{H(\tau_j) + \sum_{i=0}^{\tau_k-1} \beta(i)}{H(\tau_j)} \leq 1 + \frac{\tau_k - \tau_j \beta(\tau_j)}{\tau_j \beta(\tau_j)} = \frac{\tau_k}{\tau_j},
\]

where the last inequality is due to the non-increasing property of \( \beta(\cdot) \). The equality holds when \( \beta(0) = \cdots = \beta(\tau_k - 1) \). \( \square \)

**Lemma A.3** (Sum of state norms in a single batch). In Algorithm 1, for each batch \( b = 0, 1, \ldots, B - 1 \), the following inequality holds:

\[
\sum_{t=t_{bu}}^{t_{bu+1}-1} \| x_t \| \leq H(t_b) \| x_{t_{bu}} \| + \gamma w_{\text{max}} (t_b - 1)
\]

**Proof.** For \( t = t_b \), we have \( \| x_t \| \leq \beta(0) \| x_{t_{bu}} \| \) since \( \beta(0) = 1 \). For \( t_b < t \leq t_{bu+1} - 1 \), we have

\[
\| x_t \| \leq \beta(t - t_b) \| x_{t_{bu}} \| + \gamma w_{\text{max}}.
\]

Summing up all inequalities gives

\[
\sum_{t=t_{bu}}^{t_{bu+1}-1} \| x_t \| \leq H(t_{bu+1} - t_b) \| x_{t_{bu}} \| + \gamma w_{\text{max}} (t_{bu+1} - t_b - 1).
\]

Since Line 5 of Algorithm 1 is not satisfied, \( \tau_b = t_{bu+1} - t_b \). This completes the proof. \( \square \)

**Lemma A.4** (Weighted sum of state norms between the two consecutive Break statements). In Algorithm 1, suppose that \( \frac{\tau_b}{\tau_0} \beta(\tau_0) < 1 \). For every next batch index after the Break statement \( u = 0, \ldots, U \), the following inequality holds:

\[
\sum_{b=b_u}^{b_{u+1}-1} H(\tau_b) \| x_{t_{bu}} \| \leq \frac{1}{1 - \frac{\tau_{bu+1}}{\tau_b} \beta(\tau_b)} H(\tau_{bu}) \| x_{t_{bu}} \| + \frac{\gamma w_{\text{max}}}{1 - \beta(\tau_{bu+1})} \sum_{b=b_u}^{b_{u+1}-1} H(\tau_b).
\]

**Proof.** Since we designed \( (\tau_b)_{b\geq0} \) to have a non-decreasing \( \tau_b \) and non-increasing \( \frac{\tau_{b+1}}{\tau_b} \beta(\tau_b) \), notice that we have \( \beta(\tau_b) \leq \frac{\tau_{b+1}}{\tau_b} \beta(\tau_{b-1}) \leq \frac{\tau_{b-1}}{\tau_{b-2}} \beta(\tau_{b-2}) \leq \frac{\tau_0}{\tau_0} \beta(\tau_0) < 1 \) for every \( b \geq 1 \) since \( \beta(\cdot) \) is non-increasing.

If \( b_{u+1} = b_u + 1 \), the inequality clearly holds since \( \frac{1}{1 - \frac{\tau_{bu+1}}{\tau_{bu}} \beta(\tau_{bu})} > 0 \). Otherwise, consider the following inequality for \( b_u < b \leq b_{u+1} - 1 \):

\[
H(\tau_b) \| x_{t_{bu}} \| \leq H(\tau_b) \beta(\tau_{b-1}) \| x_{t_{bu-1}} \| + H(\tau_b) \gamma w_{\text{max}}
\]

\[
= \frac{H(\tau_b)}{H(\tau_{b-1})} \beta(\tau_{b-1}) H(\tau_{b-1}) \| x_{t_{bu-1}} \| + H(\tau_b) \gamma w_{\text{max}},
\]

\[
\leq \frac{H(\tau_b)}{H(\tau_{b-1})} H(\tau_{b-1}) \| x_{t_{bu-1}} \| + H(\tau_b) \gamma w_{\text{max}},
\]

\[
= \frac{H(\tau_b)}{H(\tau_{b-1})} H(\tau_{b-1}) \| x_{t_{bu-1}} \| + H(\tau_b) \gamma w_{\text{max}}.
\]

12
where the inequality holds since Line 5 of Algorithm 1 is not satisfied. Recursively applying this inequality, one arrives at

\[
H(\tau_b) \|x_{t_b}\| \leq \prod_{a=b_u}^{a=b} \left[ \frac{H(\tau_{a+1})}{H(\tau_a)} \beta(\tau_a) \right] \cdot H(\tau_{b_u}) \|x_{t_{b_u}}\| + \sum_{b'=b_u+1}^{b-1} \prod_{a=b'}^{a=b} \beta(\tau_a) \]

\[
\leq \prod_{a=b_u}^{a=b} \left[ \frac{H(\tau_{a+1})}{H(\tau_a)} \beta(\tau_a) \right] \cdot H(\tau_{b_u}) \|x_{t_{b_u}}\| + H(\tau_b) \gamma w_{\max} (1 + \sum_{b'=b_u+1}^{b-1} \beta(\tau_{a+1}))^{b-b'} \]

\[
\leq \prod_{a=b_u}^{a=b} \left[ \frac{\tau_{a+1}}{\tau_a} \beta(\tau_a) \right] \cdot H(\tau_{b_u}) \|x_{t_{b_u}}\| + H(\tau_b) \gamma w_{\max} \frac{1}{1-\beta(\tau_{b_u+1})} \]

where the second inequality comes from the non-increasing property of \(\beta(\cdot)\), the third inequality is by \(\beta(\tau_{b_u+1}) < 1\), the fourth inequality is due to Lemma A.2, and the last inequality comes from the non-increasing property of \(\frac{\tau_{a+1}}{\tau_{a}} \beta(\tau_{a})\). Since \(\frac{\tau_{b_u+1}}{\tau_{b_u}} \beta(\tau_{b_u}) < 1\), summing up the above inequalities for \(b_u < b \leq b_{u+1} - 1\) completes the proof.

Lemma A.5 (Next state norm after the Break statement). Define \(M_1 := L_f(1 + L_\pi)\gamma w_{\max} + L_f(\pi_{0,\max} + w_{\max}) + f_0\).
Then, for every \(u = 1, \ldots, U\), we have

\[
\|x_{t_{b_u}}\| \leq L_f(1 + L_\pi)\beta(0)\|x_{t_{b_u-1}}\| + M_1.
\]

Proof. Suppose we picked a controller \(\pi_t\) at time step \(t\). Then, by Assumption 2.6, we have

\[
\|u_t\| = \|\pi_t(x_t) - \pi_t(0)\| + \|\pi_t(0)\| \leq \|\pi_t(x_t) - \pi_t(0)\| + \|\pi_t(0)\| \leq L_\pi \|x_t\| + \pi_{0,\max}.
\]

Combining the above inequality with Assumption 2.1, one can write

\[
\|x_{t+1}\| = \|f(x_t, u_t, w_t) - f(0, 0, 0) + f(0, 0, 0)\|
\]

\[
\leq \|f(x_t, u_t, w_t) - f(0, 0, 0)\| + \|f(0, 0, 0)\| \leq L_f(\|x_t\| + \|u_t\| + \|w_t\|) + f_0
\]

\[
\leq L_f(\|x_t\| + L_\pi \|x_t\| + \pi_{0,\max} + w_{\max}) + f_0
\]

Thus, for every \(u = 1, \ldots, U\), we obtain that

\[
\|x_{t_{b_u}}\| \leq L_f(1 + L_\pi)\|x_{t_{b_u-1}}\| + L_f(\pi_{0,\max} + w_{\max}) + f_0
\]

\[
\leq L_f(1 + L_\pi)\beta(t_{b_u} - t_{b_u-1} - 1)\|x_{t_{b_u-1}}\| + \gamma w_{\max} + L_f(\pi_{0,\max} + w_{\max}) + f_0
\]

\[
= L_f(1 + L_\pi)\beta(t_{b_u} - t_{b_u-1} - 1)\|x_{t_{b_u-1}}\| + M_1
\]

\[
\leq L_f(1 + L_\pi)\beta(0)\|x_{t_{b_u-1}}\| + M_1,
\]

where the second inequality holds since Line 5 of Algorithm 1 is not satisfied during \(t_{b_u-1} \leq t \leq t_{b_u} - 1\) and the equality holds for the last inequality when \(t_{b_u} = t_{b_u-1} + 1\). This completes the proof.

Lemma A.6 (Weighted sum of state norms along the Break statements). In Algorithm 1, suppose that \(\frac{\tau_{0}}{\tau_0} \beta(\tau_0) < 1\). Define \(M_2 := L_f(1 + L_\pi)\beta(0)\frac{\pi_{0,\max}}{1-\beta(\tau_0)} + M_1\). Then, there exists a constant \(C \geq 1\) such that

\[
\sum_{u=0}^{U} H(\tau_{b_u}) \|x_{t_{b_u}}\| \leq \frac{[L_f(1 + L_\pi)\beta(0)C]^U - 1}{L_f(1 + L_\pi)\beta(0)C - 1} H(\tau_0)\|x_0\| + \frac{([L_f(1 + L_\pi)\beta(0)C]U - 1)M_2}{[L_f(1 + L_\pi)\beta(0)C - 1]^2} H(\tau_{b_0})
\]

13
Proof. Since we designed $\frac{\tau_{b+1}}{\tau_b}$ to converge, there exists $R > 0$ such that $\frac{\tau_{b+1}}{\tau_b} \leq R$ for all $b \geq 0$. Moreover, since $\lim_{b \to \infty} \frac{\tau_{b+1}}{\tau_b} = 1$ and $\beta(\tau_0) < 1$, there exists $b^* > 0$ such that

$$b \geq b^* \implies \frac{\tau_{b+1}}{\tau_b} < \frac{1}{\beta(\tau_0)}. \quad (12)$$

Accordingly, for any two batches $b' > b \geq 0$, we have

$$\frac{\tau_{b'}}{\tau_b} \left[ \frac{\beta(\tau_b)}{\beta(\tau_0)} \right]^{b'-b-1} \leq \frac{\beta(\tau_b)}{\beta(\tau_0)} \frac{\tau_{b+1}}{\tau_b} \leq \frac{R^{b'}}{\beta(\tau_0)}, \quad (13)$$

considering that $b = 0$ and $b' = b^*$ yields the largest possible upper bound due to (12). Now, define $C := \frac{R^{b^*}}{\beta(\tau_0)}$. Notice that we have $C \geq 1$ since the left-hand side of (13) is greater than equal to 1 when $b' = b + 1$. Then, for every $u = 1, \ldots, U$, one can write

$$H(\tau_{b_u}) \|x_{t_{b_u}}\| \leq L_f(1 + L_\pi)\beta(0) \frac{H(\tau_{b_u})}{H(\tau_{b_u-1})} H(\tau_{b_u-1}) \|x_{t_{b_u-1}}\| + H(\tau_{b_u}) M_1$$

$$\leq L_f(1 + L_\pi)\beta(0) \frac{H(\tau_{b_u})}{H(\tau_{b_u-1})} \prod_{a=b_u-2}^{b_u-1} \left[ \frac{H(\tau_{a+1})}{H(\tau_a)} \beta(\tau_a) \right] \cdot H(\tau_{b_u-1}) \|x_{t_{b_u-1}}\|$$

$$+ L_f(1 + L_\pi)\beta(0) \frac{\gamma w_{\max}}{1 - \beta(\tau_{b_u-1})} + H(\tau_{b_u}) M_1$$

$$\leq L_f(1 + L_\pi)\beta(0) \frac{H(\tau_{b_u})}{H(\tau_{b_u-1})} \beta(\tau_{b_u-1}) \|x_{t_{b_u-1}}\| + H(\tau_{b_u}) M_2$$

$$\leq L_f(1 + L_\pi)\beta(0) \frac{\tau_{b_u}}{\tau_{b_u-1}} \beta(\tau_{b_u-1}) \|x_{t_{b_u-1}}\| + H(\tau_{b_u}) M_2$$

$$\leq L_f(1 + L_\pi)\beta(0) C \cdot H(\tau_{b_u-1}) \|x_{t_{b_u-1}}\| + H(\tau_{b_u}) M_2$$

where the first inequality is due to Lemma A.5, the second inequality is by (10) in Lemma A.4, the fourth inequality is due to Lemma A.2, and the last inequality is by (13). Recursively applying this inequality, one arrives at

$$H(\tau_{b_u}) \|x_{t_{b_u}}\| \leq \left[ L_f(1 + L_\pi)\beta(0) C \right] u H(\tau_0) \|x_0\| + M_2 \sum_{i=1}^{u} \left[ L_f(1 + L_\pi)\beta(0) C \right]^{u-i} H(\tau_{b_u})$$

$$\leq \left[ L_f(1 + L_\pi)\beta(0) C \right] u H(\tau_0) \|x_0\| + M_2 H(\tau_{b_u}) \cdot \left[ \frac{L_f(1 + L_\pi)\beta(0) C}{L_f(1 + L_\pi)\beta(0) C - 1} \right]$$

$$< \left[ L_f(1 + L_\pi)\beta(0) C \right] u \left[ H(\tau_0) \|x_0\| + \frac{M_2 H(\tau_{b_u})}{L_f(1 + L_\pi)\beta(0) C - 1} \right],$$

where the second inequality comes from the non-decreasing property of $H(\cdot)$ and the equality holds when $H(\tau_{b_u}) = \cdots = H(\tau_{b_u})$. Notice that for $b' > b \geq 0$, the case $H(\tau_{b_u}) = H(\tau_{b_u})$ arises when $\tau_{b_u} = \tau_0$ or $\beta(\tau_{b_u}+1) = \cdots = \beta(\tau_{b_u}) = 0$. Since $L_f(1 + L_\pi)\beta(0) C > 1$, summing up the above inequality for $u = 1, \ldots, U$ completes the proof. \hfill \□

**Lemma A.7 (Sum of state norms).** In Algorithm 1, suppose that $\frac{\tau_{b_u}}{\tau_0} \beta(\tau_0) < 1$. Then, we have

$$\sum_{t=0}^{T} \|x_t\| \leq O([L_f(1 + L_\pi)\beta(0) C] U (\|x_0\| + H(\tau_{b_u}))) + \gamma w_{\max} \cdot (O(\sum_{b=0}^{B-1} H(\tau_b)) + T)$$

**Proof.** Applying Lemma A.3, A.4, and A.6 in turn, we have

$$\sum_{t=0}^{T} \|x_t\| \leq \sum_{u=0}^{U} \sum_{b=b_u}^{b_u+1-1} \sum_{t=t_{b_u}} \|x_t\|$$

$$\leq \sum_{u=0}^{U} \sum_{b=b_u}^{b_u+1-1} \left[ H(\tau_b) \|x_{t_{b_u}}\| + \gamma w_{\max}(\tau_b - 1) \right]$$
where the second inequality leverages the non-decreasing property of both \( \tau \).

where the second inequality is due to the non-decreasing property of \( \tau \).

where the first equality is due to Lemma A.1 and

Theorem A.8 (Restatement of Theorem 4.2, Asymptotic stability). In Algorithm 1, suppose that

\[
\frac{\tau_0}{\tau_0} \beta(\tau_0) < 1.
\]

Then, it holds that

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \|x_t\| \leq \gamma u_{\text{max}}.
\]

Proof. We mainly use Lemma A.1 to prove the asymptotic stability. First, we have

\[
H(\tau_{B-1}) \leq H(\tau_{B-1}) = o(\tau_{B-1}) = o(T),
\]

where the first equality is due to Lemma A.1 and \( \tau_{B-1} = T \) when there is only one batch over the entire horizon. Now, consider the following relationship between the number of batch \( B \) and the time horizon \( T \):

\[
\sum_{b=0}^{B-1} \tau_b \geq T \geq \sum_{b=0}^{B-1} \tau_b + U,
\]

where the second inequality is due to the non-decreasing property of \( \tau \). Now, if \( \sum_{b=0}^{B-1} H(\tau_b) < \infty \), clearly \( \sum_{b=0}^{B-1} H(\tau_b) = o(T) \). Otherwise, define \( H(\tau_B) = H(\tau_{B-1}) \). Then, we have

\[
\lim_{T \to \infty} \frac{\sum_{b=0}^{B-1} H(\tau_b)}{T} \leq \lim_{B \to \infty} \frac{\sum_{b=0}^{B-1} H(\tau_b)}{\tau_B + U} \leq \lim_{B \to \infty} \frac{\int_0^B H(\tau_b) db}{\tau_0 + \int_0^{B-U-1} \tau_b db + U} = \lim_{B \to \infty} \frac{H(\tau_{B-1})}{\tau_{B-U-1}} = \lim_{B \to \infty} \frac{H(\tau_{B-1})}{\tau_{B-1}} \prod_{b=B-U-1}^{B-2} \frac{\tau_b+1}{\tau_b} = 0 \cdot U = 0.
\]

where the second inequality leverages the non-decreasing property of both \( \tau_b \) and \( H(\tau_b) \), the remaining equalities leverages L’Hôpital’s rule, Lemma A.1, and \( \lim_{b \to \infty} \frac{\tau_{b+1}}{\tau_b} = 1 \). Thus, with Lemma A.7, we have

\[
\sum_{t=0}^{T} \|x_t\| \leq O([L_f(1 + L_{\pi})\beta(0)C]^U(\|x_0\| + o(T)) + \gamma u_{\text{max}} \cdot (T + o(T)).
\]

This completes the proof.

\[
\]
Algorithm 1, and the controller for batch where the first equality is because

\[ \tau \]

Theorem A.10 (Restatement of Theorem 4.3, Finite-gain stability). In Algorithm 1, suppose that \( \frac{\tau}{\tau_0} \beta(\tau_0) \) \( < \) 1. Assume that \( \lim_{t \to \infty} H(t) \) \( < \) \( \infty \). Then, Algorithm 1 achieves finite-gain \( L_1 \) stability; i.e., there exist constants \( A_1, A_2 > 0 \) such that for all \( T \) \( \in \mathbb{Z}_+ \),

\[
\sum_{t=0}^{T} \|x_t\| \leq A_1 \cdot w_{\max} T + A_2.
\]

Proof. Since \( \lim_{t \to \infty} H(t) \) \( < \) \( \infty \), there exists a constant \( q_1 \) that upper-bounds \( H(t) \); i.e., \( H(t) \) \( \leq \) \( q_1 \) for all \( t \) \( \geq \) 0. Likewise, by Lemma A.9, there exists a constant \( q_2 \) that upper-bounds \( \frac{B}{T} \). Thus, with Lemma A.7, one can write

\[
\sum_{t=0}^{T} \|x_t\| \leq O([L_f(1 + L_\tau)\beta(0)C]^U ([|x_0| + q_1]) + \gamma w_{\max} \cdot (O(q_1) + T)
\]

\[
= O([L_f(1 + L_\tau)\beta(0)C]^U ([|x_0| + q_1]) + \gamma (1 + \frac{B}{T} O(q_1)) \cdot w_{\max} T
\]

\[
\leq O([L_f(1 + L_\tau)\beta(0)C]^U ([|x_0| + q_1]) + \gamma (1 + O(q_1 q_2)) \cdot w_{\max} T.
\]

This completes the proof. \( \square \)

B. Regret Proof for Algorithm 1

Lemma B.1. In Algorithm 1, we have

\[
E_{K_{B-1:0}}[w_b(K_b)] = E_{K_{B-1:0}}[E_{k \sim p_b}[w_b'(k)]],
\]

Proof. Given \( K_{b-1}, \ldots, K_0 \), we have

\[
E_{k \sim p_b}[w_b'(k)] = \sum_{k \in P_b} p_b(k) \frac{w_b(K_b)}{p_b(k)} I(K_b\sim k) = w_b(K_b),
\]

which implies that \( w_b'(k) \) sampled from \( p_b \) is an unbiased estimator of \( w_b(K_b) \).

Thus, for all \( b = 0, 1, \ldots, B - 1 \), one can write

\[
E_{K_{B-1:0}}[w_b(K_b)] = E_{K_{b-1}}[w_b(K_b)] = E_{K_{b-1:0}} E_{K_b}[w_b(K_b) | K_{b-1:0}]
\]

\[
= E_{K_{b-1:0}} E_{K_b}[E_{k \sim p_b}[w_b'(k)] | K_{b-1:0}]
\]

\[
= E_{K_{b-1}}[E_{k \sim p_b}[w_b'(k)]] = E_{K_{B-1:0}}[E_{k \sim p_b}[w_b'(k)]]
\]

where the first equality is because \( K_{B-1}, \ldots, K_{b+1} \) does not affect the value of \( w_b(K_b) \) and the remaining equalities are by law of total expectation and (17). \( \square \)

Now, we let \( w^K_b(i) \) denote the cost incurred at batch \( b \) if one selects the controllers for batch \( 0, \ldots, b-1 \) according to Algorithm 1, and the controller for batch \( b \) to be \( i \).
Lemma B.2. In Algorithm 1, for any $i \in \mathcal{P}_b$, we have
\[ \mathbb{E}_{K_{b-1:0}}[w_b'(i)] = \mathbb{E}_{K_{b-1:0}}[w_b^K(i)] \]
and for some controller $i^b \in \mathcal{P}_b$, we have
\[ \mathbb{E}_{K_{b-1:0}} \left[ \frac{\eta_0}{2} \frac{(w_b(K_b))^2}{p_b(K_b)} \right] \leq \frac{\eta_0 N}{2} \mathbb{E}_{K_{b-1:0}}(w_b^K(i^b))^2. \]

Proof. For all $b = 0, 1, \ldots, B - 1$ and for all $i \in \mathcal{P}_b$, we have
\[ \mathbb{E}_{K_{b-1:0}}[w_b'(i)] = \mathbb{E}_{K_{b-0}}[w_b'(i)] = \mathbb{E}_{K_{b-1:0}}[\mathbb{E}_{K_b}[w_b'(i) \mid K_{b-1:0}]] \]
\[ = \mathbb{E}_{K_{b-1:0}} \left[ \sum_{K_b \in \mathcal{P}_b} p_b(K_b) \frac{w_b(K_b)}{p_b(i)} I(K_b = i) \right] \]
\[ = \mathbb{E}_{K_{b-1:0}}[w_b^K(i)] = \mathbb{E}_{K_{b-1:0}}[w_b^K(i)] \]
where the first equality is because $K_{b-1}, \ldots, K_b$ does not affect the value of $w_b'(i)$ and the last equality is because $K_{b-1}, \ldots, K_b$ does not affect the value of $w_b^K(i)$. Next, we can also obtain that
\[ \mathbb{E}_{K_{b-1:0}} \left[ \frac{\eta_0}{2} \frac{(w_b(K_b))^2}{p_b(K_b)} \right] = \mathbb{E}_{K_{b-0}} \left[ \frac{\eta_0}{2} \frac{(w_b(K_b))^2}{p_b(K_b)} \right] = \mathbb{E}_{K_{b-1:0}} \mathbb{E}_{K_b} \left[ \frac{\eta_0}{2} \frac{(w_b(K_b))^2}{p_b(i)} \mid K_{b-1:0} \right] \]
\[ = \mathbb{E}_{K_{b-1:0}} \sum_{K_b \in \mathcal{P}_b} \left[ \frac{\eta_0}{2} p_b(K_b) \frac{(w_b(K_b))^2}{p_b(K_b)} \right] = \mathbb{E}_{K_{b-1:0}} \sum_{K_b \in \mathcal{P}_b} \left[ \frac{\eta_0}{2} (w_b(K_b))^2 \right] \]
\[ \leq \frac{\eta_0 N}{2} \mathbb{E}_{K_{b-1:0}}(w_b^K(i^b))^2, \]
for the controller $i^b = \arg \max_{i \in \mathcal{P}_b} (w_b^K(i))^2$. This completes the proof. \hfill \qed

In Algorithm 1, define $\mathcal{L} := \{0 \leq b \leq B - 1, b \in \mathbb{Z}_+ : s_{b+1} \neq s_b\}$ and let $b^1, \ldots, b^{|\mathcal{L}|}$ denote the batch where Line 22 of Algorithm 1 is satisfied; i.e., $s_{b+l+1} \neq s_b$ for $l = 1, \ldots, |\mathcal{L}|$. For convenience, we let $b^0 = 0$, $b^{|\mathcal{L}|+1} = B - 1$, and $s_B = s_{B-1}$. Also, define $\mathcal{V}' := \{0 \leq b \leq B - 1, b \in \mathbb{Z}_+ : s_b \neq 0\}$.

Lemma B.3 (Restatement of Lemma 4.5). In Algorithm 1, suppose that $\beta(\tau_0) < 1$ and let $U$ denote the number of times that the Break statement is activated. Then, it holds that $|\mathcal{L}| = \tilde{O}(U)$ and $|\mathcal{V}| = O(U)$.

Proof. For every batch $b = 0, \ldots, B - 1$, we have
\[ \|x_{b+1}\| < (\alpha_b)^{s_{b+1}}\|x_b\| + \delta \tag{18} \]
by Lines 11-20. If the Break statement is not activated, since we designed $\delta \geq \frac{\gamma w_{\max}}{1 - \beta(\tau_0)}$, it yields that
\[ \|x_{b+1}\| \leq \beta(\tau_b)\|x_{b+1}\| + \gamma w_{\max} \]
\[ \leq \beta(\tau_b)(\alpha_b)^{s_{b+1}}\|x_0\| + \beta(\tau_b)\|x_{b+1}\| + \gamma w_{\max} \]
\[ \leq (\alpha_b)(\alpha_b)^{s_{b+1}}\|x_0\| + \delta < (\alpha_b)^{s_{b+1}}\|x_0\| + \delta, \]
where the second and the last inequalities are due to $\beta(\tau_b) \leq \beta(\tau_0) < 1$ and the third inequality is by the formulation of $\delta$. Then, $s_{b+1} > s_b$ cannot occur when the Break statement is not activated. Also, Line 14 avoids $s_{b+1} > s_b + 1$. As a result, starting from $s_0 = 0$, the event $s_{b+1} = s_b + 1$ can occur at most $U$ times. Accordingly, the event $s_{b+1} < s_b$ also can occur at most $U$ times, leading to $|\mathcal{L}| \leq 2U$.

Now, we observe the number of batches $\tilde{b}$ needed to stabilize the state norm; i.e., $\min \{ \tilde{b} > 0 : s_{b+\tilde{b}} < s_b \}$ when the Break statement is not activated. Starting from batch $b$ and the corresponding $s_b$, provided that the Break statement is not activated, one can write
\[ \|x_{b+\tilde{b}}\| \leq \beta(\tau_{b+\tilde{b}-1})\|x_{b+\tilde{b}-1}\| + \gamma w_{\max} \]

17
while a mix loss for $b$ when the Break statement is not activated. In other words, starting from a batch $\alpha$ upper-bounded as follows:

$$ (\text{cumulative mix loss}) B \text{ interval between two consecutive batches in }$$

More specifically, consider two sets of batches: $s \in U$ for $l = 0, \ldots, |\mathcal{L}|$, the cumulative mix loss is upper-bounded as follows:

$$ \sum_{b=0}^{B-1} \frac{1}{\eta_0} \log (E_{K_{b-1}} \exp (-\eta_0 w'_b(k))) \leq \frac{\tilde{O}(U+1)}{\eta_0} + \sum_{l=0}^{\mathcal{L}} \sum_{b=b'}^{b'+1-1} w'_{b}^{k}(i') \frac{1}{(\alpha_b)^2 s_b}$$

Proof. Given $l = 0, \ldots, |\mathcal{L}|$, we can analyze a single mix loss for $b = b' + 1, \ldots, b'+1 - 1$ as follows:

$$ -\frac{1}{\eta_0} \log (E_{K_{b-1}} \exp (-\eta_0 w'_b(k))) = -\frac{1}{\eta_0} \log (\sum_{k \in F_b} p_b(k) \exp (-\eta_0 w'_b(k))) $$

$$ = -\frac{1}{\eta_0} \log (\sum_{k \in F_b} \exp (-\eta_0 W_b(k)) \exp (-\eta_0 w'_b(k)))) $$

$$ = -\frac{1}{\eta_0} \log (\sum_{k \in F_b} \exp (-\eta_0 W_b(k)) \sum_{i \in F_b} \exp (-\eta_0 W_b(i))$$

while a mix loss for $b = b'$ is as follows:

$$ -\frac{1}{\eta_0} \log (E_{K_{b-1}} \exp (-\eta_0 w'_b(k))) = -\frac{1}{\eta_0} \log (\sum_{k \in F_{b'}} p_{b'}(k) \exp (-\eta_0 w'_{b'}(k))) $$

when the Break statement is not activated. In other words, starting from a batch $b$ where $s_b > 0$, within the number of batches on the right-hand side of (21), either the Break statement is activated or the value of $s_b$ decreases.

More specifically, consider two sets of batches: $B_1 = \{0 \leq b \leq B - 1, b \in Z_+ : \text{the Break statement activated}\}$ and $B_2 = \{0 \leq b \leq B - 1, b \in Z_+ : s_{b+1} < s_b\}$. Let $B = B_1 \cup B_2$ be the set ordered by batch numbers. Then, the batch interval between two consecutive batches in $B$ is upper-bounded by (21). Thus, considering that $|\mathcal{L}| \leq 2U$, we have

$$ |\mathcal{L}| \leq (2U - 1) \left( \frac{\log (\alpha_b + \frac{\delta}{\|x_0\|})}{-\log \beta(\tau_0)} \right), $$

which completes the proof. □

**Lemma B.4** (cumulative mix loss). In Algorithm 1, for any controller $i^l \in U^c$ for $l = 0, \ldots, |\mathcal{L}|$, the cumulative mix loss is upper-bounded as follows:
where the last equality only holds when \( b^{l+1} > b^l + 1 \). Now, notice that the batches \( b = b^l, \ldots , b^{l+1} - 1 \) share the same learning rate; i.e., \( \eta_{b^l} = \cdots = \eta_{b^{l+1} - 1} \) since the same \( \alpha_b \) yields the same \( \alpha_b \), and thus the same \( \eta_b \). Thus, in the case where \( b^{l+1} > b^l + 1 \), we have

\[
\sum_{b=b^l}^{b^{l+1}-1} - \frac{1}{\eta_0} \log(\mathbb{E}_{k \sim p_b} \exp(-\eta_b w_b^l(k))) \leq \frac{\log N}{\eta_0} - \frac{1}{\eta_0} \log(\prod_{b=b^l}^{b^{l+1}-1} \sum_{k \in \mathcal{P}_b} \exp(-\eta_b W_b^l(k))) - \frac{1}{\eta_0} \log(\sum_{k \in \mathcal{P}_{b+1}} \exp(-\eta_b W_{b+1}^l(k)))
\]

\[
\leq \frac{\log N}{\eta_0} - \frac{1}{\eta_0} \log(\sum_{k \in \mathcal{P}_{b+1}} \exp(-\eta_b W_{b+1}^l(k))),
\]

(25)

where the first inequality is by (22) and (24) and the second inequality comes from \( \mathcal{P}_b \subseteq \mathcal{P}_{b-1} \). Considering both cases (23) and (25), for any controller \( i^0, \ldots , i^{|\mathcal{L}|} \in \mathcal{U}^c \), one can write

\[
\sum_{k=0}^{B-1} - \frac{1}{\eta_0} \log(\mathbb{E}_{k \sim p_b} \exp(-\eta_b w_b^l(k))) \leq \left( |\mathcal{L}| + 1 \right) \frac{\log N}{\eta_0} - \frac{1}{\eta_0} \log(\sum_{l=0}^{|\mathcal{L}|} \exp(-\eta_b \sum_{b=b^l}^{b^{l+1}-1} w_b^l(i^l)))
\]

\[
= \frac{\tilde{O}(U+1)}{\eta_0} + \sum_{l=0}^{|\mathcal{L}|} \frac{\sum_{b=b^l}^{b^{l+1}-1} w_b^l(i^l)}{(\alpha_b)^2} \eta_0,
\]

(26)

where the first inequality considers \( W_{b+1}^l(k) = \sum_{b=b^l}^{b^{l+1}-1} w_b^l(k) \) in (25), the second inequality is because any controller \( i^l \) is an element of \( \mathcal{P}_{b+1} \), and the last equality is because the definition of \( \eta_{b} = \eta_{0}/(\alpha_{b})^{2x_b} \) and \( |\mathcal{L}| = O(U) \) by Lemma B.3. Finally, by Lemma B.2, taking the expectation of (26) with respect to \( K_{B-1,0} \) completes the proof. \( \square \)

Now, we consider the cumulative mixability gap.

**Lemma B.5** (Cumulative mixability gap). In Algorithm 1, there exists a set of controllers \( i^b \in \mathcal{P}_b \) for \( b = 0, \ldots , B-1 \) such that the cumulative mixability gap is upper-bounded as follows:

\[
\mathbb{E}_{K_{b-1,0}} \frac{B-1}{b=0} \mathbb{E}_{k \sim p_b} [w_b^l(k)] + \frac{1}{\eta_0} \log(\mathbb{E}_{k \sim p_b} \exp(-\eta_b w_b^l(k))) \leq \frac{O(U)}{2\eta_0} + \frac{\eta_0 N}{2} \sum_{b=0}^{B-1} \mathbb{E}_{K_{b-1,0}} \mathbb{E}_{K_{b-1,0}} [w_b^K(i^b)]
\]

**Proof.** Given the set \( \mathcal{V} \), we can analyze a single mixability gap for \( b \in \mathcal{V} \) and \( b \not\in \mathcal{V} \), respectively. Since \( s_b = 0 \) for \( b \not\in \mathcal{V} \), given \( K_{b-1,0}, \ldots , K_0 \), we have

\[
\mathbb{E}_{k \sim p_b} [w_b^l(k)] + \frac{1}{\eta_0} \log(\mathbb{E}_{k \sim p_b} \exp(-\eta_b w_b^l(k))) = \mathbb{E}_{k \sim p_b} [w_b^l(k)] + \frac{1}{\eta_0} \log(\mathbb{E}_{k \sim p_b} \exp(-\eta_b w_b^l(k)))
\]

\[
\leq \mathbb{E}_{k \sim p_b} [w_b^l(k)] + \frac{1}{\eta_0} (\mathbb{E}_{k \sim p_b} \exp(-\eta_b w_b^l(k)) - 1)
\]

19
We let
\[\eta_0 = \frac{1}{\eta_0} \mathbb{E}_{k \sim p_b} [(w_b'(k))^2] \]
where the first inequality uses \( \log(x) \leq x - 1 \) for all \( x \in \mathbb{R} \) and the second inequality uses \( e^x \leq 1 + x + \frac{x^2}{2} \) for all \( x \in \mathbb{R} \).
Now, for \( b \in \mathcal{V} \), given \( K_{b-1}, \ldots, K_0 \), we obtain that
\[
\mathbb{E}_{k \sim p_b} [w_b'(k)] + \frac{1}{\eta_0} \log(\mathbb{E}_{k \sim p_b} \exp(-\eta_0 w_b'(k))) \leq \mathbb{E}_{k \sim p_b} [w_b'(k)] + \frac{1}{\eta_0} (\mathbb{E}_{k \sim p_b} \exp(-\eta_0 w_b'(k)) - 1)
\]
\[
\leq \mathbb{E}_{k \sim p_b} [w_b'(k)]
\]
\[
\leq \mathbb{E}_{k \sim p_b} [w_b'(k)] + \frac{1}{\eta_0} (\mathbb{E}_{k \sim p_b} \eta_0^2 (w_b'(k))^2) - \eta_0 w_b'(k) + \frac{1}{2}
\]
where the second inequality uses \( e^x \leq 1 + x + \frac{x^2}{2} \) for all \( x \in \mathbb{R} \). Since \( |\mathcal{V}| = O(U) \) by Lemma B.3, we have inequality (28) holding at most \( O(U) \) times and (27) holding in the remaining batches among \( b = 0, \ldots, B - 1 \). Finally, by Lemma B.2, taking expectation of (27) and (28) with respect to \( K_{B-1,0} \) completes the proof.

We let \( x_t \) and \( u_t \) denote the state and action sequence in the algorithm depending on the context. We let \( x^K_t(i) \) and \( u^K_t(i) \) for \( t = t_b, \ldots, t_{b+1} - 1 \) denote the state and action sequence generated by selecting the controllers before batch \( b \) according to Algorithm 1, while selecting the controller \( i \) at batch \( b \). Accordingly, we have \( w^K_b(i) = \sum_{t=t_b}^{t_{b+1}-1} c_t(x^K_t(i), u^K_t(i)) \). We also let \( x^*_t \) and \( u^*_t \) denote the optimal state and action sequence generated by the best stabilizing controller \( i^* \) that satisfies both of Definitions 2.3 and 2.4; i.e., \( i^* = \arg \min_{i \in S} \sum_{t=0}^{T} c_t(x_t, \pi_i(x_t)) \) subject to the transition dynamics.

**Lemma B.6.** In Algorithm 1, suppose that \( \frac{1}{\eta_0} (\beta(\tau_0))^2 < \frac{1}{2 \sqrt{2}} \). For any controller \( i^* \in \mathcal{P}_b \) for \( b = 0, \ldots, B - 1 \), we have
\[
\sum_{b=0}^{B-1} \mathbb{E}_{K_{b-1,0}} (w^K_b(i^*))^2 = \exp(O(U))O(\tau_{B-1}H(\tau_{B-1})) + O(\sum_{b=0}^{B-1} (\tau_b)^2).
\]

**Proof.** By Assumption 2.2, for all \( x \in \mathbb{R}^n \) and \( u \in \mathbb{R}^m \), we have
\[
|c_t(x, u)| = |c_t(x, u) - c_t(0, 0) + c_t(0, 0)| \leq |c_t(x, u) - c_t(0, 0)| + |c_t(0, 0)|
\]
\[
\leq (L_{c_1}(||x|| + ||u||) + L_{c_2}(||x|| + ||u||)) + c_{0,\text{max}}
\]
\[
= L_{c_1}(||x|| + ||u||)^2 + L_{c_2}(||x|| + ||u||) + c_{0,\text{max}}
\]
\[
\leq 2L_{c_1}(||x||^2 + ||u||^2) + L_{c_2}(||x|| + ||u||) + c_{0,\text{max}},
\]
where the last inequality is due to Cauchy–Schwarz inequality. Thus, we can upper-bound \( (w^K_b(i^*))^2 \) for any controller \( i^* \in \mathcal{P}_b \) for \( b = 0, \ldots, B - 1 \) as follows:
\[
(w^K_b(i^*))^2 \leq (w^K_b(i^*))^2 \leq \sum_{t=t_b}^{t_{b+1}-1} c_t(x^K_t(i^*), u^K_t(i^*))^2 (t_{b+1} - t_b)
\]
\[
\leq (t_{b+1} - t_b) \sum_{t=t_b}^{t_{b+1}-1} (2L_{c_1}(||x^K_t(i^*)||^2 + ||u^K_t(i^*)||^2) + L_{c_2}(||x^K_t(i^*)|| + ||u^K_t(i^*)||) + c_{0,\text{max}})^2
\]
\[
\leq 5(t_{b+1} - t_b) \sum_{t=t_b}^{t_{b+1}-1} (4L_{c_1}^2(||x^K_t(i^*)||^4 + ||u^K_t(i^*)||^4) + L_{c_2}^2(||x^K_t(i^*)||^2 + ||u^K_t(i^*)||^2) + c_{0,\text{max}}^2)
\]
where the first and the third inequalities are due to Cauchy–Schwarz inequality.

From (9), for \( t_b < t \leq t_{b+1} - 1 \), we have

\[
\|x_t^K(i^b)\|^2 \leq 2[\beta(t - t_b)]^2 \|x_{t_b}^K(i^b)\|^2 + 2\gamma^2 u_{\text{max}}^2 \tag{31}
\]

\[
\|x_t^K(i^b)\|^4 \leq 8[\beta(t - t_b)]^4 \|x_{t_b}^K(i^b)\|^4 + 8\gamma^4 u_{\text{max}}^4, \tag{32}
\]

where the inequalities are by Cauchy-Schwarz inequality. Accordingly, we obtain that

\[
\sum_{t = t_b}^{t_{b+1}-1} \|x_t^K(i^b)\|^2 \leq 2H(t_{b+1} - t_b)\|x_{t_b}^K(i^b)\|^2 + 2\gamma^2 u_{\text{max}}^2 (t_{b+1} - t_b - 1) \tag{33}
\]

\[
\sum_{t = t_b}^{t_{b+1}-1} \|x_t^K(i^b)\|^4 \leq 8H(t_{b+1} - t_b)\|x_{t_b}^K(i^b)\|^4 + 8\gamma^4 u_{\text{max}}^4 (t_{b+1} - t_b - 1), \tag{34}
\]

where we use \( \beta(\cdot) \leq 1 \) to derive \( \sum_{t=0}^{t_{b} - t_b - 1} |\beta(t)|p \leq \sum_{t=0}^{t_{b+1} - t_b - 1} |\beta(t)| = H(t_{b+1} - t_b) \) for \( p \geq 1 \).

From (11), for \( t_b \leq t \leq t_{b+1} - 1 \), we have

\[
\|u_t^K(i^b)\|^2 \leq 2L_{\pi}^2 \|x_t^K(i^b)\|^2 + 2\pi_{0,\text{max}}^2 \tag{35}
\]

\[
\|u_t^K(i^b)\|^4 \leq 8L_{\pi}^4 \|x_t^K(i^b)\|^4 + 8\pi_{0,\text{max}}^4, \tag{36}
\]

where the inequalities are by Cauchy-Schwarz inequality. Now, we substitute (33), (34), (35), (36), and \( t_{b+1} - t_b \leq \tau_b \) into the right-hand side of (30) to upper-bound \( (w_b^K(i^b))^2 \) as follows:

\[
(w_b^K(i^b))^2 \leq 5\tau_b[32L_{c1}^2(1 + 8L_{\pi}^4)H(\tau_b)\|x_{t_b}^K(i^b)\|^4 + 2L_{c2}^2(1 + 2L_{\pi}^2)H(\tau_b)\|x_{t_b}^K(i^b)\|^2] + 5\tau_b^2[32L_{c1}^2(1 + 8L_{\pi}^4)\gamma^4 u_{\text{max}}^4 + \pi_{0,\text{max}}^4] + 2L_{c2}(1 + 2L_{\pi}^2)\gamma^2 u_{\text{max}}^2 + \pi_{0,\text{max}}^2 + c_{0,\text{max}}^2
\]

\[
= M_3\tau_bH(\tau_b)\|x_{t_b}^K(i^b)\|^4 + M_4\tau_bH(\tau_b)\|x_{t_b}^K(i^b)\|^2 + M_5\tau_b^2
\]

\[
= M_3\tau_bH(\tau_b)\|x_{t_b}^K\|^4 + M_4\tau_bH(\tau_b)\|x_{t_b}^K\|^2 + M_5\tau_b^2, \tag{37}
\]

where \( M_3, M_4, M_5 \) are constants determined by \( L_{c1}, L_{c2}, L_{\pi}, \gamma, u_{\text{max}}, \pi_{0,\text{max}}, \) and \( c_{0,\text{max}} \). The last equality comes from \( x_{t_b}^K(i^b) = x_{t_b} \) for any \( i^b \in \mathcal{P}_b \).

Meanwhile, one can upper-bound both \( \sum_{b=0}^{B-1} \tau_bH(\tau_b)\|x_{t_b}\|^4 \) and \( \sum_{b=0}^{B-1} \tau_bH(\tau_b)\|x_{t_b}\|^2 \) by successively applying Lemma A.3, A.4, and A.6 in the same fashion as presented in the proof of Lemma A.7. Since \( \frac{2}{\tau_0} 8(\beta(\tau_0))^4 < 1 \), by (31) and (32), there exists \( C_1, C_2 \geq 1 \) such that

\[
\sum_{b=0}^{B-1} \tau_bH(\tau_b)\|x_{t_b}\|^4 = O(8L_{\pi}^2(1 + L_{\pi})^4)C_1U(\|x_0\|^4 + \tau_{\beta}\epsilon(\tau_{\beta})) + 8\gamma^4 u_{\text{max}}^2 \cdot O(\sum_{b=0}^{B-1} \tau_bH(\tau_b)) \]

\[
\sum_{b=0}^{B-1} \tau_bH(\tau_b)\|x_{t_b}\|^2 = O(8L_{\pi}^2(1 + L_{\pi})^2)C_2U(\|x_0\|^2 + \tau_{\beta}\epsilon(\tau_{\beta})) + 2\gamma^2 u_{\text{max}}^2 \cdot O(\sum_{b=0}^{B-1} \tau_bH(\tau_b)).
\]

Substituting the equalities into the summation of (37) for \( b = 0, \ldots, B - 1 \) yields

\[
\sum_{b=0}^{B-1} (w_b^K(i^b))^2 = \exp(O(U))O(\tau_{\beta}\epsilon(\tau_{\beta})) + O(\sum_{b=0}^{B-1} \tau_bH(\tau_b)) + O(\sum_{b=0}^{B-1} \tau_bH(\tau_b)). \tag{38}
\]

Notice that taking expectation of \( (w_b^K(i^b))^2 \) with respect to \( K_{b-1:0} \) does not affect the inequality. Finally, \( \tau_{\beta} \leq \tau_{B-1} \) and \( H(\tau_b) = O(\tau_b) \) completes the proof.

**Lemma B.7.** In Algorithm 1, for the best stabilizing controller \( i^* \in S \), we have

\[
\mathbb{E}_{K_{B-1:0}} \sum_{b=0}^{B-1} \sum_{t=t_b}^{t_{b+1}-1} \left[ c_t(x_{t_b}^K(i^*), u_{t_b}^K(i^*)) - c_t(x_t^*, u_t^*) \right] \leq O(U) + O(\sum_{b=0}^{B-1} H(\tau_b)).
\]

21
Proof. Since \( x_t^i \) is generated by a stabilizing controller, we have

\[
\begin{align*}
\| x_t^i \| &\leq \beta(t) \| x_0 \| + \gamma w_{\text{max}} \leq \beta(0) \| x_0 \| + \gamma w_{\text{max}} \\
\| x_t^i \|^2 &\leq 2 \beta(t)^2 \| x_0 \|^2 + 2 \gamma^2 w_{\text{max}}^2 \leq 2 \beta(0)^2 \| x_0 \|^2 + 2 \gamma^2 w_{\text{max}}^2,
\end{align*}
\]

where the inequalities are by Cauchy-Schwarz inequality and the non-increasing property of \( \beta(\cdot) \). Then, by (11), (29), and (35), we have

\[
\begin{align*}
c_t(x_t^i, u_t^i) &\leq 2 L_{c1}(\| x_t^i \|^2 + \| u_t^i \|^2) + L_{c2}(\| x_t^i \| + \| u_t^i \|) + c_{0, \text{max}} \\
&\leq 2 L_{c1}(1 + 2 L_{c2}^2) \| x_t^i \|^2 + 2 \gamma^2 w_{\text{max}}^2 + L_{c2}(1 + L_{\pi}) \| x_t^i \| + \pi_{0, \text{max}} + c_{0, \text{max}} \\
&\leq 4 L_{c1}(1 + 2 L_{c2}^2) (\beta(0))^2 \| x_0 \|^2 + L_{c2}(1 + L_{\pi}) \beta(0) \| x_0 \| + 4 L_{c1}(1 + 2 L_{c2}^2) \gamma^2 w_{\text{max}}^2 \\
&\quad + L_{c2}(1 + L_{\pi}) \gamma w_{\text{max}} + 4 L_{c1} \pi_{0, \text{max}} + L_{c2} \pi_{0, \text{max}} + c_{0, \text{max}} := M_6.
\end{align*}
\]

In Algorithm 1, one can write

\[
\begin{align*}
\frac{\| x_{t_b} \|}{(\alpha_b)^{s_b}} &\leq \frac{(\alpha_b)^{s_b+1}}{(\alpha_b)^{s_b}} \| x_0 \| + \delta \leq \alpha_b \| x_0 \| + \delta \quad \text{(40)} \\
\frac{\| x_t^i \|}{(\alpha_b)^{s_b}} &\leq \frac{\beta(t) \| x_0 \| + \gamma w_{\text{max}}}{(\alpha_b)^{s_b}} \leq \beta(0) \| x_0 \| + \gamma w_{\text{max}}, \quad \text{(41)}
\end{align*}
\]

where the equalities hold for the last inequalities of (40) and (41) when \( s_b = 0 \).

By Assumption 2.2, for the best stabilizing controller \( i^* \in S \), we have

\[
\begin{align*}
\frac{1}{(\alpha_b)^{2s_b}} | c_t(x_t^K(i^*), u_t^K(i^*)) - c_t(x_t^i, u_t^i) | &\leq \frac{1}{(\alpha_b)^{2s_b}} L_{c1}(\max\{\| x_t^K(i^*) \|, \| x_t^i \| \}) \max\{\| u_t^K(i^*) \|, \| u_t^i \| \} + L_{c2}(\| x_t^K(i^*) - x_t^i \| + \| u_t^K(i^*) - u_t^i \|) \\
&\leq \frac{1}{(\alpha_b)^{2s_b}} L_{c1}(1 + L_{\pi}) \max\{\| x_t^K(i^*) \|, \| x_t^i \| \} + \pi_{0, \text{max}} + L_{c2}(1 + L_{\pi}) \| x_t^K(i^*) - x_t^i \| \\
&\quad = (1 + L_{\pi})(L_{c1}(1 + L_{\pi}) \max\{\| x_t^K(i^*) \|, \| x_t^i \| \} + L_{c2}(1 + L_{\pi}) \| x_t^K(i^*) - x_t^i \| \\
&\quad + L_{c1}(1 + L_{\pi}) \gamma w_{\text{max}} + L_{c1} \pi_{0, \text{max}} + L_{c2}(1 + L_{\pi}) \gamma w_{\text{max})) \cdot (t - t_b) \frac{\| x_t^K(i^*) - x_t^i \|}{(\alpha_b)^{s_b}} \\
&\leq L_{c1}(1 + L_{\pi})^2 \beta(t - t_b)^2 \left( \frac{\| x_{t_b} \|}{(\alpha_b)^{s_b}} + \frac{\| x_t^i \|}{(\alpha_b)^{s_b}} \right)^2 \\
&\quad + L_{c1}(1 + L_{\pi}) (t - t_b)^2 + L_{c1}(1 + L_{\pi}) \gamma w_{\text{max}} + L_{c1} \pi_{0, \text{max}} + L_{c2}(1 + L_{\pi}) \beta(t - t_b) \left( \frac{\| x_{t_b} \|}{(\alpha_b)^{s_b}} + \frac{\| x_t^i \|}{(\alpha_b)^{s_b}} \right) \\
&\leq M_7 \beta(t - t_b)^2 + M_8 \beta(t - t_b), \quad \text{(42)}
\end{align*}
\]

where \( M_7 \) and \( M_8 \) are constants determined by \( L_{c1}, L_{c2}, L_{\pi}, \pi_{0, \text{max}}, \beta(0), \delta, \gamma, w_{\text{max}} \) and \( \max_{b \in \{0, 1, \ldots, B - 1\}} \alpha_b \). Notice that \( \alpha_b \) in Line 14 of Algorithm 1 is upper-bounded by some constant by Lemma A.5. The second inequality is by (11), the third inequality is due to Definition 2.4 and by leveraging the same stabilizing controller \( i^* \) from \( t_b \) for both trajectories \( x_t^K(i^*) \) and \( x_t^i \), the fourth inequality uses \( x_{t_b}^K(i^*) = x_{t_b} \), and the fifth inequality is by (40) and (41). By combining (39) and (42), we have

\[
\left| \frac{c_t(x_t^K(i^*), u_t^K(i^*))}{(\alpha_b)^{2s_b}} - c_t(x_t^i, u_t^i) \right| \leq \frac{c_t(x_t^K(i^*), u_t^K(i^*))}{(\alpha_b)^{2s_b}} - \frac{c_t(x_t^i, u_t^i)}{(\alpha_b)^{2s_b}} - \frac{(\alpha_b)^{2s_b} - 1}{(\alpha_b)^{2s_b}} c_t(x_t^i, u_t^i)
\]
where the first inequality is due to Lemma B.4 and B.5, and the last inequality is due to Lemma B.6 and B.7. Using $\sum_{b=0}^{B-1} \sum_{t=t_b}^{t_{b+1}-1} |c_t(x^*_t, u_t^*)| \leq \sum_{b=0}^{B-1} |V_b| + \sum_{b=0}^{B-1} (M_T + M_B) H(t_{b+1} - t_b)$ completes the proof. (43)

where the first inequality uses $\beta(\cdot) \leq 1$ to derive $\sum_{t=t_b}^{t_{b+1}-1} [\beta(t - t_b)] \leq \sum_{t=t_b}^{t_{b+1}-1} [\beta(t - t_b)] = H(t_{b+1} - t_b)$ and the last equality uses $t_{b+1} - t_b \leq \tau_b$ and Lemma B.3. Taking expectation of (43) with respect to $K_{B-1.0}$ completes the proof. $\square$

**Theorem B.8** (Restatement of Theorem 4.6, Regret Bound). In Algorithm 1, suppose that $E(\frac{1}{\tau_0}(\beta(\tau_0))^2 < \frac{1}{2\sqrt{2}}$. Then, the regret bound is as follows:

$$E_{K_{B-1.0}}[\sum_{t=0}^{T} (c_t(x_t, u_t) - c_t(x^*_t, u^*_t))] = O(|U|) + O(\sum_{b=0}^{B-1} H(\tau_b)) + \frac{\hat{C}(|U| + 1)}{\eta_0} + \frac{N}{2}[\exp(O(|U|))O(\tau_{B-1/2}H(\tau_{B-1/2})) + O(\sum_{b=0}^{B-1} (\tau_b)^2)].$$

**Proof.** By Lemma B.1, we have

$$E_{K_{B-1.0}}[\sum_{t=0}^{T} (c_t(x_t, u_t) - c_t(x^*_t, u^*_t))] = E_{K_{B-1.0}}[\sum_{b=0}^{B-1} \sum_{t=t_b}^{t_{b+1}-1} (c_t(x_t, u_t) - c_t(x^*_t, u^*_t))]$$

$$= E_{K_{B-1.0}}[\sum_{b=0}^{B-1} E_{K_{B-1.0}}[w_b(K_b)]]$$

$$\leq \frac{\hat{C}(|U| + 1)}{\eta_0} + \frac{N}{2}|\sum_{b=0}^{B-1} E_{K_{B-1.0}}[w_b(K_b)]^2 + E_{K_{B-1.0}}[\sum_{t=0}^{T} [\sum_{b=0}^{B-1} (c_t(x^*_t, u^*_t))]]$$

$$\leq \frac{\hat{C}(|U| + 1)}{\eta_0} + \frac{N}{2}[\exp(O(|U|))O(\tau_{B-1/2}H(\tau_{B-1/2})) + O(\sum_{b=0}^{B-1} (\tau_b)^2)]$$

$$+ O(U) + O(\sum_{b=0}^{B-1} H(\tau_b)) + E_{K_{B-1.0}}[\sum_{t=0}^{T} (c_t(x^*_t, u^*_t)],$$

where the first inequality is due to Lemma B.4 and B.5, and the last inequality is due to Lemma B.6 and B.7. Using $U \leq |U|$ completes the proof. $\square$

**Theorem B.9** (Restatement of Theorem 4.7, Regret bound with known $|U|$). In Algorithm 1, let $\tau_0 = \left(\frac{z}{N(|U| + 1)}\right)^{1/2}$ and $\tau_b = \left(\frac{z}{N(|U| + 1)}\right)^{1/2}$ for every $b \geq 1$ with the constants $z, \nu > 0$ that satisfies $\tau_0 > 0$ and $\frac{\tau}{\tau_0}(\beta(\tau_0))^2 < \frac{1}{2\sqrt{2}}$. Also, let $\eta_0 = O\left(\frac{|U|^{1/3}}{T^{2/3}N^{1/3}}\right)$, when $T \geq \max\left\{\frac{(\omega_0^2 \exp(O(|U|)))^{1/2}}{N(|U| + 1)^{1/2}}, \frac{|U|^{1/2}}{N(|U| + 1)^{1/2}}, N(|U| + 1)\right\}$, we have

$$E_{K_{B-1.0}}[\sum_{t=0}^{T} (c_t(x_t, u_t) - c_t(x^*_t, u^*_t))] = O(T^{2/3}N^{1/3}(|U| + 1)^{1/3}) + o(T),$$

23
which implies that we achieve a sublinear regret bound. Moreover, when \( \lim_{t \to \infty} H(t) < \infty \) and \( T \geq \max\{\exp(O(|U|)), \frac{|U|^{3/2}}{(N(|U|+1))^{3/2}}, N(|U|+1)\} \), we have

\[
\mathbb{E}_{K_{B-1.0}} \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x^*_t, u^*_t)] = \tilde{O}(T^{2/3}N^{1/3}(|U|+1)^{1/3}).
\]

**Proof.** By the formulation of \((\tau_b)_{b \geq 0}\), we have

\[
\sum_{b=0}^{B-1} \left( \frac{(\nu b + z)^{1/2}}{(N(|U|+1))^{1/2}} - 1 \right) = \sum_{b=0}^{B-1} \frac{(\nu b + z)^{1/2} - 1}{(N(|U|+1))^{1/2}} \leq T
\]

where we can further use non-decreasing property of \((\cdot)^{1/2}\) to arrive at

\[
\frac{z^{1/2}}{(N(|U|+1))^{1/2}} + \int_{0}^{B} \frac{(\nu b + z)^{1/2} - 1}{(N(|U|+1))^{1/2}} \, db = O(\frac{B^2}{N(|U|+1)}) = O(T^{4/3}N^{-1/3}(|U|+1)^{-1/3}),
\]

thus we have \( B = O(T^{2/3}N^{1/3}(|U|+1)^{1/3}) \) from the first inequality and \( T = O(B^{3/2}N^{-1/2}(|U|+1)^{-1/2}) \) from the second inequality and \( T \geq N(|U|+1) \). Similarly, we can find the order of \( \sum_{b=0}^{B-1} (\tau_b)^2 \) as follows:

\[
\sum_{b=0}^{B-1} (\tau_b)^2 \leq \sum_{b=0}^{B-1} \left[ \frac{(\nu b + z)^{1/2}}{(N(|U|+1))^{1/2}} + 1 \right]^2 \leq \int_{0}^{B} \left[ \frac{(\nu b + z)}{(N(|U|+1))} + \frac{2(\nu b + z)^{1/2}}{(N(|U|+1))^{1/2}} + 1 \right] \, db = O(\frac{B^2}{N(|U|+1)}) = O(T^{4/3}N^{-1/3}(|U|+1)^{-1/3}),
\]

where the last equality is by \( B = O(T^{2/3}N^{1/3}(|U|+1)^{1/3}) \). We also have

\[
\tau_{B-1} = \left[ \frac{(\nu (B-1) + z)}{N(|U|+1)} \right]^{1/2} = O(B^{1/2}N^{-1/2}(|U|+1)^{-1/2}) = O(T^{1/3}N^{-1/3}(|U|+1)^{-1/3}).
\]

Thus, we have

\[
O(\tau_{B-1} H(\tau_{B-1})) = o((\tau_{B-1})^2) = o(T^{2/3}N^{-2/3}(|U|+1)^{-2/3}) = \frac{o(1)}{\eta_0 N},
\]

where the first equality is due to Lemma A.1. With \( T \geq \frac{o(1)}{\frac{\exp(O(U))}{(N(|U|+1))^{3/2}}} \) and \( T \geq \frac{|U|^{3/2}}{(N(|U|+1))^{3/2}} \), we have

\[
\eta_0 N \exp(O(U)) O(\tau_{B-1} H(\tau_{B-1})) = o(1) \exp(O(U)) = O(T^{2/3}N^{1/3}(|U|+1)^{1/3})
\]

(48)

\[
O(|U|) = O(T^{2/3}N^{1/3}(|U|+1)^{1/3}).
\]

(49)

With (45), (47), (48), and (49), we can apply Theorem B.8 to derive

\[
\mathbb{E}_{K_{B-1.0}} \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x^*_t, u^*_t)] = \tilde{O}(T^{2/3}N^{1/3}(|U|+1)^{1/3}) + O\left( \sum_{b=0}^{B-1} H(\tau_b) \right).
\]

Applying (16) to \( O(\sum_{b=0}^{B-1} H(\tau_b)) \) achieves a sublinear regret bound.

Moreover, when \( \lim_{t \to \infty} H(t) < \infty \), there exists a constant \( q_1 \) that upper-bounds \( H(t) \); i.e., \( H(t) \leq q_1 \) for all \( t \geq 0 \). Then, we have

\[
\sum_{b=0}^{B-1} H(\tau_b) \leq q_1 B = O(B) = O(T^{2/3}N^{1/3}(|U|+1)^{1/3}).
\]

(50)
Also, (47) and (48) can be modified to
\[ \tau_{B-1} H(\tau_{B-1}) \leq q_1 \tau_{B-1} = O\left(T^{1/3} N^{-1/3}(|\mathcal{U}| + 1)^{-1/3}\right), \]
\[ \eta_0 N \exp(O(U)) O(\tau_{B-1} H(\tau_{B-1})) = O\left(T^{2/3} N^{1/3}(|\mathcal{U}| + 1)^{1/3}\right), \] (51)
only with \( T \geq \exp(O(U)) \). Using (50) and (51) completes the proof.

\[ \square \]

C. Regret Proof for Algorithm 2

**Theorem C.1** (Restatement of Theorem 4.10, Regret bound with unknown \(|\mathcal{U}|\)). In Algorithm 2, let \( \tau_0 = \left[\left(\frac{1}{N}\right)^{1/2}\right] \) and \( \tau_b = \left[\left(\frac{(1/b^2 + 2)}{N}\right)^{1/2}\right] \) for every \( b \geq 1 \) with the constants \( z, \nu > 0 \) that satisfies \( \tau_0 > 0 \) and \( \frac{\nu}{\tau_0} (\beta(\tau_0))^{2} < \frac{1}{2\sqrt{2}} \). Also, let \( \eta_0 = O\left(\frac{1}{T^{2/3} N^{1/3}}\right) \) and \( y = \frac{1}{2} \). When \( T \geq \max\{\frac{(\nu (1/\exp(O(|\mathcal{U}|)))^{3/2}}{N^{1/2}} \cdot \frac{|\mathcal{U}|^{3/2}}{N^{1/2} (|\mathcal{U}| + 1)^{3/2}} \cdot N\}, \) we have
\[ \mathbb{E}_{\mathcal{K}_{B-1,0}} \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x^*_t, u^*_t)] = \tilde{O}(T^{2/3} N^{1/3} (|\mathcal{U}| + 1)^{1/2}) + o(T), \]
which implies that we achieve a sublinear regret bound. Moreover, when \( \lim_{t \to \infty} H(t) < \infty \) and \( T \geq \max\{\frac{\exp(O(|\mathcal{U}|))}{T |\mathcal{U}|^{3/2}} \cdot \frac{|\mathcal{U}|^{3/2}}{N^{1/2} (|\mathcal{U}| + 1)^{3/2}} \cdot N\}, \) we have
\[ \mathbb{E}_{\mathcal{K}_{B-1,0}} \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x^*_t, u^*_t)] = \tilde{O}(T^{2/3} N^{1/3} (|\mathcal{U}| + 1)^{1/2}). \]

**Proof.** By the formulation of \( (\tau_b)_{b \geq 2} \), as in (44), we can derive
\[ B = O(T^{2/3} N^{1/3}) \quad \text{and} \quad T = O(B^{3/2} N^{-1/2}) \]
when \( T \geq N \). We can also obtain
\[ \sum_{b=0}^{B-1} (\tau_b)^2 = O(T^{4/3} N^{-1/3}) \quad \text{and} \quad O(\tau_{B-1} H(\tau_{B-1})) = o(T^{2/3} N^{-2/3}) \]
similar to (45) and (47). Now, define \( \eta_{0,r} := \eta_0 (r + 1)^y = \eta_0 \sqrt{r + 1} \). Let \( \mathcal{B}_r \) denote the set of batches where \( \mu_b = r \); i.e., \( \mathcal{B}_r = \{0 \leq b \leq B - 1, b \in \mathbb{Z}_+ : \mu_b = r\} \). Then, one can write
\[ \sum_{b=0}^{B-1} (\tau_b)^2 = O\left(T^{4/3} N^{-1/3}\right) \quad \text{and} \quad O(\tau_{B-1} H(\tau_{B-1})) = o(T^{2/3} N^{-2/3}) \]
where the first equality holds by Lemma B.6, second inequality holds by \( U \leq |\mathcal{U}| \), and the last equality holds when \( T \geq \frac{\nu (1/\exp(O(|\mathcal{U}|)))^{3/2}}{N^{1/2}} \cdot \frac{|\mathcal{U}|^{3/2}}{N^{1/2} (|\mathcal{U}| + 1)^{3/2}} \cdot N \).

Recall the definition and the cardinality of \( \mathcal{L} = \{0 \leq b \leq B - 1, b \in \mathbb{Z}_+ : s_{b+1} \neq s_b\} \) and \( \mathcal{V} = \{0 \leq b \leq B - 1, b \in \mathbb{Z}_+ : s_b \neq 0\} \) in Lemma B.3. We focus on the mix loss and the mixability gap with the denominator \( \eta_{0,r} \); i.e., \( -\frac{1}{\eta_{0,r}} \log(\mathbb{E}_{k \sim \mathcal{P}_b} \exp(-\eta_b w_b(k))) \) and \( \mathbb{E}_{k \sim \mathcal{P}_b} [w_b(k)] + \frac{1}{\eta_{0,r}} \log(\mathbb{E}_{k \sim \mathcal{P}_b} \exp(-\eta_b w_b(k))) \). Considering that \( \frac{\mu_b}{\eta_{0,r}} \) still remains to be \( \frac{1}{(\alpha_b)^y} \) as in Algorithm 1, Lemma B.4 can be modified to
\[ \mathbb{E}_{\mathcal{K}_{B-1,0}} \sum_{r=0}^{U} \sum_{b \in \mathcal{B}_r} \frac{1}{\eta_{0,r}} \log(\mathbb{E}_{k \sim \mathcal{P}_b} \exp(-\eta_b w_b(k))) \leq \sum_{r=0}^{U} \frac{p_r}{\eta_{0,r}} \log N + \mathbb{E}_{\mathcal{K}_{B-1,0}} \sum_{l=0}^{\lceil \frac{1}{\eta_{0,r}} \sum_{b \in \mathcal{B}_r} \frac{1}{\eta_{0,r}} \log(\mathbb{E}_{k \sim \mathcal{P}_b} \exp(-\eta_b w_b(k))) \rceil} \sum_{b \in \mathcal{B}_r} \frac{1}{\eta_{0,r}} \log(\mathbb{E}_{k \sim \mathcal{P}_b} \exp(-\eta_b w_b(k))) \]
(53)
where \( \rho_r^l \) denotes the number of batches in \( B_r \cap L \). Similarly, considering that \( \eta_{0,r} \) now depends on the value of \( r \), Lemma B.5 can be modified to

\[
E_{K_{s=1 \ldots 0}} \sum_{b=0}^{U} \sum_{r=0}^{U} \exp(w_h^r(k)) \leq \frac{U}{N} \sum_{r=0}^{U} \rho_{r}^{b} + \frac{U}{2} \sum_{r=0}^{U} \eta_{0,r} E_{K_{s=1 \ldots 0}}(w_b^K) ^2, \tag{54}
\]

where \( \rho_r^b \) denotes the number of batches in \( B_r \cap V \). Now, our goal is to upper-bound \( \sum_{r=0}^{U} \rho_{r}^{b} \) in (54) and \( \sum_{r=0}^{U} \rho_{r}^{b} \) in (55) and (56). It is straightforward to infer that \( \rho_1^0 + \rho_1^1 + \cdots + \rho_U^U \leq 2U + 1 \) by Lemma B.3 and (26), which also leads to \( \rho_1^0 + \rho_1^1 + \cdots + \rho_r^r \leq 2r + 1 \) for \( r = 0, \ldots, U \). Similarly, we can infer that \( \rho_1^0 = 0 \) and \( \rho_1^0 + \cdots + \rho_U^U \leq (2U - 1) \rho_0^{v} \) by Lemma B.3 and (28), which also leads to \( \rho_1^0 + \cdots + \rho_r^r \leq (2r - 1) \rho_0^{v} \) for \( r = 1, \ldots, U \). Define \( M_9 := \frac{\log(a_{b}^{v} + \frac{b}{2} \log(b)})}{- \log(\beta)} \) and consider the following maximization problems to get the upper bound.

\[
\begin{align*}
l^* = \max_{\rho_0^l, \ldots, \rho_U^l} & \sum_{r=0}^{U} \rho_r^l \quad \text{s.t.} \quad \rho_0^l \leq 1 \\
& \rho_0^l + \rho_1^l \leq 3 \\
& \cdots \\
& \rho_0^l + \rho_1^l + \cdots + \rho_U^l \leq 2U + 1,
\end{align*}
\]

\[
\begin{align*}
v^* = \max_{\rho_0^v, \ldots, \rho_U^v} & \sum_{r=0}^{U} \rho_r^v \quad \text{s.t.} \quad \rho_1^v \leq M_9 \\
& \rho_1^v + \rho_2^v \leq 3M_9 \\
& \cdots \\
& \rho_1^v + \rho_2^v + \cdots + \rho_U^v \leq (2U - 1)M_9.
\end{align*}
\]

We can easily achieve an optimal point of each linear programming (LP) problem by the well-known Karush-Kuhn-Tucker (KKT) conditions. There exist positive constants \( \lambda_0, \ldots, \lambda_U, \kappa_1, \ldots, \kappa_U \) such that

\[
\begin{bmatrix}
1 \\
\frac{1}{\sqrt{2}} \\
\vdots \\
\frac{1}{\sqrt{U + 1}}
\end{bmatrix} \begin{bmatrix}
\sum_{r=0}^{U} \lambda_r \\
\sum_{r=1}^{U} \lambda_r \\
\vdots \\
\sum_{r=1}^{U} \lambda_U
\end{bmatrix} = \begin{bmatrix}
1 \\
\frac{1}{\sqrt{2}} \\
\vdots \\
\frac{1}{\sqrt{U + 1}}
\end{bmatrix} \begin{bmatrix}
\sum_{r=0}^{U} \kappa_r \\
\sum_{r=2}^{U} \kappa_r \\
\vdots \\
\sum_{r=2}^{U} \kappa_U
\end{bmatrix}, \tag{55}
\]

\[
\begin{bmatrix}
1 \\
\frac{1}{\sqrt{2}} \\
\vdots \\
\frac{1}{\sqrt{U + 1}}
\end{bmatrix} \begin{bmatrix}
\sum_{r=0}^{U} \lambda_r \\
\sum_{r=1}^{U} \lambda_r \\
\vdots \\
\sum_{r=1}^{U} \lambda_U
\end{bmatrix} = \begin{bmatrix}
1 \\
\frac{1}{\sqrt{2}} \\
\vdots \\
\frac{1}{\sqrt{U + 1}}
\end{bmatrix} \begin{bmatrix}
\sum_{r=0}^{U} \kappa_r \\
\sum_{r=2}^{U} \kappa_r \\
\vdots \\
\sum_{r=2}^{U} \kappa_U
\end{bmatrix}, \tag{56}
\]

which yields \( \lambda_U = \kappa_U = \frac{1}{\sqrt{U + 1}} \), \( \lambda_r = \kappa_r = \frac{1}{\sqrt{U + 1}} - \frac{1}{\sqrt{U + 2}} > 0 \) for \( r = 1, \ldots, U - 1 \), and \( \lambda_0 = 1 - \frac{1}{\sqrt{2}} \). Since every dual variable is positive, complementary slackness tells that there is no slack for every inequality at the optimal solution. Thus, the optimal solutions are

\[
\rho_0^l = 1, \quad \rho_r^l = 2, \quad r = 1, \ldots, U, \\
\rho_1^v = M_9, \quad \rho_r^v = 2M_9, \quad r = 2, \ldots, U,
\]

where the corresponding optimal objective values are

\[
l^* = 1 + \sum_{r=1}^{U} \frac{2}{\sqrt{r + 1}} \leq 1 + \sqrt{2} + \int_1^U \frac{1}{\sqrt{r + 1}} dr = O(\sqrt{U + 1})
\]

\[
v^* = \frac{M_9}{\sqrt{2}} + \sum_{r=2}^{U} \frac{2M_9}{\sqrt{r + 1}} \leq \frac{M_9}{\sqrt{2}} + \frac{2M_9}{\sqrt{3}} + \frac{2M_9}{\sqrt{4}} \int_2^U \frac{1}{\sqrt{r + 1}} dr = O(\sqrt{U + 1}),
\]

where we leverage the non-increasing property of \( \frac{1}{\sqrt{r + 1}} \) for the inequalities. Thus, we have both \( \frac{1}{\sqrt{v_0^{l}}} \sum_{r=0}^{U} \rho_r^{l} \log(N) \) = \( \tilde{O}(T^{2/3}N^{1/3}(U + 1)^{1/2}) \) and \( \frac{1}{\sqrt{v_0^{l}}} \sum_{r=0}^{U} \rho_r^{l} \log(N) = O(T^{2/3}N^{1/3}(U + 1)^{1/2}) \). Combining (52), (53), and (54) with Lemma B.7 and \( U \leq \|U\| \), one can write

\[
E_{K_{s=1 \ldots 0}} \sum_{t=0}^{T} \left[ c_t(x_t, u_t) - c_t(x_t^*, u_t^*) \right] = \tilde{O}(T^{2/3}N^{1/3}(U + 1)^{1/2}) + O(\|U\|) + O(\sum_{b=0}^{B-1} H(b))
\]
where the second equality holds when $T \geq \frac{|\mathcal{U}|^{3/2}}{N^{1/2}(|\mathcal{U}|+1)^{3/2}}$. Using (16) shows a sublinear regret bound. When $\lim_{t\to\infty} H(t) < \infty$, (50) and (51) are modified to

\[
\sum_{b=0}^{B-1} H(\tau_b) \leq q_1 B = O(B) = O(T^{2/3}N^{1/3}),
\]

\[
\tau_{B-1} H(\tau_{B-1}) \leq q_1 \tau_{B-1} = O(T^{1/3}N^{-1/3}),
\]

\[
\eta_0 N \exp(O(|\mathcal{U}|))O(\tau_{B-1} H(\tau_{B-1})) = O(T^{2/3}N^{1/3}(|\mathcal{U}| + 1)^{1/2}),
\]

only with $T \geq \frac{\exp(O(|\mathcal{U}|))}{(|\mathcal{U}|+1)^{3/2}}$. This completes the proof.

\[\]

**D. Regret Proof for Algorithm 3**

Algorithm 3 can easily be generalized to the situation where we have $O(U)$ number of system switches or controller pool switches. In fact, we can simply add $\tau_{|\mathcal{U}|+1}, \ldots, \tau_{|\mathcal{U}|+O(U)} \in S$ to the set of best stabilizing controllers \(\{\tau_0, \ldots, \tau_{|\mathcal{U}|}\} \subseteq S\), where \(|\mathcal{U}| = O(U)\) by Lemma B.3. Thus, it suffices to derive the regret bound of Algorithm 3, even in the context of general switched systems.

Let $x_t'$ and $u_t'$ denote the state and action sequence generated by our set of best stabilizing controllers $\{\tau_0, \ldots, \tau_{|\mathcal{U}|}\} \subseteq S$. We consider a regret with switching cost where the unit switching cost is $d \geq 1$; i.e., $\mathbb{E}_{K_{B-1:0}} \left[ \sum_{t=0}^{T} c_t(x_t, u_t) - c_t(x'_t, u'_t) \right] + d \sum_{b=1}^{B-1} I_{(K_b \neq K_{b-1})} - d \sum_{l=1}^{|\mathcal{U}|} I_{(l \neq l_{b-1})}$.

**Lemma D.1.** *In Algorithm 3, let $\tau_0 = \left(\frac{\pi}{N(|\mathcal{U}|+1)}\right)^{1/2}$ and $\tau_0 = \left(\frac{\left(\frac{\eta}{\sqrt{2}N}|\mathcal{U}|\right)^{1/2}}{\eta}\right)^{1/2}$ for every $b \geq 1$ with the constants $z, \nu > 0$ that satisfies $\tau_0 > 0$ and $\tau_b(\beta(\tau_0))^2 < \frac{1}{2\sqrt{2}}$. When $T \geq \frac{\eta(1) \exp(O(|\mathcal{U}|))^{1/2}}{(\eta z N(|\mathcal{U}|+1))^{2}}$, we have*

\[
\mathbb{E}_{K_{B-1:0}} \sum_{b=1}^{B-1} I_{(K_b \neq K_{b-1})} = O(|\mathcal{U}|) + O(\eta_0 NT).
\]

**Proof.** For all $b = 1, \ldots, B - 1$ such that $s_b = s_{b-1}$, given $K_{b-1}, \ldots, K_0$, we have

\[
Pr(K_b \neq K_{b-1}) \leq 1 - \frac{\exp(-\eta_0 W_b(K_{b-1}))}{\exp(-\eta_0 W_b(K_{b-1}))} \leq 1 - \frac{\exp(-\eta_0 W_{b-1}(K_{b-1}))}{\exp(-\eta_0 W_{b-1}(K_{b-1}))}
\]

\[
\leq \eta_0 W_{b-1}(K_{b-1}) = \eta_0 \frac{w_{b-1}(K_{b-1})}{p_{b-1}(K_{b-1})}.
\]

where the second inequality is because $\eta_b = \eta_{b-1}$ when $s_b = s_{b-1}$, the third inequality uses $\eta_0 \geq \eta_b$ for all $b \geq 0$, and the last inequality uses $1 + x \leq e^x$ for all $x \in \mathbb{R}$. Now, given a set of controllers $\bar{V} \in \mathcal{P}_b$ for $b = 0, \ldots, B - 1$, we can upper-bound $\sum_{b=0}^{B-2} w_b(\bar{V})$ by $t_{b+1} - t_b \leq \tau_b$ as follows:

\[
\sum_{b=0}^{B-2} w_b(\bar{V}) = \sum_{b=0}^{B-2} \sum_{t=t_b}^{t_{b+1}} c_t(x_t, u_t) \leq \sum_{b=0}^{B-2} \sum_{t=t_b}^{t_{b+1}} 2L_{c_1} \|x_t\|^2 + \|u_t\|^2 + L_{c_2}(\|x_t\| + \|u_t\|) + c_{0, \text{max}}
\]

\[
\leq \sum_{b=0}^{B-2} \sum_{t=t_b}^{t_{b+1}} 2L_{c_1} ((1 + 2L_{c_2}^2)\|x_t\|^2 + 2\pi_0^2 + L_{c_2}(1 + L_\pi)\|x_t\| + \pi_0^2 + c_{0, \text{max}})
\]

\[
\leq \sum_{b=0}^{B-2} 2L_{c_1} (1 + 2L_{c_2}^2) H(\tau_b) \|x_t\|^2 + L_{c_2}(1 + L_\pi) H(\tau_b) \|x_t\| + \tau_b [4L_{c_1} \pi_0^2 + L_{c_2} \pi_0^2 + c_{0, \text{max}}]
\]
where the first inequality is due to (29), the second inequality is by (11) and (35), the third inequality is due to using (38). With \( T \geq \frac{(a_1) \exp(O(|\mathcal{U}|))}{(N(M+1))^{1/2}} \), we obtain by (46) that

\[
O(\exp(O(|\mathcal{U}|))H(\tau_{w})) + O(\sum_{b=0}^{B-2} H(\tau_b)) + O(\sum_{b=0}^{B-2} \tau_b) \leq O(T).
\]

Thus, one can write

\[
\mathbb{E}_{K_{b-1:0}} \sum_{b=1}^{B-1} I(K_b \neq K_{b-1}) = \sum_{b=1}^{B-1} \mathbb{E}_{K_{b-1:0}} I(K_b \neq K_{b-1}) = \sum_{b=1}^{B-1} \mathbb{E}_{K_{b-1:0}} \mathbb{E}_{K_b} [I(K_b \neq K_{b-1}) | K_{b-1:0}]
\]

\[
= \sum_{b=1}^{B-1} \mathbb{E}_{K_{b-1:0}} Pr(K_b \neq K_{b-1} | K_{b-1:0})
\]

\[
= \sum_{b=1}^{B-1} \mathbb{E}_{K_{b-1:0}} [Pr(s_b = s_{b-1}, \mathcal{P}_b = \mathcal{P}_{b-1} | K_{b-1:0}) Pr(K_b \neq K_{b-1} | s_b = s_{b-1}, \mathcal{P}_b = \mathcal{P}_{b-1}, K_{b-1:0})
\]

\[
+ Pr(s_b \neq s_{b-1} \text{ or } \mathcal{P}_b \neq \mathcal{P}_{b-1} | K_{b-1:0}) Pr(K_b \neq K_{b-1} | s_b \neq s_{b-1} \text{ or } \mathcal{P}_b \neq \mathcal{P}_{b-1}, K_{b-1:0})]
\]

\[
= |\mathcal{L}| + U + \sum_{b=1}^{B-1} \mathbb{E}_{K_{b-1:0}} [Pr(s_b = s_{b-1}, \mathcal{P}_b = \mathcal{P}_{b-1} | K_{b-1:0}) Pr(K_b \neq K_{b-1} | s_b = s_{b-1}, \mathcal{P}_b = \mathcal{P}_{b-1}, K_{b-1:0})
\]

\[
\leq |\mathcal{L}| + U + \sum_{b=1}^{B-1} \mathbb{E}_{K_{b-1:0}} \eta_0 \frac{w_{b-1}(K_{b-1})}{p_{b-1}(K_{b-1})}
\]

\[
= |\mathcal{L}| + U + \sum_{b=1}^{B-1} \mathbb{E}_{K_{b-2:0}} \mathbb{E}_{K_{b-1}} \left[ \eta_0 \frac{w_{b-1}(K_{b-1})}{p_{b-1}(K_{b-1})} | K_{b-2:0} \right]
\]

\[
= |\mathcal{L}| + U + \sum_{b=1}^{B-1} \eta_0 \mathbb{E}_{K_{b-2:0}} \sum_{K_{b-1} \in \mathcal{P}_{b-1}} p_{b-1}(K_{b-1}) \frac{w_{b-1}(K_{b-1})}{p_{b-1}(K_{b-1})}
\]

\[
\leq |\mathcal{L}| + U + \sum_{b=1}^{B-1} \eta_0 N \mathbb{E}_{K_{b-2:0}} w_{b-1}(i^{b-1})
\]

(60)

for the controller \( r^{b-1} = \arg \max_{i \in \mathcal{P}_{b-1}} w_{b-1}(i) \). The first equality is because \( K_{b-1}, \ldots, K_{b+1} \) does not affect on \( I(K_b \neq K_{b-1}) \) and the second inequality is by (57). Taking expectation of (58) with respect to \( K_{b-1:0} \) and applying it to (60) yields

\[
\mathbb{E}_{K_{b-1:0}} \sum_{b=1}^{B-1} I(K_b \neq K_{b-1}) = |\mathcal{L}| + O(\eta_0 NT)
\]

by (59). Using \( |\mathcal{L}| = O(U) \) in Lemma B.3 and \( U \leq |\mathcal{U}| \) completes the proof.

Algorithm 3 uses the same distribution with Algorithm 1 if \( b = 0 \) or \( s_b \neq s_{b-1} \) or \( \mathcal{P}_b \neq \mathcal{P}_{b-1} \). It turns out that even if \( s_b = s_{b-1} \) and \( \mathcal{P}_b = \mathcal{P}_{b-1} \), the distribution of policy from Algorithm 1 and 3 are indeed the same, which is motivated by Anava et al. (2015). For the sake of completeness, we state the lemma in this paper.
Lemma D.2. Let \( p_b \) and \( \tilde{p}_b \) denote the distribution of policy at batch \( b = 0, \ldots, B - 1 \) resulting from Algorithm 1 and 3, respectively. Then, \( p \) and \( \tilde{p} \) are the same distribution.

Proof. For \( b = 0 \), \( p_0(k) = \tilde{p}_0(k) = \frac{1}{B} \) for all \( k \in \mathcal{P}_0 \). For all \( b = 1, \ldots, B - 1 \) such that \( s_b \neq s_{b-1} \) or \( \mathcal{P}_b \neq \mathcal{P}_{b-1} \), it holds that \( p_b = \tilde{p}_b \). Thus, it suffices to prove the induction step for \( b = 1, \ldots, B - 1 \) such that \( s_b = s_{b-1} \) and \( \mathcal{P}_b = \mathcal{P}_{b-1} \). Define \( Y_b := \sum_{k \in \mathcal{P}_b} \exp(-\eta_b W_b(k)) \) and suppose that \( p_{b-1} = \tilde{p}_{b-1} \). Thus, we have

\[
\tilde{p}_b(k) = \tilde{p}_{b-1}(k) \cdot \frac{\exp(-\eta_b W_b(k))}{\exp(-\eta_{b-1} W_{b-1}(k))} + p_b(k) \cdot \sum_{i \in \mathcal{P}_b} (1 - \frac{\exp(-\eta_b W_b(i))}{\exp(-\eta_{b-1} W_{b-1}(i))}) \cdot \tilde{p}_{b-1}(i)
\]

\[
= p_{b-1}(k) \cdot \frac{\exp(-\eta_b W_b(k))}{\exp(-\eta_{b-1} W_{b-1}(k))} + p_b(k) \cdot \sum_{i \in \mathcal{P}_b} (1 - \frac{\exp(-\eta_b W_b(i))}{\exp(-\eta_{b-1} W_{b-1}(i))}) \cdot p_{b-1}(i)
\]

\[
= \frac{\exp(-\eta_{b-1} W_{b-1}(k))}{Y_{b-1}} \cdot \frac{\exp(-\eta_b W_b(k))}{\exp(-\eta_{b-1} W_{b-1}(k))} + \frac{\exp(-\eta_b W_b(k))}{Y_b} \sum_{i \in \mathcal{P}_b} (1 - \frac{\exp(-\eta_b W_b(i))}{\exp(-\eta_{b-1} W_{b-1}(i))}) \cdot \frac{\exp(-\eta_{b-1} W_{b-1}(k))}{Y_{b-1}}
\]

\[
= \exp(-\eta_b W_b(k)) \cdot \frac{1}{Y_{b-1}} + \frac{\exp(-\eta_b W_b(k))}{Y_b} \sum_{i \in \mathcal{P}_b} (1 - \frac{\exp(-\eta_b W_b(i))}{\exp(-\eta_{b-1} W_{b-1}(i))}) \cdot \frac{1}{Y_{b-1}}
\]

\[
= \exp(-\eta_b W_b(k)) \cdot \frac{1}{Y_{b-1}} = p_b(k),
\]

where the first equality is due to the law of total probability, the second equality is due to the induction hypothesis, and the fifth equality is by \( \mathcal{P}_b = \mathcal{P}_{b-1} \). Notice that \( s_b = s_{b-1} \) yields \( \eta_b = \eta_{b-1} \) and \( W_b(k) \geq W_{b-1}(k) \), and thus \( 0 \leq \frac{\exp(-\eta_{b-1} W_{b-1}(k))}{\exp(-\eta_{b-1} W_{b-1}(k))} \leq 1 \); i.e., the probability distribution is properly defined for every batch. This completes the proof.

Theorem D.3 (Restatement of Theorem 4.11, Regret with switching costs bound with known \(|\mathcal{U}|\)). In Algorithm 3, let \( \tau_0 = \lfloor \frac{z}{N(|\mathcal{U}|+1)} \rfloor \) and \( \tau_b = \lfloor \frac{b \tau_0 + z}{N(|\mathcal{U}|+1)} \rfloor \) for every \( b \geq 1 \) with the constants \( z, \nu > 0 \) that satisfies \( \tau_0 > 0 \) and \( \frac{\tau_0}{\tau_0} (\beta(\tau_0))^2 < \frac{1}{2\nu^2} \). Also, let \( \eta_0 = O(\frac{\eta_1^{(1)}}{\mathcal{T}^{2/3} N^{1/3} d^{1/3}}) \). When \( T \geq \max\{ (\frac{\alpha(1)}{\eta_0^{(1)}}) \exp(O(|\mathcal{U}|))^{3/2} / (N(|\mathcal{U}|+1)^{1/2} \cdot \frac{|\mathcal{U}|^{3/2} \cdot \mathcal{P} \cdot \mathcal{T}^{1/3} N^{1/3} d^{1/3}}{2}), N(|\mathcal{U}|+1) d \} \), we have

\[
\mathbb{E}_{K_{b-1}} \left[ \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x_t', u_t')] + d \sum_{b=1}^{B-1} \mathcal{I}_{(K_b \neq K_{b-1})} - d \sum_{l=1}^{\mathcal{S}} \mathcal{I}_{(i_l \neq i_{l-1})} \right]
\]

\[
= \tilde{O}(T^{2/3} N^{1/3} (|\mathcal{U}| + 1)^{1/3} d^{1/3}) + o(T),
\]

which implies that we achieve a sublinear regret bound. Moreover, when \( \lim_{t \to \infty} H(t) < \infty \) and \( T \geq \max\{ (\frac{\exp(O(|\mathcal{U}|))}{\mathcal{T}^{3/2} \cdot \frac{d^{1/3}}{\mathcal{P} \cdot \mathcal{T}^{1/3} N^{1/3} d^{1/3}}}), N(|\mathcal{U}|+1) d \} \), we have

\[
\mathbb{E}_{K_{b-1}} \left[ \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x_t', u_t')] + d \sum_{b=1}^{B-1} \mathcal{I}_{(K_b \neq K_{b-1})} - d \sum_{l=1}^{\mathcal{S}} \mathcal{I}_{(i_l \neq i_{l-1})} \right] = \tilde{O}(T^{2/3} N^{1/3} (|\mathcal{U}| + 1)^{1/3} d^{1/3}).
\]

Proof. The distribution of policy is the same for Algorithm 1 and 3 by Lemma D.2. Thus, we can use Theorem B.8 with Lemma D.1 to achieve

\[
\mathbb{E}_{K_{b-1}} \left[ \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x_t', u_t')] + d \sum_{b=1}^{B-1} \mathcal{I}_{(K_b \neq K_{b-1})} - d \sum_{l=1}^{\mathcal{S}} \mathcal{I}_{(i_l \neq i_{l-1})} \right]
\]

\[
\leq \frac{\tilde{O}(|\mathcal{U}| + 1)}{\eta_0} + \frac{\eta_0 N}{2} \exp(O(|\mathcal{U}|)) O(\tau_{B-1} H(\tau_{B-1})) + O(\sum_{b=0}^{B-1} (\tau_b)^2)
\]
Online Bandit Control with Dynamic Batch Length and Adaptive Learning Rate

\[ + O\left(\sum_{b=0}^{B-1} H(\tau_b)\right) + O(d|U|) + O(d\eta_0 NT), \]  

(61)

since \( d \geq 1 \) and \( \sum_{d=1}^{L} I_{\{i|\neq I_{i-1}\}} \geq 0 \). Notice that \((\tau_b)_{b \geq 0}\) is the same for Algorithm 1 and 3. Accordingly, we still have \( B = O(T^{2/3} N^{1/3}(|U| + 1)^{1/3}) \) by (44) and \( T \geq N(|U| + 1) d \geq N(|U| + 1) \). We also still have (45) and (46). Thus, with \( T \geq \frac{\|\|d\|^3/2}{(N(|U|+1))/2} \) and \( T \geq \frac{|\|d\|^3/2}{(N(|U|+1))/2} \), we obtain that

\[ \eta_0 N \exp(O(\|U\|)) O(\tau_B - 1 H(\tau_B - 1)) = o(d^{-1/3}) \exp(O(\|U\|)) = O(T^{2/3} N^{1/3}(|U| + 1)^{1/3} d^{-1/3}). \]

\[ O(d|U|) = O(T^{2/3} N^{1/3}(|U| + 1)^{1/3} d^{1/3}). \]

Also, with \( T \geq N(|U| + 1) d \), we have

\[ O(d\eta_0 NT) = O(T^{2/3} N^{1/3}(|U| + 1)^{1/3} d^{1/3}). \]

Combining all the above equalities with (61), one can write

\[ E_{KB-1;0} \left[ \sum_{t=0}^{T} [c_t(x_t, u_t) - c_t(x'_t, u'_t)] + d \sum_{b=1}^{B-1} I_{\{K_b \neq K_{b-1}\}} - d \sum_{t=1}^{T} I_{\{I_{t} \neq I_{t-1}\}} \right] \]

\[ = O(T^{2/3} N^{1/3}(|U| + 1)^{1/3} d^{1/3}) + O(\sum_{b=0}^{B-1} H(\tau_b)). \]

Using (16) shows a sublinear regret bound. When \( \lim_{t \rightarrow \infty} H(t) < \infty \), (51) is modified to

\[ \eta_0 N \exp(O(\|U\|)) O(\tau_B - 1 H(\tau_B - 1)) = O(T^{2/3} N^{1/3}(|U| + 1)^{1/3} d^{1/3}), \]

only with \( T \geq \frac{\exp(O(\|U\|))}{d^{2/3}} \). This completes the proof. \( \square \)

E. Numerical Experiment Details

E.1. Experiments for the Linear system

In this subsection, we introduce the implementation details and present more experiments on the linear system (7) discussed in Example 1 of Section 5.

We consider three different noises for the experiments. To perform a fair comparison, the bounding constant \( w_{\text{max}} \) is set to 1.

(a) Sanity check: Gaussian noise with mean 0.3 and standard deviation 0.1, truncated to \([-0.4, 1] \]

(b) Sinusoidal noise \( w_t = \left[ \sin \left( \frac{t}{5\pi} \right), \sin \left( \frac{t}{11\pi} \right) \right] \)

(c) Uniform random walk, where \( w_0 = \text{Uniform} \left[ \frac{1}{3} - \frac{2}{3T}, \frac{1}{3} + \frac{2}{3T} \right] \) and \( w_t - w_{t-1} \) follows Uniform \( \left[ -\frac{2}{3T}, \frac{2}{3T} \right] \)

where \( T \) is time horizon. One can easily see that for uniform random walk, \(|w_T| \leq 1 \) for any \( T \). Notice that we use statistical (Gaussian) noise for the sanity check, and the rest are the adversarial disturbances.

We perform the ablation study of Algorithm 1, which means that we consider four scenarios: (fixed, dynamic) batch length and (fixed, adaptive) learning rate. For all the experiments implementing the algorithm, we use \( T = 3000, \eta_0 = 0.025, \gamma = 2.5, \alpha_0 = 1.01, \) and \( x_0 = [100, 200]^T \). For the dynamic batch length, we consider \( \tau_0 = 11 \) and \( \tau_0 = \lceil \tau_0 \cdot \left( \frac{10^5}{10^4} \right) \rceil \). It is well known that every (asymptotically) stabilizing controller in the linear system is indeed exponentially stabilizing controller (Khalil, 2015). Hence, we use \( \beta(t) = 0.99^t \) without relaxing the assumptions on stabilizing controllers. Finally, we use \( \delta = \frac{\eta_{\text{max}}}{\tau_0 \cdot \eta_0} \). Since the sinusoidal noise case is already presented in Figure 2, we only present truncated Gaussian noise case and uniform random walk case here.

In Figures 2, 4, and 5, we observe that each component of DBAR, a dynamic batch length and an adaptive learning rate, respectively and jointly improves both the stability and the regret regardless of the noise form. Specifically, a dynamic batch
length delays the time that large state norms occur during learning, and simultaneously stabilizes that state norm. This can be observed in Figures 2(d), 4(d), and 5(d) when comparing fixed and dynamic batch lengths under an adaptive learning rate. This is because we use a non-decreasing batch length, but the increasing ratio between two consecutive batch lengths is determined to converge to 1 (see Assumption 3.1). On the other hand, an adaptive learning rate effectively lowers the state norm at the time that large state norms occur without delay, since the learning rate adaptively decreases whenever the agent faces large state norm. This can be seen in 2(c), 4(c), and 5(c), the ablation study about the comparison between fixed and adaptive learning rates under a dynamic batch length. Thus, DBAR effectively stabilizes the state norm below $\gamma w_{max}$ and minimizes the regret, where the two components support each other.

E.2. Experiments for the Nonlinear system

In this subsection, we introduce the implementation details and present more experiments on the nonlinear ball-beam system introduced in Example 2 of Section 5. To study this continuous-time nonlinear system, we first derive the first-order state representation of (8) with the states $(y_1, y_2, y_3, y_4) = (x, \dot{x}, -9.81B\theta, -9.81B\dot{\theta}) \in \mathbb{R}^4$ and the action $v = -9.81Bu$:

$$
\dot{y}_1 = y_2, \quad \dot{y}_2 = 9.81B \sin\left(\frac{y_3}{9.81B}\right) + \frac{y_1 y_3^2}{B(9.81)^2} + 3w, \quad \dot{y}_3 = y_4, \quad \dot{y}_4 = v,
$$

where $w$ is a sinusoidal noise $\sin\left(\frac{t}{7}\right)$ and $w_{max} = 1$. A nested saturating control policy is known to successfully stabilize the ball-beam system if the correct parameters are given, but it does not necessarily exponentially stabilize the system (Barbu et al., 1997). This necessitates our approach of extending the notion of stabilizing controllers beyond exponential assumptions. In this experiment, we aim to learn the parameters of the best stabilizing controller. We choose a nested

Figure 4. The stability and the regret in the linear system under truncated Gaussian noise. Ablation study of the algorithm is presented.

Figure 5. The stability and the regret in the linear system under Uniform random walk. Ablation study of the algorithm is presented.
Online Bandit Control with Dynamic Batch Length and Adaptive Learning Rate

(a) Stability analysis under $\beta_1(t)$ (b) Regret analysis under $\beta_1(t)$ (c) Stability analysis under $\beta_2(t)$ (d) Regret analysis under $\beta_2(t)$

Figure 6. The stability and the regret in the noise-injected ball-beam system under sinusoidal noise and the choice of $\beta_1(t)$ or $\beta_2(t)$.

The different behaviors of $\beta$ eliminating potential destabilizing controllers not yet seen in an unstable region in advance. The stability and the regret in the noise-injected ball-beam system under sinusoidal noise and the choice of $\beta_1(t)$ or $\beta_2(t)$.

We again perform the ablation study of Algorithm 1. For the experiments implementing the algorithm, we use where $\beta$ jointly components of DBAR adaptive learning rate stabilizes the system faster than only using the adaptive learning rate. This again shows that the two

We again perform the ablation study of Algorithm 1. For the experiments implementing the algorithm, we use where $\beta$ jointly components of DBAR adaptive learning rate stabilizes the system faster than only using the adaptive learning rate. This again shows that the two

saturating control policy $v'$ determined by three positive parameters ($p, k_1, k_2$):

$$
\epsilon = \frac{1}{\sqrt{1 + y_t^2 + z_t^2}}, \quad p_1 = p, \quad p_2 = \frac{p}{\epsilon}, \quad p_3 = \frac{p}{\epsilon^2}, \quad p_4 = \frac{p}{\epsilon^3},
$$

$$
z_1 = y_1 + k_1y_2 + k_1y_3 + y_4, \quad z_2 = y_2 + k_2y_3 + y_4, \quad z_3 = y_3 + y_4, \quad z_4 = y_4,
$$

$$
v' = \sigma_p(z_4 + \sigma_{p_3}(z_2 + \sigma_{p_2}(z_2 + \sigma_{p_1}(z_1))));
$$

where $\sigma_p(z)$ is the saturating function defined as $p$ if $z > p$, $-p$ if $z < -p$, and $z$ if $|z| \leq p$. We consider the controller pool $V' = \{v' : p \in \{2, 16, 30, 44, 58, 72, 86, 100\}, k_1 \in \{2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5\}, k_2 \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5\}\}$, which has a total of 800 controllers. Among them, we do not know if a controller stabilizes the system. For simplicity, we perform forward-Euler discretization on the system with a sampling time $0.01$. The resulting discrete-time states and actions are denoted by $y_t$ and $v_t$ at $t$th sampling time. We use the cost function $c_t(y_t, v_t) = ||y_t^2||^2 + ||v_t^2||^2$ to stabilize the ball position and the beam angle towards 0.

We again perform the ablation study of Algorithm 1. For the experiments implementing the algorithm, we use $T = 5000$, $\eta_0 = 0.025$, $\gamma = 1.5$, $\alpha_0 = 1.01$, and $y^0 = [-32, 24, 5.6, 24]$. For the dynamic batch length, we consider $\tau_0 = 9$ and $\tau_6 = [\tau_0 \cdot (\frac{24 + \gamma}{40})^{0.5}]$.

Unlike the choice of $\beta(t)$ in Section E.1, we select the stabilizing controller only to satisfy (asymptotic) ISS in Definition 2.3, instead of exponential ISS. To deeply study this notion, we consider two scenarios with regards to $\beta(t)$:

$$
\beta_1(t) = \max\{1 - (0.01)^{0.6t}, 0\}, \quad \beta_2(t) = \max\{1 - (0.01)^{0.75t}, 0\}
$$

Figure 3(b) shows the stability analysis of the system under $\beta_1(t)$. For the completeness, we present the same picture in Figure 6(a). In our experiment, there are 225 controllers out of 800 controllers that induces the system to explode, starting from the initial state. However, there exist far more destabilizing controllers within this pool, since most of 575 controllers are only locally stabilizing controllers, meaning that the system is stabilized only at some initial states. Thus, with only few stabilizing controllers in the pool, Figure 6 demonstrates that our algorithm DBAR successfully stabilizes the state norm below $\gamma w_{\text{max}}$ and minimizes the regret for both $\beta_1(t)$ and $\beta_2(t)$. For $\beta_1(t)$, a dynamic batch length fails to stabilize the system, whereas an adaptive learning rate successfully learns stabilizing controllers. This is because an adaptive learning rate directly addresses large state norms, while dynamic batch length tries to address them indirectly via time delay. Although the dynamic batch length does not contribute to the improved stability and regret, adding the dynamic batch length to the adaptive learning rate stabilizes the system faster than only using the adaptive learning rate. This again shows that the two components of DBAR jointly improves the stability and the regret. For $\beta_2(t)$, both dynamic batch length and adaptive learning rate contribute to the desirable system stability and the regret, and combining the two further improves the system to learn the best stabilizing controller faster.

The different behaviors of $\beta_1(t)$ and $\beta_2(t)$ come from the amount of discarding the destabilizing controllers. $\beta_2(t)$ removes the controller with stricter criteria than $\beta_1(t)$ since $0.75 > 0.6$. This prevents the explosion of the nonlinear system by eliminating potential destabilizing controllers not yet seen in an unstable region in advance.