Safe Adaptive Learning for Linear Quadratic Regulators with Constraints

Yingying Li[1], Tianpeng Zhang[3], Subhro Das[2], Jeff Shamma[1], and Na Li[3]

[1]University of Illinois at Urbana-Champaign
[2]MIT-IBM Watson AI Lab, IBM Research
[3]Harvard University

Abstract

This paper considers single-trajectory adaptive/online learning for linear quadratic regulator (LQR) with an unknown system and constraints on the states and actions. The major challenges are two-fold: 1) how to ensure safety without restarting the system, and 2) how to mitigate the inherent tension among exploration, exploitation, and safety. To tackle these challenges, we propose a single-trajectory learning-based control algorithm that guarantees safety with high probability. Safety is achieved by robust certainty equivalence and a SafeTransit algorithm. Further, we provide a sublinear regret bound compared with the optimal safe linear policy. By this, we solve an open question in Dean et al. (2019a). When developing the regret bound, we also establish a novel estimation error bound for nonlinear policies, which can be interesting on its own. Lastly, we test our algorithm in numerical experiments.

1 Introduction

Recent years have witnessed great interest in learning-based control, and a lot of results have been developed for *unconstrained* systems (Fazel et al., 2018; Dean et al., 2018, 2019b; Mania et al., 2019; Simchowitz et al., 2018, 2020; Cohen et al., 2019). However, practical systems usually face *constraints* on the states and control inputs, especially in safety-critical applications (Campbell et al., 2010; Vasic and Billard, 2013). For example, drones are not supposed to visit certain locations to avoid collision and the thrusts of drones are usually bounded. Therefore, it is crucial to study *safe* learning-based control for *constrained* systems.

As a starting point, this paper considers LQR with linear constraints on the states and actions, i.e.,

$$D_x x_t \leq d_x, \qquad D_u u_t \leq d_u.$$

We consider a linear system $x_{t+1} = A_* x_t + B_* u_t + w_t$ with bounded disturbances $w_t \in \mathbb{W} = \{w : \|w\|_\infty \leq w_{\max}\}$ and unknown model (A_*, B_*). We aim to design an adaptive control algorithm to minimize the quadratic cost $\mathbb{E}[x_t^\top Q x_t + u_t^\top R u_t]$ with *safety* guarantees during the learning process, i.e. satisfying the constraints for any $w_t \in \mathbb{W}$.

The constraints on LQR bring great difficulties even when the model is known. Unlike unconstrained LQR, which enjoys closed-form optimal policies (Lewis et al., 2012), there is no computationally efficient method to solve the optimal policy for constrained LQR (Rawlings and Mayne, 2009).[1] Thus, most literature sacrifices optimality for computation efficiency by designing policies with certain structures, e.g. linear policies (Dean et al., 2019a; **?**), piecewise-affine policies in robust model predictive control (RMPC) (Bemporad and Morari, 1999; Rawlings and Mayne, 2009), etc. Therefore, when the model is unknown, a reasonable goal is to learn and achieve what can be obtained with perfect model information. In this paper, we adopt the optimal safe linear policy as our benchmark/target and leave the discussions on RMPC as future work.

The current literature on learning an optimal safe linear policy adopts an offline/non-adaptive learning approach, which does not improve the policies until the learning terminates (Dean et al., 2019a). To improve the control performance during learning, adaptive/online learning-based control algorithms should be designed. However, though

[1] Efficient algorithms exist for some special cases, e.g. when $w_t = 0$, the optimal controller is piecewise-affine and can be computed as in Bemporad et al. (2002).

adaptive learning for unconstrained LQR can be designed by direct conversions from offline algorithms (see e.g., (Simchowitz and Foster, 2020; Mania et al., 2019; Dean et al., 2018)), it is much more challenging for the constrained case because direct conversions may cause infeasibility and/or constraint violation for single-trajectory adaptive learning as noted in Dean et al. (2019a).

Our contributions. In this paper, we propose a single-trajectory adaptive learning algorithm for constrained LQR. Our algorithm ensures safety on a single trajectory without restarts by certainty-equivalence (CE) with robust constraint satisfaction against system uncertainties and a novel SafeTransit algorithm for safe policy updates.

Theoretically, for safety, we guarantee feasibility and constraint satisfaction with high probability. Constraint satisfaction is self-explanatary. Feasibility means our algorithm admits some feasible solution at every stage (this is nontrivial when constrained optimizations are solved). For non-asymptotic optimality, we provide a sublinear regret bound of order $\tilde{O}(T^{2/3})$ compared with the optimal safe linear policy with perfect model information for horizon T. With the results above, we solve an open question in Dean et al. (2019a), i.e., how to design safe online learning for constrained LQR with (recursive) feasibility and non-asymptotic optimality guarantees.

Interestingly, our regret bound also holds when compared against a special RMPC algorithm proposed in Mayne et al. (2005). Discussions on more general regret benchmarks are left for the future.

Technically, when developing our theoretical results, we provide a model estimation error bound for general and possible *nonlinear* policies. This is to handle the potential nonlinearity in our designed controllers when the model errors are non-negligible. Our error bound extends the existing results for linear policies in Dean et al. (2019b,a) and can be useful on its own.

Lastly, we numerically test our algorithm on a quadrotor vertical flight system to complement our theoretical results.

Related work. *Constrained LQR with linear policies* is studied in Dean et al. (2019a); **?**. Dean et al. (2019a) consider an *unknown* model and propose an offline learning method with sample complexity guarantees. In contrast, **?** study online constrained LQR with a *known* model and time-varying cost functions. However, it remains open how to design online learning algorithms to compute safe linear policies under model uncertainties.

Constrained LQR by model predictive control (MPC). MPC and its variants are popular methods for constrained control, e.g. RMPC designed for hard constraints (Mayne et al., 2005; Limon et al., 2010; Rawlings and Mayne, 2009), and stochastic MPC methods for soft constraints (Mesbah, 2016; Oldewurtel et al., 2008). With model uncertainties, robust adaptive MPC (RAMPC) algorithms are proposed to learn the model and update the policies (Zhang and Shi, 2020; Bujarbaruah et al., 2019; Köhler et al., 2019; Lu et al., 2019). Most RAMPC algorithms guarantee recursive feasibility and constraint satisfaction but lack non-asymptotic performance guarantees compared with the known-model case. There is an interesting recent paper (**?**) that provides a regret bound for learning-based MPC, but its benchmark policy is conservative since it is robustly safe for all uncertain models instead of just for the true model. In contrast, there are some recent results on non-asymptotic regret bounds by sacrificing feasibility and/or constraint satisfaction, e.g., Wabersich and Zeilinger (2020) establish a regret bound for an adaptive MPC algorithm that requires restarting the system to some safe feasible state, Muthirayan et al. (2020) provides a regret bound for an adaptive algorithm without considering state constraints.

Learning-based unconstrained LQR enjoys rich literature, so we only review the most related papers below. Firstly, our algorithm is related with the CE-based adaptive control (Dean et al., 2018; Mania et al., 2019; Cohen et al., 2019; Simchowitz and Foster, 2020), and this approach is shown to be optimal for the unconstrained LQR (Mania et al., 2019; Simchowitz and Foster, 2020). Further, similar to Agarwal et al. (2019a,b); Plevrakis and Hazan (2020), we adopt the disturbance-action policies to approximate linear policies.

Safe reinforcement learning (RL). Safety in RL has different definitions (Mihatsch and Neuneier, 2002; Garcıa and Fernández, 2015). This paper is related with RL with constraints (Marvi and Kiumarsi, 2021; Leurent et al., 2020; Fisac et al., 2018; Garcıa and Fernández, 2015; Cheng et al., 2019; Fulton and Platzer, 2018). Safe RL usually allows complex system dynamics, but there are limited results on hard constraint satisfaction with non-asymptotic optimality bounds.

Model estimation for nonlinear systems. There are model estimation bounds for general nonlinear systems (Foster et al., 2020; Sattar and Oymak, 2020), but our estimation error bound leverages the special structure of our problem: nonlinear policies on a linear system, to obtain better bounds.

Notations. For a distribution $\mathcal{D}_\eta$, we write $\eta \overset{\text{ind.}}{\sim} \bar{\eta}\mathcal{D}_\eta$ if $\eta/\bar{\eta}$ is generated with distribution $\mathcal{D}_\eta$ and is independent from other random variables in the context. Let $\|\cdot\|_F$ denote the Frobenius norm and $\|\cdot\|_p$ denote the l_p norm for $1 \leq p \leq \infty$. Define $\mathbb{B}(\hat{\theta}, r) = \{\theta : \|\theta - \hat{\theta}\|_F \leq r\}$. Let $\mathbb{1}_n$ be an all-one vector in $\mathbb{R}^n$. For any set E and $x \in E$, let $\mathbb{I}_E(x)$ denote an indicator function, i.e., $\mathbb{I}_E(x) = 1$ if $x \in E$ and $\mathbb{I}_E(x) = 0$ otherwise. Further, for vector $y \in \mathbb{R}^n$ and set $E \subseteq \mathbb{R}^n$, let $\Pi_E(y)$ denote the projection onto set E in l_2 norm, i.e., $\Pi_E(y) = \arg\min_{z \in E} \|z - y\|_2^2$.

2 Problem formulation

We consider the following constrained LQR problem,

$$\begin{aligned} &\min_{u_0,u_1,\dots} \lim_{T\to+\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[l(x_t,u_t)] \\ &\text{s.t. } x_{t+1} = A_* x_t + B_* u_t + w_t,\ \forall\, t \geq 0, \\ &\quad D_x x_t \leq d_x, D_u u_t \leq d_u,\ \forall\, \{w_k \in \mathbb{W}\}_{k\geq 0}. \end{aligned} \tag{1}$$

where $l(x,u) = x^\top Q x + u^\top R u$, Q and R are positive definite matrices, $x_t \in \mathbb{R}^n$ is the state with a given initial state x_0, $u_t \in \mathbb{R}^m$ is the action, and w_t is the disturbance. The parameters D_x, d_x, and D_u, d_u determine the constraint sets of the state and action respectively, where $d_x \in \mathbb{R}^{k_x}, d_u \in \mathbb{R}^{k_u}$. Further, the constraint sets on the state and action are assumed to be bounded with $x_{\max} = \sup_{D_x x \leq d_x} \|x\|_2$ and $u_{\max} = \sup_{D_u u \leq d_u} \|u\|_2$. Besides, denote $\theta_* := (A_*, B_*)$ and $\theta := (A, B)$ for simplicity. The model parameters θ_* are unknown but other parameters are known.

An algorithm/controller is called 'safe' if its induced states and actions satisfy the constraints for all t under any possible disturbances $w_t \in \mathbb{W}$, which is also called robust constraint satisfaction under disturbances w_t.

Notice that even with known model θ_*, the optimal policy to problem (1) cannot be computed efficiently, but there are efficient methods to compute sub-optimal policies, e.g. optimal safe linear policies by quadratic programs (Dean et al., 2019a; **?**) and piecewise affine policies by RMPC (Mayne et al., 2005; Rawlings and Mayne, 2009). In this paper, we focus on the optimal safe linear policy as our learning goal and briefly discuss RMPC in Section **??**. We aim to achieve our learning goal by designing safe adaptive learning-based control. Further, we consider single-trajectory learning, which is more challenging since the system cannot be restarted to ensure feasibility and constraint satisfaction.

For simplicity, we assume $x_0 = 0$. Our results can be generalized to x_0 in a small neighborhood around 0.[2]

Regret and benchmark. Roughly speaking, we measure the performance of our adaptive learning algorithm by comparing it with the optimal safe linear policy $u_t = -K^* x_t$ computed with perfect model information.

To formally define the performance metric, we first define a quantitative characterization of matrix stability as in e.g., Agarwal et al. (2019a,b); Cohen et al. (2019).

Definition 1. *For $\kappa \geq 1$, $\gamma \in [0,1)$, a matrix A is called (κ,γ)-stable if $\|A^t\|_2 \leq \kappa(1-\gamma)^t, \forall t \geq 0$.*[3]

Consider the following benchmark policy set:

$$\begin{aligned} \mathcal{K} = \{K : (A_* - B_* K) \text{ is } (\kappa,\gamma)\text{-stable}, \|K\|_2 \leq \kappa \\ D_x x_t^K \leq d_x, D_u u_t^K \leq d_u, \forall\, t, \forall\, \{w_k \in \mathbb{W}\}_{k\geq 0}\}, \end{aligned}$$

where x_t^K, u_t^K are generated by policy $u_t = -K x_t$.

For any safe learning algorithm/controller $\mathcal{A}$, we measure its performance by 'regret' as defined below:

$$\text{Regret} = \textstyle\sum_{t=0}^{T-1} l(x_t^{\mathcal{A}}, u_t^{\mathcal{A}}) - T \min_{K\in\mathcal{K}} J(K)$$

where $x_t^{\mathcal{A}}, u_t^{\mathcal{A}}$ are generated by the algorithm $\mathcal{A}$ and $J(K) = \lim_{T\to+\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[l(x_t^K, u_t^K)]$.

Next, we provide and discuss the assumptions.

Assumptions. Firstly, though the model θ_* is not perfectly known, we assume there is some prior knowledge, which is captured by a bounded model uncertainty set Θ_{ini} that contains θ_*. It is widely acknowledged that without such prior knowledge, hard constraint satisfaction is extremely difficult, if not impossible (Dean et al., 2019a). We also assume that Θ_{ini} is small enough so that there exists a linear controller $u_t = -K_{\text{stab}} x_t$ to stabilize any system in Θ_{ini}.

Assumption 1. *There is a known model uncertainty set $\Theta_{\text{ini}} = \{\theta : \|\theta - \hat\theta_{\text{ini}}\|_F \leq r_{\text{ini}}\}$*[4] *for some $0 < r_{\text{ini}} < +\infty$ such that (i) $\theta_* \in \Theta_{\text{ini}}$, and (ii) there exist $\kappa \geq 1, \gamma \in [0,1)$, and K_{stab} such that for any $(A,B) \in \Theta_{\text{ini}}$, $A - BK_{\text{stab}}$ is (κ,γ)-stable.*

[2] The appendix will discuss more on nonzero x_0. Here, we discuss the implication of small x_0. Remember that state 0 represents a desirable system equilibrium. With x_0 close to 0, we study how to safely optimize the performance around the equilibrium instead of safely steering a distant state back to the equilibrium. As an example of applications, we study how to safely maintain a drone around a target in the air despite wind disturbances with minimum battery consumption, instead of flying the drone to the target from a distance. In practice, one can first apply algorithms such as Mayne et al. (2005) to drive the system to around 0, then apply our algorithm to achieve optimality and safety around 0.

[3] Agarwal et al. (2019a) call this as $(\sqrt{\kappa}, \gamma)$-strong stability.

[4] The symmetry of Θ_{ini} is assumed for simplicity and not restrictive. We only need Θ_{ini} to be small and contain θ_*.

Though requiring a robustly stabilizing K_{stab} can be restrictive, this is necessary for robust constraint satisfaction during our safe adaptive learning. Besides, this assumption is common in the safe adaptive control literature for constrained LQR, e.g., (Köhler et al., 2019; Lu et al., 2019). Lastly, K_{stab} can be computed by, e.g., linear matrix inequalities (LMIs) (see Caverly and Forbes (2019) as a review).

Further, we need to assume a feasible linear policy exists for our constrained LQR (1), otherwise our regret benchmark is not well-defined. Here, we impose a slightly stronger assumption of strict feasibility to allow approximation errors in our control design. This assumption can also be verified by LMIs (Caverly and Forbes, 2019).

Assumption 2. *There exists $K_F \in \mathcal{K}$ and $\epsilon_{F,x} > 0, \epsilon_{F,u} > 0$ such that $D_x x_t^{K_F} \leq d_x - \epsilon_{F,x}\mathbb{1}_{k_x}$ and $D_u u_t^{K_F} \leq d_u - \epsilon_{F,u}\mathbb{1}_{k_u}$ for all $t \geq 0$ under all $w_k \in \mathbb{W}$.*

Lastly, we impose assumptions on disturbance w_t. We assume a certain anti-concentration property around 0 as in (Abeille and Lazaric, 2017), which essentially requires the random vector w to have large enough probability at a distance from 0 in all directions and is essential for our estimation error bound for general policies.

Definition 2 (Anti-concentration)**.** *A random vector $w \in \mathbb{R}^n$ satisfies (s,p)-anti-concentration for some $s > 0, p \in (0,1)$ if $\mathbb{P}(\lambda^\top w \geq s) \geq p$ for any $\|\lambda\|_2 = 1$.*

Assumption 3. *$w_t \in \mathbb{W}$ is i.i.d., σ_{sub}^2-sub-Gaussian, zero mean, and (s_w, p_w)-anti-concentration.*[5]

Remark 1 (On bounded disturbances)**.** *Here, we assume bounded disturbances because we aim to achieve constraint satisfaction despite any disturbances, which is generally impossible for unbound disturbances. Nevertheless, we can allow Gaussian disturbances by considering chance constraints as discussed in Oldewurtel et al. (2008).*

2.1 Preliminaries

2.1.1 Approximation of optimal safe linear polices with disturbance-action policies. This section reviews a computation-efficient method for approximating the optimal safe linear policy when the model θ_* is known **?**. The method is based on the disturbance-action policy (DAP) defined below.

$$u_t = -K_{\text{stab}} x_t + \textstyle\sum_{k=1}^{H} M[k] w_{t-k}, \tag{2}$$

where $\boldsymbol{M} = \{M[1], \ldots, M[H]\}$ denote the policy parameters and $H \geq 1$ denotes the policy's memory length.

Roughly, **?** show that the optimal safe linear policy can be approximated by the optimal safe DAP for large H, and the optimal safe DAP can be solved efficiently by a quadratic program (QP). More details are as follows.[6]

Firstly, Proposition 1 shows that the state and action can be approximated by affine functions of DAP parameters $\mathbf{M}$.

Proposition 1 (Agarwal et al. (2019a))**.** *Under a time-invariant DAP $\mathbf{M}$, the state x_t and action u_t can be approximated by affine functions on $\mathbf{M}$: $\tilde{x}_t(\mathbf{M};\theta_*) = \sum_{k=1}^{2H} \Phi_k^x(\mathbf{M};\theta_*) w_{t-k}$ and $\tilde{u}_t(\mathbf{M};\theta_*) = \sum_{k=1}^{2H} \Phi_k^u(\mathbf{M};\theta_*) w_{t-k}$, where $A_{K*} = A_* - B_* K_{\text{stab}}$, $\Phi_k^x(\mathbf{M};\theta_*) = A_{K*}^{k-1}\mathbb{I}_{(k\leq H)} + \sum_{i=1}^{H} A_{K*}^{i-1} B_* M[k-i]\mathbb{I}_{(1\leq k-i\leq H)}$ and $\Phi_k^u(\mathbf{M};\theta_*) = -K_{stab}\Phi_k^x(\mathbf{M};\theta_*) + M[k]\mathbb{I}_{(k\leq H)}$. The approximation errors are $O((1-\gamma)^H)$.*

Secondly, we review the polytopic *safe policy set* (SPS) in **?** by imposing constraints on the approximate states and actions and by tightening the constraints to allow for approximation errors. Specifically, define constraint functions on $\tilde{x}_t, \tilde{u}_t$, i.e., $g_i^x(\mathbf{M};\theta_*) = \sup_{w_k \in \mathbb{W}} D_{x,i}^\top \tilde{x}_t(\mathbf{M};\theta_*) = \sum_{k=1}^{2H} \|D_{x,i}^\top \Phi_k^x(\mathbf{M};\theta_*)\|_1 w_{\max}$ for $1 \leq i \leq k_x$ and $g_j^u(\mathbf{M};\theta_*) = \sup_{w_k \in \mathbb{W}} D_{u,j}^\top \tilde{u}_t(\mathbf{M};\theta_*) = \sum_{k=1}^{2H} \|D_{u,j}^\top \Phi_k^u(\mathbf{M};\theta_*)\|_1 w_{\max}$ for $1 \leq j \leq k_u$. The SPS in **?** is defined as follows.

$$\begin{aligned}\Omega(\theta_*, \epsilon_x, \epsilon_u) = \{\mathbf{M} \in \mathcal{M}_H \colon & g_i^x(\mathbf{M};\theta_*) \leq d_{x,i} - \epsilon_*^x, 1 \leq i \leq k_x, \\ & g_j^u(\mathbf{M};\theta_*) \leq d_{u,j} - \epsilon_*^u, 1 \leq j \leq k_u.\},\end{aligned} \tag{3}$$

where $\mathcal{M}_H = \{\mathbf{M} : \|M[k]\|_\infty \leq 2\sqrt{n}\kappa^2(1-\gamma)^{k-1}, \forall\, 1 \leq k \leq H\}$ is introduced for stability and technical simplicity, $\epsilon_*^x(H) = O((1-\gamma)^H)$ and $\epsilon_*^u(H) = O((1-\gamma)^H)$ are constraint-tightening terms to account for the approximation errors when the model θ_* is available. Notice that (3) defines a polytopic set of the policy parameter $\mathbf{M}$.

[5] By $\mathbb{W} = \{w : \|w\|_\infty \leq w_{\max}\}$, we have $\sigma_{sub} \leq \sqrt{n} w_{\max}$.

[6] There are other methods to approximately compute the optimal safe linear policy under similar ideas: let u_t be affine on history disturbances (see e.g., Dean et al. (2019a); Mesbah (2016)).

Finally, we review the QP reformulation for the optimal safe DAP $\mathbf{M}^*$ when the model θ_* is available.

$$\mathbf{M}^* = \underset{\mathbf{M}\in\Omega(\theta_*,\epsilon_*^x,\epsilon_*^u)}{\arg\min} \; f(\mathbf{M};\theta_*) = \mathbb{E}[l(\tilde{x}_t(\mathbf{M},\theta_*),\tilde{u}_t(\mathbf{M},\theta_*)] \tag{4}$$

Notice that $f(\mathbf{M};\theta_*)$ is a quadratic convex function of $\mathbf{M}$. Further, **?** showed that the optimal safe DAP $\mathbf{M}^*$ approximates the optimal safe linear policy K^* with error $J(\mathbf{M}^*) - J(K^*) \leq O((1-\gamma)^H)$.

2.1.2 Safe slowly varying policies. Notice that the SPS in (3) only considers a *time-invariant* policy $\mathbf{M}$. Nevertheless, **?** show that with *additional constraint-tightening* terms $\epsilon_v^x(\Delta_M), \epsilon_v^u(\Delta_M)$ to allow for small policy variation Δ_M, the SPS can also be used to guarantee the safety of *slowly varying* policy sequences, which is called a slow-variation trick.

Lemma 1 (Slow-variation trick (**?**)). *Consider a slowly varying DAP sequence $\{\mathbf{M}_t\}_{t\geq 0}$ with $\|\mathbf{M}_t - \mathbf{M}_{t-1}\|_F \leq \Delta_M$, where Δ_M is called the policy variation budget. $\{\mathbf{M}_t\}_{t\geq 0}$ is safe to implement if $\mathbf{M}_t \in \Omega(\theta_*, \epsilon_x, \epsilon_u)$ for all $t \geq 0$, where $\epsilon_x \geq \epsilon_*^x + \epsilon_v^x(\Delta_M)$, $\epsilon_u \geq \epsilon_*^u + \epsilon_v^u(\Delta_M)$, and $\epsilon_v^x(\Delta_M), \epsilon_v^u(\Delta_M) = O(\sqrt{H}\Delta_M)$.*

3 Safe Adaptive Control Algorithm

This section introduces our safe adaptive control algorithm for constrained LQR in Algorithm 1.

Algorithm overview. The high-level algorithm structure is standard in model-based adaptive learning-based control (Mania et al., 2019; Simchowitz and Foster, 2020). That is, first solve a near-optimal policy ($\mathbf{M}_\dagger^e$) by the certainty equivalence (CE) principle (Line 3), then collect data by implementing this policy for a while (Line 5-6), then update the model uncertainty set (Θ^{e+1}) with the newly collected data (Line 7), then update the policy with the new model estimation (Line 8), collect more data with the updated policy (Line 10-11), and so on.

However, the constraints in our problem bring additional challenges on safety and the explore-exploit tradeoff under the safety constraints. To address these challenges, we design three parts in Algorithm 1 that are different or irrelevant in the unconstrained case in Mania et al. (2019); Simchowitz and Foster (2020). Part i): instead of CE, we adopt *robust CE* to ensure robust constraint satisfaction despite system uncertainties (Line 3, 8, and Subroutine `RobustCE`). Part ii): we design a *SafeTransit* algorithm to ensure safe policy updates (Line 4, 9, and Algorithm 2). Part iii): we include a *pure-exploitation* phase at each episode to improve the explore-exploit tradeoff (Line 8-11). In the following, we will explain Parts i)-ii) in detail and leave the discussion of Part iii) after our regret analysis in Section 4.

(i) Robust CE and Approximate DAP. We first explain Subroutine `ApproxDAP`. With an estimated model $\hat{\theta}$, we approximate DAP by

$$u_t = -K_{\text{stab}}x_t + \textstyle\sum_{k=1}^H M[k]\hat{w}_{t-k} + \eta_t, \tag{5}$$

where we let $\hat{w}_t = \Pi_{\mathbb{W}}(x_{t+1} - \hat{A}x_t - \hat{B}u_t)$ approximate the true disturbance and add an excitation noise $\eta_t \overset{\text{ind.}}{\sim} \bar{\eta}\mathcal{D}_\eta$ in (5) to encourage exploration. For $\hat{w}$, its projection on $\mathbb{W}$ benefits constraint satisfaction but also introduces policy nonlinearity with history states. For η_t, it has an excitation level $\bar{\eta}$, i.e., $\|\eta_t\|_\infty \leq \bar{\eta}$, and zero mean. The distribution $\mathcal{D}_\eta$ is (s_η, p_η)-anti-concentrated for some s_η, p_η. Examples of $\mathcal{D}_\eta$ include truncated Gaussian, uniform distribution, etc.

Next, we explain the Subroutine `RobustCE`. Given a model uncertainty set $\Theta = \mathbb{B}(\hat{\theta}, r) \cap \Theta_{\text{ini}}$, where $\hat{\theta}$ is the estimated model and r is the uncertainty radius, we define a robustly safe policy set (RSPS) below to ensure safety despite model uncertainty in Θ and excitation noise η_t:

$$\begin{aligned}\Omega(\hat{\theta}; \epsilon_{rob}^x, \epsilon_{rob}^u), \text{for } \epsilon_{rob}^x &= \epsilon_*^x + \epsilon_{unc}^x(r,\bar{\eta}) + \epsilon_v^x(\Delta_M),\\ \epsilon_{rob}^u &= \epsilon_*^u + \epsilon_{unc}^u(r,\bar{\eta}) + \epsilon_v^u(\Delta_M).\end{aligned} \tag{6}$$

Compared with SPS (3) with a known model θ_*, RSPS approximates the true model by $\hat{\theta}$ and adds additional constraint-tightening terms $\epsilon_{unc}^x(r,\bar{\eta})$ and $\epsilon_{unc}^u(r,\bar{\eta})$ to allow for additional model uncertainties in Θ and excitation noises with excitation level $\bar{\eta}$. Besides, RSPS (6) also includes constraint-tightening terms $\epsilon_v^x(\Delta_M), \epsilon_v^u(\Delta_M)$ to allow for small policy variation Δ_M as discussed in Section 2.1.2. Formulas of the additional constraint-tightening terms are provided below. The proof is by perturbation analysis.

Lemma 2 (RSPS). *Let $\epsilon_{unc}^x(r,\bar{\eta}) = O(r+\bar{\eta})$ and $\epsilon_{unc}^u(r,\bar{\eta}) = O(r+\bar{\eta})$. Then, set $\Omega(\hat{\theta}; \epsilon_{rob}^x, \epsilon_{rob}^u)$ is RSPS despite uncertainties Θ and $\bar{\eta}$ and policy variation Δ_M.*

Based on RSPS (6), we can compute a robust CE policy by the QP below with cost function estimated by $\hat{\theta}$:

$$\min_{\mathbf{M}\in\Omega(\hat{\theta},\epsilon_{rob}^x,\epsilon_{rob}^u)} f(\mathbf{M};\hat{\theta}) \tag{7}$$

Algorithm 1: Safe Adaptive Control

```
Input: $\Theta_{\text{ini}}, \mathcal{D}_\eta, K_{\text{stab}}$. $T^e, H^e, \bar{\eta}^e, \Delta_M^e, T_D^e, \forall e$.
Initialize: $\hat{\theta}^0 = \hat{\theta}_{\text{ini}}, r^0 = r_{\text{ini}}, \Theta^0 = \Theta_{\text{ini}}$. Define $w_t = \hat{w}_t = 0$ for $t < 0$, $t_1^0 = 0$.
for Episode $e = 0, 1, 2, \ldots$ do
    // Phase 1: exploration & exploitation
    $(\mathbf{M}_\dagger^e, \Omega_\dagger^e) \leftarrow$RobustCE$(\Theta^e, H^e, \bar{\eta}^e, \Delta_M^e)$.
    If $e > 0$, run Algo. 2 to safely update the policy to $\mathbf{M}_\dagger^e$ with inputs $(\mathbf{M}^{e-1}, \Omega^{e-1}, \Theta^e, 0, \Delta_M^{e-1})$,
     $(\mathbf{M}_\dagger^e, \Omega_\dagger^e, \Theta^e, \bar{\eta}^e, \Delta_M^e), T^e$ and output $t_1^e$.
    for $t = t_1^e, \ldots, t_1^e + T_D^e - 1$ do
        Implement ApproxDAP$(\boldsymbol{M}_\dagger^e, \hat{\theta}^e, \bar{\eta}^e)$.
    // Model update by least square estimation
    Estimate $\hat{\theta}^{e+1}$ by LSE with projection on $\Theta_{\text{ini}}$: $\tilde{\theta}^{e+1} = \arg\min_\theta \sum_{k=t_1^e}^{t_1^e+T_D^e-1} \|x_{k+1} - Ax_k - Bu_k\|_2^2$,
     $\hat{\theta}^{e+1} = \Pi_{\Theta_{\text{ini}}}(\tilde{\theta}^{e+1})$. Update the model uncertainty set: $\Theta^{e+1} = B(\hat{\theta}^{e+1}, r^{e+1}) \cap \Theta_{\text{ini}}$ with radius
     $r^{e+1} = \tilde{O}(\frac{\sqrt{n^2+nm}}{\sqrt{T_D^e \bar{\eta}^e}})$ by Cor. 1.
    // Phase 2: pure exploitation $(\bar{\eta} = 0)$
    $(\mathbf{M}^e, \Omega^e) \leftarrow$RobustCE$(\Theta^{e+1}, H^e, 0, \Delta_M^e)$.
    Run Algo. 2 to safely update the policy to $\mathbf{M}^e$ with output $t_2^e$, inputs $(\mathbf{M}_\dagger^e, \Omega_\dagger^e, \Theta^e, \bar{\eta}^e, \Delta_M^e)$,
     $(\mathbf{M}^e, \Omega^e, \Theta^{e+1}, 0, \Delta_M^e), t_1^e + T_D^e$.
    for $t = t_2^e, \ldots, T^{(e+1)} - 1$ do
        Implement ApproxDAP$(\boldsymbol{M}^e, \hat{\theta}^{e+1}, 0)$.
```

Algorithm 2: SafeTransit

```
Input: $(\mathbf{M}, \Omega, \Theta, \bar{\eta}, \Delta_M)$, $(\mathbf{M}', \Omega', \Theta', \bar{\eta}', \Delta_M')$, $t_0$.
Set $\bar{\eta}_{\text{min}} = \min(\bar{\eta}, \bar{\eta}')$, $\hat{\theta}_{\text{min}} = \hat{\theta}\mathbb{I}_{(r \le r')} + \hat{\theta}'\mathbb{I}_{(r > r')}$.
Set an intermediate policy as $\mathbf{M}_{\text{mid}} \in \Omega \cap \Omega'$.
// Step 1: slowly move from $\mathbf{M}$ to $\mathbf{M}_{\text{mid}}$
Define $W_1 = \max(\lceil \frac{\|\mathbf{M} - \mathbf{M}_{\text{mid}}\|_F}{\min(\Delta_M, \Delta_M')} \rceil, H')$.
for $t = t_0, \ldots, t_0 + W_1 - 1$ do
    Set $\mathbf{M}_t = \mathbf{M}_{t-1} + \frac{1}{W_1}(\mathbf{M}_{\text{mid}} - \mathbf{M})$.
    Run ApproxDAP$(\boldsymbol{M}_t, \hat{\theta}_{\text{min}}, \bar{\eta}_{\text{min}})$.
// Step 2: slowly move from $\mathbf{M}_{\text{mid}}$ to $\mathbf{M}'$
Define $W_2 = \lceil \frac{\|\mathbf{M}' - \mathbf{M}_{\text{mid}}\|_F}{\Delta_M'} \rceil$.
for $t = t_0 + W_1, \ldots, t_0 + W_1 + W_2 - 1$ do
    Set $\mathbf{M}_t = \mathbf{M}_{t-1} + \frac{1}{W_2}(\mathbf{M}' - \mathbf{M}_{\text{mid}})$.
    Run ApproxDAP$(\boldsymbol{M}_t, \hat{\theta}', \bar{\eta}')$.
Output: Termination stage $t_1 = t_0 + W_1 + W_2$.
```

```
Subroutine ApproxDAP ($\boldsymbol{M}, \hat{\theta}, \bar{\eta}$):
    Implement (5) with $\eta_t \overset{\text{ind.}}{\sim} \bar{\eta}\mathcal{D}_\eta$.
    Observe $x_{t+1}$ and record $\hat{w}_t = \Pi_{\mathbb{W}}(x_{t+1} - \hat{A}x_t - \hat{B}u_t)$.
```

$\mathbf{M}=(0,1)$[.2][c] M1eps-converted-to.pdf $\mathbf{M}'=(-2,1)$[.2][c] M2eps-converted-to.pdf First $\mathbf{M}$ then $\mathbf{M}'$[.2][c] M3.eps-converted-to.pdf Illustration of Algo. 2[.21][c]

Figure 1: Figure a-c considers system $x_{t+1} = 0.5x_t + u_t + w_t$ for $w_t \sim \text{Unif}[-1, 1]$. Figure a-b show that $\mathbf{M}$ and $\mathbf{M}'$ are safe to implement in a time-invariant fashion. Figure c shows that directly switching from $\mathbf{M}$ to $\mathbf{M}'$ violates the constraint. Figure d illustrates the construction of the safe policy path in Algo. 2.

(ii) SafeTransit Algorithm. Notice that at the start of each phase in Algo. 1, we compute a new policy to implement in this phase. However, directly changing the old policy to the new one may cause constraint violation even though both old and new policies are safe time-invariant policies. This is illustrated in Figure 1 a-c. An intuitive explanation behind this phenomenon will be discussed in the appendix. Here, we focus on how to address this issue.

```
Subroutine RobustCE($\Theta, H, \bar{\eta}, \Delta_M$):
    Construct the robustly safe policy set:
    $\Omega = \Omega(\hat{\theta}, \epsilon^x_{rob}, \epsilon^u_{rob})$ for $(\epsilon^x_{rob}, \epsilon^u_{rob})$ defined in (6).
    Compute the optimal policy $\mathbf{M}$ to (7).
    return policy $\mathbf{M}$ and robustly safe policy set $\Omega$.
```

To address this, we design Algorithm 2 to ensure safe policy updates at the start of each phase. The high-level idea of Algorithm 2 is based on the slow-variation trick reviewed in Section 2.1.2. That is, we construct a policy path connecting the old policy to the new policy such that this policy path is contained in some robustly safe policy set with an additional constraint tightening term to allow slow policy variation, then by slowly varying the policies along this path, we are able to safely transit to the new policy.

Next, we discuss the construction of this policy path,[7] which is illustrated in Figure 1d. We follow the notations in Algorithm 2, i.e. the old policy is $\mathbf{M}$ in an old RSPS Ω and the new policy is $\mathbf{M}'$ in Ω'. Notice that the straight line from $\mathbf{M}$ to $\mathbf{M}'$ does not satisfy the requirements of the slow variation trick because some parts of the line are outside both RSPSs. To address this, Algorithm 2 introduces an intermediate policy $\mathbf{M}_{\text{mid}}$ in $\Omega \cap \Omega'$, and slowly moves the policy from the old one $\mathbf{M}$ to the intermediate one $\mathbf{M}_{\text{mid}}$ (Step 1), then slowly moves from $\mathbf{M}_{\text{mid}}$ to the new policy $\mathbf{M}'$ (Step 2). In this way, all the path is included in at least one of the robustly safe policy sets, which allows safe transition from the old policy to the new policy. The choice of $\mathbf{M}_{\text{mid}}$ is not unique. In practice, we recommend selecting $\mathbf{M}_{\text{mid}}$ with a shorter path length for quicker policy transition. $\mathbf{M}_{\text{mid}}$ can be computed efficiently since the set $\Omega \cap \Omega'$ is a polytope. The existence of $\mathbf{M}_{\text{mid}}$ can be guaranteed if the first `RobustCE` program in Algorithm 1 (Phase 1 of episode 0) is strictly feasible. This is usually called recursive feasibility and will be formally proved in Theorem 2.

Remark 2 (More discussions on $\mathbf{M}_{\text{mid}}$). *If RSPSs are monotone, e.g., if $\Omega \subseteq \Omega'$, then we can let $\mathbf{M}_{mid} = \mathbf{M}$ and the path constructed by Algorithm 2 reduces to the straight line from $\mathbf{M}$ to $\mathbf{M}'$. Hence, a non-trivial $\mathbf{M}_{mid}$ is only relevant for non-monotone RSPS, which can be caused by the non-monotone model uncertainty sets generated by LSE (even though the error bound r of LSE decreases with more data, the change in the point estimator $\hat{\theta}$ may cause $\Theta' \not\subseteq \Theta$). Though one can enforce decreasing uncertainty sets by taking joints over all the history uncertainty sets, this approach leads to an increasing number of constraints when determining the RSPS in* `RobustCE`*, thus demanding high computation for large episode e.*

Remark 3 (Single trajectory and computation comparison). *Though Algo. 1 is implemented by episodes, it is still a single trajectory since no system starts are needed when new episodes start. Compared with the projected-gradient-descent algorithm in* **?***, our algorithm only solves constrained QP once in a while, but* **?** *solve constrained QP for projection at every stage. In this sense, our algorithm reduces the computational burden.*

Remark 4 (Model Estimation). *As for the model updates, we can use all the history data in practice, though we only use part of history in* `ModelEst` *for simpler analysis.* `ModelEst` *also projects the estimated model onto Θ_{ini} to ensure bounded estimation.*

Remark 5 (Safe algorithm comparison). *Some constrained control methods construct safe* state *sets and safe* action *sets for the current state, e.g., control barrier functions (Ames et al., 2016), reachability-set-based methods (Akametalu et al., 2014), regulation maps (Kellett and Teel, 2004), etc. In contrast, this paper constructs safe* policy *sets in the space of policy parameters $\mathbf{M}$. This is possible because our policy structure (linear on history disturbances) allows a transformation from linear constraints on the states and actions to polytopic constraints on the policy parameters.*

[7] This construction is not unique, other methods such as MPC can also work.

4 Theoretical Analysis

In this section, we provide theoretical guarantees of our algorithms including model estimation errors, feasibility, constraint satisfaction, and a regret bound.

For technical simplicity, we assume $K_{\text{stab}} = 0$ for the theoretical analysis. This is without loss of generality and will only change the regret's order on the dimensionalities.

4.1 Estimation Error Bound

The estimation error bounds for linear policies have been studied in the literature Dean et al. (2018). However, due to the projection in our disturbance approximation in (5), the policies implemented by our algorithms can be sometimes nonlinear. To cope with this, we provide an error bound below for general policies.

Theorem 1 (Estimation error bound)**.** *Consider actions* $u_t = \pi_t(x_0, \{w_k, \eta_k\}_{k=0}^{t-1}) + \eta_t$, *where* $\|\eta_t\|_\infty \le \bar\eta$ *is generated as discussed after* (5) *and policies* $\pi_t(\cdot)$ *ensure bounded states and actions, i.e.* $\|(x_t^\top, u_t^\top)\|_2 \le b_z$ *for all* $t \ge 0$. *Let* $\tilde\theta_T = \min_{A,B} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|_2^2$ *denote the model estimation. For any* $0 < \delta < 1/3$, *for* $T \ge O(\log(1/\delta) + (m+n)\log(b_z/\bar\eta))$, *with probability (w.p.)* $1 - 3\delta$, *we have* $\|\tilde\theta_T - \theta_*\|_2 \le O\left(\sqrt{n+m}\frac{\sqrt{\log(b_z/\bar\eta + 1/\delta)}}{\sqrt{T}\bar\eta}\right)$.

Theorem 1 holds for both linear and nonlinear policies as long as the induced states and actions are bounded, which can be guaranteed by the stability of the policies. Further, the error bound in Theorem 1 is $\tilde O(\frac{\sqrt{m+n}}{\bar\eta\sqrt{T}})$, which coincides with the error bound for linear policies in terms of $T, \bar\eta, n, m$ in Dean et al. (2018, 2019a).

Based on Theorem 1, we obtain a formula for the estimation error bound r^e in Line 7 of Algorithm 1.

Corollary 1 (Formula of r^e)**.** *Suppose* $H^0 \ge \log(2\kappa)/\log((1-\gamma)^{-1})$, $T^{e+1} \ge t_1^e + T_D^e$ *and* T_D^e *satisfies the condition on* T *in Theorem 1. For any* $0 < p < 1$ *and* $e \ge 1$, *with probability at least* $1 - \frac{p}{2e^2}$, *we have* $\|\hat\theta^e - \theta_*\|_F \le r^e$, *where*

$$r^e = O\left(\frac{(n+\sqrt{mn})\sqrt{\log(\sqrt{mn}/\bar\eta^{(e-1)} + e^2/p)}}{\sqrt{T_D^{(e-1)}}\bar\eta^{(e-1)}}\right). \tag{8}$$

Corollary 1 considers the $\|\cdot\|_F$ norm because Algorithm 1 projects matrix $\tilde\theta^e$ onto Θ_{ini} and the $\|\cdot\|_F$ norm is more convenient to analyze and implement when matrix projections are involved. Due to the change of norms, the error bound has an additional $\sqrt{n}$ factor.

4.2 Feasibility and Constraint Satisfaction

This section provides feasibility and constraint satisfaction guarantees of our adaptive control algorithm.

Theorem 2 (Feasibility)**.** *Algorithms 1 and 2 output feasible policies for all* t *under the following conditions.*

(i) (Strict initial feasibility) There exists $\epsilon_0 > 0$ *such that* $\Omega(\hat\theta^0, \epsilon^{x.0} + \epsilon_0, \epsilon^{u,0}) \ne \emptyset$, *where* $\epsilon_x^0, \epsilon_u^0$ *are defined by* (7) *with initial parameters* $r^0, \bar\eta^0, H^0, \Delta_M^0$.

(ii) (Monotone parameters) $\bar\eta^e, H^e, T_D^e, \Delta_M^e$ *are selected s.t.* $(H^e)^{-1}, \sqrt{H^e}\Delta_M^e$, $\bar\eta^e$, r^e *are all non-increasing with* e, *and* $r^1 \le \frac{\epsilon_0}{c_1\sqrt{mn}}$, *where* r^e *is defined in* (8), c_1 *is defined in Lemma 2.*

Further, under Assumption 2, condition (i) is satisfied if (iii) $\epsilon_x^0 + \epsilon_0 \le \epsilon_{F,x} - \epsilon_{unc}^x(r_{\text{ini}}, 0) - \epsilon_P$, $\quad \epsilon_u^0 \le \epsilon_{F,u} - \epsilon_P$, *where* $\epsilon_P = O(\sqrt{mn}(1-\gamma)^{H^0})$.

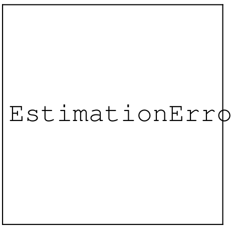
Condition (i) requires the initial policy set $\Omega_\dagger^0$ to contain a policy that strictly satisfies the constraints on $g_i^x(\mathbf{M}; \hat\theta^0)$. Condition (ii) requires monotonic parameters in later phases, where the non-increasing estimation error r^e requires an increasing number of exploration stages T_D^e. Conditions (i) and (ii) together establish the *recursive feasibility*: if our algorithm is (strictly) feasible at the initial stage, then the algorithm is feasible in the future under proper parameters.

The condition (iii) in Theorem 2 guarantees strict *initial feasibility*, which is based on the ϵ_F-strictly safe policy $u_t = -K_F x_t$ in Assumption 2. The term ϵ_P captures the difference between the policy K_F and its approximate DAP. Consider (iii) requires large enough H^0, T_D^0 and small enough $\bar\eta^0, \Delta_M^0$. Further, consider (iii) implicitly requires a small enough initial uncertainty radius: $\epsilon_{unc}^x(r_{\text{ini}}, 0) \le \epsilon_{F,x}$. If the initial uncertainty set is too large but a safe policy is available, one can first explore the system with the safe policy to reduce the model uncertainty until the strict initial feasibility is obtained and then apply our algorithm.

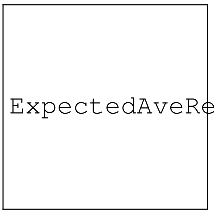
Theorem 3 (Constraint Satisfaction). *Under the conditions in Theorem 2 and Corollary 1, when $T^{e+1} \geq t_2^e$, we have $u_t \in \mathbb{U}$ for all $t \geq 0$ w.p.1 and $x_t \in \mathbb{X}$ for all $t \geq 0$ w.p. $1-p$, where p is chosen in Corollary 1.*

The control constraint satisfaction is always ensured by the projection onto $\mathbb{W}$ in (5). Besides, we can show that the state constraints are satisfied if the true model is inside the confidence sets Θ^e for all $e \geq 0$, whose probability is at least $1-p$ by Corollary 1.

Model estimation error[.23][c]



Average expected regret[.23][c]



Altitude

error's range[.23][c] Thrust error's range[.23][c]

Figure 2: Figure (a) plots the model estimation error $\|\hat{\theta}^e - \theta_*\|_F$ for different episode e. Figure (b) plots average expected regret $\mathbb{E}\,\text{Regret}/t$ for different stage t. The solid lines are the mean and the shades are 90% confidence bounds. Figures c-d plot the altitude error $z_t - z^{ref}$ and thrust error $\upsilon_t - \upsilon^{ref}$ of our algorithm. The solid lines are the median and the shades are the error range over 50 trials.

4.3 Regret Bound

Next, we show that our algorithm can achieve a $\tilde{O}(T^{2/3})$ regret bound together with feasibility and constraint satisfaction under proper conditions. Further, we explain the reasons behind the pure exploitation phase.

Theorem 4 (Regret bound). *Suppose $\epsilon_\theta(r_{\text{ini}}) \leq \epsilon_{F,x}/4$. For any $0 < p < 1/2$, with parameters $T^1 \geq \tilde{O}((\sqrt{n}m + n\sqrt{m})^3)$, $T^{e+1} = 2T^e$, $T_D^e = (T^{e+1} - T^e)^{2/3}$, $\Delta_M^e = O(\frac{\epsilon_F^x}{\sqrt{mn}H^{(e)}(T^{e+1})^{1/3}})$, $\bar{\eta}^e = O(\min(\frac{\epsilon_F^x}{\sqrt{m}}, \epsilon_F^u))$, and $H^e \geq O(\log(\max(T^{e+1}, \frac{\sqrt{n}}{\min(\epsilon_F)})))$, Algorithm 1 is feasible and satisfies $\{u_t \in \mathbb{U}\}_{t\geq 0}$ w.p.1 and $\{x_t \in \mathbb{X}\}_{t \geq 0}$ w.p. $(1-p)$. Further, w.p. $(1-2p)$, we have*
$\text{Regret} \leq \tilde{O}((n^2m^2 + n^{2.5}m^{1.5})\sqrt{mn + k_x + k_u}T^{2/3})$.

On r_{ini}. As discussed after Theorem 2, the initial feasibility requires a small enough r_{ini}, otherwise more exploration is needed before applying our algorithm. For technical simplicity, Theorem 4 assumes $\epsilon_\theta(r_{\text{ini}}) \leq \epsilon_{F,x}/4$ and establishes other conditions accordingly.

On parameters. Theorem 4 provides choices of parameters that ensure feasibility, safety, and the regret bound. Here, we choose exponentially increasing episode lengths T^e, and explore for $(T^e)^{2/3}$ stages at each episode with a small enough constant excitation level $\bar{\eta}^e$. Compared with the *first-explore-then-exploit* algorithms, episodic updates allow early improvements in system performance and adaptation to possible changes in the environment. Finally, we select large enough memory lengths $H^e \geq O(\log(T^e))$ and small enough variation budgets $\Delta_M^e \leq O((T^{e+1})^{-1/3})$.

On regret. Though our regret bound $\tilde{O}(T^{2/3})$ is worse than the $\tilde{O}(\sqrt{T})$ regret bound for *unconstrained* LQR, interestingly, Dean et al. (2018) show that $\tilde{O}(T^{2/3})$ is optimal for CE-based methods with *robust* stability. Thus, we conjecture that $\tilde{O}(T^{2/3})$ is also optimal for CE-based methods with *robust* safety, but its formal proof is left for the future.

Discussions on the pure exploitation phase. Algorithm 1 includes a pure exploitation phase with no excitation noise at each episode, which is not present in the unconstrained algorithms (Dean et al., 2018; Mania et al., 2019; Simchowitz and Foster, 2020). This phase is motivated by our regret analysis: the algorithm can still work without this phase but will generate a worse regret bound. Specifically, consider a robust CE policy (7) with respect to $\bar{\eta}$ and uncertainty radius r, the regret of this policy per stage can be roughly bounded by $\tilde{O}(\bar{\eta} + r)$ in the supplementary (we omit Δ_M here for simplicity). With no pure exploitation phases, the regret in episode e can be roughly bounded by $\tilde{O}(T^e(\bar{\eta}^e + r^e))$, where $r^e = \tilde{O}(\frac{1}{\sqrt{T^{(e-1)}}\bar{\eta}^{(e-1)}})$ by Corollary 1 (hiding n, m). Therefore, the total regret can be bounded by $\sum_e(\frac{1}{\sqrt{T^{(e-1)}}\bar{\eta}^{(e-1)}} + \bar{\eta}^e)T^e \approx \sum_e(\frac{1}{\sqrt{T^e}\bar{\eta}^e} + \bar{\eta}^e)T^e$, which is minimized at $\frac{1}{\sqrt{T^e}\bar{\eta}^e} = \bar{\eta}^e = (T^e)^{-1/4}$, leading to a worse regret bound $\tilde{O}(T^{3/4})$. Lastly, with a constant $\bar{\eta}^e$, our algorithm suffers slightly larger stage regret during the Phase 1 (exploration & exploitation) compared with diminishing $\eta^e = (T^e)^{-1/4} \to 0$, but the performance still improves by refining the models and reducing Δ_M^e.

5 Numerical Experiments

In this section, we numerically test our safe adaptive control algorithm on a quadrotor vertical flight problem. Specifically, we consider the affine dynamical model for vertical flight with linear air drag force as in **?**: $\ddot{z} = \upsilon/m - g - I^a \dot{z}/m + d$, where z is the quadrotor's altitude, υ is the motor thrust, $m = 1$kg is the mass, $g = 9.8$m/s^2 is the gravitational acceleration, $I^a = 0.25$kg/s is the drag coefficient of the air resistance, $-I^a \dot{z}$ models air drag force, and d represents other system disturbances. The unknown system parameters are the mass due to the unknown load and the air drag coefficient. The constraints on the altitude are $0 \le z \le 1.7$m and the constraints on the thrust are $0 \le \upsilon \le 12$N. The control task is to safely maintain the quadrotor around a target altitude $z^{\text{ref}} = 0.7$m. The corresponding velocity and control input at this equilibrium point is $\dot{z}^{\text{ref}} = 0, \upsilon^{\text{ref}} = g$. We consider a quadratic cost function to measure the deviation from the desirable equilibrium point: $0.1(z - z^{\text{ref}})^2 + 0.1(\dot{z} - \dot{z}^{\text{ref}})^2 + 0.1(\upsilon - \upsilon^{\text{ref}})^2$. Consider unknown mass and air drag coefficient. The initial estimation are 0.83kg $\le \hat{m} \le 2$kg, and 0.1kg/s $\le \hat{I}^a \le 0.33$kg/s. We generate the system disturbances i.i.d. from distribution Bern$[-0.2, 0.2]$ with $w_{\max} = 0.2$. We select $K_{stab} = (2/3, 1)$ to robustly stabilize the system. Let time discretization be $\Delta_c = 1$s.

Figure 2a shows that the model estimation error decreases as $\bar{\eta}$ increases, which is consistent with Corollary 1. Figure 2b shows that the average regret first decreases then increases as $\bar{\eta}$ increases. This is because for small $\bar{\eta}$, the regret benefits from faster estimation error reduction, but for large $\bar{\eta}$, the regret suffers from larger constraint-tightening terms due to large excitation noises. This shows the tradeoff between exploration and exploitation under safety constraints. Figure 2c,d show that our algorithm ensure safety by satisfying constraint satisfaction. Further, as learning continues, the ranges decrease due to better model estimation.

Appendices

This appendices include the proofs for the theoretical results and more discussions for the paper.

- Appendix A provides necessary lemmas that will be used throughout the appendices and defines the constraint-tightening factors promised in Lemma 2.
- Appendix B focuses on the estimation error and provides proofs for Theorem 1 and Corollary 1.
- Appendix C studies feasibility and provides a proof for Theorem 2.
- Appendix D focuses on the constraint satisfaction and provides a proof for Theorem 3, which also includes a proof for Lemma 2.
- Appendix E analyzes the regret and proves Theorem 4.
- Appendix F provides additional discussions on RMPC and non-zero x_0.
- Appendix G provides proofs to the technical lemmas used in the appendices A-E.

Additional Notations. Define $\mathbb{X} = \{x : D_x x \le d_x\}$ and $\mathbb{U} = \{u : D_u u \le d_u\}$. Let $\upsilon_{\min}(A)$ and $\upsilon_{\max}(A)$ denote the minimum and the maximum eigenvalue of a symmetric matrix A respectively. For two symmetric matrices X and Y, we write $X \le Y$ if $Y - X$ is positive semi-definite, we write $X < Y$ if $Y - X$ is positive definite. For two vectors $x, y \in \mathbb{R}^n$, we write $x \le y$ is $(y - x)_i \ge 0$ for $1 \le i \le n$, i.e. x is smaller than y elementwise. Consider a σ-algebra $\mathcal{F}_t$ and a random vector $y_t \in \mathbb{R}^n$, we write $y_t \in \mathcal{F}_t$ if the random vector y_t is measurable in $\mathcal{F}_t$. We let I_n denote the identity matrix in $\mathbb{R}^{n \times n}$. Denote an aggregated vector $z_t = (x_t^\top, u_t^\top)^\top$ for notational simplicity. Define $z_{\max} = \sqrt{x_{\max}^2 + u_{\max}^2}$ as the maximum l_2 norm of z_t for any x_t, u_t satisfying the constraints in (1). We use "a.s." as an abbreviation for "almost surely". Finally, for memory lengths $H_1 < H_2$, notice that the set $\mathcal{M}_{H_1}$ can be viewed as a subset of $\mathcal{M}_{H_2}$ since we can append 0 matrices to any $\mathbf{M} \in \mathcal{M}_{H_1}$ to generate a corresponding matrix in $\mathcal{M}_{H_2}$, so we will slightly abuse the notation and write $\mathbf{M} \in \mathcal{M}_{H_2}$ for any $\mathbf{M} \in \mathcal{M}_{H_1}$ with $H_1 < H_2$.

A Preparations: State Approximation and Constraint-tightening Terms

This section provides results that will be useful throughout the rest of the appendices. Specifically, the first subsection provides a state approximation lemma and a constraint-decomposition corollary for the approximate DAP, and the second subsection defines and discusses the constraint-tightening terms in Lemma 2 and (7) based on the upper bounds of the constraint decomposition terms.

A.1 State Approximation and Constraint Decompositions

We consider a more general form of approximate DAP below than that in (2), i.e., an approximate DAP with time-varying policy matrices $\mathbf{M}_t$, time-varying excitation levels $\bar{\eta}_t$, and time-varying model estimations $\hat{\theta}_t$.

$$u_t = \sum_{t=1}^{H_t} M_t[k]\hat{w}_{t-k} + \eta_t, \quad \hat{w}_t = \Pi_{\mathbb{W}}(x_{t+1} - \hat{\theta}_t z_t), \quad \|\eta_t\|_\infty \le \bar{\eta}_t, \quad t \ge 0, \tag{9}$$

where $\mathbf{M}_t \in \mathcal{M}_{H_t}$ and $\{H_t\}_{t\ge 0}$ is non-decreasing.

When implementing the time-varying approximate DAP (9) to the system $x_{t+1} = A_* x_t + B_* u_t + w_t$, we have the following state approximation lemma.

Lemma 3 (State approximation under time-varying approximate DAP)**.** *When implementing the time-varying approximate DAP* (9) *to the system $x_{t+1} = A_* x_t + B_* u_t + w_t$, we have the following state approximation result:*

$$x_t = A_*^{H_t} x_{t-H_t} + \sum_{k=2}^{2H_t}\sum_{i=1}^{H_t} A_*^{i-1} B_* M_{t-i}[k-i]\hat{w}_{t-k}\mathbb{1}_{(1\le k-i\le H_{t-i})} + \sum_{i=1}^{H_t} A_*^{i-1} w_{t-i} + \sum_{i=1}^{H_t} A_*^{i-1} B_* \eta_{t-i}$$

The lemma above is a straightforward extension from Proposition 1 reviewed in Section 2.1 for the case with perfect model information, thus the proof is omitted.

To simplify the exposition, we introduce the following notations for time-varying DAP.

$$\tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t};\theta) = A^{k-1}\mathbb{1}_{(k\le H_t)} + \sum_{i=1}^{H_t} A^{i-1} B M_{t-i}[k-i]\mathbb{1}_{(1\le k-i\le H_t)}, \quad \forall\, 1 \le k \le 2H_t, \tag{10}$$

$$\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1};\theta) = \sup_{\hat{w}_k\in\mathbb{W}} D_{x,i}^\top \sum_{k=1}^{2H_t} \tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t-1};\theta)\hat{w}_{t-k} = \sum_{k=1}^{2H_t} \|D_{x,i}^\top \tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t-1};\theta)\|_1 w_{\max}, \tag{11}$$

where $1 \le i \le k_x$ and we define $\mathbf{M}_t = \mathbf{M}_0$ for $t \le 0$ for notational simplicity. Notice that when $\mathbf{M}_t = \mathbf{M}$ and $H_t = H$ (the time-invariant case), the definitions of $\tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t};\theta)$ and $\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1};\theta)$ above reduce to the definitions of $\Phi_k^x(\mathbf{M};\theta)$ and $g_i^x(\mathbf{M};\theta)$ respectively in Section 2.1.

Based on Lemma 3 and the notations defined above, we can obtain the following corollary on the decompositions of the state constraints $D_x x_t$ and action constraints $D_u u_t$. The decompositions are crucial when defining our constraint-tightening terms and developing the constraint satisfaction guarantees.

Corollary 2 (Constraint decomposition)**.** *When implementing the time-varying approximate DAP* (9) *to the system $x_{t+1} = A_* x_t + B_* u_t + w_t$, for each $1 \le i \le k_x$ and $1 \le j \le k_u$, we have the following decompositions:*

$$\begin{aligned}
D_{x,i}^\top x_t \le\ & \underbrace{g_i^x(\mathbf{M}_t;\hat{\theta}_t^g)}_{\textit{estimated state constraint function}} + \underbrace{(g_i^x(\mathbf{M}_t;\theta_*) - g_i^x(\mathbf{M}_t;\hat{\theta}_t^g)) + \sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})}_{\textit{model estimation errors}} \\
& + \underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\textit{excitation errors on the state}} + \underbrace{D_{x,i}^\top A_*^{H_t} x_{t-H_t}}_{\textit{history truncation errors}} + \underbrace{(\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1};\theta_*) - g_i^x(\mathbf{M}_t;\theta_*))}_{\textit{policy variation errors}}, \\
D_{u,j}^\top u_t \le\ & \underbrace{g_j^u(\mathbf{M}_t)}_{\textit{action constraint function}} + \underbrace{D_{u,j}^\top \eta_t}_{\textit{excitation error on the action}}.
\end{aligned}$$

where $\hat{\theta}_t^g$ is an estimated model used to approximate the state constraint function, and we allow $\hat{\theta}_t^g \ne \hat{\theta}_t$ for generality.

Proof. The proof is by the definitions of $g_i^x, g_j^u, \tilde{g}_i^x$, and Lemma 3. For the action constraints, by the definition (9), we have

$$D_{u,j}^\top u_t = \sum_{t=1}^{H_t} D_{u,j}^\top M_t[k]\hat{w}_{t-k} + D_{u,j}^\top \eta_t \leq \sum_{t=1}^{H_t} \|D_{u,j}^\top M_t[k]\|_1 w_{\max} + D_{u,j}^\top \eta_t = g_j^u(\mathbf{M}_t) + D_{u,j}^\top \eta_t$$

where the inequality is because $\hat{w}_{t-k} \in \mathbb{W}$ and the Hölder's inequality. The state constraints can be similarly proved: notice that we apply the Hölder's inequality on $\hat{w}_{t-k}$ instead of w_{t-k}. □

A.2 The Constraint-tightening Terms

This subsection provides the definitions of the factors c_1, c_2, c_3 in the constraint-tightening terms introduced in Lemma 2. Further, this subsection provides explanations on all the constraint-tightening terms in (7) by showing that each constraint-tightening term serves as an upper bound on an error term in the constraint decompositions in Corollary 2.

Definition and explanation of $\epsilon_\theta(r)$. The next lemma formally shows that the constraint-tightening term $\epsilon_\theta(r)$ is an upper bound on the model estimation errors in the state constraint decomposition in Corollary 1, where r is the model estimation error bound.

Lemma 4 (Definition of $\epsilon_\theta(r)$)**.** *Consider implementing the time-varying approximate DAP* (9) *to the system* $x_{t+1} = A_* x_t + B_* u_t + w_t$. *For a fixed t, suppose* $\hat{\theta}_{t-k}, \hat{\theta}_t^g \in \Theta_{ini}$, $\|\hat{\theta}_t^g - \theta_*\|_F \leq r$, *and* $\|\hat{\theta}_{t-k} - \theta_*\|_F \leq r$ *for all* $1 \leq k \leq H_t$. *Further, suppose* $x_{t-k} \in \mathbb{X}, u_{t-k} \in \mathbb{U}$ *for all* $1 \leq k \leq H_t$. *Then, we have*

$$g_i^x(\mathbf{M}_t; \theta_*) - g_i^x(\mathbf{M}_t; \hat{\theta}_t^g) \leq \epsilon_{\hat{\theta}}(r), \quad \sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1}(w_{t-k} - \hat{w}_{t-k}) \leq \epsilon_{\hat{w}}(r),$$

$$\underbrace{(g_i^x(\mathbf{M}_t; \theta_*) - g_i^x(\mathbf{M}_t; \hat{\theta}_t^g)) + \sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})}_{\textit{model estimation errors}} \leq \epsilon_\theta(r)$$

where $\epsilon_{\hat{w}}(r) = \|D_x\|_\infty z_{\max}\kappa/\gamma \cdot r = O(r)$, $\epsilon_{\hat{\theta}}(r) = 5\kappa^4\kappa_B\|D_x\|_\infty w_{\max}/\gamma^3\sqrt{mn}r = O(\sqrt{mn}r)$, *and* $\epsilon_\theta(r) = \epsilon_{\hat{\theta}}(r) + \epsilon_{\hat{w}}(r) = O(\sqrt{mn})r$. *We can let* $c_1 = \|D_x\|_\infty z_{\max}\kappa/\gamma + 5\kappa^4\kappa_B\|D_x\|_\infty w_{\max}/\gamma^3$.

The proof of Lemma 4 is based on the perturbation analysis and deferred to Appendix G.1. When proving Lemma 4, we also establish the following lemma.

Lemma 5 (Disturbance approximation error)**.** *Consider* $\hat{w}_t = \Pi_{\mathbb{W}}(x_{t+1} - \hat{\theta} z_t)$ *and* $x_{t+1} = \theta_* z_t + w_t$. *Suppose* $\|z_t\|_2 \leq b_z$ *and* $\|\theta_* - \hat{\theta}\|_F \leq r$, *then*

$$\|w_t - \hat{w}_t\|_2 \leq b_z r$$

Proof. By non-expansiveness of projection, we have $\|w_t - \hat{w}_t\|_2 \leq \|x_{t+1} - \theta_* z_t - (x_{t+1} - \hat{\theta} z_t)\|_2 = \|(\hat{\theta} - \theta_*) z_t\|_2 \leq b_z r$. □

Definition and explanation of $\epsilon_{\eta,x}(\bar{\eta})$ and $\epsilon_{\eta,u}(\bar{\eta})$ The next lemma formally shows that the terms $\epsilon_{\eta,x}(\bar{\eta})$ and $\epsilon_{\eta,u}(\bar{\eta})$ bounds the excitation errors on the state and action constraint decompositions in Corollary 2.

Lemma 6 (Definition of $\epsilon_\eta(\bar{\eta})$)**.** *Consider implementing the time-varying approximate DAP* (9) *to the system* $x_{t+1} = A_* x_t + B_* u_t + w_t$. *For a fixed t, suppose* $\|\eta_t\|_\infty \leq \bar{\eta}$ *for all* $0 \leq k \leq H_t$. *Then,*

$$\underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\textit{excitation errors on state}} \leq \epsilon_{\eta,x}(\bar{\eta}), \qquad \underbrace{D_{u,j}^\top \eta_t}_{\textit{excitation errors on actions}} \leq \epsilon_{\eta,u}(\bar{\eta}),$$

where $\epsilon_{\eta,x} = \|D_x\|_\infty \kappa\kappa_B/\gamma\sqrt{m}\bar{\eta} = O(\sqrt{m}\bar{\eta})$, $\epsilon_{\eta,u} = \|D_u\|_\infty \bar{\eta} = O(\bar{\eta})$, *and we define* $\epsilon_\eta = (\epsilon_{\eta,x}, \epsilon_{\eta,u})$, $c_2 = \|D_x\|_\infty \kappa\kappa_B/\gamma$, $c_3 = \|D_u\|_\infty$.

Proof. The proof is provided below.

$$\begin{aligned}
\|D_x \sum_{i=1}^{H_t} A_*^{i-1} B_* \eta_{t-i}\|_\infty &\leq \|D_x\|_\infty \sum_{i=1}^{H_t} \|A_*^{i-1} B_*\|_\infty \|\eta_{t-i}\|_\infty \leq \|D_x\|_\infty \sqrt{m} \sum_{i=1}^{H_t} \|A_*^{i-1} B_*\|_2 \|\eta_{t-i}\|_\infty \\
&\leq \|D_x\|_\infty \sqrt{m} \sum_{i=1}^{H_t} \kappa(1-\gamma)^{i-1} \kappa_B \|\eta_{t-i}\|_\infty \leq \|D_x\|_\infty \sqrt{m} \kappa \kappa_B / \gamma \bar{\eta} \\
\|D_u \eta_t\|_\infty &\leq \|D_u\|_\infty \|\eta_t\|_\infty \leq \|D_u\|_\infty \bar{\eta}
\end{aligned}$$

□

Definition of $\epsilon_H(H)$ The term $\epsilon_H(H)$ has been introduced in Li et al. (2020) for the known-model case to bound the history truncation errors in the state constraint decomposition. Here, we slightly improve its dependence on the problem dimensions and include our proof below.

Lemma 7 (Definition of ϵ_H). *For any $x_{t-H_t} \in \mathbb{X}$, we have*

$$\underbrace{D_{x,i}^\top A_*^{H_t} x_{t-H_t}}_{\text{history truncation errors}} \leq \epsilon_H(H_t) = \|D_x\|_\infty \kappa x_{\max} (1-\gamma)^{H_t} = O((1-\gamma)^{H_t}).$$

Proof.

$$\begin{aligned}
\|D_x A_*^{H_t} x_{t-H_t}\|_\infty &\leq \|D_x\|_\infty \|A_*^{H_t} x_{t-H_t}\|_\infty \leq \|D_x\|_\infty \|A_*^{H_t} x_{t-H_t}\|_2 \\
&\leq \|D_x\|_\infty \|A_*^{H_t}\|_2 \|x_{t-H_t}\|_2 \leq \|D_x\|_\infty \kappa (1-\gamma)^{H_t} x_{\max}.
\end{aligned}$$

□

Definition of $\epsilon_v(\Delta_M, H)$ The error term $\epsilon_v(\Delta_M, H)$ has also been introduced in Li et al. (2020) for the known-model case to bound the policy variation error. Here, we also slightly improve its dependence on the problem dimensions and the memory length in the next lemma. The proof is based on the perturbation analysis and will be provided in Appendix G.2.

Lemma 8 (Definition of $\epsilon_v(\Delta_M, H)$). *Under the conditions in Lemma 3, suppose $\Delta_M \geq \max_{1 \leq k \leq H_t} \frac{\|\mathbf{M}_t - \mathbf{M}_{t-k}\|_F}{k}$, then we have*

$$\underbrace{(\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1}; \theta_*) - g_i^x(\mathbf{M}_t; \theta_*))}_{\text{policy variation errors}} \leq \epsilon_v(\Delta_M, H_t)$$

where $\epsilon_v(\Delta_M, H_t) = \|D_x\|_\infty w_{\max} \kappa \kappa_B / \gamma^2 \sqrt{mnH_t} \Delta_M = O(\sqrt{mnH_t} \Delta_M)$.

B Estimation Error Bounds

This section provides a proof for Theorem 1 and a proof for Corollary 1. When proving Corollary 1, we also establishes a.s. upper bounds on the state and action trajectories of our algorithm.

B.1 Proof of Theorem 1

Our proof of Theorem 1 relies on a recently developed least square estimation error bound for general time series satisfying a block matingale small-ball (BMSB) condition Simchowitz et al. (2018). The general error bound and the definition of BMSB are included below for completeness. In the literature Dean et al. (2018, 2019a), only linear policies are considered and shown to satisfy the BMSB condition. Our contribution is to show that even for general policies, BMSB still holds as long as the corresponding states and actions are bounded (which is usually the case if certain stability properties are satisfied). By general policies, we allow time-varying policies, nonlinear policies, policies that depend on all the history, etc., (i.e. we consider $u_t = \pi_t(x_0, \{w_k, \eta_k\}_{k=0}^{t-1}) + \eta_t$). More rigorous discussions are provided below.

Definition 3 (Block Martingale Small-Ball (BMSB) (Definition 2.1 Simchowitz et al. (2018)))**.** *Let $\{X_t\}_{t\geq 1}$ be an $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process taking values in $\mathbb{R}^d$. We say that it satisfies the (k, Γ_{sb}, p)-block martingale small-ball (BMSB) condition for $\Gamma_{sb} > 0$ if, for any fixed $\lambda \in \mathbb{R}^d$ such that $\|\lambda\|_2 = 1$ and for any $j \geq 0$, one has $\frac{1}{k}\sum_{i=1}^{k} \mathbb{P}(|\lambda^\top X_{j+i}| \geq \sqrt{\lambda^\top \Gamma_{sb} \lambda} \mid \mathcal{F}_j) \geq p$ almost surely.*

Theorem 5 (Theorem 2.4 in Simchowitz et al. (2018))**.** *Fix $\epsilon \in (0,1)$, $\delta \in (0, 1/3)$, $T \geq 1$, and $0 < \Gamma_{sb} < \bar{\Gamma}$. Consider a random process $\{X_t, Y_t\}_{t\geq 1} \in (\mathbb{R}^d \times \mathbb{R}^n)^T$ and a filtration $\{\mathcal{F}_t\}_{t\geq 1}$. Suppose the following conditions hold,*

1. *$Y_t = \theta_* X_t + \eta_t$, where $\eta_t \mid \mathcal{F}_t$ is σ_{sub}^2-sub-Gaussian and mean zero,*
2. *$\{X_t\}_{t\geq 1}$ is an $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process satisfying the (k, Γ_{sb}, p)-block martingale small-ball (BMSB) condition,*
3. *$\mathbb{P}(\sum_{t=1}^{T} X_t X_t^\top \not\leq T\bar{\Gamma}) \leq \delta$.*

Define the (ordinary) least square estimator as $\tilde{\theta} = \arg\min_{\theta\in\mathbb{R}^{n\times d}} \sum_{t=1}^{T} \|Y_t - \theta X_t\|_2^2$. Then if

$$T \geq \frac{10k}{p^2}\left(\log(\frac{1}{\delta}) + 2d\log(10/p) + \log\det(\bar{\Gamma}\Gamma_{sb}^{-1})\right),$$

we have

$$\|\tilde{\theta} - \theta_*\|_2 \leq \frac{90\sigma_{sub}}{p}\sqrt{\frac{n + d\log(10/p) + \log\det(\bar{\Gamma}\Gamma_{sb}^{-1}) + \log(1/\delta)}{T\upsilon_{\min}(\Gamma_{sb})}}$$

with probability at least $1 - 3\delta$.

Next, we present a proof for our Theorem 1 by verifying the conditions in Theorem 5 for general nonlinear policies.

Proof of Theorem 1. Condition 1 is straightforward: $x_{t+1} = \theta_* z_t + w_t$, and $w_t \mid \mathcal{F}_t = w_t$ which is mean 0 and σ_{sub}^2-sub-Gaussian by Assumption 3. Condition 3 is also straightforward. Notice that $\upsilon_{\max}(z_t z_t^\top) \leq \text{trace}(z_t z_t^\top) = \|z_t\|_2^2 \leq b_z^2$. Therefore, we can define $\bar{\Gamma} = b_z^2 I_{n+m}$, and then $\mathbb{P}(\sum_{t=1}^{T} z_t z_t^\top \not\leq T\bar{\Gamma}) = 0 \leq \delta$.

The tricky part is Condition 2. Next, we will show the BMSB condition holds for our system. Then, by Theorem 5, we complete the proof.

Lemma 9 (Verification of BMSB condition)**.** *Define filtration $\mathcal{F}_t = \{w_0, \ldots, w_{t-1}, \eta_0, \ldots, \eta_t\}$. Under the conditions in Theorem 1,*

$$\{z_t\}_{t\geq 0} \text{ satisfies the } (1, s_z^2 I_{n+m}, p_z)\text{-BMSB condition,}$$

where $p_z = \min(p_w, p_\eta)$, $s_z = \min(s_w/4, \frac{\sqrt{3}}{2} s_\eta \bar{\eta}, \frac{s_w s_\eta}{4b_u}\bar{\eta})$.

Proof of Lemma 9. Define filtration $\mathcal{F}_t^m = \mathcal{F}(w_0, \ldots, w_{t-1}, \eta_0, \ldots, \eta_{t-1})$. Notice that the policy in Theorem 1 can be written as $u_t = \pi_t(\mathcal{F}_t^m) + \eta_t$. Note that $z_t \in \mathcal{F}_t$ is by definition. Next,

$$z_{t+1} \mid \mathcal{F}_t = \begin{bmatrix} x_{t+1} \\ u_{t+1} \end{bmatrix} \mid \mathcal{F}_t = \begin{bmatrix} \theta_* z_t + w_t \mid \mathcal{F}_t \\ \pi_{t+1}(\mathcal{F}_{t+1}^m) + \eta_{t+1} \mid \mathcal{F}_t \end{bmatrix},$$

where $\mathcal{F}_{t+1}^m = \mathcal{F}(w_0, \ldots, w_t, \eta_0, \ldots, \eta_t)$.

Notice that conditioning on $\mathcal{F}_t$, the variable $\theta_* z_t$ is determined, but the variable $\pi_{t+1}(\mathcal{F}_{t+1}^m)$ is still random due to the randomness of w_t. For the rest of the proof, we will always condition on $\mathcal{F}_t$, and omit the conditioning notation, i.e., $\cdot \mid \mathcal{F}_t$, for notational simplicity.

Consider any $\lambda = (\lambda_1^\top, \lambda_2^\top)^\top \in \mathbb{R}^{m+n}$, where $\lambda_1 \in \mathbb{R}^n$, $\lambda_2 \in \mathbb{R}^m$, $\|\lambda\|_2^2 = \|\lambda_1\|_2^2 + \|\lambda_2\|_2^2 = 1$. Define $k_0 = \max(2/\sqrt{3}, 4b_u/s_w)$. We consider three cases: (i) when $\|\lambda_2\|_2 \leq 1/k_0$ and $\lambda_1^\top \theta_* z_t \geq 0$, (ii) when $\|\lambda_2\|_2 \leq 1/k_0$ and $\lambda_1^\top \theta_* z_t < 0$, (iii) when $\|\lambda_2\|_2 > 1/k_0$. We will show in all three cases,

$$\mathbb{P}(|\lambda^\top z_{t+1}| \geq s_z) \geq p_z$$

Consequently, by Definition 2.1 in Simchowitz et al. (2018), we have $\{z_t\}$ is $(1, s_z^2 I, p_z)$-BMSB.

Case 1: when $\|\lambda_2\|_2 \leq 1/k_0$ and $\lambda_1^\top \theta_* z_t \geq 0$

$$\begin{aligned}\lambda_1^\top w_t &\le \lambda_1^\top(w_t+\theta_* z_t) \le |\lambda_1^\top(w_t+\theta_* z_t)|\\ &= |\lambda^\top z_{t+1} - \lambda_2^\top u_{t+1}| \le |\lambda^\top z_{t+1}| + |\lambda_2^\top u_{t+1}| \le |\lambda^\top z_{t+1}| + \|\lambda_2\|_2 b_u\\ &\le |\lambda^\top z_{t+1}| + b_u/k_0 \le |\lambda^\top z_{t+1}| + s_w/4\end{aligned}$$

where the last inequality uses $k_0 \ge 4b_u/s_w$.

Further, notice that $k_0 \ge 2/\sqrt{3}$, so $\|\lambda_2\|_2^2 \le 1/k_0^2 \le 3/4$, thus, $\|\lambda_1\|_2^2 \ge 1/4$, which means $\|\lambda_1\|_2 \ge 1/2$. Therefore,

$$\mathbb{P}(\lambda_1^\top w_t \ge s_w/2) = \mathbb{P}(\frac{\lambda_1^\top w_t}{\|\lambda_1\|_2} \ge \frac{s_w}{2\|\lambda_1\|_2}) \ge \mathbb{P}(\frac{\lambda_1^\top w_t}{\|\lambda_1\|_2} \ge s_w) = p_w$$

Then,

$$\begin{aligned}\mathbb{P}(|\lambda^\top z_{t+1}| \ge s_z) &\ge \mathbb{P}(|\lambda^\top z_{t+1}| \ge s_w/4) = \mathbb{P}(|\lambda^\top z_{t+1}| + s_w/4 \ge s_w/2)\\ &\ge \mathbb{P}(\lambda_1^\top w_t \ge s_w/2) \ge p_w\end{aligned}$$

which completes case 1.

Case 2: when $\|\lambda_2\|_2 \le 1/k_0$ and $\lambda_1^\top \theta_* z_t < 0$.

$$\begin{aligned}\lambda_1^\top w_t &\ge \lambda_1^\top(w_t+\theta_* z_t) \ge -|\lambda_1^\top(w_t+\theta_* z_t)|\\ &= -|\lambda^\top z_{t+1} - \lambda_2^\top u_{t+1}| \ge -|\lambda^\top z_{t+1}| - |\lambda_2^\top u_{t+1}| \ge -|\lambda^\top z_{t+1}| - \|\lambda_2\|_2 b_u\\ &\ge -|\lambda^\top z_{t+1}| - b_u/k_0 \ge -|\lambda^\top z_{t+1}| - s_w/4\end{aligned}$$

where the last inequality uses $k_0 \ge 4b_u/s_w$.

Further, notice that $k_0 \ge 2/\sqrt{3}$, so $\|\lambda_2\|_2^2 \le 1/k_0^2 \le 3/4$, thus, $\|\lambda_1\|_2^2 \ge 1/4$, which means $\|\lambda_1\|_2 \ge 1/2$. Therefore,

$$\mathbb{P}(\lambda_1^\top w_t \le -s_w/2) = \mathbb{P}(\frac{\lambda_1^\top w_t}{\|\lambda_1\|_2} \le -\frac{s_w}{2\|\lambda_1\|_2}) \ge \mathbb{P}(\frac{\lambda_1^\top w_t}{\|\lambda_1\|_2} \le -s_w) = \mathbb{P}(\frac{-\lambda_1^\top w_t}{\|\lambda_1\|_2} \ge s_w) = p_w$$

by $s_w/(2\|\lambda_1\|_2) \le s_w$, and thus $-s_w/(2\|\lambda_1\|_2) \ge -s_w$, and Assumption 3.

Consequently,

$$\begin{aligned}\mathbb{P}(|\lambda^\top z_{t+1}| \ge s_z) &\ge \mathbb{P}(|\lambda^\top z_{t+1}| \ge s_w/4) = \mathbb{P}(-|\lambda^\top z_{t+1}| - s_w/4 \le -s_w/2)\\ &\ge \mathbb{P}(\lambda_1^\top w_t \le -s_w/2) \ge p_w\end{aligned}$$

which completes case 2.

Case 3: when $\|\lambda_2\|_2 > 1/k_0$. Define $v = \bar\eta s_\eta/k_0 = \min(\sqrt{3}\bar\eta s_\eta/2, s_w\bar\eta s_\eta/(4b_u))$. Define

$$\begin{aligned}\Omega_1^\lambda &= \{w_t \in \mathbb{R}^n \mid \lambda_1^\top(w_t+\theta_* z_t) + \lambda_2^\top(\pi_{t+1}(\mathcal{F}_{t+1}^m)) \ge 0\}\\ \Omega_2^\lambda &= \{w_t \in \mathbb{R}^n \mid \lambda_1^\top(w_t+\theta_* z_t) + \lambda_2^\top(\pi_{t+1}(\mathcal{F}_{t+1}^m)) < 0\}\end{aligned}$$

Notice that $\mathbb{P}(w_t \in \Omega_1^\lambda) + \mathbb{P}(w_t \in \Omega_2^\lambda) = 1$.

$$\begin{aligned}\mathbb{P}(|\lambda^\top z_{t+1}| \ge s_z) &\ge \mathbb{P}(|\lambda^\top z_{t+1}| \ge v) = \mathbb{P}(\lambda^\top z_{t+1} \ge v) + \mathbb{P}(\lambda^\top z_{t+1} \le -v)\\ &\ge \mathbb{P}(\lambda^\top z_{t+1} \ge v, w_t \in \Omega_1^\lambda) + \mathbb{P}(\lambda^\top z_{t+1} \le -v, w_t \in \Omega_2^\lambda)\\ &\ge \mathbb{P}(\lambda_2^\top \eta_{t+1} \ge v, w_t \in \Omega_1^\lambda) + \mathbb{P}(\lambda_2^\top \eta_{t+1} \le -v, w_t \in \Omega_2^\lambda)\\ &= \mathbb{P}(\lambda_2^\top \eta_{t+1} \ge v)\mathbb{P}(w_t \in \Omega_1^\lambda) + \mathbb{P}(\lambda_2^\top \eta_{t+1} \le -v)\mathbb{P}(w_t \in \Omega_2^\lambda)\\ &\ge p_\eta\end{aligned}$$

where the last inequality is because of the following arguments. Notice that

$$\begin{aligned}\mathbb{P}(\lambda_2^\top \eta_{t+1} \geq v) &= \mathbb{P}(\lambda_2^\top \eta_{t+1}/\|\lambda_2\|_2 \geq v/\|\lambda_2\|_2)\\ &= \mathbb{P}(\lambda_2^\top \tilde{\eta}_{t+1}/\|\lambda_2\|_2 \geq v/(\|\lambda_2\|_2 \bar{\eta}))\\ &\geq \mathbb{P}(\lambda_2^\top \tilde{\eta}_{t+1}/\|\lambda_2\|_2 \geq k_0 v/(\bar{\eta}))\\ &= \mathbb{P}(\lambda_2^\top \tilde{\eta}_{t+1}/\|\lambda_2\|_2 \geq s_\eta) \geq p_\eta\end{aligned}$$

Then,

$$\mathbb{P}(\lambda_2^\top \eta_{t+1} \leq -v) = \mathbb{P}(-\lambda_2^\top \eta_{t+1} \geq v) \geq p_\eta$$

This completes the proof of Case 3. □

Finally, we apply Theorem 5. Notice that $d = m+n$ in our problem, and $\log\det(\bar{\Gamma}\Gamma_{sb}^{-1}) = 2(m+n)\log(b_z/s_z) = O((m+n)\log(b_z/\bar{\eta}))$ as $\bar{\eta} \to 0$, $\upsilon_{\min}(\Gamma_{sb}) = s_z^2 = O(1/\bar{\eta}^2)$ as $\bar{\eta} \to 0$, and $p = p_z$ here. Therefore, for T large enough, we have:

$$\|\tilde{\theta}_T - \theta_*\|_2 \leq O\left(\sqrt{n+m}\frac{\sqrt{\log(b_z/\bar{\eta} + 1/\delta)}}{\sqrt{T}\bar{\eta}}\right).$$

□

B.2 Proof of Corollary 1

Corollary 1 follows directly from Theorem 1. We only need to verify the boundedness of the states and actions. In the following, we will show that $u_t \in \mathbb{U}$ for all t and $\|x_t\|_2 \leq O(\sqrt{mn})$ for all t. Notice that though we can further show a much smaller bound $\|x_t\|_2 \leq x_{\max}$ with probability $(1-p)$ in Theorem 3, Theorem 1 requires an almost sure bound and thus we provide a larger bound $\|x_t\|_2 \leq O(\sqrt{mn})$ here.

In the following, we show that $u_t \in \mathbb{U}$ for all t and $\|x_t\|_2 \leq O(\sqrt{mn})$ for all t.

Lemma 10 (Action constraint satisfaction)**.** *When applying Algorithm 1, $u_t \in \mathbb{U}$ for all t and for any $w_k \in \mathbb{W}$.*

Proof. Notice that $u_t = \sum_{k=1}^{H^{(e-1)}} M_t[k]\hat{w}_{t-k} + \eta_t$. Hence, for any $1 \leq j \leq k_u$, we have

$$D_{u,j}^\top u_t = D_{u,j}^\top \sum_{k=1}^{H^{(e-1)}} M_t[k]\hat{w}_{t-k} + D_{u,j}^\top \eta_t \leq \sum_{k=1}^{H^{(e-1)}} \|D_{u,j}^\top M_t[k]\|_1 w_{\max} + \|D_u\|_\infty \|\eta_t\|_\infty = g_j^u(\mathbf{M}_t) + \|D_u\|_\infty \|\eta_t\|_\infty$$

Our goal is to show that $g_j^u(\mathbf{M}_t) + \|D_u\|_\infty\|\eta_t\|_\infty \leq d_{u,j}$ for all j and for all $t \geq 0$. This is straightforward when $t_1^{(e)} \leq t \leq t_1^{(e)} + T_D^{(e)} - 1$ and $t_2^{(e)} \leq t \leq T^{(e+1)} - 1$. For example, when $t_1^{(e)} \leq t \leq t_1^{(e)} + T_D^{(e)} - 1$, we have $\mathbf{M}_t = \mathbf{M}_\dagger^{(e)}$ and $\|\eta_t\|_\infty \leq \bar{\eta}^{(e)}$, which leads to $g_j^u(\mathbf{M}_\dagger^{(e)}) + \|D_u\|_\infty\|\eta_t\|_\infty = g_j^u(\mathbf{M}_\dagger^{(e)}) + c_3\|\bar{\eta}^{(e)}\|_\infty \leq d_{u,j}$ by RobustCE and Lemma 6. Similar results can be shown for $t_2^{(e)} \leq t \leq T^{(e+1)} - 1$.

Next, we focus on the safe policy transition stages. It suffices to show that $u_t = \sum_{k=1}^{H^{(e-1)}} M_t[k]\hat{w}_{t-k} + \eta_t$ in all stages of Algorithm 2. In the following, we will adopt the notations in Algorithm 2. In Step 1 of Algorithm 2, we have $\|\eta_t\|_\infty \leq \bar{\eta}_{\min} \leq \bar{\eta}$ and $\mathbf{M}_t \in \Omega$ by the convexity of Ω. Therefore, we have $g_j^u(\mathbf{M}_t) + \|D_u\|_\infty\|\eta_t\|_\infty = g_j^u(\mathbf{M}_t) + c_3\|\bar{\eta}\|_\infty \leq d_{u,j}$, where we used the definition of Ω in RobustCE. In Step 2, we have $\|\eta_t\|_\infty \leq \bar{\eta}'$ and $\mathbf{M}_t \in \Omega'$ by the convexity of Ω'. Therefore, we have $g_j^u(\mathbf{M}_t) + \|D_u\|_\infty\|\eta_t\|_\infty = g_j^u(\mathbf{M}_t) + c_3\|\bar{\eta}'\|_\infty \leq d_{u,j}$, where we used the definition of Ω' by RobustCE with input $\bar{\eta}'$. □

Lemma 11 (Almost sure upper bound on x_t)**.** *Consider DAP policy $u_t = \sum_{k=1}^{H_t} M_t[k]\hat{w}_{t-k} + \eta_t$, where $\mathbf{M}_t \in \mathcal{M}_{H_t}$, $\{H_t\}_{t\geq 0}$ is non-decreasing, and $\|\eta_t\|_\infty \leq \eta_{\max}$. Suppose $H_0 \geq \log(2\kappa)/\log((1-\gamma)^{-1})$ and $\eta_{\max} \leq w_{\max}/\kappa_B$. Let $\{x_t, u_t\}_{t\geq 0}$ denote the trajectory generated by this policy on the system with parameter θ_* and disturbance w_t. Then, there exists $b_x = 4\sqrt{n}\kappa w_{\max}/\gamma + 4\sqrt{mn}\kappa^3\kappa_B w_{\max}/\gamma^2 = O(\sqrt{mn})$ such that*

$$\|x_t\|_2 \leq b_x, \quad \forall\, t \geq 0, \quad \forall\, w_k,\, \hat{w}_k \in \mathbb{W}.$$

This lemma is a natural extension of Lemma 2 in Li et al. (2020) and the proof is deferred to Appendix G.3.

Proof of Corollary 1. By letting $\delta^{(e)} = \frac{p}{6e^2}$ for $e \geq 1$, we have that $\|\tilde{\theta}^{(e)} - \theta_*\|_2 \leq O(\frac{(\sqrt{m+n})\sqrt{\log(\sqrt{mn}/\bar{\eta}^{(e-1)}) + \log(e)}}{\sqrt{T_D^{(e-1)}}\bar{\eta}^{(e-1)}})$ w.p. $1 - p/(2e^2)$. Notice that $\|\hat{\theta}^{(e)} - \theta_*\|_F \leq \|\tilde{\theta}^{(e)} - \theta_*\|_F \leq \sqrt{n}\|\tilde{\theta}^{(e)} - \theta_*\|_2$, which completes the proof. □

C Feasibility

This appendix provides a proof for Theorem 2. We will first establish the recursive feasibility and then prove the initial feasibility. For notational simplicity, we define $\Omega_0 := \Omega(\hat{\theta}^{(0)}, \epsilon_x^{(0)} + \epsilon_0, \epsilon_u^{(0)})$.

Proof of recursive feasibility: To show that Algorithm 1 and 2 are feasible at all stages, we need to show that $\Omega_\dagger^{(e)}, \Omega^{(e)}, \Omega_\dagger^{(e)} \cap \Omega^{(e)}, \Omega_\dagger^{(e+1)} \cap \Omega^{(e)}$ are all non-empty for $e \geq 0$. Notice that it suffices to show that $\Omega_0 \subseteq \Omega_\dagger^{(e)}$ and $\Omega_0 \subseteq \Omega^{(e)}$ for all $e \geq 0$.

Consider $\Omega_\dagger^{(e)}$ for $e \geq 0$. Notice that $\Omega_0 \subseteq \Omega_\dagger^{(0)}$ by definition, so we will focus on $e \geq 1$ below. We first consider the action constraints. For any $\mathbf{M} \in \Omega_0$, we have

$$g_j^u(\mathbf{M}) \leq d_{u,j} - \epsilon_{\eta,u}(\bar{\eta}^{(0)}), \quad \forall\, 1 \leq j \leq k_u.$$

Since $\bar{\eta}^{(0)} \geq \bar{\eta}^{(e)}$ by condition (ii) of Theorem 2, we have $\epsilon_{\eta,u}(\bar{\eta}^{(0)}) \geq \epsilon_{\eta,u}(\bar{\eta}^{(e)})$, so $\mathbf{M}$ satisfies the action constraints in $\Omega_\dagger^{(e)}$:

$$g_j^u(\mathbf{M}) \leq d_{u,j} - \epsilon_{\eta,u}(\bar{\eta}^{(e)}) \quad \forall\, 1 \leq j \leq k_u.$$

Next, we consider the state constraints. Notice that $\hat{\theta}^{(e)} \in \Theta_{ini}$ by `ModelEst`, so $\|\hat{\theta}^{(e)} - \hat{\theta}^{(0)}\|_F \leq r^{(0)}$ for $e \geq 1$. By Lemma 4, for any $\mathbf{M} \in \Omega_0$, we have

$$\begin{aligned}
g_i^x(\mathbf{M}; \hat{\theta}^{(e)}) &\leq g_i^x(\mathbf{M}; \hat{\theta}^{(0)}) + \epsilon_\theta(r^{(0)}) \\
&\leq d_{x,i} - \epsilon_x^{(0)} - \epsilon_0 + \epsilon_\theta(r^{(0)}) \\
&= d_{x,i} - \epsilon_\theta(r^{(0)}) - \epsilon_{\eta,x}(\bar{\eta}^{(0)}) - \epsilon_H(H^{(0)}) - \epsilon_v(\Delta_M^{(0)}, H^{(0)}) - \epsilon_0 + \epsilon_\theta(r^{(0)}) \\
&= d_{x,i} - \epsilon_{\eta,x}(\bar{\eta}^{(0)}) - \epsilon_H(H^{(0)}) - \epsilon_v(\Delta_M^{(0)}, H^{(0)}) - \epsilon_0
\end{aligned}$$

Further, since $r^{(e)} \leq r^{(1)} \leq \epsilon_0/(c_1\sqrt{mn})$ by condition (ii) of Theorem 2, we have $\epsilon_\theta(r^{(e)}) \leq \epsilon_\theta(r^{(1)}) \leq \epsilon_0$, so

$$\begin{aligned}
g_i^x(\mathbf{M}; \hat{\theta}^{(e)}) &\leq d_{x,i} - \epsilon_{\eta,x}(\bar{\eta}^{(0)}) - \epsilon_H(H^{(0)}) - \epsilon_v(\Delta_M^{(0)}, H^{(0)}) - \epsilon_0 \\
&\leq d_{x,i} - \epsilon_{\eta,x}(\bar{\eta}^{(0)}) - \epsilon_H(H^{(0)}) - \epsilon_v(\Delta_M^{(0)}, H^{(0)}) - \epsilon_\theta(r^{(e)}) \\
&\leq d_{x,i} - \epsilon_{\eta,x}(\bar{\eta}^{(e)}) - \epsilon_H(H^{(e)}) - \epsilon_v(\Delta_M^{(e)}, H^{(e)}) - \epsilon_\theta(r^{(e)})
\end{aligned}$$

by condition (ii) of Theorem 2. So $\mathbf{M} \in \Omega_\dagger^{(e)}$ for $e \geq 1$. Similarly, we can show $\mathbf{M} \in \Omega^{(e)}$ for $e \geq 0$. This completes the recursive feasibility.

Proof of initial feasibility. By Lemma 4 and Corollary 2 in Li et al. (2020), we can construct $\mathbf{M}_F$ with length $H^{(0)}$ based on K_F in Assumption 2 such that

$$\begin{aligned}
g_i^x(\mathbf{M}_F; \theta_*) &\leq d_{x,i} - \epsilon_{F,x} + \epsilon_P(H^{(0)}) \\
g_i^u(\mathbf{M}_F) &\leq d_{u,j} - \epsilon_{F,u} + \epsilon_P(H^{(0)}).
\end{aligned}$$

where ϵ_P corresponds to $\epsilon_1 + \epsilon_3$ in Li et al. (2020).[8]

Therefore, by Lemma 4, we have

$$g_i^x(\mathbf{M}_F; \hat{\theta}^{(0)}) \leq d_{x,i} - \epsilon_{F,x} + \epsilon_P(H^{(0)}) + \epsilon_\theta(r^{(0)}) \quad g_i^u(\mathbf{M}_F) \leq d_{u,j} - \epsilon_{F,u} + \epsilon_P(H^{(0)}).$$

Therefore, $\mathbf{M}_F \in \Omega_0$ if (**??**) in Theorem 2 holds.

[8] Notice that $\epsilon_1 + \epsilon_3 = O(n\sqrt{m}H(1-\gamma)^H)$ in Li et al. (2020), but we improve the bound to $\epsilon_P = O(\sqrt{mn}(1-\gamma)^H)$. Specifically, $\epsilon_3 = O(\sqrt{n}(1-\gamma)^H)$ remains unchanged, but we can show $\epsilon_1(H) = O(\sqrt{mn}(1-\gamma)^H)$. This is because the proof of Lemma 1 in Li et al. (2020) shows that $\epsilon_1 = O(b_x(1-\gamma)^H)$, where $\|x_t\|_2 \leq b_x$ a.s.. In Lemma 11, we show $b_x = O(\sqrt{mn})$ in this paper, so we have $\epsilon_1(H) = O(\sqrt{mn}(1-\gamma)^H)$.

D Constraint Satisfaction

This section provides a proof for the constraint satisfaction guarantee in Theorem 3. Notice that the control constraint satisfaction has already been established in Lemma 10. Hence, we will focus on state constraint satisfaction in this appendix. Firstly, we present and prove a general state constraint satisfaction lemma for time-varying approximate DAPs, which is more general than Lemma 2. Secondly, we prove the state constraint satisfaction of our algorithms by showing that our algorithms satisfy the conditions in the general state constraint satisfaction lemma.

D.1 A General State Constraint Satisfaction Lemma

This subsection provides a general state constraint satisfaction lemma for time-varying approximate DAPs, which includes Lemma 2 as a special case.

Lemma 12 (General Constraint Satisfaction Lemma)**.** *Consider the time-varying approximate DAPs in* (9)*, where* $\mathbf{M}_t \in \mathcal{M}_{H_t}$ *for non-decreasing* $\{H_t\}_{t\geq 0}$*,* $\hat{\theta}_t \in \Theta_{\text{ini}}$*. Define*

$$\begin{aligned}
\epsilon_{H,t} &= (1-\gamma)^{H_t} \cdot \|D_x\|_\infty \kappa x_{\max} \\
\epsilon_{v,t} &= \sqrt{mnH_t}\Delta_{M,t} \cdot \|D_x\|_\infty w_{\max}\kappa\kappa_B/\gamma^2, \quad \Delta_{M,t} = \max_{1\leq k\leq H_t} \frac{\|\mathbf{M}_t - \mathbf{M}_{t-k}\|_F}{k}, \\
\epsilon_{\theta,t} &= c_1 \max_{0\leq k\leq H_t} \|\hat{\theta}_{t-k} - \theta_*\|_F \\
\epsilon_{\eta,x,t} &= c_2\sqrt{m} \max_{1\leq k\leq H_t} \bar{\eta}_{t-k},
\end{aligned}$$

where c_1, c_2 *are defined in Lemma 4 and Lemma 6, and we let* $\mathbf{M}_t = \mathbf{M}_0$*,* $\bar{\eta}_t = 0$*,* $H_t = H_0$*,* $\hat{\theta}_t = \hat{\theta}_0$*,* $w_t = \hat{w}_t = x_t = 0$*, for* $t \leq -1$*.*

For any $t \geq 0$*, if* $x_s \in \mathbb{X}, u_s \in \mathbb{U}$ *for all* $s \leq t-1$ *and*

$$g_i^x(\mathbf{M}_t; \hat{\theta}_t) \leq d_{x,i} - \epsilon_{H,t} - \epsilon_{\eta,x,t} - \epsilon_{\theta,t} - \epsilon_{v,t}, \quad \forall\, 1 \leq i \leq k_x, \tag{12}$$

then $x_t \in \mathbb{X}$*.*

Consequently, if (12) *holds and* $u_t \in \mathbb{U}$ *for all* $t \geq 0$*, then* $x_t \in \mathbb{X}$ *for all* $t \geq 0$*.*

Proof. Consider stage $t \geq 0$. By Lemma 3, for any $1 \leq i \leq k_x$, we have

$$\begin{aligned}
D_{x,i}^\top x_t &= D_{x,i}^\top A_*^{H_t} x_{t-H_t} \\
&\quad + \sum_{k=2}^{2H_t}\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* M_{t-i}[k-i]\hat{w}_{t-k}\mathbb{1}_{(1\leq k-i\leq H_{t-i})} + \sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} w_{t-i} + \sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i} \\
&= D_{x,i}^\top A_*^{H_t} x_{t-H_t} \\
&\quad + \sum_{k=1}^{2H_t} D_{x,i}^\top \Big(A_*^{i-1}\mathbb{1}_{k\leq H_t} + \sum_{i=1}^{H_t} A_*^{i-1} B_* M_{t-i}[k-i]\mathbb{1}_{(1\leq k-i\leq H_{t-i})}\Big)\hat{w}_{t-k} + \sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1}(w_{t-i} - \hat{w}_{t-i}) \\
&\quad + \sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i} \\
&= D_{x,i}^\top A_*^{H_t} x_{t-H_t} + \sum_{k=1}^{2H_t} D_{x,i}^\top \Phi_k^x(\mathbf{M}_{t-H_t:t-1}; \theta_*)\hat{w}_{t-k} + \sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1}(w_{t-i} - \hat{w}_{t-i}) \\
&\quad + \sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i} \\
&\leq \|D_x\|_\infty \kappa(1-\gamma)^{H_t} x_{\max} + g_i^x(\mathbf{M}_{t-H_t:t-1}; \theta_*) + \|D_x\|_\infty \kappa/\gamma \max_{1\leq k\leq H_t} \|\hat{\theta}_{t-k} - \theta_*\|_2 z_{\max} \\
&\quad + \|D_x\|_\infty \kappa\kappa_B/\gamma\sqrt{m} \max_{1\leq k\leq H_t} \bar{\eta}_{t-k}
\end{aligned}$$

$$\begin{aligned}
&\leq \epsilon_{H,t} + \mathring{g}_i^x(\mathbf{M}_t;\theta_*) + \epsilon_{v,t} + \|D_x\|_\infty \kappa/\gamma \max_{1\leq k\leq H_t} \|\hat{\theta}_{t-k} - \theta_*\|_2 z_{\max} + \epsilon_{\eta,x,t} \\
&\leq \epsilon_{H,t} + \mathring{g}_i^x(\mathbf{M}_t;\hat{\theta}_t) + \epsilon_{\hat{\theta}}(\|\theta_* - \hat{\theta}_t\|_F) + \epsilon_{v,t} + \|D_x\|_\infty \kappa/\gamma \max_{1\leq k\leq H_t} \|\hat{\theta}_{t-k} - \theta_*\|_2 z_{\max} + \epsilon_{\eta,x,t} \\
&\leq \epsilon_{H,t} + \mathring{g}_i^x(\mathbf{M}_t;\hat{\theta}_t) + \epsilon_{\theta,t} + \epsilon_{v,t} + \epsilon_{\eta,x,t} \\
&\leq d_{x,i}
\end{aligned}$$

where we used Lemma 5, Lemma 4, $x_s \in \mathbb{X}, u_s \in \mathbb{U}$ for all $s \leq t-1$, and (12). The last inequality guarantees $x_t \in \mathbb{X}$. Therefore, the proof can be completed by induction. □

D.2 Proof of Theorem 3

Define an event

$$\mathcal{E}_{\text{safe}} = \{\theta_* \in \bigcap_{e=0}^{N-1} \Theta^{(e)}\}. \tag{13}$$

Notice that

$$\mathbb{P}(\mathcal{E}_{\text{safe}}) = 1 - \mathbb{P}(\mathcal{E}_{\text{safe}}^c) \geq 1 - \sum_{e=0}^{N} \mathbb{P}(\theta_* \notin \Theta^{(e)}) \geq 1 - \sum_{e=1}^{N} p/(2e^2) \geq 1-p$$

where we used Corollary 1 and $\theta_* \in \Theta^{(0)} = \Theta_{\text{ini}}$. In the following, we will condition on the event $\mathcal{E}_{\text{safe}}$ and show $x_t \in \mathbb{X}$ for all $t \geq 0$ under this event. By Lemma 12, we only need to show (12) for any t.

We discuss three possible cases based on the value of t. We introduce some notations for our case-by-case discussion: let $W_1^{(e)}, W_2^{(e)}$ denote the W_1, W_2 defined in Algorithm 2 during the transition in Phase 1, and let $\tilde{W}_1^{(e)}, \tilde{W}_2^{(e)}$ denote the W_1, W_2 defined in Algorithm 2 during the transition in Phase 2.

(Case 1: when $T^{(e)} \leq t \leq T^{(e)} + W_1^{(e)} - 1$**.)** In this case, $\mathbf{M}_t \in \Omega^{(e-1)}$, so

$$g_i^x(\mathbf{M}_t;\hat{\theta}^{(e)}) \leq d_{x,i} - \epsilon_H(H^{(e-1)}) - \epsilon_v(\Delta_M^{(e-1)}) - \epsilon_\theta(r^{(e)})$$

Notice that $H_t = H^{(e-1)}$, so $\epsilon_{H,t} = \epsilon_H(H^{(e-1)})$. Further, by our algorithm design, $\Delta_{M,t} \leq \Delta_M^{(e-1)}$. Since $\tilde{W}_1^{(e-1)} \geq H^{(e-1)}$ and $\eta_k = 0$ for $t_1^{(e-1)} + T_D^{(e-1)} \leq k \leq t$, we have $\epsilon_{\eta,x,t} = \max_{1\leq k\leq H_t} c_2\sqrt{m}\bar{\eta}_{t-k} = 0$. Next, since $r_\theta^{(e)} \leq r_\theta^{(e-1)}$ by Condition 2 of Theorem 2 for $e \geq 1$, $\epsilon_{\theta,t} \leq \epsilon_\theta(r^{(e)})$. So we satisfy (12).

(Case 2: when $T^{(e)} + W_1^{(e)} \leq t \leq t_1^{(e)} + T_D^{(e)} + \tilde{W}_1^{(e)} - 1$**.)** We have $\mathbf{M}_t \in \Omega_\dagger^{(e)}$. So

$$g_i^x(\mathbf{M}_t;\hat{\theta}^{(e)}) \leq d_{x,i} - \epsilon_H(H^{(e)}) - \epsilon_v(\Delta_M^{(e)}) - \epsilon_{\eta,x}(\bar{\eta}^{(e)}) - \epsilon_\theta(r^{(e)})$$

Next, $H_t = H^{(e)}$, since $W_1^{(e)} \geq H^{(e)}$, we have$\epsilon_{\theta,t} \leq \epsilon_w(r^{(e)})$, and $\epsilon_{v,t} = \epsilon_v(\Delta_M^{(e)})$. Since we take minimum over potential $\bar{\eta}$ in Step 1 of Algorithm 2 and $W_1^{(e)} \geq H^{(e)}$, we have $\epsilon_{\eta,x,t} \leq \epsilon_{\eta,x}(\bar{\eta}^{(e)})$. So we satisfy (12).

(Case 3: when $t_1^{(e)} + T_D^{(e)} + \tilde{W}_1^{(e)} \leq t \leq T^{(e+1)} - 1$**.)** We have $\mathbf{M}_t \in \Omega^{(e)}$. So

$$g_i^x(\mathbf{M}_t;\hat{\theta}^{(e+1)}) \leq d_{x,i} - \epsilon_H(H^{(e)}) - \epsilon_v(\Delta_M^{(e)}) - \epsilon_\theta(r^{(e+1)})$$

Next, $H_t = H^{(e)}$, since $W_1^{(e)} \geq H^{(e)}$, we have $\epsilon_{\theta,t} \leq \epsilon_\theta(r^{(e+1)})$ by $r^{(e+1)} \leq r_\theta^{(e)}$, and $\epsilon_{v,t} \leq \epsilon_v(\Delta_M^{(e)})$. Since we take minimum over potential $\bar{\eta}$ in Step 1 of Algorithm 2 and $\tilde{W}_1^{(e)} \geq H^{(e)}$, we have $\epsilon_{\eta,x,t} = 0$. So we satisfy (12).

In conclusion, we satisfy (12) for all $t \geq 0$. By Lemma 12 and Lemma 10, we can show state constraint satisfaction under $\mathcal{E}_{\text{safe}}$.

E Regret Analysis

In this section, we provide a proof for Theorem 4. Specifically, we first prove the regret bound and then verify the conditions for feasibility and constraint satisfaction. Before the formal proof, we note that the statement in Theorem 4 has the following typos.

E.1 Proof of the Regret Bound

Our proof of the regret bound relies on decomposing the regret into several parts and bounding each part.

Firstly, we decompose the T stages into two parts and decompose the regret accordingly. For $e \geq 0$, define

$$\mathcal{T}_1^{(e)} = \{T^{(e)} \leq t \leq t_2(e) + H^{(e)} - 1\}, \quad \mathcal{T}_2^{(e)} = \{t_2(e) + H^{(e)} \leq t \leq T^{(e+1)} - 1\}.$$

Then, decompose the regret by the stage decomposition below:

$$\text{Regret} = \sum_{t=0}^{T-1} (l(x_t, u_t) - J^*) = \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_1^{(e)}} (l(x_t, u_t) - J^*)}_{\text{First term}} + \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (l(x_t, u_t) - J^*)}_{\text{Second term}} \tag{14}$$

The first term can be bounded straightforwardly by the fact that the single-stage regret is bounded and the total number of stages in $\mathcal{T}_1^{(e)}$ for all e can be bounded by $O(T^{2/3})$.

Lemma 13 (Regret Bound of the First Term). *When the event $\mathcal{E}_{\mathit{safe}}$ defined in* (13) *happens, under the conditions in Theorem 4, we have*

$$\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_1^{(e)}} (l(x_t, u_t) - J^*) \leq O(T^{2/3})$$

Proof. When $\mathcal{E}_{\text{safe}}$ is true, by Theorem 3, we have $x_t \in \mathbb{X}$ and $u_t \in \mathbb{U}$, thus $\|x_t\|_2 \leq x_{\max}$ and $\|u\|_2 \leq u_{\max}$ and $l(x_t, u_t) - J^* \leq \|Q\|_2 x_{\max}^2 + \|R\|_2 u_{\max}^2 = O(1)$.

Next, we bound the number of stages in $\mathcal{T}_1^{(e)}$. Under the conditions in Theorem 4, the number of the stages in $\mathcal{T}_1^{(e)}$ is $T_D^{(e)} + H^{(e)}$ plus the safe policy transition stages in Phase 1 and Phase 2. Since $\mathcal{M}_{H^{(e)}}$ is a bounded set, the number of stages in SafeTransit between any two policies in $\mathcal{M}_{H^{(e)}}$ can be bounded by $O(\max(1/\Delta_M^{(e)}, H^{(e)})) = \tilde{O}(\sqrt{mn}(T^{(e+1)})^{1/3})$, where we used $H^{(e)} = O(\log(T^{(e+1)}))$ and $\Delta_M^{(e)} = O(\frac{\epsilon_F^x}{\sqrt{mnH^{(e)}}(T^{(e+1)})^{1/3}})$. Further, by $T^{(e+1)} = 2T^{(e)}$, $T_D^{(e)} = (T^{(e+1)} - T^{(e)})^{2/3}$, we have $T_D^{(e)} = O((T^{(e+1)})^{2/3})$. Consequently, the total number of stages in $\mathcal{T}_1^{(e)}$ can be bounded by $O((T^{(e+1)})^{2/3})$ (notice that $T^{(e+1)})^{1/3} \geq \sqrt{mn}$ by our condition of $T^{(1)}$ in Theorem 4).

Finally, with the help of the algebraic fact in Lemma 14, we are able to bound the total regret in all episodes by $O((T^{(e+1)})^{2/3})$. □

Lemma 14 is a technical fact that will be used throughout our regret proof.

Lemma 14 (An algebraic fact). *When $T^{(e)} = 2^{e-1}T^{(1)}$, and $T^{(N)} \geq T > T^{(N-1)}$, $N \leq O(\log T)$. Further, for any $\alpha > 0$, we have*

$$\sum_{e=1}^{N} (T^{(e)})^\alpha = O(T^\alpha)$$

Proof. By $T \geq T^{(N-1)} \geq 2^{(N-2)}$, we have $\log T \geq (N-2)\log(2)$, so $N \leq O(\log T)$. Further, $\sum_{e=1}^{N}(T^{(e)})^\alpha = \sum_{e=1}^{N}(2^{e-1})^\alpha (T^{(1)})^\alpha \leq O((2^N)^\alpha (T^{(1)})^\alpha) \leq O(T^\alpha)$. □

The second term in (14) is more complicated to bound, so we further decompose it into four parts as follows.

$$\begin{aligned}\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (l(x_t, u_t) - J^*) &= \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (l(x_t, u_t) - l(\hat{x}_t, \hat{u}_t))}_{\text{Part i}} + \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (l(\hat{x}_t, \hat{u}_t) - f(\mathbf{M}^{(e)}; \theta_*))}_{\text{Part ii}} \\ &+ \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}^{(e)}; \theta_*) - f(\mathbf{M}^*_{H^{(e)}}; \theta_*))}_{\text{Part iii}} + \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - J^*)}_{\text{Part iv}},\end{aligned}$$

where we introduced auxiliary states $\hat{x}_t$ and actions $\hat{u}_t$ defined as

$$\hat{x}_t = \sum_{k=1}^{2H^{(e)}} \Phi_k^x(\mathbf{M}^{(e)}; \theta_*) w_{t-k}, \quad \hat{u}_t = \sum_{k=1}^{H^{(e)}} M^{(e)}[k] w_{t-k},$$

which are basically the approximate states and the actions generated by the disturbance-action policy $\mathbf{M}^{(e)}$ computed in Phase 2 of Algorithm 1 when the actual disturbances w_{t-k} are known. We also introduce an auxiliary policy $\mathbf{M}^*_{H^{(e)}}$ in Part iii, which is defined as the optimal DAP policy in (2) with a memory length $H = H^{(e)}$ under a known model, i.e. $\mathbf{M}^*_{H^{(e)}} = \arg\min_{\mathbf{M} \in \Omega_*^{(e)}} f(\mathbf{M}; \theta_*)$, where $\Omega_*^{(e)}$ is defined by (3) with $H = H^{(e)}$.

The rest of the proof is to bound Parts i-iv. Establishing the bound on Part iii is the major part of the proof and the bound on Part iii is the dominating term in our regret bound, so we will present our bound on Part iii first. Then, we will establish bounds on Parts i, ii, and iv.

E.1.1 Bound on Part iii

Notice that $\mathbf{M}^{(e)}$ is the solution to the CCE program in (7) and $\mathbf{M}^*_{H^{(e)}}$ is the solution to the optimal DAP program in (2). Further, the CCE program (7) can be viewed as a slightly perturbed version of the optimal DAP program (2) due to model estimation errors and constraint-tightening terms. Therefore, we can bound Part iii by the perturbation analysis.

Specifically, we establish the following general perturbation bound. This bound is not only useful for our bound on Part iii but also helps the discussions after Theorem 4 on the reasons for including the pure exploitation phases.

Lemma 15 (Perturbation analysis for CCE)**.** *Consider a fixed memory length* $H \geq \log(2\kappa)/\log((1-\gamma)^{-1})$ *and* $\theta_1, \theta_2 \in \Theta_{\text{ini}}$. *Consider two CCE programs* $\mathbf{M}_1 = \arg\min_{\mathbf{M} \in \Omega(\theta_1, \epsilon_{x1}, \epsilon_{u1})} f(\mathbf{M}; \theta_1)$ *and* $\mathbf{M}_2 = \arg\min_{\mathbf{M} \in \Omega(\theta_2, \epsilon_{x2}, \epsilon_{u2})} f(\mathbf{M}; \theta_2)$. *Suppose there exists* $\epsilon_g > 0$ *such that* $\Omega(\theta_1, \epsilon_{x1} + \epsilon_g, \epsilon_{u1} + \epsilon_g) \cap \Omega(\theta_2, \epsilon_{x2} + \epsilon_g, \epsilon_{u2} + \epsilon_g)$ *is non-empty. Then, we have*

$$f(\mathbf{M}_1, \theta_2) - f(\mathbf{M}_2, \theta_2) \leq O(mnr + (\sqrt{mn} + \sqrt{k_x + k_u}) n \sqrt{mH} \max(|\epsilon_{x1} - \epsilon_{x2}|, |\epsilon_{u1} - \epsilon_{u2}|))/\epsilon_g$$

where $\|\theta_1 - \theta_2\|_F \leq r$.

Proof. Notice that both the objective functions and the constraints are different in the two CCE programs above, so we introduce an auxiliary policy $\mathbf{M}_3 = \arg\min_{\mathbf{M} \in \Omega(\theta_1, \epsilon_{x1}, \epsilon_{u1})} f(\mathbf{M}; \theta_2)$ to discuss the perturbation bounds on cost differences and constraint differences separately. We will first bound $f(\mathbf{M}_1, \theta_2) - f(\mathbf{M}_3, \theta_2)$ and then bound $f(\mathbf{M}_3, \theta_2) - f(\mathbf{M}_2, \theta_2)$ below.

Perturbation on the cost functions.

$$\begin{aligned} f(\mathbf{M}_1, \theta_2) - f(\mathbf{M}_3, \theta_2) &= f(\mathbf{M}_1, \theta_2) - f(\mathbf{M}_1, \theta_1) + f(\mathbf{M}_1, \theta_1) - f(\mathbf{M}_3, \theta_1) + f(\mathbf{M}_3, \theta_1) - f(\mathbf{M}_3, \theta_2) \\ &\leq f(\mathbf{M}_1, \theta_2) - f(\mathbf{M}_1, \theta_1) + f(\mathbf{M}_3, \theta_1) - f(\mathbf{M}_3, \theta_2) \\ &\leq O(mn\|\theta_1 - \theta_2\|_F) \end{aligned}$$

where the first inequality is because $\mathbf{M}_1$ and $\mathbf{M}_3$ are in the same set and $\mathbf{M}_1$ minimizes the cost $f(\mathbf{M}, \theta_1)$ in this set, and the second inequality is because of the following perturbation lemma on the cost functions.

Lemma 16 (Perturbation bound on f with respect to θ)**.** *For any* $H \geq \log(2\kappa)/\log((1-\gamma)^{-1})$, $\mathbf{M} \in \mathcal{M}_H$, *any* $\theta, \hat{\theta} \in \Theta_{\text{ini}}$ *, we have*

$$|f(\mathbf{M}; \theta) - f(\mathbf{M}; \hat{\theta})| \leq O(mnr)$$

where $\|\theta - \hat{\theta}\|_F \leq r$.

Proof. We let $\tilde{x}_t(\theta)$ and $\tilde{x}_t(\theta_2)$ denote the approximate states defined by Proposition 1, and we omit $\mathbf{M}$ in this proof for notational simplicity. Notice that

$$\begin{aligned} \|\tilde{x}(\theta) - \tilde{x}(\hat{\theta})\|_2 &= \|\sum_{k=1}^{2H} (\Phi_k^x(\theta) - \Phi_k^x(\hat{\theta})) w_{t-k}\|_2 \leq \sum_{k=1}^{2H} \|(\Phi_k^x(\theta) - \Phi_k^x(\hat{\theta})) w_{t-k}\|_2 \\ &\leq \sum_{k=1}^{H} \|(A^{k-1} - \hat{A}^{k-1}) w_{t-k}\|_2 + \sum_{k=1}^{2H} \sum_{i=1}^{H} \|(A^{i-1}B - \hat{A}^{i-1}\hat{B}) M[k-i] \mathbb{I}_{(1 \leq k-i \leq H)} w_{t-k}\|_2 \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{H} \|(A^{k-1} - \hat{A}^{k-1})\|_2 \sqrt{n} w_{\max} + \sum_{k=1}^{2H}\sum_{i=1}^{H} \|(A^{i-1}B - \hat{A}^{i-1}\hat{B})\|_2 \sqrt{m} \|M[k-i] \mathbb{I}_{(1\leq k-i \leq H)} w_{t-k}\|_\infty \\
&\leq \sum_{k=1}^{H} \sqrt{n} w_{\max} O(k(1-\gamma)^{k-1} \|\theta - \hat{\theta}\|_F) + \sum_{k=1}^{2H}\sum_{i=1}^{H} O(\sqrt{m} \|(A^{i-1}B - \hat{A}^{i-1}\hat{B})\|_2 \|M[k-i]\|_\infty \mathbb{I}_{(1\leq k-i\leq H)}) \\
&\leq O(\sqrt{n} r_\theta) + \sum_{k=1}^{2H}\sum_{i=1}^{H} \|\theta - \hat{\theta}\|_F \sqrt{mn} O((1-\gamma)^{k-i-1}(i-1)(1-\gamma)^{i-2} \mathbb{I}_{(i\geq 2)} \mathbb{I}_{(1\leq k-i\leq H)}) \\
&= O(\sqrt{n}\|\theta - \hat{\theta}\|_F) + O(\sum_{i=1}^{H}\sum_{j=1}^{H} \|\theta - \hat{\theta}\|_F \sqrt{mn}(i-1)(1-\gamma)^{i-2}(1-\gamma)^{j-1}) \\
&\leq O(\sqrt{n}\|\theta - \hat{\theta}\|_F) + O(\sqrt{mn}\|\theta - \hat{\theta}\|_F) = O(\sqrt{mn}\|\theta - \hat{\theta}\|_F)
\end{aligned}
$$

where we used Lemma 25 in the third and fourth inequality.

□

Perturbation on the constraints. Our bound on $f(\mathbf{M}_3, \theta_2) - f(\mathbf{M}_2, \theta_2)$ is established based on Proposition 2 and Lemma 9 in Li et al. (2020). The major difference is that Li et al. (2020) only considers changes in the right-hand-side of the constraint inequalities but in our setting, the left-hand-side of the constraint inequalities also change. To handle this difference, we introduce two auxiliary sets Ω_1 and Ω_2 with the same left-hand-sides in the constraint inequalities such that $\Omega_1 = \Omega(\theta_1, \epsilon_{x1}, \epsilon_{u1})$ and $\Omega_2 = \Omega(\theta_2, \epsilon_{x2}, \epsilon_{u2})$, which is achieved by adding inactive inequalitiy constraints. Specifically, define

$$\Omega_1 = \Omega(\theta_1, \epsilon_{x1}, \epsilon_{u1}) \cap \Omega(\theta_2, \epsilon_{x1} - \epsilon_{\hat{\theta}}(r), \epsilon_{u1}), \quad \Omega_2 = \Omega(\theta_2, \epsilon_{x2}, \epsilon_{u2}) \cap \Omega(\theta_1, \epsilon_{x2} - \epsilon_{\hat{\theta}}(r), \epsilon_{u2}).$$

Notice that the constraints in $\Omega(\theta_2, \epsilon_{x1} - \epsilon_{\hat{\theta}}(r), \epsilon_{u1})$ and $\Omega(\theta_1, \epsilon_{x2} - \epsilon_{\hat{\theta}}(r), \epsilon_{u2})$ are inactive due to Lemma 4. Further, notice that Ω_1, Ω_2 are both polytopes.

Another difference between our setting and that in Proposition 2 of Li et al. (2020) is that Ω_1 may not be a subset of Ω_2 and vice versa, so we need to generalize Proposition 2 to this setting as follows.

Lemma 17 (Extension from Proposition 2 Li et al. (2020))**.** *Consider two polytopes: $\Omega_1 = \{x : Cx \leq h - \Delta_1\}, \Omega_2 = \{x : Cx \leq h - \Delta_2\}$, where Δ_1, Δ_2 are two vectors. Define $\Delta_0 = \min(\Delta_1, \Delta_2)$ elementwise. Define $\Delta_3 = \max(\Delta_1, \Delta_2)$ elementwise. Suppose the l_2-diameter of Ω_0 is d_{Ω_0}, $f(x)$ is L-Lipschitz continuous, and there exists $x_F \in \Omega_3$, then we have*

$$|\min_{\Omega_1} f(x) - \min_{\Omega_2} f(x)| \leq \frac{2L d_{\Omega_0} \|\Delta_1 - \Delta_2\|_\infty}{\min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}} (h - \Delta_3 - Cx_F)_i}$$

The proof is deferred to Appendix G.5.

Now, we are ready to bound $f(\mathbf{M}_3, \theta_2) - f(\mathbf{M}_2, \theta_2)$. By our definitions and discussions above, we have $f(\mathbf{M}_3, \theta_2) = \min_{\mathbf{M}\in\Omega_1} f(\mathbf{M}; \theta_2)$ and $f(\mathbf{M}_2, \theta_2) = \min_{\mathbf{M}\in\Omega_2} f(\mathbf{M}; \theta_2)$. Further, notice that Ω_1, Ω_2 are both polytopes and can be written in the form of $\{\vec{P} : C\vec{P} \leq h - \Delta_1\}, \{\vec{P} : C\vec{P} \leq h - \Delta_2\}$ by introducing auxiliary variables to represent the absolute values (see Lemma 10 in Li et al. (2020) for more details). Therefore, we can apply Lemma 17 to obtain the bound on $f(\mathbf{M}_3, \theta_2) - f(\mathbf{M}_2, \theta_2)$ by bounding the corresponding constants $d_{\Omega_0}, L, \|\Delta_1 - \Delta_2\|_\infty$, and $\min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}} (h - \Delta_3 - Cx_F)_i$. Same as the proof of Lemma 9, we can show the l_2-diameter of Ω_0 is $d_{\Omega_0} = O(\sqrt{mn} + \sqrt{k_x + k_u})$ and $\|\Delta_1 - \Delta_2\|_\infty = \max(|\epsilon_{x1} - \epsilon_{x2}|, |\epsilon_{u1} - \epsilon_{u2}|)$. Further, the Lipschitz factor L can be obtained from the gradient bound $L = G_f = O(\sqrt{n^2 mH})$ as provided below, whose proof is provided in Appendix G.4.

Lemma 18 (Gradient bound of $f(\mathbf{M}; \theta)$)**.** *For any $H \geq 1$, $\mathbf{M} \in \mathcal{M}_H$, $\theta \in \Theta_{ini}$, we have $\|\nabla f(\mathbf{M}; \theta)\|_F \leq G_f = O(\sqrt{n^2 mH})$.*

Next, since $\Omega(\theta_1, \epsilon_{x1} + \epsilon_g, \epsilon_{u1} + \epsilon_g) \cap \Omega(\theta_2, \epsilon_{x2} + \epsilon_g, \epsilon_{u2} + \epsilon_g)$ is non-empty, there exists $\vec{P}_F \in \Omega_3$ such that $\min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}} (h - \Delta_3 - Cx_F)_i \geq \epsilon_0$. Lastly, by applying the constants above to Lemma 17, we can show $f(\mathbf{M}_3, \theta_2) - f(\mathbf{M}_2, \theta_2) \leq O((\sqrt{mn} + \sqrt{k_x + k_u})\sqrt{n^2 mH}) \max(|\epsilon_{x1} - \epsilon_{x2}|, |\epsilon_{u1} - \epsilon_{u2}|)/\epsilon_g$.

The proof of Lemma 15 is completed by combining the bounds on $f(\mathbf{M}_1, \theta_2) - f(\mathbf{M}_3, \theta_2)$ and $f(\mathbf{M}_3, \theta_2) - f(\mathbf{M}_2, \theta_2)$.

□

Based on Lemma 15, we can show the following bound on Part iii when $\mathcal{E}_{\text{safe}}$ is true.

$$\text{Part iii} = \sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}^{(e)}; \theta_*) - f(\mathbf{M}^*_{H^{(e)}}; \theta_*)) \leq O((n^2 m^2 + n^{2.5} m^{1.5})\sqrt{mn + k_x + k_u} T^{2/3}) \tag{15}$$

The proof for (15) is provided below. For each $e \geq 0$, notice that $\mathbf{M}^{(e)} = \arg\min_{\mathbf{M} \in \Omega(\hat{\theta}^{(e+1)}, \hat{\epsilon}_x^{(e)}, 0)} f(\mathbf{M}; \theta^{(e+1)})$, where we define $\hat{\epsilon}_x^{(e)} = \epsilon_\theta(r^{(e+1)}) + \epsilon_H(H^{(e)}) + \epsilon_v(\Delta_M^{(e)}, H^{(e)})$; and $\mathbf{M}^*_{H^{(e)}} = \arg\min_{\mathbf{M} \in \Omega(\theta_*, \epsilon_H(H^{(e)}), 0)} f(\mathbf{M}; \theta_*)$. Therefore, we can apply Lemma 15. We first verify the conditions of Lemma 15. Notice that $\theta^{(e+1)}, \theta_* \in \Theta_{\text{ini}}$, $H^{(e)}$ is large enough. Further, we have $\epsilon_g = min(\epsilon_{F,x}, \epsilon_{F,u})/4 > 0$ such that $\Omega(\hat{\theta}^{(e+1)}, \hat{\epsilon}_x^{(e)} + \epsilon_g, \epsilon_g) \cap \Omega(\theta_*, \epsilon_H(H^{(e)}) + \epsilon_g, \epsilon_g)$ is not empty due to the feasibility conditions in Theorem 2. Therefore, when $\mathcal{E}_{\text{safe}}$ is true, by Lemma 15, under our choices of parameters in Theorem 4, we have

$$\begin{aligned} f(\mathbf{M}^{(e)}; \theta_*) - f(\mathbf{M}^*_{H^{(e)}}; \theta_*) &\leq O(mnr^{(e+1)} + (\sqrt{mn} + \sqrt{k_x + k_u}) n \sqrt{mH^{(e)}} (\epsilon_\theta(r^{(e+1)}) + \epsilon_v(\Delta_M^{(e)}, H^{(e)}))) \\ &\leq \tilde{O}(mn(n\sqrt{m} + m\sqrt{n})(T^{(e)})^{-1/3} + (\sqrt{mn} + \sqrt{k_x + k_u}) n\sqrt{m}\sqrt{mn}(n\sqrt{m} + m\sqrt{n})(T^{(e+1)})^{-1/3}) \\ &\leq \tilde{O}\left((n^2 m^2 + n^{2.5} m^{1.5})\sqrt{mn + k_c}(T^{(e+1)})^{-1/3}\right) \end{aligned}$$

Consequently, by Lemma 14, we prove the bound (15).

E.1.2 Bound on Part i

When $\mathcal{E}_{\text{safe}}$ is true, we are able to show

$$\text{Part i} = \sum_e \sum_{t \in \mathcal{T}_2^{(e)}} l(x_t, u_t) - l(\hat{x}_t, \hat{u}_t) \leq \tilde{O}(n\sqrt{m}\sqrt{m+n} T^{2/3}). \tag{16}$$

The proof is provided below.

Firstly, under $\mathcal{E}_{\text{safe}}$, we have $x_t \in \mathbb{X}$ and $u_t \in \mathbb{U}$, so $\|x_t\|_2 \leq O(1)$, $\|u_t\|_2 \leq O(1)$. Further, by the definitions of $\hat{x}_t, \hat{u}_t$ and the proof of Theorem 3, we can also verify that $\hat{x}_t \in \mathbb{X}$ and $\hat{u}_t \in \mathbb{U}$ based on the proof of Lemma 10 and Lemma 12. Therefore, we have$\|\hat{x}_t\|_2 \leq O(1)$, $\|\hat{u}_t\|_2 \leq O(1)$

Next, by Lemma 3, for $t \in \mathcal{T}_2^{(e)}$, we can write $x_t, \hat{x}_t, u_t, \hat{u}_t$ as

$$\begin{aligned} x_t &= A_*^{H_t} x_{t-H_t} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} A_*^{i-1} B_* M_{t-i}[k-i] \hat{w}_{t-k} \mathbb{1}_{(1 \leq k-i \leq H_{t-i})} + \sum_{i=1}^{H_t} A_*^{i-1} w_{t-i} \\ \hat{x}_t &= A_*^{H_t} x_{t-H_t} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} A_*^{i-1} B_* M_{t-i}[k-i] w_{t-k} \mathbb{1}_{(1 \leq k-i \leq H_{t-i})} + \sum_{i=1}^{H_t} A_*^{i-1} w_{t-i} \\ u_t &= \sum_{t=1}^{H_t} M_t[k] \hat{w}_{t-k}, \quad \hat{u}_t = \sum_{t=1}^{H_t} M_t[k] w_{t-k}. \end{aligned}$$

Hence, we can bound $\|x_t - \hat{x}_t\|_2$ and $\|u_t - \hat{u}_t\|_2$ by Lemma 5 below:

$$\begin{aligned} &\|x_t - \hat{x}_t\|_2 \leq O((1-\gamma)^{H^{(e)}} + \sqrt{mn} r_\theta^{(e+1)}) = \tilde{O}(nm\sqrt{m+n}(T^{(e+1)})^{-1/3}) \\ &\|u_t - \hat{u}_t\|_2 \leq O(\sqrt{mn} r_\theta^{(e+1)}) = \tilde{O}(nm\sqrt{m+n}(T^{(e+1)})^{-1/3}) \end{aligned}$$

Consequently, by applying Lemma 14 and the quadratic structure of $l(x, u)$, we can bound the Part i by

$$\sum_e \sum_{t \in \mathcal{T}_2^{(e)}} (l(x_t, u_t) - l(\hat{x}_t, \hat{u}_t)) \leq \sum_e T^{(e+1)} \tilde{O}(nm\sqrt{m+n}(T^{(e+1)})^{-1/3}) \leq \tilde{O}(nm\sqrt{m+n} T^{2/3}).$$

E.1.3 Bound on Part ii

Lemma 19 (Bound on Part ii)**.** *With probability* $1-p$*, Part ii* $\leq \tilde{O}(mn\sqrt{T})$.

Notice that this part is not a dominating term in the regret bound. The proof relies on a martingale concentration analysis and is very technical, so we defer it to Appendix G.6.

E.1.4 Bound on Part iv

In the following, we will show that

$$\text{Part iv} = \sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - J^*) = \tilde{O}(n\sqrt{m}\sqrt{mn+k_c}\sqrt{n}) \tag{17}$$

The proof is provided below.

Remember that J^* is generated by the optimal safe linear policy K^*. By Lemma 4 and Corollary 2 in Li et al. (2020), for a memory length $H^{(e)}$, we can define $\mathbf{M}_{H^{(e)}}(K^*) \in \Omega(\theta_*, -\epsilon_P^{(e)}, 0)$, where $\epsilon_P^{(e)} = \sqrt{mn}(1-\gamma)^{H^{(e)}}$ corresponds to $\epsilon_1 + \epsilon_3$ in Li et al. (2020) with $H = H^{(e)}$.[9] Further, by Lemma 6 in Li et al. (2020), we have

$$f(\mathbf{M}_{H^{(e)}}(K^*); \theta_*) - J^* = \lim_{T \to +\infty} \frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{M}_{H^{(e)}}(K^*); \theta_*) - \mathbb{E}(l(x_t^*, u_t^*)) \leq O(n^2 m (H^{(e)})^2 (1-\gamma)^{H^{(e)}})$$

In addition, we have

$$\begin{aligned} f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - f(\mathbf{M}_{H^{(e)}}(K^*); \theta_*) &\leq f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - \min_{\mathbf{M} \in \Omega(\theta_*, -\epsilon_P^{(e)}, 0)} f(\mathbf{M}; \theta_*) \\ &\leq \tilde{O}(n\sqrt{m}\sqrt{mn + k_x + k_u}(\epsilon_P^{(e)} + \epsilon_H(H^{(e)}))) \\ &\leq \tilde{O}(n\sqrt{m}\sqrt{mn+k_c}\sqrt{mn}(1-\gamma)^{H^{(e)}}) \end{aligned}$$

By combining the bounds above and by choosing $H^{(e)} \geq \log(T^{(e+1)}) / \log((1-\gamma)^{-1})$, we have

$$\begin{aligned} f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - J^* &= f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - f(\mathbf{M}_{H^{(e)}}(K^*); \theta_*) + (\mathbf{M}_{H^{(e)}}(K^*); \theta_*) - J^* \\ &\leq \tilde{O}(n\sqrt{m}\sqrt{mn+k_c}\sqrt{mn}/T^{(e+1)}), \end{aligned}$$

which directly leads to the bound (17) on Part iv.

Completing the proof of the regret bound.

By combining (16), (15), (17), and Lemma 19, we obtain our regret bound in Theorem 4. Notice that (16), (15), (17) all condition on $\mathcal{E}_{\text{safe}}$, and $\mathcal{E}_{\text{safe}}$ holds w.p. $1-p$. But Lemma 19 conditions on a different event and that event also holds with probability $1-p$. Putting them together, we have that our regret bound holds w.p. $1-2p$.

E.2 Condition Verification for Feasibility and Constraint Satisfaction

In this subsection, we briefly show that there exist parameters characterized by Theorem 4 that satisfy the conditions for feasibility and constraint satisfaction in Theorem 2 and Theorem 3, which include: $T_D^{(e)}$ satisfying the condition in Theorem 1 (Corollary 1's condition), condition (**??**), condition (ii) of Theorem 2, and $T^{(e+1)} \geq t_2^{(e)}$.

Firstly, in Corollary 1, we need $T_D^{(e)} \geq O(\log(2e^2/p) + (m+n)\log(1/\bar{\eta}))$ for $e \geq 0$. By our choices, we have $T_D^{(e)} = (T^{(\min(e,1))})^{2/3}$ and $T^{(e+1)} = 2T^{(e)}$, so $T_D^{(e)}$ increases exponentially. Therefore, $T_D^{(e)} \geq O(\log(2e^2/p) + (m+n)\log(1/\bar{\eta}))$ can be guaranteed if $T_D^{(1)} \geq O(\log(1/p) + (m+n)\log(1/\bar{\eta}))$ with some sufficiently large constant factor, which requires $T^{(1)} \geq O((m+n)^{3/2})$.

[9] As discussed in footnote 8, our $\epsilon_P^{(e)}$ has smaller dependence on n, m, H compared with Li et al. (2020).

Secondly, for condition (**??**), we set $\epsilon_0 = \epsilon_{F,x}/4$ and let $\epsilon_P + \epsilon_H(H^{(0)}) \leq \epsilon_{F,x}/12$, $\epsilon_{\eta,x} \leq \epsilon_{F,x}/12$, $\epsilon_v(\Delta_M^{(0)}, H^{(0)}) \leq \epsilon_{F,x}/12$, and $\epsilon_P \leq \epsilon_{F,u}/4$, $\epsilon_{\eta,u} \leq \epsilon_{F,u}/4$. These conditions require $H^{(0)} \geq O(\log(\sqrt{mn}/\min(\epsilon_F)))$, $\bar{\eta}^{(e)} = O(\min(\frac{\epsilon_F^x}{\sqrt{m}}, \epsilon_F^u))$, and $\Delta_M^{(e)} = O(\frac{\epsilon_F^x}{\sqrt{mnH^{(e)}}(T^{(e+1)})^{1/3}})$.

Thirdly, for the condition (ii) of Theorem 3, the monotonicity for $H^{(e)}, \sqrt{H^{(e)}}\Delta_M^{(e)}$ are satisfied and $\bar{\eta}^{(e)}$ is a constant, so its monotonicity condition is also satisfied. With exponentially increasing $T_D^{(e)}$, the decreasing $r^{(e)}$ is also satisfied. We only need to verify that $r^{(1)} \leq r_{\text{ini}}$. This requires $T^{(1)} \geq \tilde{O}((\sqrt{n}m + n\sqrt{n})^3)$.

Lastly, for $T^{(e+1)} \geq t_2^{(e)}$, notice that Phase 1 only takes $(T^{(e+1)} - T^{(e)})^{2/3}$ stages, and the safe transitions only takes $\tilde{O}((T^{(e+1)})^{1/3})$ stages, so $T^{(e+1)} \geq t_2^{(e)}$ for all e for large enough initial $T^{(1)}$.

F More Discussions

In this appendix, we briefly introduce RMPC in Mayne et al. (2005) and show that its infinite-horizon averaged cost can be captured by $J(\mathbb{K})$ for some safe linear policy $\mathbb{K}$. Therefore, algorithms with small regret compared with optimal safe linear policies can also achieve comparable performance with RMPC in Mayne et al. (2005) for long horizons, which further motivates our choice of regret benchmarks as safe linear policies. Further, we discuss the implementation of our algorithm for non-zero x_0.

F.1 A brief review of RMPC in Mayne et al. (2005)

RMPC is a popular method to handle constrained system with disturbances and/or other system uncertainties. Since we will include RMPC in the benchmark policy class, we assume the model θ_* is available here, but RMPC can also handle model uncertainties. Many different versions of RMPC have been proposed in the literature, (see Rawlings and Mayne (2009) for a review). In this appendix, we will focus on a tube-based RMPC defined in Mayne et al. (2005). The RMPC method in Mayne et al. (2005) enjoys desirable theoretical guarantees, such as robust exponential stability, recursive feasibility, constraint satisfaction, and is thus commonly adopted. RMPC usually considers $x_0 \neq 0$. When considering RMPC for regulation problems, one goal of RMPC is to quickly and safely steer the states to a neighborhood of origin (due to the system disturbances, one cannot steer the state to the origin exactly).

Next, we briefly introduce the tube-based RMPC scheme. In most tube-based RMPC schemes (not just Mayne et al. (2005)), it is required to know a linear static controller $u_t = -\mathbb{K}x_t$ such that this controller is strictly safe if the system starts from the origin. A disturbance-invariant set for the closed-loop system $x_{t+1} = Ax_t - B\mathbb{K}x_t + w_t$ is also needed.

Definition 4. *Ξ is called a disturbance-invariant set for $x_{t+1} = Ax_t - B\mathbb{K}x_t + w_t$ if for any $x_t \in \Xi$, and $w_t \in \mathbb{W}$, we have $x_{t+1} \in \Xi$.*

For computational purposes, a polytopic approximation of disturbance-invariant set is usually employed. Further, the implementation of RMPC also requires the knowledge of a terminal set X_f such that for any $x_0 \in X_f$, implementing the controller $u_t = -\mathbb{K}x_t$ is safe, as well as a terminal cost function $V_f(x) = x^\top P x$ satisfying certain conditions (see Mayne et al. (2005) for more details).

RMPC scheme in Mayne et al. (2005). Now, we are ready to define the tube-based RMPC proposed in Mayne et al. (2005). At each stage t, consider a planning window $t+k|t$ for $0 \leq k \leq W$, RMPC in Mayne et al. (2005) solves the following optimization:

$$\begin{aligned}
\min_{x_{t|t}, u_{t+k|t}} \quad & \sum_{k=0}^{W-1} l(x_{t+k|t}, u_{t+k|t}) + V_f(x_{t+W|t}) \\
\text{s.t.} \quad & x_{t+k+1|t} = A_* \bar{x}_{t+k|t} + B_* u_{t+k|t}, \quad k \geq 0 \\
& x_{t|t} \in x_t \oplus \Xi \\
& x_{t+k|t} \in \mathbb{X} \ominus \Xi, \forall 0 \leq k \leq W-1 \\
& u_{t+k|t} \in \mathbb{U} \ominus \mathbb{K}\Xi, \forall 0 \leq k \leq W-1 \\
& x_{t+W|t} \in X_f \subseteq \mathbb{X} \ominus \Xi
\end{aligned} \qquad \text{(RMPC Mayne et al. (2005))}$$

Then, implement control:

$$u_t = -\mathbb{K}(x_t - x_{t|t}^*) + u_{t|t}^*.$$

Notice that $x^*_{t|t}, u^*_{t|t}$ are functions of x_t. Further, by Bemporad et al. (2002), u_t is a piece-wise affine (PWA) function of the state x_t when Ξ is a polytope. Define the set of feasible initial values as

$$X_N = \{x_0 : \text{(RMPC Mayne et al. (2005)) is feasible when } x_t = x_0\}.$$

The RMPC scheme in Mayne et al. (2005) is a variant of the traditional RMPC schemes by allowing more freedom when choosing $x_{t|t}$, i.e., in the scheme above, $x_{t|t}$ is also an optimization variable as long as $x_{t|t} \in x_t \oplus \Xi$, but in traditional RMPC schemes, $x_{t|t} = x_t$ is fixed. With this adjustment, the RMPC scheme in Mayne et al. (2005) enjoys robust exponential stability.

Theorem 6 (Theorem 1 in Mayne et al. (2005))**.** *The set Ξ is robustly exponentially stable for the closed-loop system with* (RMPC Mayne et al. (2005)) *for $w_k \in \mathbb{W}$ with an attraction region X_N, i.e., there exists $c > 0, \gamma_1 \in (0, 1)$, such that for any $x_0 \in X_N$, for any $w_k \in \mathbb{W}$.*

$$\text{dist}(x_t, \Xi) \leq c\gamma_1^t \text{dist}(x_0, \Xi).$$

Theorem 6 suggests that (RMPC Mayne et al. (2005)) can quickly reduce the distance between x_t and Ξ, i.e. it can drive a large initial state $x_0 \neq 0$ quickly to a neighborhood around Ξ, which is also a neighborhood around the origin.

Based on the robust exponential stability, we can build a connection between the infinite horizon averaged cost of RMPC and that of the safe linear policy $\mathbb{K}$.

Theorem 7 (Connection between RMPC in Mayne et al. (2005) and linear control's infinite-horizon costs)**.** *Consider (RMPC Mayne et al. (2005)) defined above with $\mathbb{K}$ satisfying the requirements in Mayne et al. (2005). For any $x_0 \in X_N$, the infinite-horizon averaged cost of RMPC in Mayne et al. (2005) equals the infinite-horizon averaged cost of $\mathbb{K}$, i.e.*

$$J(\text{RMPC in Mayne et al. (2005)}) = J(\mathbb{K}),$$

The proof is deferred to the end of this appendix.

Notice that $\mathbb{K}$ is a pre-fixed safe linear policy, so by Theorem 7, we have $J(K^*) \leq J(\text{RMPC in Mayne et al. (2005)})$, where K^* is our regret benchmark, i.e., the optimal safe linear policy. This suggests that RMPC in Mayne et al. (2005) achieves similar or worse performance than the optimal safe linear policy in the long run. Since our adaptive control algorithm enjoys a sublinear regret compared to the optimal safe linear policy, Theorem 7 suggests that our algorithm achieves the same regret bound even if we include RMPC in Mayne et al. (2005) to the benchmark policy set. Further, if $\mathbb{K} \neq K^*$, our adaptive algorithm can even achieve better performance than RMPC in Mayne et al. (2005) at around the equilibrium point 0.

Nevertheless, one major strength of RMPC in Mayne et al. (2005) compared with our algorithm is that RMPC can guarantee safety for large nonzero x_0 and can drive a large state exponentially to a small neighborhood of 0. Therefore, an interesting and natural idea is to combine RMPC in Mayne et al. (2005) with our algorithm to achieve the strengths of both methods: quickly and safely drive a large initial state to a neighborhood around 0, and learning to optimize the performance around 0.[10] We leave more studies on this combination as future work.

Remark 6. *Since our proof relies on the robust exponential stability property of RMPC in Mayne et al. (2005), for other RMPC schemes without this property, we still cannot include them to our benchmark policy class and generate a sublinear regret. We leave the regret analysis compared with other RMPC schemes without robust exponential stability as future work. Further, we note that there are a few papers on the regret analysis with RMPC as the benchmark, e.g., Wabersich and Zeilinger (2018); Muthirayan et al. (2020). However, Wabersich and Zeilinger (2018) allows constraint violation during the learning process and allows restarts when policies are updated, and Muthirayan et al. (2020) does not consider state constraints and the proposed algorithm involves an intractable oracle. In conclusion, the regret analysis with RMPC as the benchmark is largely under-explored and is an important direction for future research.*

Proof of Theorem 7. To prove Theorem 7, we introduce some necessary results from the existing literature and some lemmas based on these existing results.

Firstly, we review the structure of constrained LQR's solution proved in Bemporad et al. (2002).

[10] Though RMPC in Mayne et al. (2005) requires a known model, there are standard approaches to extend RMPC to handle model uncertainties, e.g., Köhler et al. (2019); Lu et al. (2019).

Proposition 2 (Corollary 2 and Theorem 4 and Section 4.4 in Bemporad et al. (2002))**.** *Consider (CLQR) with p.d. quadratic costs and polytopic constraints below:*

$$\begin{aligned} \min_{u_{t+k|t}} \ & \sum_{k=0}^{W-1} l(x_{t+k|t}, u_{t+k|t}) + x_{t+W|t}^\top P x_{t+W|t}^\top \\ s.t. \ & x_{t+k+1|t} = A_* x_{t+k|t} + B_* u_{t+k|t}, \quad k \geq 0 \\ & D_x x_{t+k|t} \leq d_x, \quad \forall 0 \leq k \leq W-1 \\ & D_u u_{t+k|t} \leq d_u, \quad \forall 0 \leq k \leq W-1 \\ & D_{term} x_{t+W|t} \leq d_{term} \\ & x_{t|t} = x \end{aligned} \tag{CLQR}$$

Denote the optimal policy as $\pi_{CLQR}(x) = u_{t|t}^*$, and denote the feasible region as X_N. Then, X_N is convex, and $\pi_{CLQR}(x)$ is continuous and PWA on a finite number of closed convex polytopic regions. that is,

$$\pi_{CLQR}(x) = K_i x + b_i, \quad G_i x \leq h_i, \quad i = 0, 1, \ldots, N_{clqr}.$$

Further, the number of different gain matrices can bounded by a constant $\bar{N}_{clqr-gain}$ that only depends on the dimensionality of the problem.

Based on Proposition 2, we have that $\pi_{CLQR}(x)$ is Lispchitz continous with Lipschitz factor $L_{CLQR} = \max_i \|K_i\|_2$ since $\pi_{CLQR}(x)$ is continuous and piecewise-affine with respect to x.

Next, we will use the exponential convergence results of RMPC in Mayne et al. (2005).

Proposition 3 (See the proof of Theorem 1 in Mayne et al. (2005))**.** *There exists $c_1 > 0$ and $\rho \in (0,1)$ such that for any $x_0 \in X_N$, and for any admissible disturbances w_k, we have*

$$\|x_{t|t}^*(x_t)\|_2 \leq c_1 \rho^t \|x_{0|0}^*(x_0)\|_2.$$

Based on this, we can also show the exponential decay of $u_{t|t}^*(x_t)$.

Lemma 20. *There exists $c_2 > 0$ and $\rho \in (0,1)$ such that for any $x_0 \in X_N$, and for any admissible disturbances w_k, $u_{t|t}^*(x_{t|t}^*)$ is Lipschitz continous with a finite factor denoted as L_{rmpc} on a convex feasible set. Further, we have $\|u_{t|t}^*(x_t)\|_2 \leq c_2 \rho^t$, where $c_2 = L_{rmpc} c_1 x_{\max}$.*

Proof. First of all, we point out that for the (RMPC Mayne et al. (2005)) optimization, when $x_{t|t}^*$ is fixed, then $u_{t|t}^*$ can be viewed as $u_{t|t}^* = \pi_{CLQR}(x_{t|t}^*)$ for a (CLQR) problem with the same polytopic constraints and strongly convex quadratic cost functions with (RMPC Mayne et al. (2005)). Therefore, $u_{t|t}^*(x_{t|t}^*)$ is Lipschitz continous with a finite factor denoted as L_{rmpc} on a convex feasible set.

Further, notice that $u_{t|t}^*(0) = 0$. Therefore,

$$\|u_{t|t}^*(x_{t|t}^*)\|_2 = \|u_{t|t}^*(x_{t|t}^*) - u_{t|t}^*(0)\|_2 \leq L_{rmpc} \|x_{t|t}^*\|_2 \leq L_{rmpc} c_1 \rho^t \|x_{0|0}^*(x_0)\|_2 \leq c_2 \rho^t$$

where $c_2 = L_{rmpc} c_1 x_{\max}$. □

Lastly, a technical lemma of a standard results. The proof is very straightforward.

Lemma 21. *Consider $y^+ = A_{\mathbb{K}} y + w$, where $y_0 = x_0 \in \mathbb{X}$ and $p = -\mathbb{K} y$. Since $\mathbb{K}$ is (κ, γ) strongly convex, both y and p are bounded by*

$$\|y_t\|_2 \leq \|w\|_2 \kappa^2 / \gamma + \kappa^2 x_{\max} = y_{\max}, \|p_t\|_2 \leq \|w\|_2 \kappa^3 / \gamma + \kappa^2 x_{\max} = p_{\max}.$$

Now, we are ready for the proof of Theorem 7.

Proof of Theorem 7. The closed-loop system of (RMPC Mayne et al. (2005)) is

$$x_{t+1} = A_* x_t + B_* \pi_{RMPC}(x_t) + w_t = A_* x_t - B_* \mathbb{K} x_t + B_*(\mathbb{K} x^*_{t|t}(x_t) + u^*_{t|t}(x_t)) + w_t.$$

Consider a possibly unsafe system:

$$y_{t+1} = A_* y_t + B_* p_t + w_t, \quad p_t = -\mathbb{K} y_t$$

with the same sequence of disturbances and $y_0 = x_0$.

The dynamics of the error $e_t = x_t - y_t$ is

$$e_{t+1} = A_{\mathbb{K}} e_t + v_t$$

where $A_{\mathbb{K}} = A_* - B_* \mathbb{K}$, and $v_t = B_*(\mathbb{K} x^*_{t|t}(x_t) + u^*_{t|t}(x_t))$. Notice that by Proposition 3 and Lemma 20, we have

$$\|v_t\|_2 \leq \|B_*\|_2 (\kappa c_1 \rho^t x_{\max} + c_2 \rho^t) = c_3 \rho^t,$$

where $c_3 = \|B_*\|_2(\kappa c_1 x_{\max} + c_2)$.

Therefore,

$$\begin{aligned}\|e_t\|_2 &= \|v_{t-1} + A_{\mathbb{K}} v_{t-2} + A_{\mathbb{K}}^{t-1} v_0\|_2 \\ &\leq c_3 \rho^{t-1} + \kappa^2 (1-\gamma) c_3 \rho^{t-2} + \dots \\ &\leq c_3 \kappa^2 t \max(\rho, 1-\gamma)^{t-1} = c_4 t \rho_0^{t-1}\end{aligned}$$

where $\rho_0 = \max(\rho, 1-\gamma) \in (0,1)$ and $c_4 = c_3 \kappa^2$. Further,

$$\|u_t - p_t\|_2 = \| - \mathbb{K} e_t + v_t\|_2 \leq \kappa c_4 t \rho_0^{t-1} + c_3 \rho^t \leq c_5 t \rho_0^{t-1},$$

where $c_5 = c_4 \kappa + c_3/\rho$.

Therefore, the stage cost difference is

$$\begin{aligned}|l(x_t, u_t) - l(y_t, p_t)| &\leq \|Q\|_2 \|e_t\|_2 (x_{\max} + y_{\max}) + \|R\|_2 \|u_t - p_t\|_2 \|u_{\max} + p_{\max}\|_2 \\ &\leq \|Q\|_2 (x_{\max} + y_{\max}) c_4 t \rho_0^{t-1} + \|R\|_2 \|u_{\max} + p_{\max}\|_2 c_5 t \rho_0^{t-1} = c_6 t \rho_0^{t-1}\end{aligned}$$

where $c_6 = \|Q\|_2 (x_{\max} + y_{\max}) c_4 + \|R\|_2 \|u_{\max} + p_{\max}\|_2 c_5$.

Therefore,

$$\left| \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} l(x_t, u_t) - l(y_t, p_t) \right| \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left| l(x_t, u_t) - l(y_t, p_t) \right| \leq \frac{1}{T} c_6 / (1 - \rho_0)^2$$

By taking $T \to +\infty$, we have $\lim_{T \to +\infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} l(x_t, u_t) - l(y_t, p_t) = 0$. Since $\lim_{T \to +\infty} \frac{1}{T} \mathbb{E}\, l(y_t, p_t) = J(\mathbb{K})$, we have $\lim_{T \to +\infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} l(x_t, u_t) = J(\mathbb{K})$. □

□

G Additional Proofs

G.1 Proof of Lemma 4

The proof relies on the following two lemmas.

Lemma 22 (Definition of $\epsilon_{\hat{w}}$). *Under the conditions in Lemma 6,*

$$\sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1} (w_{t-k} - \hat{w}_{t-k}) \leq \epsilon_{\hat{w}}(r)$$

Proof.

$$\begin{aligned}
\|D_x \sum_{k=1}^{H_t} A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})\|_\infty &\leq \|D_x\|_\infty \sum_{k=1}^{H_t} \|A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})\|_\infty \\
&\leq \|D_x\|_\infty \sum_{k=1}^{H_t} \|A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})\|_2 \\
&\leq \|D_x\|_\infty \sum_{k=1}^{H_t} \kappa(1-\gamma)^{k-1} r z_{\max} \\
&\leq \|D_x\|_\infty \kappa/\gamma z_{\max} r = \epsilon_{\hat{w}}(r)
\end{aligned}$$

□

Lemma 23 (Definition of $\epsilon_{\hat{\theta}}$)**.** *For any* $\mathbf{M} \in \mathcal{M}$*, any* $\hat{\theta}, \theta \in \Theta^{(0)}$ *such that* $\|\hat{\theta} - \theta\|_F \leq r$*, we have*

$$|g_i^x(\mathbf{M}; \hat{\theta}) - g_i^x(\mathbf{M}; \theta)| \leq \epsilon_{\hat{\theta}}(r)$$

where $\epsilon_{\hat{\theta}}(r) = c_{\hat{\theta}} r \sqrt{mn}$*.*

Proof. Firstly, we show that it suffices to prove an upper bound of a simpler quantity.

$$\begin{aligned}
|g_i^x(\mathbf{M}; \hat{\theta}) - g_i^x(\mathbf{M}; \theta)| &= |\sum_{k=1}^{2H} \|D_{x,i}^\top \Phi_k^x(\mathbf{M}; \hat{\theta})\|_1 - \|D_{x,i}^\top \Phi_k^x(\mathbf{M}; \theta)\|_1 | w_{\max} \\
&\leq \sum_{k=1}^{2H} | \|D_{x,i}^\top \Phi_k^x(\mathbf{M}; \hat{\theta})\|_1 - \|D_{x,i}^\top \Phi_k^x(\mathbf{M}; \theta)\|_1 | w_{\max} \\
&\leq \sum_{k=1}^{2H} \|D_{x,i}^\top \Phi_k^x(\mathbf{M}; \hat{\theta}) - D_{x,i}^\top \Phi_k^x(\mathbf{M}; \theta)\|_1 w_{\max} \\
&\leq \sum_{k=1}^{2H} \|D_x\|_\infty \|\Phi_k^x(\mathbf{M}; \hat{\theta}) - \Phi_k^x(\mathbf{M}; \theta)\|_\infty w_{\max}
\end{aligned}$$

thus, it suffices to bound $\sum_{k=1}^{2H} \|\Phi_k^x(\mathbf{M}; \hat{\theta}) - \Phi_k^x(\mathbf{M}; \theta)\|_\infty$. To bound this, we need several small lemmas below.

Lemma 24. *When* $\|\theta - \hat{\theta}\|_F \leq r$*, we have* $\max(\|\hat{A} - A\|_2, \|\hat{B} - B\|_2) \leq \max(\|\hat{A} - A\|_F, \|\hat{B} - B\|_F) \leq r$

This is quite straightforward so the proof is omitted.

Lemma 25. *For any* $k \geq 0$*, any* $\hat{\theta}, \theta \in \Theta^{(0)}$ *such that* $\|\hat{\theta} - \theta\|_F \leq r$*, we have*

$$\begin{aligned}
\|A^k - \hat{A}^k\|_2 &\leq k\kappa^2 (1-\gamma)^{k-1} r \mathbb{1}_{(k \geq 1)} \\
\|A^k B - \hat{A}^k \hat{B}\|_2 &\leq k\kappa^2 \kappa_B (1-\gamma)^{k-1} r \mathbb{1}_{(k \geq 1)} + \kappa(1-\gamma)^k r
\end{aligned}$$

Proof. When $k = 0$, $\|A^0 - \hat{A}^0\|_2 = 0$. When $k \geq 1$,

$$\begin{aligned}
\|\hat{A}^k - A^k\|_2 &= \|\sum_{i=0}^{k-1} \hat{A}^{k-i-1}(\hat{A} - A)A^i\|_2 \\
&\leq \sum_{i=0}^{k-1} \|\hat{A}^{k-i-1}\|_2 \|\hat{A} - A\| \|A^i\|_2 \\
&\leq \sum_{i=0}^{k-1} \kappa(1-\gamma)^{k-i-1} \epsilon \kappa (1-\gamma)^i
\end{aligned}$$

$$
\begin{aligned}
&= k\kappa^2 r(1-\gamma)^{k-1} \\
\|\hat{A}^k\hat{B} - A^kB\|_2 &\le \|\hat{A}^k\hat{B} - A^k\hat{B}\|_2 + \|A^k\hat{B} - \hat{A}^k\hat{B}\|_2 \\
&\le k\kappa^2\kappa_B r(1-\gamma)^{k-1}\mathbb{1}_{(k\ge1)} + \kappa(1-\gamma)^k r
\end{aligned}
$$

□

Now, we can bound $\sum_{k=1}^{2H}\|\Phi_k^x(\mathbf{M};\hat{\theta}) - \Phi_k^x(\mathbf{M};\theta)\|_\infty$. For any $1 \le k \le 2H$,

$$
\begin{aligned}
&\|\Phi_k^x(\mathbf{M};\hat{\theta}) - \Phi_k^x(\mathbf{M};\theta)\|_\infty \\
&= \|\hat{A}^{k-1}\mathbb{1}_{(k\le H)} + \sum_{i=1}^{H}\hat{A}^{i-1}\hat{B}M_{t-i}[k-i]\mathbb{1}_{(1\le k-i\le H)} - A^{k-1}\mathbb{1}_{(k\le H)} - \sum_{i=1}^{H}A^{i-1}BM_{t-i}[k-i]\mathbb{1}_{(1\le k-i\le H)}\|_\infty \\
&\le \|\hat{A}^{k-1} - A^{k-1}\|_\infty\mathbb{1}_{(k\le H)} + \sum_{i=1}^{H}\|(\hat{A}^{i-1}\hat{B} - A^{i-1}B)M_{t-i}[k-i]\|_\infty\mathbb{1}_{(1\le k-i\le H)} \\
&\le \sqrt{n}\|\hat{A}^{k-1} - A^{k-1}\|_2\mathbb{1}_{(k\le H)} + \sqrt{m}\sum_{i=1}^{H}\|\hat{A}^{i-1}\hat{B} - A^{i-1}B\|_2 2\sqrt{n}\kappa^2(1-\gamma)^{k-i-1}\mathbb{1}_{(1\le k-i\le H)}
\end{aligned}
$$

There are two terms in the last right-hand-side of the inequality above. We sum each term over k below.

$$
\sum_{k=1}^{2H}\sqrt{n}\|\hat{A}^{k-1} - A^{k-1}\|_2\mathbb{1}_{(k\le H)} \le \sum_{k=1}^{2H}\sqrt{n}(k-1)\kappa^2(1-\gamma)^{k-2}r\mathbb{1}_{(2\le k\le H)} \le \sqrt{n}\kappa^2 r/\gamma^2
$$

$$
\begin{aligned}
&\sum_{k=1}^{2H}\sqrt{m}\sum_{i=1}^{H}\|\hat{A}^{i-1}\hat{B} - A^{i-1}B\|_2 2\sqrt{n}\kappa^2(1-\gamma)^{k-i-1}\mathbb{1}_{(1\le k-i\le H)} \\
&\le \sum_{k=1}^{2H}\sqrt{m}\sum_{i=1}^{H}(i-1)\kappa^2\kappa_B(1-\gamma)^{i-2}r\mathbb{1}_{(i\ge2)}2\sqrt{n}\kappa^2(1-\gamma)^{k-i-1}\mathbb{1}_{(1\le k-i\le H)} \\
&\quad + \sum_{k=1}^{2H}\sqrt{m}\sum_{i=1}^{H}\kappa(1-\gamma)^{i-1}r2\sqrt{n}\kappa^2(1-\gamma)^{k-i-1}\mathbb{1}_{(1\le k-i\le H)} \\
&= 2\sqrt{mn}\kappa^4\kappa_B r\sum_{i=1}^{H}\sum_{j=1}^{H}(i-1)(1-\gamma)^{i-2}(1-\gamma)^{j-1} + 2\sqrt{mn}\kappa^3 r\sum_i\sum_j(1-\gamma)^{i-1}(1-\gamma)^{j-1} \\
&= 2\sqrt{mn}\kappa^4\kappa_B r/\gamma^3 + 2\sqrt{mn}\kappa^3 r/\gamma^2
\end{aligned}
$$

□

G.2 Proof of Lemma 8

For notational simplicity, we omit the subscript t in H_t in this proof. Remember that $g_i^x(\mathbf{M}_{t-H:t-1};\theta) = \sum_{s=1}^{2H}\|D_{x,i}^\top\Phi_s^x(\mathbf{M}_{t-H:t-1};\theta)\|_1$

$$
\begin{aligned}
|\tilde{g}_i^x(\mathbf{M}_{t-H:t-1};\theta) - g_i^x(\mathbf{M};\theta)| &= \left|\sum_{k=1}^{2H}\|D_{x,i}^\top\tilde{\Phi}_k^x(\mathbf{M}_{t-H:t-1};\theta)\|_1 - \|D_{x,i}^\top\Phi_k^x(\mathbf{M}_t;\theta)\|_1\right| w_{\max} \\
&\le \sum_{k=1}^{2H}\left|\|D_{x,i}^\top\Phi_k^x(\mathbf{M}_{t-H:t-1};\theta^*)\|_1 - \|D_{x,i}^\top\Phi_k^x(\mathbf{M}_t;\theta)\|_1\right| w_{\max} \\
&\le \sum_{k=1}^{2H}\|D_{x,i}^\top(\tilde{\Phi}_k^x(\mathbf{M}_{t-H:t-1};\theta) - \Phi_k^x(\mathbf{M}_t;\theta))\|_1 w_{\max} \\
&\le \sum_{k=1}^{2H}\|D_x\|_\infty\|\tilde{\Phi}_k^x(\mathbf{M}_{t-H:t-1};\theta) - \Phi_k^x(\mathbf{M}_t;\theta)\|_\infty w_{\max}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{2H} \|D_x\|_\infty \| \sum_{i=1}^{H} A^{i-1} B (M_{t-i}[k-i] - M_t[k-i]) \|_\infty \mathbb{1}_{(1\leq k-i \leq H)} w_{\max} \\
&\leq \sum_{k=1}^{2H} \|D_x\|_\infty \sum_{i=1}^{H} \|A^{i-1} B\|_\infty \|M_{t-i}[k-i] - M_t[k-i]\|_\infty \mathbb{1}_{(1\leq k-i \leq H)} w_{\max} \\
&\leq \|D_x\|_\infty \sqrt{m} w_{\max} \sum_{k=1}^{2H} \sum_{i=1}^{H} \kappa (1-\gamma)^{i-1} \kappa_B \|M_{t-i}[k-i] - M_t[k-i]\|_\infty \mathbb{1}_{(1\leq k-i \leq H)} \\
&= \|D_x\|_\infty \sqrt{m} w_{\max} \kappa \kappa_B \sum_{i=1}^{H} \sum_{j=1}^{H} (1-\gamma)^{i-1} \|M_{t-i}[j] - M_t[j]\|_\infty \\
&\leq \|D_x\|_\infty \sqrt{m} w_{\max} \kappa \kappa_B \sqrt{nH} \sum_{i=1}^{H} (1-\gamma)^{i-1} \|\mathbf{M}_{t-i} - \mathbf{M}_t\|_F \\
&\leq \|D_x\|_\infty \sqrt{mnH} w_{\max} \kappa \kappa_B \sum_{i=1}^{H} (1-\gamma)^{i-1} i \Delta_M \\
&\leq \|D_x\|_\infty \sqrt{mnH} w_{\max} \kappa \kappa_B / \gamma^2 \Delta_M
\end{aligned}
$$

where the third last inequality is because $M[j] \in \mathbb{R}^{m\times n}$

$$
\sum_{j=1}^{H} \|M[j]\|_\infty \leq \sum_{j=1}^{H} \|M[j]\|_2 \sqrt{n} \leq \sum_{j=1}^{H} \|M[j]\|_F \sqrt{n} \leq \|\mathbf{M}\|_F \sqrt{n} \sqrt{H}
$$

G.3 Proof of Lemma 11

For notational simplicity, we define $y_t = \sum_{i=1}^{H_t} A_*^{i-1} w_{t-i} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} A_*^{i-1} B_* M_{t-i}[k-i] \hat{w}_{t-k} \mathbb{1}_{1\leq k-i \leq H_t} + \sum_{i=1}^{H_t} A_*^{i-1} B_* \eta_{t-i}$. Since A_* is (κ, γ)-stable, we have

$$
\begin{aligned}
\|y_t\|_2 &\leq \sum_{i=1}^{H_t} \|A_*^{i-1}\|_2 \|w_{t-i}\|_2 + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} \|A_*^{i-1} B_* M_{t-i}[k-i] \hat{w}_{t-k}\|_2 \mathbb{1}_{1\leq k-i \leq H_t} + \sum_{i=1}^{H_t} \|A_*^{i-1} B_* \eta_{t-i}\|_2 \\
&\leq \sum_{i=1}^{H_t} \kappa (1-\gamma)^{i-1} \sqrt{n} w_{\max} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} \|A_*^{i-1} B_*\|_2 \|M_{t-i}[k-i] \hat{w}_{t-k}\|_2 \mathbb{1}_{1\leq k-i \leq H_t} + \sum_{i=1}^{H_t} \|A_*^{i-1} B_*\|_2 \|\eta_{t-i}\|_2 \\
&\leq \kappa \sqrt{n} w_{\max}/\gamma + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} \kappa (1-\gamma)^{i-1} \kappa_B \sqrt{m} \|M_{t-i}[k-i] \hat{w}_{t-k}\|_\infty \mathbb{1}_{1\leq k-i \leq H_t} + \sum_{i=1}^{H_t} \kappa (1-\gamma)^{i-1} \kappa_B \sqrt{n} \eta_{\max} \\
&\leq \kappa \sqrt{n} w_{\max}/\gamma + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} \kappa (1-\gamma)^{i-1} \kappa_B \sqrt{m} 2\sqrt{n} \kappa^2 (1-\gamma)^{k-i-1} w_{\max} \mathbb{1}_{1\leq k-i \leq H_t} + \kappa \kappa_B/\gamma \sqrt{n} \eta_{\max} \\
&\leq \kappa \sqrt{n} w_{\max}/\gamma + \kappa \kappa_B/\gamma \sqrt{n} \eta_{\max} + \kappa^3 \kappa_B 2\sqrt{mn} w_{\max} \sum_{i=1}^{H_t} \sum_{j=1}^{H_t} (1-\gamma)^{i-1} (1-\gamma)^{j-1} \\
&\leq \sqrt{n} (\kappa w_{\max} + \kappa \kappa_B \eta_{\max})/\gamma + \kappa^3 \kappa_B 2\sqrt{mn} w_{\max}/\gamma^2 \\
&\leq 2\sqrt{n} \kappa w_{\max}/\gamma + \kappa^3 \kappa_B 2\sqrt{mn} w_{\max}/\gamma^2 \leq c_{bx} \sqrt{mn}
\end{aligned}
$$

Remember that $x_t = A_*^{H_t} x_{t-H_t} + y_t$ and and $\|x_t\|_2 = 0 \leq b_x$ for $t \leq 0$. We prove the bound on x_t by induction. Suppose at $t \geq 0$, $\|x_{t-H_t}\|_2 \leq b_x$, then

$$
\begin{aligned}
\|x_t\|_2 &\leq \|A_*^{H_t}\|_2 \|x_{t-H_t}\|_2 + \|y_t\|_2 \leq \kappa (1-\gamma)^{H_t} b_x + 2\sqrt{n} \kappa w_{\max}/\gamma + \kappa^3 \kappa_B 2\sqrt{mn} w_{\max}/\gamma^2 \\
&\leq b_x/2 + 2\sqrt{n} \kappa w_{\max}/\gamma + \kappa^3 \kappa_B 2\sqrt{mn} w_{\max}/\gamma^2 = b_x
\end{aligned}
$$

where the last inequality is by $\kappa(1-\gamma)^{H_t} \leq 1/2$ when $H_t \geq \log(2\kappa)/\log((1-\gamma)^{-1})$. This completes the proof.

G.4 Proof of Lemma 18

Proof. We omit θ in this proof for simplicity of notations.

For any $H \geq 1$, define $\mathcal{M}_{out,H} = \{\mathbf{M} \in \mathbb{R}^{mnH} : \|M[k]\|_\infty \leq 4\kappa^2\sqrt{n}(1-\gamma)^{k-1}\}$. Notice that $\mathcal{M}_H \subseteq interior(\mathcal{M}_{out,H})$. Therefore, for any $\mathbf{M} \in \mathcal{M}_H$,

$$\begin{aligned}\|\nabla f(\mathbf{M};\theta)\|_F &= \sup_{\Delta\mathbf{M}\neq 0, \mathbf{M}+\Delta\mathbf{M}\in\mathcal{M}_{out,H}} \frac{\langle \nabla f(\mathbf{M};\theta), \Delta\mathbf{M}\rangle}{\|\Delta\mathbf{M}\|_F} \\ &\leq \sup_{\Delta\mathbf{M}\neq 0, \mathbf{M}+\Delta\mathbf{M}\in\mathcal{M}_{out,H}} \frac{f(\mathbf{M}+\Delta\mathbf{M}) - f(\mathbf{M})}{\|\Delta\mathbf{M}\|_F}\end{aligned}$$

For $\mathbf{M}, \mathbf{M}' \in \mathcal{M}_{out,H}$, we bound the following.

$$\begin{aligned}\|\tilde{x} - \tilde{x}'\|_2 &\leq \sum_{k=1}^{2H} \|(\Phi_k^x(\mathbf{M}) - \Phi_k^x(\mathbf{M}'))w_{t-k}\|_2 \\ &\leq \sum_{k=1}^{2H} \|\sum_{i=1}^{H^{(e)}} A^{i-1}B(M[k-i] - M'[k-i])\mathbb{1}_{(1\leq k-i\leq H)} w_{t-k}\|_2 \\ &\leq \sum_{j=1}^{H} O(\sqrt{n})\|M[j] - M'[j]\|_2 \\ &\leq \sum_{j=1}^{H} O(\sqrt{n})\|M[j] - M'[j]\|_F \\ &\leq O(\sqrt{n}\sqrt{H})\|\mathbf{M} - \mathbf{M}'\|_F \\ \|\tilde{u} - \tilde{u}'\|_2 &\sum_{k=1}^{H} \|M[k] - M'[k]\|_2\sqrt{n}w_{\max} \leq O(\sqrt{n}\sqrt{H})\|\mathbf{M} - \mathbf{M}'\|_F\end{aligned}$$

where the third inequality uses $\theta \in \Theta_{ini}$.

Further, even though we make $\mathcal{M}_{out,H}$ larger, but we don't change the dimension, so by Lemma 24, $\|\tilde{x}\|_2 \leq \sqrt{mn}$. Further, even when we don't have additional conditions on $\mathbf{M}$, we still have $\|\tilde{u}\|_2 \leq O(\sqrt{mn})$. Therefore, for $\mathbf{M}, \mathbf{M}' \in \mathcal{M}_{out,H}$,

$$|f(\mathbf{M}) - f(\mathbf{M}')| \leq O(\sqrt{mn}\sqrt{n}\sqrt{H})\|\mathbf{M} - \mathbf{M}'\|_F$$

Therefore,

$$\begin{aligned}\|\nabla f(\mathbf{M};\theta)\|_F &\leq \sup_{\Delta\mathbf{M}\neq 0, \mathbf{M}+\Delta\mathbf{M}\in\mathcal{M}_{out,H}} \frac{f(\mathbf{M}+\Delta\mathbf{M}) - f(\mathbf{M})}{\|\Delta\mathbf{M}\|_F} \\ &\leq \sup_{\Delta\mathbf{M}\neq 0, \mathbf{M}+\Delta\mathbf{M}\in\mathcal{M}_{out,H}} \frac{O(\sqrt{mn}\sqrt{n}\sqrt{H})\|\Delta\mathbf{M}\|_F}{\|\Delta\mathbf{M}\|_F} \leq O(n\sqrt{m}\sqrt{H})\end{aligned}$$

□

G.5 Proof of Lemma 17

Proof. Notice that Ω_1 and Ω_3 satisfies the conditions in Proposition 2 in Li et al. (2020). Therefore,

$$|\min_{\Omega_1} f(x) - \min_{\Omega_3} f(x)| \leq \frac{L d_{\Omega_0}\|\Delta_1 - \Delta_3\|_\infty}{\min_{\{i:(\Delta_1)_i > (\Delta_3)_i\}}(h - \Delta_1 - Cx_F)_i}$$

Notice that

$$(\Delta_3)_i = \begin{cases}(\Delta_1)_i, & \text{if } (\Delta_1)_i \geq (\Delta_2)_i \\ (\Delta_2)_i, & \text{if } (\Delta_1)_i < (\Delta_2)_i\end{cases}$$

therefore, $\|\Delta_1 - \Delta_3\|_\infty \leq \|\Delta_1 - \Delta_2\|_\infty$. Further, $\{i : (\Delta_3)_i > (\Delta_1)_i\} = \{i : (\Delta_2)_i > (\Delta_1)_i\} \subseteq \{i : (\Delta_1)_i \neq (\Delta_2)_i\}$. So $\min_{\{i:(\Delta_3)_i>(\Delta_1)_i\}}(h - \Delta_1 - Cx_F)_i \geq \min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}}(h - \Delta_1 - Cx_F)_i \geq \min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}}(h - \Delta_3 - Cx_F)_i$. Therefore,

$$|\min_{\Omega_1} f(x) - \min_{\Omega_3} f(x)| \leq \frac{Ld_{\Omega_0}\|\Delta_1 - \Delta_3\|_\infty}{\min_{\{i:(\Delta_1)_i>(\Delta_3)_i\}}(h - \Delta_1 - Cx_F)_i} \leq \frac{Ld_{\Omega_0}\|\Delta_1 - \Delta_2\|_\infty}{\min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}}(h - \Delta_3 - Cx_F)_i}$$

Similarly,

$$|\min_{\Omega_2} f(x) - \min_{\Omega_3} f(x)| \leq \frac{Ld_{\Omega_0}\|\Delta_2 - \Delta_3\|_\infty}{\min_{\{i:(\Delta_2)_i>(\Delta_3)_i\}}(h - \Delta_2 - Cx_F)_i} \leq \frac{Ld_{\Omega_0}\|\Delta_1 - \Delta_2\|_\infty}{\min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}}(h - \Delta_3 - Cx_F)_i}$$

which completes the bound. □

G.6 Proof of Lemma 19

In this subsection, we provide a proof for our bound on Part ii by martingale concentration inequalities.

Lemma 26. *In our Algorithm 1,* $\mathbf{M}^{(e)} \in \mathcal{F}(w_0, \ldots, w_{t_1^{(e)}+T_D^{(e)}-1}, \eta_0, \ldots, \eta_{t_1^{(e)}+T_D^{(e)}-1}) = \mathcal{F}^m_{t_1^{(e)}+T_D^{(e)}} \subseteq \mathcal{F}_{t_2^{(e)}-H^{(e)}}$.

Proof. By definition, we have the following fact: $\mathbf{M}^{(e)} \in \mathcal{F}(\hat{\theta}^{(e+1)}) = \mathcal{F}(\{z_k, x_{k+1}\}_{k=t_1^{(e)}}^{t_1^{(e)}+T_D^{(e)}-1}) = \mathcal{F}(w_0, \ldots, w_{t_1^{(e)}+T_D^{(e)}-1}, \eta_0, \ldots, \eta_{t}$ $\mathcal{F}^m_{t_1^{(e)}+T_D^{(e)}}$. By $\tilde{W}_1^{(e)} \geq H^{(e)}$, we have $t_1^{(e)} + T_D^{(e)} + H^{(e)} \leq t_2^{(e)}$, and since $\mathcal{F}^m_t \subseteq \mathcal{F}_t$, we have the last claim. □

Lemma 27. *When* $t \in \mathcal{T}_2^{(e)}$, $w_{t-2H^{(e)}} \perp\!\!\!\perp \mathcal{F}_{t_2^{(e)}-H^{(e)}}$

Proof. When $t \in \mathcal{T}_2^{(e)}$, $t \geq t_2^{(e)} + H^{(e)}$, so $t - 2H^{(e)} \geq t_2^{(e)} - H^{(e)}$. Since $\mathcal{F}_t$ contains up to w_{t-1}, we have $w_{t-2H^{(e)}} \perp\!\!\!\perp \mathcal{F}_{t_2^{(e)}-H^{(e)}}$. □

Lemma 28. *In our Algorithm 1, when* $t \in \mathcal{T}_2^{(e)}$, *we have* $\mathbb{E}[l(\hat{x}_t, \hat{u}_t) \mid \mathcal{F}_{t_2^{(e)}-H^{(e)}}] = f(\mathbf{M}^{(e)}; \theta_*)$.

Proof. By our lemmas above, $\mathbf{M}^{(e)} \in \mathcal{F}_{t_2^{(e)}-H^{(e)}}$, but $w_{t-2H^{(e)}} \perp\!\!\!\perp \mathcal{F}_{t_2^{(e)}-H^{(e)}}$. Then, by our definition of $\hat{x}_t, \hat{u}_t$ and $f(\mathbf{M}; \theta_*)$, we have the result. □

Definition 5 (Martingale)**.** $\{X_t\}_{t\geq 0}$ *is a martingale wrt* $\{\mathcal{F}_t\}_{t\geq 0}$ *if (i)* $\mathbb{E}\,|X_t| < +\infty$, *(ii)* $X_t \in \mathcal{F}_t$, *(iii)* $\mathbb{E}(X_{t+1} \mid \mathcal{F}_t) = X_t$ *for* $t \geq 0$.

Proposition 4 (Azuma-Hoeffding Inequality)**.** $\{X_t\}_{t\geq 0}$ *is a martingale with respect to* $\{\mathcal{F}_t\}_{t\geq 0}$. *If (i)* $X_0 = 0$, *(ii)* $|X_t - X_{t-1}| \leq \sigma$ *for any* $t \geq 1$, *then, for any* $\alpha > 0$, *any* $t \geq 0$,

$$\mathbb{P}(|X_t| \geq \alpha) \leq 2\exp\left(-\alpha^2/(2t\sigma^2)\right)$$

Corollary 3. $\{X_t\}_{t\geq 0}$ *is a martingale wrt* $\{\mathcal{F}_t\}_{t\geq 0}$. *If (i)* $X_0 = 0$, *(ii)* $|X_t - X_{t-1}| \leq \sigma$ *for any* $t \geq 1$, *then, for any* $\delta \in (0, 1)$,

$$|X_t| \leq \sqrt{2t}\sigma\sqrt{\log(2/\delta)}$$

w.p. at least $1 - \delta$.

Proof. The proof is by letting $\alpha = \sqrt{2t\sigma^2\log(2/\delta)}$ in Proposition 4. □

Lemma 29. *Define* $q_t = l(\hat{x}_t, \hat{u}_t) - f(\mathbf{M}^{(e)}; \theta_*)$. *Then,* $|q_t| \leq O(mn)$ *w.p.1.*

Proof. We can show that $\|\hat{x}_t\|_2 \leq O(\sqrt{mn})$ a.s. and $\hat{u}_t \in \mathbb{U}$ a.s. by the proofs of Lemmas 10 and 11. Therefore, we have $|l(\hat{x}_t, \hat{u}_t)| = O(mn)$. Since $f(\mathbf{M}^{(e)}; \theta_*) = \mathbb{E}[l(\hat{x}_t, \hat{u}_t) \mid \mathcal{F}_{t_2^{(e)}-H^{(e)}}]$, we have $|f(\mathbf{M}^{(e)}; \theta_*)| = O(mn)$. This completes the proof. □

Notations and definitions. Define, for $0 \le h \le 2H^{(e)} - 1$, that

$$\mathcal{T}_{2,h}^{(e)} = \{t \in \mathcal{T}_2^{(e)} : t \equiv h \mod (2H^{(e)})\} =: \{t_h^{(e)} + 2H^{(e)}, \dots, t_h^{(e)} + 2H^{(e)} k_h^{(e)}\} \tag{18}$$

Lemma 30. *$t_h^{(e)} \ge t_2^{(e)} - H^{(e)}$ and $k_h^{(e)} \le T^{(e+1)}/(2H^{(e)})$*

Proof. Notice that $t_h^{(e)} + 2H^{(e)} \ge t_2^{(e)} + H^{(e)}$, so the first inequality holds. Besides, notice that $2H^{(e)}k_h^{(e)} \le t_h^{(e)} + 2H^{(e)}k_h^{(e)} \le T^{(e+1)}$, so the second inequality holds. □

Define

$$\tilde{q}_{h,j}^{(e)} = q_{t_h^{(e)} + j(2H^{(e)})} \quad \forall 1 \le j \le k_h^{(e)} \tag{19}$$

$$S_{h,j}^{(e)} = \sum_{s=1}^{j} \tilde{q}_{h,s}^{(e)} \quad \forall 0 \le j \le k_h^{(e)}, \tag{20}$$

$$\mathcal{F}_{h,j}^{(e)} = \mathcal{F}_{t_h^{(e)} + j(2H^{(e)})} \quad \forall 0 \le j \le k_h^{(e)}, \tag{21}$$

where we define $\sum_{s=1}^{0} a_s = 0$. By Lemma 30, we have $\mathcal{F}_{h,0}^{(e)} = \mathcal{F}_{t_h^{(e)}} \supseteq \mathcal{F}_{t_2^{(e)} - H^{(e)}}$.

Lemma 31. *$S_{h,j}^{(e)}$ is a martingale wrt $\mathcal{F}_{h,j}^{(e)}$ for $j \ge 0$. Further, $S_{k,0}^{(e)} = 0$, $|S_{h,j+1}^{(e)} - S_{h,j}^{(e)}| \le O(mn)$.*

Proof. Since $|q_t| \le O(mn)$, $\mathbb{E}\,|S_{h,j}^{(e)}| \le O(Tmn) < +\infty$. Notice that, for $t \in \mathcal{T}_2^{(e)}$, $w_{t-1}, \dots, w_{t-2H^{(e)}} \in \mathcal{F}_t$. and $\mathbf{M}^{(e)} \in \mathcal{F}_t$, so $q_t \in \mathcal{F}_t$, so $S_{h,j}^{(e)} \in \mathcal{F}_{h,j}^{(e)}$. Next, $\mathbb{E}[S_{h,j+1}^{(e)} \mid \mathcal{F}_{h,j}^{(e)}] = S_{h,j}^{(e)} + \mathbb{E}[q_{h,j+1}^{(e)} \mid \mathcal{F}_{h,j}^{(e)}] = S_{h,j}^{(e)}$. So this is done. The rest is by definition, and q_t's bound. □

Lemma 32. *Consider our choice of $H^{(e)}$ in Theorem 3. Let $\delta = \frac{p}{2\sum_{e=0}^{N-1} H^{(e)}}$, w.p. $1-\delta$, we have $|S_{h,k_h^{(e)}}^{(e)}| \le \tilde{O}\left(\sqrt{k_h^{(e)}}mn\right)$.*

Proof. By Lemma 31, we can apply Corollary 3, and obtain the bound, where we used $\log(2/\delta) = \tilde{O}(1)$. □

Lemma 33. *Consider our choice of $H^{(e)}$ in Theorem 3. For any e, w.p. $1 - 2H^{(e)}\delta$, where $\delta = \frac{p}{2\sum_{e=0}^{N-1} H^{(e)}}$,*

$$|\sum_{h=0}^{2H^{(e)}-1} S_{h,k_h^{(e)}}^{(e)}| \le \tilde{O}\left(\sqrt{T^{(e+1)}}mn\right)$$

Proof. Define event

$$\mathcal{E}_h^{(e)} = \{|S_{h,k_h^{(e)}}^{(e)}| \le \tilde{O}\left(\sqrt{k_h^{(e)}}mn\right)\}$$

When $\cap_h \mathcal{E}_h^{(e)}$ holds,

$$|\sum_{t \in \mathcal{T}_2^{(e)}} q_t| = |\sum_{h=0}^{2H^{(e)}-1} S_{h,k_h^{(e)}}^{(e)}| \tilde{O}(mn\sqrt{\sum_h k_h^{(e)}}\sqrt{2H^{(e)}}) \le \tilde{O}(mnT^{(e+1)})$$

where we used Lemma 30 and Cauchy Schwartz.

Then, we have

$$\mathbb{P}(\cap_h \mathcal{E}_h^{(e)}) = 1 - \mathbb{P}(\cup_h (\mathcal{E}_h^{(e)})^c) \ge 1 - \sum_h \mathbb{P}((\mathcal{E}_h^{(e)})^c) \ge 1 - 2H^{(e)}\delta$$

□

Now, we can prove Lemma 19. By Lemma 33, w.p. $1-p$, we have $|\sum_{h=0}^{2H^{(e)}-1} S_{h,k_h^{(e)}}^{(e)}| \le \tilde{O}\left(\sqrt{T^{(e+1)}}mn\right)$ for all e. Then, by Lemma 14, we completed the proof.

References

Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, pages 5582–5588. IEEE, 2019a.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019b.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, volume 32, pages 10154–10164. Curran Associates, Inc., 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. *arXiv preprint arXiv:2001.09254*, 2020.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{t}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

Mark Campbell, Magnus Egerstedt, Jonathan P How, and Richard M Murray. Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4649–4672, 2010.

Milos Vasic and Aude Billard. Safety issues in human-robot interactions. In *2013 ieee international conference on robotics and automation*, pages 197–204. IEEE, 2013.

Frank L. Lewis, Draguna L. Vrabie, and Vassilis L. Syrmos. *Optimal Control*. John Wiley & Sons, third edition, 2012.

James Blake Rawlings and David Q Mayne. *Model predictive control: Theory and design*. Nob Hill Pub., 2009.

Alberto Bemporad, Manfred Morari, Vivek Dua, and Efstratios N Pistikopoulos. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1):3–20, 2002.

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

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

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

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

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

Frauke Oldewurtel, Colin N Jones, and Manfred Morari. A tractable approximation of chance constrained stochastic mpc based on affine disturbance feedback. In *2008 47th IEEE conference on decision and control*, pages 4731–4736. IEEE, 2008.

Kunwu Zhang and Yang Shi. Adaptive model predictive control for a class of constrained linear systems with parametric uncertainties. *Automatica*, 117:108974, 2020.

Monimoy Bujarbaruah, Xiaojing Zhang, Marko Tanaskovic, and Francesco Borrelli. Adaptive mpc under time varying uncertainty: Robust and stochastic. *arXiv preprint arXiv:1909.13473*, 2019.

Johannes Köhler, Elisa Andina, Raffaele Soloperto, Matthias A Müller, and Frank Allgöwer. Linear robust adaptive model predictive control: Computational complexity and conservatism. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1383–1388. IEEE, 2019.

Xiaonan Lu, Mark Cannon, and Denis Koksal-Rivet. Robust adaptive model predictive control: Performance and parameter estimation. *International Journal of Robust and Nonlinear Control*, 2019.

Kim P Wabersich and Melanie N Zeilinger. Performance and safety of bayesian model predictive control: Scalable model-based rl with guarantees. *arXiv preprint arXiv:2006.03483*, 2020.

Deepan Muthirayan, Jianjun Yuan, and Pramod P Khargonekar. Regret guarantees for online receding horizon control. *arXiv preprint arXiv:2010.07269*, 2020.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M Kakade, and Karan Singh. Online control with adversarial disturbances. In *36th International Conference on Machine Learning, ICML 2019*, pages 154–165. International Machine Learning Society (IMLS), 2019a.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pages 10175–10184, 2019b.

Orestis Plevrakis and Elad Hazan. Geometric exploration for online control. *arXiv preprint arXiv:2010.13178*, 2020.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49(2):267–290, 2002.

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

Zahra Marvi and Bahare Kiumarsi. Safe reinforcement learning: A control barrier function optimization approach. *International Journal of Robust and Nonlinear Control*, 31(6):1923–1940, 2021.

Edouard Leurent, Denis Efimov, and Odalric-Ambrym Maillard. Robust-adaptive control of linear systems: beyond quadratic costs. In *NeurIPS 2020-34th Conference on Neural Information Processing Systems*, 2020.

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

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

Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods. In *AAAI Conference on Artificial Intelligence*, 2018.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *arXiv preprint arXiv:2002.08538*, 2020.

Ryan James Caverly and James Richard Forbes. Lmi properties and applications in systems, stability, and control theory. *arXiv preprint arXiv:1903.08599*, 2019.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2016.

Anayo K Akametalu, Jaime F Fisac, Jeremy H Gillula, Shahab Kaynama, Melanie N Zeilinger, and Claire J Tomlin. Reachability-based safe learning with gaussian processes. In *53rd IEEE Conference on Decision and Control*, pages 1424–1431. IEEE, 2014.

Christopher M Kellett and Andrew R Teel. Smooth lyapunov functions and robustness of stability for difference inclusions. *Systems & Control Letters*, 52(5):395–405, 2004.

Yingying Li, Subhro Das, and Na Li. Online optimal control with affine constraints. *arXiv preprint arXiv:2010.04891*, 2020.

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