Regret and Safety Guarantees for Adaptive Linear Quadratic Control with Constraints

Anonymous Author(s)

Abstract

We study the adaptive control of an unknown linear system with safety constraints on both the states and actions subject to quadratic cost functions. The challenges of this problem arise from the tension between safety, exploration, performance, and computation. To address these challenges, we propose a polynomial-time algorithm that guarantees feasibility and constraint satisfaction with high probability under certain conditions. Our algorithm is implemented on a single trajectory and does not require restarts. Further, we analyze the regret of our algorithm compared to the optimal safe linear controller with known model information. The proposed algorithm can achieve a $\tilde{O}(T^{2/3})$ regret, where T is the number of stages and $\tilde{O}(\cdot)$ absorbs some polynomial terms of T.

1 Introduction

Reinforcement learning (RL) and learning-based control have attracted a lot of attention and witnessed many successes in recent years [43, 34, 33]. However, when applying learning-based algorithms to physical systems, how to ensure safety during the learning process becomes a major concern. This limits the application of RL to safety critical systems, such as autonomous driving [8], robotics [47], power systems [41], etc. To ensure safety, one has to be conservative in the face of uncertainties, but too much conservativeness may degrade system performance and slow down learning. The inherent tension among safety, exploration, and performance/exploitation imposes significant challenges to the algorithm design.

As an attempt to address the challenges above, we focus on a linear quadratic regulator (LQR) problem with safety constraint sets on both states and actions and *unknown* system dynamics. The goal is to learn a system model without violating safety constraints throughout the process and achieve near-optimal online performance. The problem has been analyzed with tools from both control [51, 7, 21] and learning community [49, 15, 35] but many questions still remain open.

From the control community, robust model predictive control (RMPC) is usually adopted to handle constraint satisfaction in the presence of process noises and/or model uncertainties [5, 30, 22, 40, 27]. There is also a growing interest recently in the adaptive version of RMPC (RAMPC) that actively explores the system to reduce the model uncertainties for less conservativeness and better performance [51, 7, 21, 27]. However, most RMPC and RAMPC focus on stability, feasibility, and constraint satisfaction guarantees, with fewer results on the optimality performance, especially the non-asymptotic performance analysis.

From the learning community, there are many tools to analyze non-asymptotic performance with the development of non-asymptotic estimation rates, regret analysis, and perturbation bounds. Consequently, there is an emerging interest in applying learning tools on the constrained LQR problem for non-asymptotic performance guarantees. However, most existing work on this line sacrifices safety in some sense, or suffers a large computational burden. For example, [49] analyzes the Bayesian regret of adaptive MPC but allows constraint violation during the process and requires restarting the system back to some safe state every time when the model is updated. In contrast, [15] guarantees constraint satisfaction all the time but does not allow policy updates during the learning, i.e., they consider a non-adaptive learning scheme. Moreover, [35] considers adaptive learning with input constraint satisfaction but does not consider state constraints, and their approach requires an oracle that may be computationally intractable.

Therefore, an important question is: *How to design tractable adaptive control algorithms without sacrificing constraint satisfaction but still achieve non-asymptotic online performance guarantees?*

Our contributions. In this paper, we address this question by designing a polynomial-time adaptive control algorithm for constrained LQR. Our algorithm can ensure constraint satisfaction at all stages without restarting. To ensure safety during the learning process, we adopt two steps: (i) we construct confidence sets of the model estimation and only consider policies that are robustly safe for all the potential systems in the confidence sets, (ii) we develop a novel safe transition algorithm to guarantee constraint satisfaction during the updates of the policies and confidence sets without requiring restarts.

Further, we analyze the non-asymptotic performance of our safe adaptive control algorithm by bounding the policy regret. For simplicity, we first consider safe linear static policies as our benchmark class. We show that our algorithm can achieve $\tilde{O}(T^{2/3})$ policy regret with high probability, while satisfying constraints during the learning process. This shows that our algorithm can balance exploration and exploitation under safety constraints. We then discuss how to generalize our results to broader benchmark policy classes that include linear dynamical policies as studied in [12] and certain types of robust model predictive control [30] in the supplementary file.

One key step in deriving our regret bound is that we establish a tight bound on the model estimation error (confidence bound) caused by exploring the linear system with a possibly *nonlinear* control policy. The nonlinearity is due to the projections introduced for constraint satisfaction. Current estimation error bounds either assume linear exploration policies [12] or consider general nonlinear systems without leveraging the special structures of our problem [18, 42]. Interestingly, our established bound has the same rate as the bound for linear policies, indicating the tightness of our bound. Since nonlinear policies are commonly used for constrained optimal linear control, they can be used broadly for analyzing other learning-based control algorithms. Our estimation error bounds are novel and a contribution in their own right.

Related work. *Learning-based control without constraints.* This has been actively studied in recent years [16, 13, 28, 38, 12, 46, 45, 11]. Our paper utilizes the disturbance-action policies developed for unconstrained control in [2, 3] to tackle constraint satisfaction. Further, our design relies on certainty equivalence, which has been shown to be optimal for learning-based control without constraints [28, 44]. The robust stability guarantee is also studied in [12, 14, 10].

Safe reinforcement learning. Safety in RL refers to different requirements [32, 20]. This paper is relevant to RL with state and action constraints [29, 23, 17, 20, 9, 19]. Many papers consider soft constraints by only aiming for sublinear number of stages when constraint violation happen [39, 50, 49].

Safe online control. There is an orthogonal line of work that considers known system dynamics but time-varying costs for constrained optimal control [24, 36]. This paper builds upon the results in [24] and extends the ideas to handle unknown systems.

Constrained control. Without disturbances, it is known that the optimal controller for the linearly constrained linear quadratic regulator is piecewise affine (PWA) [6]. With disturbances, the problem is much more challenging. Current methods include RMPC and its variants [26, 25, 40, 30], stochastic MPC and its variants [31, 37], system level synthesis [15], etc.

Notations. Let $\mathcal{D}_\eta$ denote a distribution, then we write $\eta \sim \bar{\eta}\mathcal{D}_\eta$ if $\eta = \bar{\eta}\tilde{\eta}$ and $\tilde{\eta}$ follows distribution $\mathcal{D}_\eta$. By $\|.\|_F$ we denote the Frobenius norm. Define $\mathbb{B}(\hat{\theta}, r) = \{\theta : \|\theta - \hat{\theta}\|_F \leq r\}$. We write $(x, y) > 0$ if $x > 0$ and $y > 0$.

2 Problem formulation

Consider the following optimal control with a constrained linear system and bounded disturbances.

$$\begin{aligned} \min_{\pi} \ & J(\pi) = \lim_{T \to +\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\, l(x_t, u_t) \\ \text{s.t. } & x_{t+1} = A_* x_t + B_* u_t + w_t, \ \forall\, t \geq 0, \\ & D_x x_t \leq d_x, \ D_u u_t \leq d_u, \ \forall \{w_t : \|w_t\|_\infty \leq w_{\max}\}, \end{aligned} \tag{1}$$

where π denotes a control policy, $J(\pi)$ denotes the infinite-horizon averaged cost of π, $l(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t$ for positive definite Q and R, $x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^m, d_x \in \mathbb{R}^{k_x}, d_u \in \mathbb{R}^{k_u}$, x_0 is given. We define some shorthand notations: $\mathbb{X} := \{x : D_x x \leq d_x\}$, $\mathbb{U} := \{u : D_u u \leq d_u\}$, $\mathbb{W} := \{w \in \mathbb{R}^n : \|w\|_\infty \leq w_{\max}\}$, $\theta_* := (A_*, B_*)$, $\theta := (A, B)$. We consider bounded $\mathbb{X}$ and $\mathbb{U}$, i.e., there exist $x_{\max}, u_{\max}$ such that $\|x\|_2 \leq x_{\max}, \forall x \in \mathbb{X}$, $\|u\|_2 \leq u_{\max}, \forall u \in \mathbb{U}$. For simplicity, we consider $x_0 = 0$.[1] For ease of illustration and discussion, we define three versions of safety.

Definition 1 (Safety)**.** *We call an algorithm to be (i)* safe *if $x_t \in \mathbb{X}, u_t \in \mathbb{U}$ for all t and all $w_k \in \mathbb{W}$ when implementing the algorithm on the true system θ_*; (ii)* ϵ-strictly safe*, where $\epsilon = (\epsilon_x, \epsilon_u) > 0$, if $D_x x_t \leq d_x - \epsilon_x \mathbb{1}_{k_x}, D_u u_t \leq d_u - \epsilon_u \mathbb{1}_{k_u}$ for all t and all $w_k \in \mathbb{W}$ when implementing the algorithm on the true system θ_*; (iii)* robustly safe *on a model uncertainty set Θ if $x_t \in \mathbb{X}, u_t \in \mathbb{U}$ for all t and all $w_k \in \mathbb{W}$ when implementing the algorithm on any system $\theta \in \Theta$.*

To demonstrate theoretical rigor, we introduce a quantitative version of matrix stability.

Definition 2. *For $\kappa \geq 1$, $\gamma \in [0, 1)$, a matrix A is called (κ, γ)-stable*[2] *if $\|A^t\|_2 \leq \kappa(1-\gamma)^t, \forall t \geq 0$.*

In this work, we consider that the system parameters $\theta_* = (A_*, B_*)$ are unknown but the constraints $((\mathbb{X}, \mathbb{U}, \mathbb{W})$ and cost functions $(Q, R,)$ are known and x_t can be observed. Though the true model (A_*, B_*) is unknown, we assume some prior knowledge on the system dynamics is available, i.e., a model uncertainty set Θ_{ini} that satisfies the following assumption.

Assumption 1. *There is a known model uncertainty set $\Theta_{\text{ini}} = \{\theta : \|\theta - \hat{\theta}_{\text{ini}}\|_F \leq r_{\text{ini}}\}$*[3] *for some $0 < r_{\text{ini}} < +\infty$ such that (i) $\theta_* \in \Theta_{\text{ini}}$, and (ii) there exist $\kappa \geq 1, \gamma \in [0, 1)$ such that for any $(A, B) \in \Theta_{\text{ini}}$, A is (κ, γ)-stable.*

Remark 1. *Condition (i) is standard in the literature [12, 30]. Condition (ii) is a strong assumption and imposed for technical simplicity. One can relax condition (ii) to the following: (iii): there exists K such that $A - BK$ is (κ, γ)-stable for any $(A, B) \in \Theta_{ini}$. Condition (iii) is a common assumption in the robust constrained control literature [21, 27] and K can be computed by solving linear matrix inequalities if it exists [21].*

We note that although A_* is stable, implementing zero control may not be safe, i.e., violating the constraints, so it calls for a more careful control design to ensure constraint satisfaction.

Next, we impose assumptions on disturbance w_t. We introduce anti-concentration property [1].

Definition 3 (Anti-concentration)**.** *A random vector $X \in \mathbb{R}^n$ is said to satisfy (s, p)-anti-concentration properties for some $s > 0, p \in (0, 1)$ if for any $\lambda \in \mathbb{R}^n$, $\|\lambda\|_2 = 1$, $\mathbb{P}(\lambda^\top X \geq s) \geq p$.*

Notice that this definition essentially requires X has positive probability on all directions and there is a positive lower bound on the probability of each direction.

Assumption 2. *$w_t \in \mathbb{W}$ is i.i.d., σ_{sub}^2-sub-Gaussian, zero mean, and (s_w, p_w)-anti-concentration.*[4]

Regret definition and Benchmark Policy. In this paper, we aim to design an adaptive algorithm $\mathcal{A}$ that learns the system parameters and updates the control policies in an online fashion to improve the system performance and guarantee constraint satisfaction for all $t \geq 0$ under any $w_t \in \mathbb{W}$. Solving (1) efficiently with constraints $\mathbb{X}$ and $\mathbb{U}$ is still an open problem even when the model θ_* is known. People reformulate (1) as a tractable problem by limiting the policy class. For example, [15] considers linear controllers with memory, and robust MPC considers piece-wise affine controllers without memory [30]. In this paper, we will consider disturbance-action policy (DAP) control as introduced later in Section 3.

To evaluate the performance of the online algorithm $\mathcal{A}$, besides ensuring safety, we analyze the regret of the algorithm $\mathcal{A}$ while bench-marking it with some optimal safe policy obtained by assuming θ_* to be known. Given the difficulty in designing safe policies even with known θ_*, we consider static linear policy class as our benchmark policies in this paper. But our results can be extended to other

[1] In the supplementary file, we discuss the scenarios where $x_0 \neq 0$. Roughly, if x_0 is sufficiently small such that it admits a safe linear controller, then our algorithm can directly be applied. If x_0 is large, we leverage RMPC in [30] to steer the state to be small enough. For too large x_0, the constrained control can be infeasible.

[2] In some literature, e.g. [2], this property is called $(\sqrt{\kappa}, \gamma)$-strong stability.

[3] Here, Θ_{ini} is symmetric on all directions, which may not be the case in practice. This is not restrictive and only assumed for technical simplicity. What we really need is that $\Theta^{(0)}$ is a compact set containing θ_*.

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

benchmarks such as linear policy with memory and a certain type of RMPC with PWA controllers. We discuss these extensions in the supplementary while leaving more general benchmarks as future work. In particular, we consider $u_t = -Kx_t$ as benchmark. Define

$$\mathcal{K} = \{K : (A_* - B_*K) \text{ is } (\kappa, \gamma) \text{ stable and } \|K\|_2 \le \kappa, x_t \in \mathbb{X}, u_t \in \mathbb{U}, \forall \{w_k \in \mathbb{W}\}_{k\ge 0}\}.$$

Let $J^* = \min_{K\in\mathcal{K}} J(K)$ denote the optimal control cost provided by policy $\mathcal{K}$ when the model is known. We measure the performance of online algorithm $\mathcal{A}$ by policy regret, which is defined as

$$\text{Regret} = \sum_{t=0}^{T-1} l(x_t^{\mathcal{A}}, u_t^{\mathcal{A}}) - TJ^*$$

To make the regret well-defined, we need to assume that $\mathcal{K}$ is not empty. For technical reasons, we will impose a stronger assumption that there exists a strictly safe controller.

Assumption 3. *There exists $K_F \in \mathcal{K}$ and $\epsilon_F = (\epsilon_{F,x}, \epsilon_{F,u}) > 0$ such that K_F is ϵ_F-strictly safe.*

A sufficient condition to verify the existence of K_F is by LMI reformulation. In the supplementary file, we provide one such reformulation.

3 Preliminaries: constrained control with known model

Our adaptive control is built upon disturbance-action policy (DAP) [2]. To introduce our algorithm, we provide a review on DAP and its application to constrained control with *known* model, which will be critical for developing our online algorithm.

Definition 4 (Disturbance-action control (DAP))**.** *Consider memory length $H \ge 1$ and policy parameters $\mathbf{M} = \{M[k]\}_{k=1}^{H}$ for $M[k] \in \mathbb{R}^{m\times n}$, DAP selects $u_t = \sum_{k=1}^{H} M[k]w_{t-k}$, where $w_t = x_{t+1} - A_*x_t - B_*u_t$ can be computed when θ_* is known and define $w_t = 0$ for $t < 0$.*[5]

As in [24], we will work with a convex polytopic constraint set $\mathcal{M}_H$ on admissible $\mathbf{M}$ for technical simplicity and without loss of generality. $\mathcal{M}_H = \{\mathbf{M} : \|M[k]\|_\infty \le 2\sqrt{n}\kappa^2(1-\gamma)^{k-1},\ \forall\, 1 \le k \le H\}$. Notice that u_t is a linear function with policy $\mathbf{M}$. Further, the next proposition shows that x_t can be approximated by $\tilde{x}_t(\mathbf{M}; \theta_*)$, which is an affine function on $\mathbf{M}$.

Proposition 1 ([2])**.** *When implementing time-invariant policy $\mathbf{M}$, we have $x_t = A_*^H x_{t-H} + \tilde{x}_t(\mathbf{M}; \theta_*) = A_*^H x_{t-H} + \sum_{k=1}^{2H} \Phi_k^x(\mathbf{M}; \theta_*) w_{t-k}$, where $\Phi_k^x(\mathbf{M}; \theta_*) = A_*^{k-1}\mathbb{1}_{(k\le H)} + \sum_{i=1}^{H} A_*^{i-1} B_* M[k-i]\mathbb{1}_{(1\le k-i\le H)}$.*

Definition of $f(\mathbf{M}; \theta_*)$. For fixed $\mathbf{M}$, define $f(\mathbf{M}; \theta_*)$ by the expected cost of approximate state and action: $f(\mathbf{M}; \theta_*) = \mathbb{E}_{w_k}\left[l(\tilde{x}_t(\mathbf{M}, \theta_*), u_t(\mathbf{M}))\right]$, which is a convex quadratic function of $\mathbf{M}$.

Guarantee safety of time-invariant DAP by tightening constraints. Define policy-constraint functions $g_i^x(\cdot; \theta_*)$ and $g_j^u(\cdot)$ based on worst-case constraints on the approximate states and actions, i.e., $g_i^x(\mathbf{M}; \theta_*) := \sup_{w_k\in\mathbb{W}} D_{x,i}^\top \tilde{x}_t(\mathbf{M}; \theta_*) = \sum_{k=1}^{2H} \|D_{x,i}^\top \Phi_k^x(\mathbf{M}; \theta_*)\|_1 w_{\max}$ and $g_j^u(\mathbf{M}) := \sup_{w_k\in\mathbb{W}} D_{u,j}^\top u_t(\mathbf{M}) = \sum_{k=1}^{H} \|D_{u,j}^\top M[k]\|_1 w_{\max}$.

Then, it can be shown that the *actual* states satisfy constraints for any admissible disturbances if the *approximate* states satisfy constraints *tightened* by an error term, denoted as $\epsilon_H(H)$, which allows for *approximation errors*. That is, if $\sup_{w_k\in\mathbb{W}} D_{x,i}^\top A_*^H x_{t-H} \le \epsilon_H(H)$, then

$$g_i^x(\mathbf{M}; \theta_*) \le d_{x,i} - \epsilon_H(H) \quad \Rightarrow \quad D_{x,i}^\top x_t \le d_{x,i},\ \forall \{w_k \in \mathbb{W}\}. \tag{2}$$

Similarly, if $g_j^u(\mathbf{M}) \le d_{u,j}$, then $D_{u,j}^\top u_t(\mathbf{M}) \le d_{u,j}$ for all $\{w_k \in \mathbb{W}\}$.[6] Notice that the constraints above describe a polytopic feasible set for policies.

Optimal safe DAP. The optimal safe DAP control $\mathbf{M}_H^* \in \mathcal{M}_H$ can be obtained by solving

$$\min_{\mathbf{M}\in\mathcal{M}_H} f(\mathbf{M}; \theta_*), \quad \text{subject to } g_i^x(\mathbf{M}; \theta_*) \le d_{x,i} - \epsilon_H(H), \forall\, i,\ g_j^u(\mathbf{M}) \le d_{u,j},\ \forall\, j. \tag{3}$$

Further, it has been shown in [24] that with sufficiently large H, $J(\mathbf{M}_H^*) - J^* \le \tilde{O}(1/T)$. Therefore, to reach $o(T)$ regret, it suffices to learn the optimal DAP control when the model is unknown.

[5]DAP is also called the finite-impulse response [15] or affine-disturbance policy [31].

[6]There is no constraint tightening since there is no approximation error on control.

Safety of time-varying DAPs by a slow-variation trick and more constraint tightening. Though time-invariant DAP is sufficient to minimize infinite-horizon averaged cost when the model is known, this paper studies adaptive control with the unknown model, which will update policies based on refined model estimations and calls for safety guarantees of time-varying DAPs.

It is known that even if every $\mathbf{M}_t$ is safe to implement in a time-invariant fashion and even when the model is known, the sequence $\{\mathbf{M}_t\}_{t\geq 0}$ may still violate the state constraints due to policy switching [15, 24]. When the model is known, [24] tackles this challenge by a *slow-variation* trick and establishes the following lemma.

Lemma 1 (Constraint satisfaction of slowly varying DAPs)**.** *If the sequence $\{\mathbf{M}_t\}_{t\geq 0}$ varies slowly, i.e.$\|\mathbf{M}_t - \mathbf{M}_{t-1}\|_F \leq \Delta_M$, where Δ_M is called the policy variation budget, and if each $\mathbf{M}_t$ satisfies*

$$g_i^x(\mathbf{M}_t; \theta_*) \leq d_{x,i} - \epsilon_H(H) - \epsilon_v(\Delta_M, H), \quad g_j^u(\mathbf{M}_t) \leq d_{u,j},$$

where $\epsilon_v(\Delta_M) = O(\Delta_M \sqrt{H})$ accounts for the error caused by ignoring the policy variation, then it is safe to implement the time-varying policies $\{\mathbf{M}_t\}_{t\geq 0}$ when the model is known.

Lemma 1 indicates that a sequence of time-varying DAPs is safe to implement with a known model if the variation between the neighboring policies is small enough, and if every individual DAP is $\epsilon_H(H) + \epsilon_v(\Delta_M, H)$-strictly safe to implement in a time-invariant fashion.

4 Safe online learning for robust constrained-satisfaction control

---

Algorithm 1: Safe online learning-based control by cautious certainty equivalence (CCE)

Input: $\Theta_{\text{ini}}, T^{(1)} \geq 1, T^{(e+1)} = 2T^{(e)}$ for $e \geq 1$. $H^{(e)}$. $\bar{\eta}^{(e)}$, $\Delta_M^{(e)}$, $T_D^{(e)}$ for $e \geq 0$.

Initialize: $\hat{\theta}^{(0)} = \hat{\theta}_{\text{ini}}, r_\theta^{(0)} = r_{\text{ini}}, \Theta^{(0)} = \Theta_{\text{ini}}$. Define $w_t = \hat{w}_t = 0$ for $t < 0$, $t_1^{(0)} = 0$.

for Episode $e = 0, 1, 2, \dots$ **do**

  [Phase 1: safe exploration & exploitation] Compute a polytopic robustly safe policy set $\Omega_\dagger^{(e)} = \Omega(\Theta^{(e)}, H^{(e)}, \bar{\eta}^{(e)}, \Delta_M^{(e)})$ by (5), and compute the cautious certainty equivalent control $\mathbf{M}_\dagger^{(e)} = \arg\min_{\mathbf{M}\in\Omega_\dagger^{(e)}} \mathring{f}(\mathbf{M}; \hat{\theta}^{(e)})$.

  Run *Algorithm 2* to safely transit from $\mathbf{M}_*^{(e-1)}$ to $\mathbf{M}_\dagger^{(e)}$ when $e \geq 1$. Let $t_1^{(e)}$ be the output.

  for $t = t_1^{(e)}, \dots, t_1^{(e)} + T_D^{(e)} - 1$ **do**

    Implement (4) with policy $\mathbf{M}_\dagger^{(e)}$, noise $\eta_t \overset{\text{i.i.d.}}{\sim} \bar{\eta}\mathcal{D}_\eta$, and estimate $\hat{w}_t$ by $\hat{\theta}^{(e)}$.

  [Model Updates] Estimate $\hat{\theta}^{(e+1)}$ by ordinary least square with projection onto Θ_{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 confidence radius $r^{(e+1)} = \tilde{O}(\frac{\sqrt{n^2+nm}}{\sqrt{T_D^{(e)}}\bar{\eta}^{(e)}})$ according to Corollary 1.

  [Phase 2: pure exploitation] Compute a new robustly safe policy set with the updated model and no excitation: $\Omega^{(e)} = \Omega(\Theta^{(e+1)}, H^{(e)}, 0, \Delta_M^{(e)})$. Compute a new cautious certainty equivalence control: $\mathbf{M}_*^{(e)} = \arg\min_{\mathbf{M}\in\Omega^{(e)}} \mathring{f}(\mathbf{M}; \hat{\theta}^{(e+1)})$.

  Run *Algorithm 2* to switch from $\mathbf{M}_\dagger^{(e)}$ to $\mathbf{M}_*^{(e)}$. Set $t_2^{(e)}$ as the output.

  for $t = t_2^{(e)}, \dots, T^{(e+1)} - 1$ **do**

    Implement (4) with policy $\mathbf{M}_\dagger^{(e)}$, noise $\eta_t = 0$ and estimate $\hat{w}_t$ by $\hat{\theta}^{(e+1)}$.

In this section, we elaborate on the different steps of the proposed Algorithm 1. The algorithm is based on cautious certainty-equivalence DAP control. The main difficulty is to ensure constraint satisfaction while allowing exploration for model improvement and achieving low regrets. To achieve this, we divide T stages into episodes,[7] and divide each episode e into two major phases.

- *Phase 1: Safe exploration & exploitation*: During this phase, compute a near-optimal robustly safe control $\mathbf{M}_\dagger^{(e)}$ based on cautious certainty equivalence (CCE) under the estimated model

[7] We consider single trajectory adaptive control and require *no* restarts at the start of each episode.

uncertainty set $\Theta^{(e)} := \mathbb{B}(\hat{\theta}^{(e)}, r^{(e)}) \cap \Theta_{\text{ini}}$ where $\hat{\theta}^{(e)}$ is the estimated model and $r^{(e)}$ is a confidence radius. Implement $u_t = \sum_{k=1}^{H} \mathbf{M}_{\dagger}^{(e)}[k]\hat{w}_{t-k} + \eta_t$ where noise η_t is introduced to excite the system for model estimation and $\hat{w}_{t-k}$ is the approximated disturbance computed using the estimated model $\hat{\theta}^{(e)}$ along with the measured states and control inputs. Note that $\mathbf{M}_{\dagger}^{(e)}$ is selected to guarantee robust constraint satisfaction for all $\theta \in \Theta^{(e)}$ under this controller and to allow safe transitions from the previous policy.

- At the end of the phase, use the new data to update the model estimation by least square estimator and refine the confidence radius by Corollary 1. Update the model uncertainty set to be $\Theta^{(e+1)} := \mathbb{B}(\hat{\theta}^{(e+1)}, r^{(e+1)}) \cap \Theta_{\text{ini}}$.

- *Phase 2: Pure exploitation*: This phase is similar to Phase 1 but uses the new model uncertainty set $\Theta^{(e+1)}$ and removes the excitation noises η_t, i.e., $\eta_t = 0$. The CCE controller computed for this phase is denoted as $\mathbf{M}_{*}^{(e)}$.

As discussed in Section 3, though policies $\mathbf{M}_{\dagger}^{(e)}$ and $\mathbf{M}_{*}^{(e)}$ are designed to be safe when being implemented in a time-invariant fashion, switching between them may violate the safety constraints. Since there are two transitions in our online algorithm, to ensure safety during the transition, two safe transition phases are introduced, one for the transition from $\mathbf{M}_{*}^{(e-1)}$ to $\mathbf{M}_{\dagger}^{(e)}$ and one for the transition from $\mathbf{M}_{\dagger}^{(e)}$ to $\mathbf{M}_{*}^{(e)}$ respectively. To allow such safe transitions, we require $\mathbf{M}_{\dagger}^{(e)}$ and $\mathbf{M}_{*}^{(e)}$ to satisfy more conservative constraints and utilize the slow-variation trick reviewed in Section 3. In the following, we will provide more details on the cautious certainty equivalence and safe transition.

Cautious Certainty Equivalence with Robust Constraint Satisfaction. For simplicity of exposition, we drop the index of episode (e) in the notations of this subsection without causing any confusion. When the true model is unknown and only an uncertainty set $\Theta = \mathbb{B}(\hat{\theta}, r) \cap \Theta_{\text{ini}}$ is known to contain the true model, we implement DAP with approximated disturbances computed by estimated model $\hat{\theta}$ and inject an excitation noise to encourage exploration for model estimation updates, i.e.,

$$u_t = \sum_{k=1}^{H} M[k]\hat{w}_{t-k} + \eta_t, \text{ where } \|\eta_t\|_\infty \leq \bar{\eta}, \text{ and } \hat{w}_t = \Pi_{\mathbb{W}}(x_{t+1} - \hat{A}x_t - \hat{B}u_t). \tag{4}$$

Here, the projection onto $\mathbb{W}$ is important for robust constraint satisfaction but it introduces nonlinearity into the policy. The excitation η_t can follow any distribution i.i.d. such that $\mathbb{E}\,\eta_t = 0$, $\|\eta_t\|_\infty \leq \bar{\eta}$, and $\eta_t/\bar{\eta}$ satisfies (s_η, p_η) anti-concentration for some s_η, p_η. For example, η_t can follow truncated Gaussian or uniform distribution. Lastly, notice that implementing (4) requires to specify $\mathbf{M}, \bar{\eta}, \hat{\theta}$.

To ensure the safety of (4) without knowing the true model, we rely on robust constraint satisfaction, which requires safe implementation on all possible models in the uncertainty set Θ. This can be achieved by tightening constraints with an error term depending on the size of the uncertainty set, denoted by $\epsilon_\theta(r)$. Besides, to ensure safe exploration with the excitation noises η_t, we have to further tighten the constraints by error term $\epsilon_\eta(\bar{\eta})$. In summary, we construct a robustly safe policy set below,

$$\begin{aligned}\Omega(\Theta, H, \bar{\eta}, \Delta_M) = \{\mathbf{M} \in \mathcal{M}_H : g_i^x(\mathbf{M}; \hat{\theta}) \leq d_{x,i} - \epsilon_\theta(r) - \epsilon_{\eta,x}(\bar{\eta}) - \epsilon_H(H) - \epsilon_v(\Delta_M),\ \forall\, i \\ g_j^u(\mathbf{M}) \leq d_{u,j} - \epsilon_{\eta,x}(\bar{\eta}),\ \forall\, j\}.\end{aligned} \tag{5}$$ [8]

In (5), the error term $\epsilon_H(H)$ is needed even when the model is known (see 2), and $\epsilon_v(\Delta_M)$ allows for safe policy variation with variation budget Δ_M (see Lemma 1), which is necessary to ensure safety during policy updates (switching). The major challenge is to construct $\epsilon_\theta(r)$ and $\epsilon_\eta(\bar{\eta})$ so that they are large enough to guarantee robust safety despite model estimation errors and excitation noises, but not too large to degrade performances or even cause empty policy sets. The construction of $\epsilon_\theta(r)$ and $\epsilon_\eta(\bar{\eta})$ are very technical and rely on the perturbation results of our systems. Therefore, we defer the details of construction to the appendix.

We define *cautious certainty equivalent* control (CCE) as the solution to the following optimization, where the cost function is based on estimated model $\hat{\theta}$ and constraints guarantee robust safety on Θ.

$$\min_{\mathbf{M}} f(\mathbf{M}; \hat{\theta}), \quad \text{subject to } \mathbf{M} \in \Omega(\Theta, H, \bar{\eta}, \Delta_M). \tag{6}$$

Safe Transitions of Policies. Suppose we implement (4) with robustly safe policy $\mathbf{M} \in \Omega(\Theta, H, \bar{\eta}, \Delta_M)$ with excitation level $\bar{\eta}$, and estimated model $\hat{\theta} \in \Theta$ before stage t_0, we want to switch to a new robustly safe policy $\mathbf{M}' \in \Omega' = \Omega(\Theta', H', \bar{\eta}', \Delta'_M)$ with new $\bar{\eta}'$, and new $\hat{\theta}' \in \Theta'$ while ensuring constraint satisfaction all the time. As discussed in Section 3, switching policies may lead to constraint violation even when the model is known. Here, we suffer extra challenges due to that the estimated models, excitation levels, H, and Δ_M are changing as well. To address these challenges, built upon the slow variation trick reviewed in Section 3, we provide a way, as shown in Algorithm 2, to safely transit all updated parameters at the same time. We briefly explain the rationale behind our algorithm below.

Firstly, we note that our choices of W_1 and W_2 in Algorithm 2 ensure slow enough variation of the policies. Secondly, Algorithm 2 adopts an auxiliary policy $\mathbf{M}_{\text{mid}} \in \Omega \cap \Omega'$ to serve as a middleground when transiting from $\mathbf{M}$ to $\mathbf{M}'$. This guarantees $\mathbf{M}_t \in \Omega$ in Step 1 and $\mathbf{M}_t \in \Omega'$ in Step 2. To see this, notice that $\mathbf{M}_t$ in Step 1 is a convex combination of $\mathbf{M}$ and $\mathbf{M}_{\text{ini}}$, which both belong to Ω. The same applies to Step 2. Therefore, every $\mathbf{M}_t$ is in some robustly safe policy sets, which ensures constraint satisfaction. In practice, it is desirable to choose $\mathbf{M}_{\text{mid}}$ that is close to $\mathbf{M}$ and $\mathbf{M}'$. Thirdly, in Step 1, we use the smaller excitation level $\bar{\eta}_{\min}$ and the better estimated model $\hat{\theta}_{\min}$. This is because the approximated disturbances and excitation noises used in Step 1 will affect the state constraints in Step 2. Thus, the disturbance approximation errors and excitation levels in Step 1 should be small enough for both Ω and Ω'. We can further show that the effects of the history are dominated by the recent H' steps, so we let $W_1 \geq H'$ to provide small history errors for Step 2.

Remark 2. *I) In Line 8 of Algorithm 1, we only use a segment of data from the current episode. This is for the simplicity of theoretical analysis. In practice, one should use all the data collected so far to construct a better estimation. II) Notice that $\Omega(\Theta, H, \bar{\eta}, \Delta_M)$ defines a polytopic set on $\mathbf{M}$ and $f(\mathbf{M}; \hat{\theta})$ is a convex quadratic function. Then, solving the CCE controller only requires solving a convex quadratic program with linear constraints, which admits polynomial-time solvers. III)We also note that Algorithm 2 is not the unique way to guarantee safe transitions. Other methods may also work, e.g., model predictive control.*

Algorithm 2: Safe Transition Algorithm

Input: $\mathbf{M}_{t_0-1} = \mathbf{M} \in \Omega = \Omega(\Theta, H, \bar{\eta}, \Delta_M)$, new policy $\mathbf{M}' \in \Omega' = \Omega(\Theta', H', \bar{\eta}', \Delta'_M)$, $H \leq H'$.

Set $\bar{\eta}_{\min} = \min(\bar{\eta}, \bar{\eta}')$, $\hat{\theta}_{\min} = \hat{\theta}\mathbb{1}_{(r_\theta \leq r'_\theta)} + \hat{\theta}'\mathbb{1}_{(r_\theta > r'_\theta)}$. Find an auxiliary policy $\mathbf{M}_{\text{mid}} \in \Omega \cap \Omega'$.

Step 1: safe transition from $\mathbf{M}$ to $\mathbf{M}_{mid}$. 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**

  Slowly update $\mathbf{M}_t$ from $\mathbf{M}$ towards $\mathbf{M}_{mid}$ by $\mathbf{M}_t = \mathbf{M}_{t-1} + \frac{1}{W_1}(\mathbf{M}_{\text{mid}} - \mathbf{M})$.

  Implement (4) with policy $\mathbf{M}_t$, noise $\eta_t \overset{\text{i.i.d.}}{\sim} \bar{\eta}_{\min}\mathcal{D}_\eta$, and estimate $\hat{w}_t$ by $\hat{\theta}_{\min}$.

Step 2: safe transition from $\mathbf{M}_{mid}$ to $\mathbf{M}'$. Define $W_2 = \max(\lceil \frac{\|\mathbf{M}' - \mathbf{M}_{\text{mid}}\|_F}{\Delta'_M} \rceil)$.

for $t = t_0 + W_{s_1}, \ldots, t_0 + W_1 + W_2 - 1$ **do**

  Slowly update $\mathbf{M}_t$ from $\mathbf{M}_{\text{mid}}$ towards $\mathbf{M}'$ by $\mathbf{M}_t = \mathbf{M}_{t-1} + \frac{1}{W_2}(\mathbf{M}' - \mathbf{M}_{\text{mid}})$.

  Implement (4) with policy $\mathbf{M}_t$, noise $\eta_t \overset{\text{i.i.d.}}{\sim} \bar{\eta}'\mathcal{D}_\eta$, and estimate $\hat{w}_t$ by $\hat{\theta}'$.

Output: $t_1 = t_0 + W_1 + W_2$

---

5 Theoretical analysis

5.1 Estimation error decay rate. Here we provide a decay rate for our model estimation error. The major technical difficulty comes from the nonlinearity of control policies caused by the projection in (4). To address this issue, we provide a general estimation error bound for general either linear or nonlinear policies $u_t = \pi_t(x_0, \{w_k, \eta_k\}_{k=0}^{t-1}) + \eta_t$ with bounded states and actions by taking advantage of the anti-concentration properties of w_t and η_t. This generalizes the existing results on linear policies [12]. Since nonlinear policies are commonly used for constrained linear optimal control, our result can be used broadly for other learning-based control algorithms.[9]

Theorem 1 (General estimation error bound). *Consider $x_{t+1} = A_* x_t + B_* u_t + w_t$, where $u_t = \pi_t(x_0, \{w_k, \eta_k\}_{k=0}^{t-1}) + \eta_t$, and $\|\eta_t\|_\infty \leq \bar{\eta}$. Suppose $\eta_t / \bar{\eta}$ are i.i.d., zero mean, with an (s_η, p_η)*

[9] There are also recent results on estimation error rates for general nonlinear systems such as [18], but our problem has special structures and thus enjoys better rates than the general nonlinear system case.

anti-concentration property, and independent from $\{w_t\}$*. Suppose* $\|x_t\|_2 \leq b_x$*,* $\|u_t\|_2 \leq b_u$ *for some* b_x, b_u *for all* t*. Define* $\tilde{\theta} = \min_\theta \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|_2^2$*. For any* $0 < \delta < 1/3$*, if* $T \geq \frac{10}{p_z^2}\left(\log(1/\delta) + 2(m+n)\log(10/p_z) + 2(n+m)\log(\sqrt{b_x^2+b_u^2}/s_z)\right)$*, then*

$$\|\tilde{\theta} - \theta_*\|_2 \leq \frac{90\sigma_{sub}}{p_z}\frac{\sqrt{n + (n+m)\log(10/p_z) + 2(n+m)\log(\sqrt{b_x^2+b_u^2}/s_z) + \log(1/\delta)}}{\sqrt{T}s_z}$$

holds with probability $1-3\delta$*, where* $p_z = \min(p_w, p_\eta)$*,* $s_z = \min(s_w/4, \frac{\sqrt{3}}{2}s_\eta\bar{\eta}, \frac{s_w s_\eta}{4b_u}\bar{\eta})$*.*

Theorem 1 shows a decay rate $\tilde{O}(\frac{\sqrt{m+n}}{\bar{\eta}\sqrt{T}})$ as $\bar{\eta} \to 0$ for the model estimation error without assuming linear policies. Interestingly, the rate is the same with the decay rate for linear policies in the literature with respect to the number of samples T, excitation level $\bar{\eta}$, and dimension $n+m$ [12, 15].

Corollary 1 (Estimation errors in Algorithm 1)**.** *Select* $\bar{\eta}^{(e)}$ *such that* $\bar{\eta}^{(e-1)}/2 \leq \bar{\eta}^{(e)} \leq \bar{\eta}^{(e-1)} \leq \min(\frac{s_w}{2\sqrt{3}s_\eta}, \frac{1}{\sqrt{3}s_\eta}, \frac{2u_{\max}}{s_w s_\eta})$ *for* $e \geq 1$*. Suppose* $t_1^{(e)} + T_D^{(e)} \leq T^{(e+1)}$ *for* $e \geq 1$*. For any* $0 < p < 1$*, with a sufficiently large* $T^{(1)} \geq \tilde{O}(m+n)$*, then* $\|\hat{\theta}^{(e)} - \theta_*\|_F \leq O(\frac{(n+\sqrt{mn})\sqrt{\log(mn/\bar{\eta}^{(e-1)})+\log(e)}}{\sqrt{T_D^{(e-1)}}\bar{\eta}^{(e-1)}})$ *with probability at least* $1 - \frac{p}{2e^2}$ *for any* $e \geq 1$*.*

In Corollary 1, we consider $\|\cdot\|_F$ norm because we project $\tilde{\theta}^{(e)}$ onto Θ_{ini} in Algorithm 1 and the projection of matrices is non-expansive on $\|\cdot\|_F$. Due to this change, the estimation error bound has an additional $\sqrt{n}$ factor.

5.2. Feasibility and constraint satisfaction. Now we provide feasibility and constraint satisfaction guarantees under proper algorithm inputs. For simplicity, we assume r_{ini} is small enough, which is commonly assumed in the literature [11] and can be replaced by implementing a safe exploration policy for sufficiently long to reduce the model estimation error. We discuss more on how to remove this assumption in the appendix.

Assumption 4 (Assumption on r_{ini})**.** r_{ini} *is small enough such that* $\epsilon_\theta(r_{ini}) \leq \frac{\epsilon_{F,x}}{4}$*.*

Theorem 2 (General conditions for feasibility)**.** *The policies computed by Algorithm 1 and Algorithm 2 are well-defined for all stages (feasibility) if the following conditions hold.*

$$T_D^{(e)} \geq T_D^{(e-1)}, \bar{\eta}^{(e)} \leq \bar{\eta}^{(e-1)}, H^{(e)} \geq H^{(e-1)}, \Delta_M^{(e)} \leq \sqrt{\frac{H^{(e-1)}}{H^{(e)}}}\Delta_M^{(e-1)}, r^{(e)} \leq r^{(e-1)}, \forall\, e \geq 1, \quad \text{(I)}$$

$$\epsilon_H(H^{(0)}) + \epsilon_{P,x}(H^{(0)}) + \epsilon_v(\Delta_M^{(0)}, H^{(0)}) + \epsilon_{\eta,x}(\bar{\eta}^{(0)}) \leq \epsilon_{F,x}/2 \quad \text{(II)}$$

$$\epsilon_{\eta,u}(\bar{\eta}^{(0)}) + \epsilon_{P,u}(H^{(0)}) \leq \epsilon_{F,u}, \quad \text{(III)}$$

and $H^{(0)} \geq \log(2\kappa)/\log((1-\gamma)^{-1})$*, where* $\epsilon_H(H) = c_1(1-\gamma)^H, \epsilon_{\eta,x}(\bar{\eta}) = c_2\sqrt{m}\bar{\eta}, \epsilon_{\eta,u}(\bar{\eta}) = c_3\bar{\eta}, \epsilon_\theta(r) = c_4\sqrt{mn}r, \epsilon_v(\Delta_M, H) = c_5\sqrt{mnH}\Delta_M, \epsilon_{P,x}(H) = c_6\sqrt{n}(1-\gamma)^H, \epsilon_{P,u}(H) = c_7\sqrt{n}(1-\gamma)^H$*, and* $c_1, \ldots, c_7$ *are poly*$(\|D_x\|_\infty, \|D_u\|_\infty, \kappa, \kappa_B, \gamma^{-1}, w_{\max}, x_{\max}, u_{\max})$*.*

Conditions (I) in Theorem 2 requires monotonicity of algorithm parameters, allowing us to verify the feasibility based on the initial conditions. Condition (II) and (III) require large enough $H^{(0)}$, small enough $\bar{\eta}^{(0)}$ and $\Delta_M^{(0)}$. Notice that $r^{(1)} \leq r^{(0)} = r_{\text{ini}}$ in (I) requires $\tilde{O}(\sqrt{n^2+nm}(\sqrt{T_D^{(e-1)}}\bar{\eta}^{(e-1)})^{-1}) \leq r_{\text{ini}}$, which suggests that $\bar{\eta}^{(0)}$ should not be too small. This reflects the trade-off between exploration and safety. Finally, $\epsilon_P = (\epsilon_{P,x}, \epsilon_{P,u})$ are introduced due to our proof techniques.

Theorem 3 (Constraint Satisfaction)**.** *Under the conditions in Theorem 2, Corollary 1, and suppose* $t_2^{(e)} \leq T^{(e+1)}$*, then* $u_t \in \mathbb{U}$ *for all* $t \geq 0$ *w.p. 1. Further,* $x_t \in \mathbb{X}$ *holds for all* $t \geq 0$ *with probability at least* $1-p$*, where* p *is defined in Corollary 1.*

The w.p. 1 control constraint satisfaction is ensured by the projection onto $\mathbb{W}$ in (4). Besides, we can show that the state constraints are satisfied if the true model is inside the confidence sets $\Theta^{(e)}$ for all $e \geq 0$, whose probability is at least $1-p$ by Corollary 1.

5.3. Regret guarantees. Next, we provide a $\tilde{O}(T^{2/3})$ regret bound while guaranteeing feasibility and constraint satisfaction. Further, we explain the reasons behind the pure exploitation phase.

Theorem 4 (Regret bound)**.** *Consider any* $0 < p < 1/2$. *Let* $T_D^{(e)} = (T^{(e+1)} - T^{(e)})^{2/3}$, $T^{(1)} \geq \tilde{O}((\sqrt{nm}+n)^3)$. *Set* $\Delta_M^{(e)} = O(\frac{\epsilon_F^x}{\sqrt{mnH^{(0)}}}(T^{(e+1)})^{-1/3})$, $H^{(e)} \geq O(\log(\max(T^{(e+1)}, \frac{\sqrt{n}}{\min(\epsilon_F)})))$, $\bar{\eta}^{(e)} = \eta_{\max} \leq \min\left(O(\frac{\epsilon_F^x}{\sqrt{m}}), O(\epsilon_F^u), \frac{s_w}{2\sqrt{3}s_\eta}, \frac{1}{\sqrt{3}s_\eta}, \frac{2u_{\max}}{s_w s_\eta}\right)$. *Then Algorithm 1 is feasible and satisfies* $\{u_t \in \mathbb{U}\}_{t\geq 0}$ *a.s. and* $\{x_t \in \mathbb{X}\}_{t \geq 0}$ *w.p.* $1-p$. *Further, with probability at least* $1-2p$,

$$\text{Regret} \leq \tilde{O}((n^2m^{1.5} + n^{2.5}m)\sqrt{mn+k_c}T^{2/3})$$

Though our regret bound $\tilde{O}(T^{2/3})$ is worse than the $\tilde{O}(\sqrt{T})$ regret bound for *unconstrained* LQR, it is the same with the robust learning of unconstrained LQR (see [12]). This motivates future work on fundamental lower bounds of learning-based control with safety/robustness guarantees.

Proof ideas. For illustrational purposes, we only consider $H^{(e)} = O(\log(T))$ below. Proofs for the general case are provided in the supplementary file. The proof heavily relies on the following perturbation error bound, which will be formally proved in the supplementary.

Lemma 2 (Cost error bound for cautious certainty equivalence)**.** *Consider model uncertainty set* $\Theta = \{\theta : \|\theta - \hat{\theta}\|_F \leq r\}$ *containing the true model* θ_**, exploration level* $\bar{\eta}$*, and variation budget* Δ_M*. Consider a large enough* H*. Consider a cautious certainty equivalent control* $\mathbf{M}_{cce} = \arg\min_{\mathbf{M}\in\Omega(\hat{\theta},\epsilon,H)} f(\mathbf{M};\hat{\theta})$ *for* $\epsilon = \epsilon_c(H, \bar{\eta}, r) + \epsilon_v(\Delta_M, H)$*. Then,*

$$f(\mathbf{M}_{cce};\theta_*) - J^* \leq \tilde{O}(r + \Delta_M + \bar{\eta})$$

where $\tilde{O}(\cdot)$ *hides polynomial factors of problem dimensions for illustration purposes.*

With parameters provided in Theorem 4, we can show that Algorithm 2 only takes $\tilde{O}(T^{1/3})$ stages, and CCE with active exploration only takes $\tilde{O}(T^{2/3})$ stages. Further, we can show that single-stage regret is bounded. Hence, implementing Algorithm 2 and Phase 1 of Algorithm 1 contribute $\tilde{O}(T^{1/3})$ and $\tilde{O}(T^{2/3})$ regrets respectively. The remaining part is to bound the regret by pure exploitation. Roughly speaking, by Lemma 2, the regret of pure exploitation in Algorithm 1 at episode e can be bounded by $\tilde{O}(T^{(e)}(r^{(e+1)} + \Delta_M^{(e)}))$, since we consider no exploration noises. According to Corollary 1, from the active exploration phase, our estimation error is updated to $r^{(e+1)} = \tilde{O}(\frac{1}{T_D^{(e)}\eta_{\max}}) = \tilde{O}(\frac{1}{(T^{(e)})^{1/3}})$. Since we select $\Delta_M^{(e)} = \tilde{O}(\frac{1}{(T^{(e)})^{1/3}})$, we are able to prove $\tilde{O}((T^{(e)})^{2/3})$ regret at episode e, which sums up to $\tilde{O}(T^{2/3})$ regret in total.

More discussions on the choices $\eta^{(e)}$ and the pure exploitation phase. Our Algorithm 1 includes a pure exploitation phase with no excitation noises and an active exploration phase with a constant $\bar{\eta}^{(e)}$ for $O((T^{(e)})^{1/3})$ to achieve $\tilde{O}(T^{2/3})$ regret. However, in most literature that considers certainty-equivalence-based learning, the exploration level $\bar{\eta}^{(e)}$ decreases with e and there is no full-exploitation phase. In fact, our first attempt when designing algorithms also considered decreasing $\bar{\eta}^{(e)}$ and no exploitation phase, however, such design can only achieve $\tilde{O}(T^{3/4})$ regret, which is worse than $\tilde{O}(T^{2/3})$. Intuitive explanations are provided below. Suppose we implement CCE with exploration level $\bar{\eta}^{(e)}$ throughout episode e, by Lemma 2, the regret at episode e is roughly $\tilde{O}(T^{(e)}(\bar{\eta}^{(e)} + r^{(e)})$ (we ignore $\Delta_M^{(e)}$ here for simplicity). By Corollary 1, $r^{(e)} = \tilde{O}(\frac{1}{\sqrt{T^{(e-1)}}\bar{\eta}^{(e-1)}})$ since our exploration phase's length is $T_D^{(e-1)} = O(T^{(e-1)})$. Now, by summing the regret over e and reorganizing terms, we have $\sum_e(\frac{1}{\sqrt{T^{(e-1)}}\bar{\eta}^{(e-1)}} + \bar{\eta}^{(e)})T^{(e)} \approx \sum_e(\frac{1}{\sqrt{T^{(e)}}\bar{\eta}^{(e)}} + \bar{\eta}^{(e)})T^{(e)}$. To minimize the regret in each episode, we select $\frac{1}{\sqrt{T^{(e)}}\bar{\eta}^{(e)}} = \bar{\eta}^{(e)}$, which leads to $\bar{\eta}^{(e)} = (T^{(e)})^{-1/4}$, and a regret bound $\tilde{O}(T^{3/4})$.

Compared with monotonically decreasing $\bar{\eta}^{(e)}$, our algorithm suffers slightly larger stage regret during the active exploration phase because our $\bar{\eta}^{(e)} \not\to 0$. Nevertheless, our active exploration only takes a short period ($O((T^{(e)})^{2/3}$ stages), and by constantly refining the models and reducing $\Delta_M^{(e)}$, the performance during the active exploration phase still improves over time and the regret only constitutes a small part to the overall regret.

Appendices

Roadmap

In this supplementary file, we include the proofs for the theoretical results and the discussions and extensions of the paper.

- Appendix A provides a list of notations and definitions to be used in the appendices.
- Appendix B.1 supplements the discussions in Section 4 and provides formal definitions of the error terms used in the robustly safe policy set (5).
- Appendix C provides the proofs of our estimation error bounds in Theorem 1 and Corollary 1.
- Appendix D provides proofs of feasibility (Theorem 2) and constraint satisfaction (Theorem 3). Appendix D.3 discusses how to remove Assumption 4.
- Appendix E proves our regret bound in Theorem 4.
- Appendix F discusses how to generalize our benchmark policy classes to include linear dynamical policies considered in [15] and one version of robust model predictive control proposed in [30]. Appendix F also discusses how to handle non-zero initial value $x_0 \neq 0$.
- Appendix G contains a list of proofs for the technical lemmas used in Appendix B.1 - F.

A Notations and Definitions

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 $X \leq Y$ 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 $x \leq y$ is $(y-x)_i \geq 0$ for $1 \leq i \leq n$, i.e. x is smaller than y elementwise. Consider a σ-algebra $\mathcal{F}_t$ and a random vector $z_t \in \mathbb{R}^n$, we write $z_t \in \mathcal{F}_t$ if the random vector z_t is measurable in $\mathcal{F}_t$. We let I_n denote the identity matrix in $\mathbb{R}^{n\times n}$.

Define $z_{\max} = \sqrt{x_{\max}^2 + u_{\max}^2}$. Further, we let $\hat{\theta}_t$ denote the estimated model used to approximate $\hat{w}_t$. We use "a.s." as an abbreviation for "almost surely". We use "w.p." as an abbreviation for "with probability".

B DAP with Model Uncertainties and Important Error Terms

In this appendix, we provide useful lemmas when implementing DAP with model uncertainties including formal definitions of error terms used in our definition of robustly safe policy set (5). This appendix supplements our discussions in **Cautious Certainty Equivalence with Robust Constraint Satisfaction** in Section 4 as well as Section 3. This appendix also contains useful lemmas for our proofs of the theoretical results in Section 5.

In the following, we first provide a note on how to handle time-varying memory lengths of DAP policies. This will be useful for our theoretical analysis since Algorithm 1 considers time-varying DAP memory length $H^{(e)}$. Then, we establish a state representation when implementing DAP with model uncertainties. This is crucial when we discuss the impact of model uncertainties and excitation noises. Last but not least, we formally define error terms used in our robustly safe policy set (5), i.e. $\epsilon_\theta(r)$, $\epsilon_{\eta,x}(\bar{\eta})$, $\epsilon_{\eta,u}(\bar{\eta})$, $\epsilon_H(H)$, $\epsilon_v(\Delta_M, H)$. Most of the technical proofs are deferred to Appendix G.

B.1 DAP with time-varying memory lengths

Consider policy $\mathbf{M}_1 \in \mathcal{M}_{H_1}$ and $\mathbf{M}_2 \in \mathcal{M}_{H_2}$ for $H_1 < H_2$ as an example. Notice that set $\mathcal{M}_{H_1}$ can be viewed a subset of $\mathcal{M}_{H_2}$ by defining $\mathbf{M}_1$ with an increased memory length in the following way:

$$M_1[k] = \begin{cases} M_1[k], & \text{if } 1 \leq k \leq H_1, \\ 0, & \text{if } H_1 + 1 \leq k \leq H_2. \end{cases}$$

and notice that the policy defined above is contained by $\mathcal{M}_{H_2}$. We will abuse the notation and still use $\mathbf{M}_1$ to denote the increased-memory-length verion of $\mathbf{M}_1$. Based on our discussion above, we can analyze both $\mathbf{M}_1$ and $\mathbf{M}_2$ in the set $\mathcal{M}_{H_2}$.

Consequently, when we consider a sequence of policies $\{\mathbf{M}_t \in \mathcal{M}_{H_t}\}_{t\geq 0}$ with non-decreasing memory lengths $\{H_t\}_{t\geq 0}$. At each time t, we can conduct all the theoretical analysis in the set $\mathcal{M}_{H_t}$.

B.2 State Representation Lemma

Firstly, we establish a formula for states x_t when implementing DAP with model uncertainties, i.e., (4). This is a generalization of Proposition 1 which considers a known model.

Lemma 3 (State representation under time-varying DAP with model uncertainties)**.** *Consider a sequence of time-varying DAP policies $\{\mathbf{M}_t\}_{t\geq 0}$, where $\mathbf{M}_t \in \mathcal{M}_{H_t}$ and $\{H_t\}_{t\geq 0}$ is non-decreasing. Consider the implementation of DAP policies $\{\mathbf{M}_t\}_{t\geq 0}$ with disturbance $\hat{w}_t$ estimated by time-varying estimated models $\hat{\theta}_t$ and excitation noises with time-varying excitation levels $\|\eta_t\|_\infty \leq \bar{\eta}_t$ as detailed below.*

$$u_t = \sum_{t=1}^{H_t} M_t[k]\hat{w}_{t-k} + \eta_t, \quad \hat{w}_t = \Pi_{\mathbb{W}}(x_{t+1} - \hat{\theta}_t z_t), \quad \|\eta_t\|_\infty \leq \bar{\eta}_t, \quad t \geq 0. \tag{7}$$

Suppose the true system has parameter θ_*, then, we can represent the state x_t by

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

We introduce the following notations that will be helpful when handling time-varying DAP policies.

$$\tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t};\theta) = A^{k-1}\mathbb{1}_{(k\leq H_t)} + \sum_{i=1}^{H_t} A^{i-1} B M_{t-i}[k-i]\mathbb{1}_{(1\leq k-i\leq H_t)}, \quad \forall\, 1\leq k\leq 2H_t, \tag{8}$$

$$\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1};\theta) = \sup_{\hat{w}_k\in\mathbb{W}} D_{x,i}^\top \sum_{k=1}^{2H_t} \tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t-1};\theta)\hat{w}_{t-k} = \sum_{k=1}^{2H_t} \|D_{x,i}^\top \tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t-1};\theta)\|_1 w_{\max}, \tag{9}$$

where $1\leq i\leq k_x$ and we define $\mathbf{M}_t = \mathbf{M}_0$ for $t\leq 0$ for notational simplicity.

Notice that when $H_t = H$, $\mathbf{M}_t = \mathbf{M}$, and $\theta_* = \theta$, we have $\tilde{\Phi}_k^x(\mathbf{M}_{t-H_t:t-1};\theta) = \Phi_k^x(\mathbf{M};\theta_*)$ and $\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1};\theta) = g_i^x(\mathbf{M};\theta_*)$, where $\Phi_k^x(\mathbf{M};\theta_*)$ and $g_i^x(\mathbf{M};\theta_*)$ are defined in Section 3 for a time-invariant policy with a known model.

Next, we provide an upper bound of $D_{x,i}^\top x_t$ based on Lemma 3. This upper bound will be helpful when proving state constraint satisfaction.

Corollary 2 (State constraint decomposition)**.** *Under the conditions in Lemma 3, we can provide an upper bound on $D_{x,i}^\top x_t$ as follows.*

$$\begin{aligned} D_{x,i}^\top x_t \leq\; & \underbrace{g_i^x(\mathbf{M}_t;\hat{\theta}_t^g)}_{\text{estimated state constraint function}} + \underbrace{(g_i^x(\mathbf{M}_t;\theta_*) - g_i^x(\mathbf{M}_t;\hat{\theta}_t^g)) + \sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})}_{\text{model estimation errors}} \\ & + \underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\text{excitation noises}} + \underbrace{D_{x,i}^\top A_*^{H_t} x_{t-H_t}}_{\text{history truncation errors}} + \underbrace{(\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1};\theta_*) - g_i^x(\mathbf{M}_t;\theta_*))}_{\text{policy variation errors}}, \end{aligned}$$

where the upper bound consists of an estimated state constraint function, a term caused by model estimation errors, a term caused by excitation noises, a term caused by truncation at H_t-step history, and a term caused by policy variation; $\hat{\theta}_t^g$ is an estimated model used to approximate the state constraint function, and we allow $\hat{\theta}_t^g \neq \hat{\theta}_t$ for generality.

B.3 Definitions of Error Terms

Definition of $\epsilon_\theta(r)$.

Lemma 4 (Definition of $\epsilon_\theta(r)$)**.** *Consider the policies* (7) *defined in Lemma 3 and suppose the conditions of Lemma 3 hold. 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*

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

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 $\epsilon_\theta(r) = \epsilon_{\hat{\theta}}(r) + \epsilon_{\hat{w}}(r) = O(\sqrt{mn}r)$.

To prove Lemma 4, we establish the following lemma.

Lemma 5 (Disturbance estimation bound)**.** *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,

$$\|w_t - \hat{w}_t\|_2 \leq \|x_{t+1} - \theta_* z_t - (x_{t+1} - \hat{\theta} z_t)\|_2 = \|(\hat{\theta} - \theta_*) z_t\|_2 \leq b_z r.$$

□

The proof of Lemma 4 is deferred to Appendix G.

Definition of $\epsilon_\eta(\bar{\eta})$

Lemma 6 (Definition of $\epsilon_\eta(\bar{\eta})$)**.** *Consider the policies* (7) *defined in Lemma 3 and suppose the conditions of Lemma 3 hold. For a fixed t, suppose $\|\eta_t\|_\infty \leq \bar{\eta}$ for all $0 \leq k \leq H_t$. Then,*

$$\underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\textit{excitation noises' impact on states}} \leq \epsilon_{\eta,x}(\bar{\eta}), \qquad \underbrace{D_{u,j}^\top \eta_t}_{\textit{excitation noises' impact on actions}} \leq \epsilon_{\eta,u}(\bar{\eta}),$$

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 $\epsilon_\eta = (\epsilon_{\eta,x}, \epsilon_{\eta,u})$.

Proof.

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

□

Definition of $\epsilon_H(H)$ The error term $\epsilon_H(H)$ has been introduced in [24] for the known-model case. 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*

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

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 [24] for the known-model case. Here, we slightly improve its dependence on the problem dimensions and the memory length and include our proof in Appendix G.

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*

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

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

C Estimation Error Bounds

In this appendix, we provide proofs of Theorem 1 and Corollary 1.

C.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 [45]. The general error bound and the definition of BMSB are included below for completeness. In the literature [12, 15], 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. (see (13)). More rigorous discussions are provided below.

Definition 5 (Block Martingale Small-Ball (BMSB) (Definition 2.1 [45]))**.** *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 [45])**.** *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

$$\tilde{\theta} = \underset{\theta\in\mathbb{R}^{n\times d}}{\arg\min} \sum_{t=1}^T \|Y_t - \theta X_t\|_2^2. \tag{10}$$

Then if

$$T \geq \frac{10k}{p^2}\left(\log(\frac{1}{\delta}) + 2d\log(10/p) + \log\det(\bar{\Gamma}\Gamma_{sb}^{-1})\right), \tag{11}$$

we have

$$\|\tilde{\theta} - \theta_*\|_2 \leq \frac{90\sigma_{sub}}{p}\sqrt{\frac{n + d\log(10/p) + \log\det(\bar{\Gamma}\Gamma_{sb}^{-1}) + \log(1/\delta)}{T\upsilon_{\min}(\Gamma_{sb})}} \tag{12}$$

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

Next, we will present a proof for our Theorem 1.

Proof of Theorem 1. To use Theorem 5, we need to verify the three conditions.

Condition 1 is straightforward. $x_{t+1} = \theta_* z_t + w_t$, and $w_t \mid \mathcal{F}_t$ is w_t which is mean 0 and σ_{sub}^2-sub-Gaussian by Assumption 2.

Condition 3 is also straightforward. Notice that

$$\upsilon_{\max}(z_t z_t^\top) \leq \text{trace}(z_t z_t^\top) = \|z_t\|_2^2 \leq b_x^2 + b_u^2.$$

Therefore, we can define $\bar{\Gamma} = (b_x^2 + b_u^2)I_{n+m}$, and then $\mathbb{P}(\sum_{t=1}^T z_t z_t^\top \not\preceq T\bar{\Gamma}) = 0 \leq \delta$.

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)**.** *Consider* $x_{t+1} = A_* x_t + B_* u_t + w_t$*, where* $u_t = \pi_t(\mathcal{F}_t^m) + \eta_t$*, and* $\mathcal{F}_t^m = \mathcal{F}(w_0, \ldots, w_{t-1}, \eta_0, \ldots, \eta_{t-1})$*. Consider* w_t *i.i.d. and* (s_w, p_w)*-anti-concentration. Consider* $\eta_t \overset{i.i.d.}{\sim} \bar{\eta}\mathcal{D}_\eta$ *and* $\eta_t/\bar{\eta}$ *satisfies the* (s_η, p_η)*-anti-concentration property. Suppose* w_t *is* σ_{sub}^2*-subGaussian and has zero mean. Consider* η_t, w_t *to be independent for all* t*. Consider general policies:*

$$u_t = \pi_t(\mathcal{F}_t^m) + \eta_t, \quad t \geq 0. \tag{13}$$

Suppose we have $\|x_t\|_2 \leq b_x$ *and* $\|u_t\|_2 \leq b_u$ *for some* b_x, b_u *for all* t *under policies* (13)*. Define* $\tilde{\theta} = \min_\theta \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|_2^2$*. Define* $\mathcal{F}_t = \{w_0, \ldots, w_{t-1}, \eta_0, \ldots, \eta_t\}$*. Then we have*

$$\{z_t\}_{t\geq 0} \text{ satisfies the } (1, s_z^2 I_{n+m}, p_z)\text{-BMSB condition,} \tag{14}$$

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. 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 [45], we have $\{z_t\}$ is $(1, s_z^2 I, p_z)$-BMSB.

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

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

$$\le |\lambda^\top z_{t+1}| + b_u/k_0 \le |\lambda^\top z_{t+1}| + s_w/4$$

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 2.

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

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

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. □

□

C.2 Proof of Corollary 1

We prove Corollary 1 by verifying that Algorithm 1 satisfies the conditions in Theorem 1. The most tricky part is to provide almost surely bounds on the generated trajectories x_t and u_t. We are able to show that $u_t \in \mathbb{U}$ almost surely, but we cannot show $x_t \in \mathbb{X}$ almost surely. Nevertheless, we are able to show that $\|x_t\|_2 \leq O(\sqrt{mn})$ almost surely by leveraging the condition $\mathbf{M} \in \mathcal{M}_H$ for proper H. In the following, we first show $u_t \in \mathbb{U}$ almost surely. Then, we show $\|x_t\|_2 \leq O(\sqrt{mn})$ almost surely. Finally, we prove Corollary 1.

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

The proof of Lemma 10 is very technical heavily relies on the projection onto $\mathbb{W}$ when estimating $\hat{w}_t$. We defer the proof to Appendix G.

Lemma 11 (Almost surely 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 [24].

Proof of Corollary 1. The proof is by verifying the conditions in Theorem 1. Firstly, in episode $e-1$, $\eta_t = \bar{\eta}^{(e-1)}\tilde{\eta}_t$. Next, by Lemma 11, $b_x = O(\sqrt{mn})$. We also show that $u_t \in \mathbb{U}$ in Lemma 10. So $b_u = u_{\max}$. Further, by $\bar{\eta}^{(e-1)} \leq \eta_{\max} \leq \frac{s_w}{2\sqrt{3}s_\eta}$, we have $s_z = c_9\bar{\eta}^{(e-1)}$, where $c_9 = \min(\frac{\sqrt{3}}{2}s_\eta, \frac{s_w s_\eta}{4u_{\max}})$.

Next, we show that when $T^{(1)} \geq \frac{10}{p_z^2}(\log(24/p) + 2(m+n)\log(10/p_z) + 2(n+m)\log(\sqrt{2b_x^2 + 2u_{\max}^2}/\eta_{\max}))$ and $T^{(e)} = 2^{e-1}T^{(1)}$, the condition $T^{(e)} \geq \frac{10}{p_z^2}(\log(6e^2/p) + 2(m+n)\log(10/p_z) + 2(n+m)\log(\sqrt{b_x^2 + u_{\max}^2}/(c_9\bar{\eta}^{(e-1)})))$ is satisfied. We prove this by induction. At $e=1$, this holds. At $e=2$, this also holds by $\bar{\eta}^{(1)} \geq \eta_{\max}/2$. Suppose at $e \geq 2$, this holds, consider $e+1$,

$$\begin{aligned}
T^{(e+1)} &= 2T^{(e)} \\
&\geq \frac{10}{p_z^2}\left(2\log(6e^2/p) + 2(m+n)\log(10/p_z) + 4(n+m)\log(\sqrt{b_x^2 + u_{\max}^2}/(c_9\bar{\eta}^{(e-1)}))\right) \\
&\geq \frac{10}{p_z^2}\left(\log(6(e+1)^2/p) + 2(m+n)\log(10/p_z) + 2(n+m)\log(\sqrt{b_x^2 + u_{\max}^2}/(c_9\bar{\eta}^{(e)}))\right)
\end{aligned}$$

where the last inequality is because $2\log(e^2) \geq \log((e+1)^2)$ when $e \geq 2$, and $2\log(1/(c_9\bar{\eta}^{(e-1)})) \geq \log(1/(c_9\bar{\eta}^{(e)}))$ when $\bar{\eta}^{(e)} \geq \bar{\eta}^{(e-1)}/2$ and $\bar{\eta}^{(e)} \leq \eta_{\max} \leq 1/(2c_9)$ for all e. This completes the induction.

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

D Feasibility and Constraint Satisfaction

In this Appendix, we prove feasibility (Theorem 2) and constraint satisfaction (Theorem 3) of our Algorithm 1 and Algorithm 2. Then, we discuss how to remove Assumption 4 while still achieving feasibility and constraint satisfaction.

For ease of notation, we define a new representation of policy sets by

$$\Omega_H(\theta, \epsilon) = \{\mathbf{M} \in \mathcal{M}_H : g_i^x(\mathbf{M}; \theta) \leq d_{x,i} - \epsilon_x, g_j^u(\mathbf{M}; \theta) \leq d_{u,j} - \epsilon_u, \forall\, i, j\} \tag{15}$$

Notice that, our robustly safe policy set $\Omega(\Theta, H, \bar{\eta}, \Delta_M)$ defined in (5) satisfies

$$\Omega(\Theta, H, \bar{\eta}, \Delta_M) = \Omega_H(\hat{\theta}, \epsilon), \text{ where } \epsilon_x = \epsilon_\theta(r) + \epsilon_{\eta,x}(\bar{\eta}) + \epsilon_H(H) + \epsilon_v(\Delta_M, H), \epsilon_u = \epsilon_{\eta,u}(\bar{\eta}). \tag{16}$$

D.1 Feasility (Proof of Theorem 2)

Firstly, we note that feasibility is guaranteed if $\Omega_\dagger^{(e)} \neq \varnothing$ (Line 3), $\Omega_\dagger^{(e)} \cap \Omega_*^{(e-1)} \neq \varnothing$ (Line 4), $\Omega_*^{(e)} \neq \varnothing$ (Line 8), and $\Omega_\dagger^{(e)} \cap \Omega_*^{(e)} \neq \varnothing$ (Line 9) for all $e \geq 0$. Therefore, it suffices to construct a policy $\mathbf{M}_F$ that belongs to $\Omega_\dagger^{(e)} \cap \Omega_*^{(e)}$ for all $e \geq 0$. Therefore, it suffices to construct a policy $\mathbf{M}_F$ that belongs to all the sets above. For the rest of the proof, we first construct a policy $\mathbf{M}_F$ and then prove that $\mathbf{M}_F$ belongs to $\Omega_\dagger^{(e)} \cap \Omega_*^{(e)}$ for all $e \geq 0$.

Construct $\mathbf{M}_F$. We use the ϵ_F-strictly safe linear controller K_F in Assumption 3 to construct $\mathbf{M}_F$ that approximates K_F. The construction method is from [2]. Further, [24] establishes the constraint satisfaction property of $\mathbf{M}_F$. For completeness, we review these results below and slightly revise the results to adapt to the setting considered in this paper.

Lemma 12 (Construction of strictly safe $\mathbf{M}_F$ ([2], Corollary 1 of [24])). *For any $K \in \mathcal{K}$, one can define $\mathbf{M} \in \mathcal{M}_{H^{(0)}}$ by $M[k] = -K(A_* - B_*K)^{k-1}$ for $1 \leq k \leq H^{(0)}$. Further, if K is ϵ_F strictly safe on the true system θ_*, then $\mathbf{M} \in \Omega_{H^{(0)}}(\theta_*, \epsilon_F - \epsilon_P(H^{(0)}))$, where $\epsilon_P(H^{(0)}) = (\epsilon_{P,x}(H^{(0)}), \epsilon_{P,u}(H^{(0)}))$, $\epsilon_{P,x}(H^{(0)}) = 2\kappa^2/\gamma w_{\max}\sqrt{n}(1-\gamma)^{H^{(0)}} + \epsilon_H(H^{(0)}) = O(\sqrt{n}(1-\gamma)^{H^{(0)}})$, $\epsilon_{P,u}(H^{(0)}) = 2\kappa^2/\gamma w_{\max}\sqrt{n}(1-\gamma)^{H^{(0)}} = O(\sqrt{n}(1-\gamma)^{H^{(0)}})$.*

Based on Lemma 12, we construct $\mathbf{M}_F \in \mathcal{M}_{H^{(0)}}$ based on K_F. Since K_F is ϵ_F-strictly safe on the true system by Assumption 3, we have $\mathbf{M}_F \in \Omega_{H^{(0)}}(\theta_*, \epsilon_F - \epsilon_P(H^{(0)}))$.

Show $\mathbf{M}_F \in \Omega_\dagger^{(e)} \cap \Omega_*^{(e)}$ for all $e \geq 0$. We first prove $\mathbf{M}_F \in \Omega_\dagger^{(e)}$. Notice that $\Omega_\dagger^{(e)} = \Omega_{H^{(e)}}(\hat{\theta}^{(e)}, \epsilon_\dagger^{(e)})$, where $\epsilon_{\dagger,x}^{(e)} = \epsilon_\theta(r^{(e)}) + \epsilon_{\eta,x}(\bar{\eta}^{(e)}) + \epsilon_H(H^{(e)}) + \epsilon_v(\Delta_M^{(e)}, H^{(e)})$, $\epsilon_{\dagger,u}^{(e)} = \epsilon_{\eta,u}(\bar{\eta}^{(e)})$. Further, by our construction before, we have $\mathbf{M}_F \in \Omega_{H^{(0)}}(\theta_*, \epsilon_F - \epsilon_P(H^{(0)}))$. Since $H^{(e)} \geq H^{(0)}$, by our discussion in Appendix B.1, we can view $\mathbf{M}_F \in \mathcal{M}_{H^{(0)}}$ as a policy in $\mathcal{M}_{H^{(e)}}$. Further, by Lemma 4, we further have the following bounds.

$$\begin{aligned}
g_i^x(\mathbf{M}_F; \hat{\theta}^{(e)}) &= g_i^x(\mathbf{M}_F; \theta_*) + g_i^x(\mathbf{M}_F; \hat{\theta}^{(e)}) - g_i^x(\mathbf{M}_F; \theta_*) \\
&\leq g_i^x(\mathbf{M}_F; \theta_*) + \epsilon_{\hat{\theta}}(r_{\text{ini}}) \\
&\leq d_{x,i} - \epsilon_{F,x} + \epsilon_{P,x}(H^{(0)}) + \epsilon_{\hat{\theta}}(r_{\text{ini}}) \\
&\leq d_{x,i} - \epsilon_{F,x} + \epsilon_{P,x}(H^{(0)}) + \epsilon_\theta(r_{\text{ini}}) \\
&\leq d_{x,i} - \epsilon_\theta(r^{(0)}) - \epsilon_{\eta,x}(\bar{\eta}^{(0)}) - \epsilon_H(H^{(0)}) - \epsilon_v(\Delta_M^{(0)}, H^{(0)}) \\
&\leq d_{x,i} - \epsilon_\theta(r^{(e)}) - \epsilon_{\eta,x}(\bar{\eta}^{(e)}) - \epsilon_H(H^{(e)}) - \epsilon_v(\Delta_M^{(e)}, H^{(e)}) \\
g_j^u(\mathbf{M}_F) &\leq d_{u,j} - \epsilon_{F,u} + \epsilon_{P,u}(H^{(0)}) \\
&\leq d_{u,j} - \epsilon_{\eta,u}(\bar{\eta}^{(0)}) \\
&\leq d_{u,j} - \epsilon_{\eta,u}(\bar{\eta}^{(e)}),
\end{aligned}$$

where the first inequality is by Lemma 4 and $\hat{\theta}^{(e)}, \theta_* \in \Theta_{\text{ini}}$, the second inequality is by $\mathbf{M}_F \in \Omega_{H^{(0)}}(\theta_*, \epsilon_F - \epsilon_P(H^{(0)}))$, the third inequality is by $\epsilon_{\hat{\theta}}(r) \le \epsilon_\theta(r)$, the fourth inequality is by Assumption 4 and Condition (II) in Theorem 2, the fifth inequality is by Condition (I) in Theorem 2, and the reasons behind the inequalities for $g_j^u(\mathbf{M}_F)$ are similar. Therefore, we have shown that $\mathbf{M}_F \in \Omega_\dagger^{(e)}$.

The proof of $\mathbf{M}_F \in \Omega^{(e)}$ is similar.

In conclusion, we have proved Theorem 2.

D.2 Constraint Satisfaction (Proof of Theorem 3)

The control constraint satisfaction has already been proved in Lemma 10. Hence, we will focus on state constraint satisfaction here. Define an event

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

Notice that

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

where we used Corollary 1 and $\theta_* \in \Theta^{(0)} = \Theta_{\text{ini}}$. In the following, we will condition on event $\mathcal{E}_{\text{safe}}$ and show $x_t \in \mathbb{X}$ for all $t \ge 0$ under this event. We prove this by induction. When $t = 0$, notice that $x_0 = 0 \in \mathbb{X}$, which is a consequence of Assumption 3 when considering $t = 0$ and $x_0 = 0$. Further, for $s < t = 0$, $x_s = 0 \in \mathbb{X}$ by our definition. Next, we suppose at stage $t \ge 1$, we have $x_s \in \mathbb{X}$ for all $s < t$. We will show $x_t \in \mathbb{X}$ below. 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 $T^{(e)} \le t \le T^{(e)} + W_1^{(e)} - 1$ for $e \ge 1$.) In this case, $\mathbf{M}_t \in \Omega^{(e-1)} \subseteq \mathcal{M}_{H^{(e-1)}}$, so

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

In this case, we define $\hat{\theta}_t^g = \hat{\theta}^{(e)}$.

In the following, we will verify the conditions of Lemma 4, 6, 7, 8. Then, we will prove $x_t \in \mathbb{X}$ by Corollary 2.

Firstly, consider Lemma 4. Here, $H_t = H^{(e-1)}$. Since $t_2^{(e-1)} \le T^{(e)}$ (our condition in Theorem 3) and $W_1 \ge H^{(e)} \ge H^{(e-1)}$ according to Algorithm 2 and Algorithm 1, we have stage $t - k \ge t_1^{(e-1)} + T_D^{(e-1)}$ for all $0 \le k \le H^{(e-1)}$. Let $\hat{\theta}_{t-k}$ denote the estimated model used to approximate $\hat{w}_{t-k}$. Then, $\hat{\theta}_{t-k} = \hat{\theta}^{(e)}$ for $0 \le k \le H^{(e-1)}$. Therefore, conditioning on $\mathcal{E}_{\text{safe}}$, we have $\|\hat{\theta}_{t-k} - \theta_*\|_F = \|\hat{\theta}_t^g - \theta_*\|_F \le r^{(e)}$. Besides, we have shown that $u_t \in \mathbb{U}$ for all t and we suppose $x_{t-k} \in \mathbb{X}$ as our induction condition for $k \ge 1$. Hence, we satisfy the conditions in Lemma 4 and thus

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

Secondly, consider Lemma 6. For $0 \le k \le H^{(e-1)}$, we have shown $t_1^{(e-1)} + T_D^{(e-1)} \le t - k \le T^{(e)} + W_1^{(e)} - 1$ in the discussion above, so we have $\eta_{t-k} = 0$ by our algorithm design. So by

Lemma 6, we have

$$\underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\text{excitation noises}} = 0.$$

Thirdly, consider Lemma 7. By our induction condition, $x_{t-H_t} \in \mathbb{X}$, so $D_{x,i}^\top A_*^{H_t} x_{t-H_t} \leq \epsilon_H(H^{(e-1)})$.

Fourth, consider Lemma 8. By our algorithm design and by $t_1^{(e-1)} + T_D^{(e-1)} \leq t-k \leq T^{(e)} + W_1^{(e)} - 1$, we have $\Delta_M^{(e-1)} \geq \max_{1 \leq k \leq H^{(e-1)}} \frac{\|\mathbf{M}_t - \mathbf{M}_{t-k}\|_F}{k}$. Therefore,

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

In conclusion, by applying Corollary 2 and the discussions above, we have

$$\begin{aligned}
D_{x,i}^\top x_t \leq & \underbrace{g_i^x(\mathbf{M}_t; \hat{\theta}_t^g)}_{\text{estimated state constraint function}} + \underbrace{(g_i^x(\mathbf{M}_t; \theta_*) - g_i^x(\mathbf{M}_t; \hat{\theta}_t^g)) + \sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})}_{\text{model estimation errors}} \\
& + \underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\text{excitation noises}} + \underbrace{D_{x,i}^\top A_*^{H_t} x_{t-H_t}}_{\text{history truncation errors}} + \underbrace{(\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1}; \theta_*) - g_i^x(\mathbf{M}_t; \theta_*))}_{\text{policy variation errors}} \\
& \leq d_{x,i} - \epsilon_H(H^{(e-1)}) - \epsilon_v(\Delta_M^{(e-1)}, H^{(e-1)}) - \epsilon_\theta(r^{(e)}) + \epsilon_\theta(r^{(e)}) + \epsilon_H(H^{(e-1)}) + \epsilon_v(\Delta_M^{(e-1)}, H^{(e-1)}) \\
& \leq d_{x,i}
\end{aligned}$$

for all i. Therefore, we have shown $x_t \in \mathbb{X}$.

(Case 2: when $T^{(e)} + W_1^{(e)} \leq t \leq t_1^{(e)} + T_D^{(e)} + \tilde{W}_1^{(e)} - 1$.**)** We have $\mathbf{M}_t \in \Omega_\dagger^{(e)}$. Hence, we have $H_t = H^{(e)}$, $\hat{\theta}_t^g = \theta^{(e)}$, and

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

The proof of $x_t \in \mathbb{X}$ is very similar to Case 1. We still verify the conditions of Lemma 4, 6, 7, 8 and then apply Corollary 2.

Firstly, consider Lemma 4. Here, $H_t = H^{(e)}$. Since $W_1^{(e)} \geq H^{(e)}$ by Line 2 of Algorithm 2, we have stage $t - k \geq T^{(e)}$ for all $0 \leq k \leq H^{(e)}$. Then, when $t - k < t_1^{(e)} + T_D^{(e)}$, we have $\hat{\theta}_{t-k} = \hat{\theta}^{(e)}$, and when $t - k \geq t_1^{(e)} + T_D^{(e)}$, we have $\hat{\theta}_{t-k} = \hat{\theta}^{(e+1)}$. Conditioning on $\mathcal{E}_{\text{safe}}$, we have $\|\hat{\theta}_{t-k} - \theta_*\|_F \leq r^{(e)}$. We also have $\|\hat{\theta}_t^g - \theta_*\|_F \leq r^{(e)}$ by $\hat{\theta}_t^g = \hat{\theta}^{(e)}$. Besides, we have shown that $u_t \in \mathbb{U}$ for all t and we suppose $x_{t-k} \in \mathbb{X}$ as our induction condition for $k \geq 1$. Hence, we satisfy the conditions in Lemma 4 and thus

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

Secondly, consider Lemma 6. For $0 \leq k \leq H^{(e)}$, when $t-k < t_1^{(e)} + T_D^{(e)}$, we have $\eta_{t-k} = \bar{\eta}^{(e)}$, and when $t-k \geq t_1^{(e)} + T_D^{(e)}$, we have $\eta_{t-k} = 0$. So by Lemma 6, we have

$$\underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\text{excitation noises' impact on states}} \leq \epsilon_{\eta,x}(\bar{\eta}^{(e)}).$$

Thirdly, consider Lemma 7. By our induction condition, $x_{t-H_t} \in \mathbb{X}$, so $D_{x,i}^\top A_*^{H_t} x_{t-H_t} \leq \epsilon_H(H^{(e)})$.

Fourth, consider Lemma 8. When $t < t_1^{(e)} + T_D^{(e)}$, we have $\Delta_M^{(e)} \geq \max_{1 \leq k \leq H^{(e)}} \frac{\|\mathbf{M}_t - \mathbf{M}_{t-k}\|_F}{k}$. When $t \geq t_1^{(e)} + T_D^{(e)}$, we have $\Delta_M^{(e)} \geq \min(\Delta_M^{(e)}, \Delta_M^{(e+1)}) \geq \max_{1 \leq k \leq H^{(e)}} \frac{\|\mathbf{M}_t - \mathbf{M}_{t-k}\|_F}{k}$. Therefore,

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

In conclusion, by applying Corollary 2 and the discussions above, we have

$$\begin{aligned} D_{x,i}^\top x_t \leq & \underbrace{g_i^x(\mathbf{M}_t; \hat{\theta}_t^g)}_{\text{estimated state constraint function}} + \underbrace{(g_i^x(\mathbf{M}_t; \theta_*) - g_i^x(\mathbf{M}_t; \hat{\theta}_t^g)) + \sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})}_{\text{model estimation errors}} \\ & + \underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\text{excitation noises}} + \underbrace{D_{x,i}^\top A_*^{H_t} x_{t-H_t}}_{\text{history truncation errors}} + \underbrace{(\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1}; \theta_*) - g_i^x(\mathbf{M}_t; \theta_*))}_{\text{policy variation errors}} \\ \leq & \; d_{x,i} \end{aligned}$$

for all i. Therefore, we have shown $x_t \in \mathbb{X}$.

(Case 3: when $t_1^{(e)} + T_D^{(e)} + \tilde{W}_1^{(e)} \leq t \leq T^{(e+1)} - 1$.) We have $\mathbf{M}_t \in \Omega^{(e)}$. Hence, we have $H_t = H^{(e)}$, $\hat{\theta}_t^g = \theta^{(e+1)}$, and

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

The proof of $x_t \in \mathbb{X}$ is very similar to Case 1 and 2. We still verify the conditions of Lemma 4, 6, 7, 8 and then apply Corollary 2.

Firstly, consider Lemma 4. Here, $H_t = H^{(e)}$. Since $\tilde{W}_1^{(e)} \geq H^{(e)}$ by Line 2 of Algorithm 2, we have stage $t-k \geq t_1^{(e)} + T_D^{(e)}$ for all $0 \leq k \leq H^{(e)}$. By Step 1 and 2 in Algorithm 2, we have $\hat{\theta}_{t-k} = \hat{\theta}^{(e+1)}$. Conditioning on $\mathcal{E}_{\text{safe}}$, we have $\|\hat{\theta}_{t-k} - \theta_*\|_F \leq r^{(e+1)}$. We also have $\|\hat{\theta}_t^g - \theta_*\|_F \leq r^{(e+1)}$ by $\hat{\theta}_t^g = \hat{\theta}^{(e+1)}$. Besides, we have shown that $u_t \in \mathbb{U}$ for all t and we suppose $x_{t-k} \in \mathbb{X}$ as our induction condition for $k \geq 1$. Hence, we satisfy the conditions in Lemma 4 and thus

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

Secondly, consider Lemma 6. We have $\eta_{t-k} = 0$ so $\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i} = 0$.

Thirdly, consider Lemma 7. By our induction condition, $x_{t-H_t} \in \mathbb{X}$, so $D_{x,i}^\top A_*^{H_t} x_{t-H_t} \leq \epsilon_H(H^{(e)})$.

Fourth, consider Lemma 8 We have $\Delta_M^{(e)} \geq \max_{1 \leq k \leq H^{(e)}} \frac{\|\mathbf{M}_t - \mathbf{M}_{t-k}\|_F}{k}$. Therefore,

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

In conclusion, by applying Corollary 2 and the discussions above, we have

$$\begin{aligned} D_{x,i}^\top x_t \le \quad & \underbrace{g_i^x(\mathbf{M}_t; \hat{\theta}_t^g)}_{\text{estimated state constraint function}} \quad + \underbrace{(g_i^x(\mathbf{M}_t;\theta_*) - g_i^x(\mathbf{M}_t;\hat{\theta}_t^g)) + \sum_{k=1}^{H_t} D_{x,i}^\top A_*^{k-1}(w_{t-k} - \hat{w}_{t-k})}_{\text{model estimation errors}} \\ & + \underbrace{\sum_{i=1}^{H_t} D_{x,i}^\top A_*^{i-1} B_* \eta_{t-i}}_{\text{excitation noises}} + \underbrace{D_{x,i}^\top A_*^{H_t} x_{t-H_t}}_{\text{history truncation errors}} + \underbrace{(\tilde{g}_i^x(\mathbf{M}_{t-H_t:t-1};\theta_*) - g_i^x(\mathbf{M}_t;\theta_*))}_{\text{policy variation errors}} \\ \le\ & d_{x,i} - \epsilon_H(H^{(e)}) - \epsilon_v(\Delta_M^{(e)}, H^{(e)}) - \epsilon_\theta(r^{(e+1)}) + \epsilon_\theta(r^{(e+1)}) + \epsilon_H(H^{(e)}) + \epsilon_v(\Delta_M^{(e)}, H^{(e)}) \\ \le\ & d_{x,i} \end{aligned}$$

for all i. Therefore, we have shown $x_t \in \mathbb{X}$.

In summary, we have shown that $x_t \in \mathbb{X}$ for all $t \ge 0$ by induction.

D.3 A Warm-up Scheme to Remove Assumption 4

Similar to [11], we can adopt a warm-up scheme to remove Assumption 4. This warm-up scheme requires the knowledge of a ϵ_s-strictly safe controller $u_t = K_s x_t + \eta_t$ for some $\epsilon_s > 0$ and η_t sufficiently small. By implementing $u_t = K_s x_t + \eta_t$ for sufficiently long, we can reduce the model estimation error according to Theorem 1. Since Assumption 4 only requires $r^{(0)}$ to be smaller than a constant, the warm-up scheme also takes a finite number of steps, so including it will not affect our regret bound's dependence on T. We define the reduced model uncertainty set as $\Theta^{(0)} = \{\theta : \|\theta - \hat{\theta}^{(0)}\|_F \le r^{(0)}\} \cap \Theta_{\text{ini}}$. Notice that we only have $\theta_* \in \Theta^{(0)}$ with high probability due to Theorem 1, so the feasibility of our algorithm is guaranteed with high probability, and the probability of state constraint satisfaction should also be adjusted accordingly. The control constraint satisfaction is not affected.

If the warm-up scheme is run separately and Algorithm 1 can be restarted from $x_0 = 0$ after the warm-up scheme, the discussion above is sufficient. However, if we do not allow restarts after the warm-up scheme, then we have to design a safe transition algorithm to transit from $u_t = K_s x_t + \eta_t$ to DAP controller $\mathbf{M}_\dagger^{(0)}$. We briefly sketch a method to achieve this goal. We design another DAP controller $\hat{\mathbf{M}}(K_s)$ by $\hat{M}[k] = -K(\hat{A} - \hat{B}K)^{k-1}$ for $1 \le k \le H^{(0)}$, where $\hat{\theta}$ can be selected as $\hat{\theta}^{(0)}$ and select a sufficiently long $H^{(0)}$. It can be shown that switching from $u_t = K_s x_t + \eta_t$ to $\hat{\mathbf{M}}(K_s)$ directly will only incur additional errors $O((1-\gamma)^{H^{(0)}} + r^{(0)})$. Since K_s is ϵ_s-strictly safe, by selecting a large enough $H^{(0)}$ and by running the warm-up scheme long enough to induce a small enough $r^{(0)}$, one can directly switch from $u_t = K_s x_t + \eta_t$ to $\hat{\mathbf{M}}(K_s)$ without violating constraints. Then, let $\mathbf{M}_*^{(-1)} = \hat{\mathbf{M}}(K_s)$ and run Algorithm 1 with Line 4 activated for $e = 0$. This allows safe transitions and only takes a finite number of steps so the regret bound order is not affected.

E Regret Analysis

In this appendix, we provide a proof of Theorem 4. It can be verified that the parameters in Theorem 4 satisy the conditions in Theorem 2 and 3, so we have feasibility and constraint satisfaction. Next, we focus on the regret bound. We consider the following regret decomposition. For $e \ge 0$, define

$$\begin{aligned} \mathcal{T}_1^{(e)} &= \{T^{(e)} \le t \le t_2(e) + H^{(e)} - 1\} \\ \mathcal{T}_2^{(e)} &= \{t_2(e) + H^{(e)} \le t \le T^{(e+1)} - 1\} \end{aligned}$$

Then, we have

$$\text{Regret} = \sum_{t=0}^{T-1} (l(x_t, u_t) - J^*) = \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_1^{(e)}} (l(x_t, u_t) - J^*)}_{\text{First term}} + \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (l(x_t, u_t) - J^*)}_{\text{Second term}} \quad (17)$$

Before the details of the proof, we establish a useful technical result for regret bounds.

Lemma 13. *When $T^{(e)} = 2^{e-1}T^{(1)}$, and $T^{(N)} \geq T > T^{(N-1)}$, $N \leq O(\log T)$. Further, for any $\alpha > 0$,*

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

Proof. By $T \geq T^{(N-1)} \geq 2^{(N-2)}$, we have $\log T \geq (N-2)\log(2)$, so $N \leq O(\log T)$.

$$\sum_{e=1}^{N} (T^{(e)})^{\alpha} = \sum_{e=1}^{N} (2^{e-1})^{\alpha}(T^{(1)})^{\alpha} \leq O((2^N)^{\alpha}(T^{(1)})^{\alpha}) \leq O(T^{\alpha})$$

□

E.1 Bound the first item in regret decomposition (17)

Lemma 14 (Regret Bound of the First Term)**.** *When $\mathcal{E}_{safe}$ is true, 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. Under the conditions, when $\mathcal{E}_{\text{safe}}$ is true, we have $x_t \in \mathbb{X}$ and $u_t \in \mathbb{U}$, so $l(x_t, u_t) - J^* \leq O(1)$. Further, under the conditions in Theorem 4, we have that the number of stages in $\mathcal{T}_1^{(e)}$ is $O((T^{(e+1)})^{2/3})$. Therefore, by Lemma 13, we have the proof. □

E.2 Bound the second item in regret decomposition (17)

We further decompose the second item in (17) into three parts.

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

where we define the following. When $t \in \mathcal{T}_2^{(e)}$,

$$\tilde{x}_t = \sum_{k=1}^{2H^{(e)}} \Phi_k^x(\mathbf{M}_*^{(e)}; \theta_*) w_{t-k}, \tag{18}$$

$$\tilde{u}_t = \sum_{k=1}^{H^{(e)}} M_*^{(e)}[k] w_{t-k}, \tag{19}$$

and define $\Omega_*^{(e)} = \Omega_{H^{(e)}}(\theta_*; \epsilon_*^{(e)})$, where $\epsilon_*^{(e)} = (\epsilon_H(H^{(e)}), 0)$. Define

$$\mathbf{M}_{H^{(e)}}^* = \arg\min_{\mathbf{M} \in \Omega_*^{(e)}} f(\mathbf{M}; \theta_*)$$

E.2.1 Helpful lemmas

We summarize some useful technical results here. The proofs are deferred to Appendix G.

Lemma 15 (Perturbation bound on f with respect to θ)**.** *For any $H \geq 1$, $\mathbf{M} \in \mathcal{M}_H$, any $\theta, \hat{\theta} \in \Theta_{ini}$ with $\|\theta - \hat{\theta}\|_F \leq r_\theta$, when $H \geq \log(2\kappa)/\log((1-\gamma)^{-1})$,*

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

Lemma 16 (Gradient bound of $f(\mathbf{M};\theta)$)**.** *For any $H \geq 1$, $\mathbf{M} \in \mathcal{M}_H$, $\theta \in \Theta_{ini}$,*

$$\|\nabla f(\mathbf{M};\theta)\|_F \leq G_f$$

and $G_f = O(\sqrt{n^2 mH})$.

Lemma 17. *For any $\theta, \theta' \in \Theta_{ini}$, for any H, any $\epsilon, \epsilon' = O(1)$, $\Omega(\theta, \epsilon, H) \cap \Omega(\theta', \epsilon', H)$ can be converted into standard linear inequality constraints set, with diameter $O(\sqrt{mn} + \sqrt{k_c})$.*

The proof is a direct generalization of Lemma 9 in [24].

Lemma 18 (Cost different lemma for linearly constrained convex optimization)**.** *Consider two polytopes, $\Omega_1 = \{x : Cx \leq h - \Delta_1\}, \Omega_2 = \{x : Cx \leq h - \Delta_2\}$. Define $\Delta_0 = \min(\Delta_1, \Delta_2)$ elementwise. Define $\Delta_3 = \max(\Delta_1, \Delta_2)$ elementwise. Suppose L_2-diameter of Ω_0 is d_{Ω_0}. Suppose $f(x)$ is L-Lipschitz continuous. Suppose there exists $x_F \in \Omega_3$, then*

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

E.2.2 Bound Part iii

This part is the dominating part in the regret bound.

Define

$$\mathbf{M}_\alpha^{(e)} = \arg\min_{\mathbf{M} \in \Omega^{(e)}} f(\mathbf{M}_*^{(e)};\theta_*)$$

We divide Part iii into two parts.

$$\underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_{H^{(e)}}^*;\theta_*))}_{\text{Part iii}} = \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_\alpha^{(e)};\theta_*))}_{\text{Part iii-A}} \tag{20}$$

$$+ \underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}_\alpha^{(e)};\theta_*) - f(\mathbf{M}_{H^{(e)}}^*;\theta_*))}_{\text{Part iii-B}} \tag{21}$$

Lemma 19 (Bound on Part iii-A)**.** *When $H^{(e)} \geq \log(2\kappa)/(\log((1-\gamma)^{-1}))$, $\mathcal{E}_{safe}$ is true, when we first do estimation updates then do full exploitation (the same holds for the reversed order), under the conditions in Corollary 1, for $e \geq 0$,*

$$\underbrace{\sum_{e=0}^{N-1} \sum_{t \in \mathcal{T}_2^{(e)}} (f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_\alpha^{(e)};\theta_*))}_{\text{Part iii-A}} \leq O(mnr_\theta^{(e+1)}) = \tilde{O}(mn\sqrt{mn+n^2}(T^{(e+1)})^{-1/3})$$

Proof.

$$f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_\alpha^{(e)};\theta_*) = f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_*^{(e)};\hat{\theta}^{(e+1)})$$
$$+ f(\mathbf{M}_*^{(e)};\hat{\theta}^{(e+1)}) - f(\mathbf{M}_\alpha^{(e)};\hat{\theta}^{(e+1)}) + f(\mathbf{M}_\alpha^{(e)};\hat{\theta}^{(e+1)}) - f(\mathbf{M}_\alpha^{(e)};\theta_*)$$

Since

$$\mathbf{M}_\alpha^{(e)} = \arg\min_{\mathbf{M} \in \Omega^{(e)}} f(\mathbf{M}_*^{(e)};\theta_*)$$

and since

$$\mathbf{M}_*^{(e)} = \arg\min_{\mathbf{M} \in \Omega^{(e)}} f(\mathbf{M}_*^{(e)};\theta^{(e+1)})$$

we have

$$f(\mathbf{M}_*^{(e)};\hat{\theta}^{(e+1)}) - f(\mathbf{M}_\alpha^{(e)};\hat{\theta}^{(e+1)}) \le 0$$

So

$$\begin{aligned}
f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_\alpha^{(e)};\theta_*) &= f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_*^{(e)};\hat{\theta}^{(e+1)}) \\
&\quad + f(\mathbf{M}_*^{(e)};\hat{\theta}^{(e+1)}) - f(\mathbf{M}_\alpha^{(e)};\hat{\theta}^{(e+1)}) + f(\mathbf{M}_\alpha^{(e)};\hat{\theta}^{(e+1)}) - f(\mathbf{M}_\alpha^{(e)};\theta_*) \\
&\le O(mn\|\hat{\theta}^{(e+1)} - \theta_*\|_F) = O(mnr_\theta^{(e+1)})
\end{aligned}$$

when $\mathcal{E}_{\text{safe}}$ is true.

Under the conditions in corollary 1,

$$f(\mathbf{M}_*^{(e)};\theta_*) - f(\mathbf{M}_\alpha^{(e)};\theta_*) \le \tilde{O}(mn\sqrt{mn+n^2}(T^{(e+1)})^{-1/3})$$

□

Lemma 20 (Bound Part iii-B)**.** *Suppose $\mathcal{E}_{safe}$ is true, then we have*

$$f(\mathbf{M}_\alpha^{(e)};\theta_*) - f(\mathbf{M}^*_{H^{(e)}};\theta_*) \le \tilde{O}\left((n^2m^{1.5}+n^{2.5}m)\sqrt{mn+k_c}(T^{(e+1)})^{-1/3}\right)$$

Proof. Remember that

$$\mathbf{M}_\alpha^{(e)} = \arg\min_{\mathbf{M}\in\Omega^{(e)}} f(\mathbf{M}_*^{(e)};\theta_*)$$

and

$$\mathbf{M}^*_{H^{(e)}} = \arg\min_{\mathbf{M}\in\Omega_*^{(e)}} f(\mathbf{M};\theta_*)$$

As auxiliary sets, we define two sets below:

$$\bar{\Omega}_*^{(e)} = \Omega(\theta_*;\tilde{\epsilon}^{(e+1)} - \epsilon_{\hat{\theta}}(r_\theta^{(e+1)}), H^{(e)}), \quad \bar{\Omega}_e = \Omega(\hat{\theta}^{(e+1)};\epsilon_*^{(e)} - \epsilon_{\hat{\theta}}(r_\theta^{(e+1)}), H^{(e)}).$$

Notice that when $\mathcal{E}_{\text{safe}}$ is true, by Lemma 13 in technical report.pdf, $\mathbf{M}_\alpha^{(e)} \in \bar{\Omega}_*^{(e)} = \Omega(\theta_*;\tilde{\epsilon}^{(e+1)} - \epsilon_{\hat{\theta}}(r_\theta^{(e+1)}), H^{(e)})$, and $\mathbf{M}^*_{H^{(e)}} \in \bar{\Omega}_e = \Omega(\hat{\theta}^{(e+1)};\epsilon_*^{(e)} - \epsilon_{\hat{\theta}}(r_\theta^{(e+1)}), H^{(e)})$. Then, define

$$\Omega_1 = \Omega^{(e)} \cap \bar{\Omega}_*^{(e)}, \quad \Omega_2 = \Omega_*^{(e)} \cap \bar{\Omega}_e$$

Notice that

$$\Omega_1 \cap \Omega_2 = \Omega^{(e)} \cap \Omega_*^{(e)}.$$

Now, we have

$$\mathbf{M}_\alpha^{(e)} = \arg\min_{\mathbf{M}\in\Omega_1} f(\mathbf{M}_*^{(e)};\theta_*), \quad \mathbf{M}^*_{H^{(e)}} = \arg\min_{\mathbf{M}\in\Omega_2} f(\mathbf{M};\theta_*)$$

Since Ω_1 and Ω_2 can be converted to linear constraints sets, we can apply Lemma 18. By Lemma 17, $d_{\Gamma_0} = O(\sqrt{mn}+\sqrt{k_c})$, by Lemma 16, $G_f = O(n\sqrt{m}H^{(e)})$. By our choices of parameters, we have $\min_{\{i:(\Delta_1)_i\neq(\Delta_2)_i\}}(h-\Delta_3 - Cx_F)_i \ge \min(\epsilon_{F,x},\epsilon_{F,u})/4$. Further,

$$\begin{aligned}
\|\Delta_1 - \Delta_2\|_\infty &\le \|\tilde{\epsilon}^{(e+1)} - \epsilon_*^{(e)}\|_\infty \\
&\le \epsilon_{\hat{w}}^{(e+1)} + \epsilon_{\hat{\theta}}^{(e+1)} + \epsilon_v^{(e)} \\
&\le O(\sqrt{mn}r_\theta^{(e+1)} + \sqrt{mnH^{(e)}}\Delta_M^{(e)}) \\
&\le \tilde{O}((T^{(e+1)})^{-1/3}(\sqrt{m^2n^2+n^3m} + \sqrt{mnH^{(e)}})
\end{aligned}$$

Therefore,

$$\begin{aligned}
&f(\mathbf{M}^{(e)}{}_\alpha;\theta_*) - f(\mathbf{M}^*_{H^{(e)}};\theta_*) \\
&\quad \le \tilde{O}\left((\sqrt{mn}+\sqrt{k_c})(n\sqrt{m}H^{(e)})((T^{(e+1)})^{-1/3}(\sqrt{m^2n^2+n^3m}+\sqrt{mnH^{(e)}})\right)
\end{aligned}$$

$$\leq \tilde{O}\left((\sqrt{mn} + \sqrt{k_c})(n\sqrt{m})((T^{(e+1)})^{-1/3}\sqrt{m^2n^2 + n^3m}\right)$$
$$\leq \tilde{O}\left((n^2m^{1.5} + n^{2.5}m)\sqrt{mn + k_c}(T^{(e+1)})^{-1/3}\right)$$

□

Theorem 6 (Bound on Part iii)**.** *Under the conditions in Theorem 3, when $\mathcal{E}_{safe}$ is true,*

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

Proof. By summing over Lemma 20 and 19 and Lemma 13. □

E.2.3 Bound Part iv

Theorem 7. *By our choice of $H^{(e)}$ in Theorem 3, we have*

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

Proof. We know $K^* \in \mathcal{K}$ and K^* is safe. Then, by Lemma 4 in [24], for $H^{(e)}$, we can define $\mathbf{M}_{H^{(e)}}(K^*) \in \Omega(\theta_*; -(\epsilon_H^{(e)}, 0) - \epsilon_P^{(e)}) =: \Omega_\beta^{(e)}$. Further, by Lemma 6 in [24],

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

Define

$$\mathbf{M}_\beta^{(e)} = \arg\min_{\Omega_\beta^{(e)}} f(\mathbf{M}; \theta_*)$$

. Then, we have

$$f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - J^* = f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - f(\mathbf{M}_{H^{(e)}}(K^*); \theta_*) + f(\mathbf{M}_{H^{(e)}}(K^*); \theta_*) - J^*$$
$$\leq f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - f(\mathbf{M}_\beta^{(e)}; \theta_*) + \tilde{O}(n^2m(1-\gamma)^{H^{(e)}})$$

Now, we can apply Lemma 18, by noticing that $\|\Delta_1 - \Delta_2\|_\infty = 2\max(2\epsilon_H^{(e)} + \epsilon_{P,x}^{(e)}, \epsilon_{P,u}^{(e)}) = O(\sqrt{n}(1-\gamma)^{H^{(e)}})$, we have

$$f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - J^* \leq f(\mathbf{M}^*_{H^{(e)}}; \theta_*) - f(\mathbf{M}_\beta^{(e)}; \theta_*) + \tilde{O}(n^2m(1-\gamma)^{H^{(e)}})$$
$$\leq \tilde{O}(n\sqrt{m}\sqrt{mn + k_c}\sqrt{n}(1-\gamma)^{H^{(e)}}) = \tilde{O}(n\sqrt{m}\sqrt{mn + k_c}\sqrt{n}(T^{(e+1)})^{-1})$$

by our choice of $H^{(e)}$ in Theorem 3. Therefore,

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

□

E.2.4 Bound Part ii

Lemma 21 (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.

E.2.5 Bound Part i

Lemma 22. *When $\mathcal{E}_{safe}$ is true, $x_t, \tilde{x}_t \in \mathbb{X}$, and $u_t, \tilde{u}_t \in \mathbb{U}$*

Proof. this is by Theorem 3's proof. $\tilde{x}_t$ is very straightforward to verify. □

Lemma 23. *When $\mathcal{E}_{safe}$ is true, under conditions in Theorem 3,*

$$\|x_t - \tilde{x}_t\|_2 = \tilde{O}(n\sqrt{m}\sqrt{m+n}(T^{(e+1)})^{-1/3}), \quad \|u_t - \tilde{u}_t\|_2 = \tilde{O}(n\sqrt{m}\sqrt{m+n}(T^{(e+1)})^{-1/3})$$

Proof. Notice that here $\eta_t = 0$,

$$\|x_t - \tilde{x}_t\|_2 \le O((1-\gamma)^{H^{(e)}} x_{\max} + \sqrt{mn} z_{\max} r_\theta^{(e+1)}) = \tilde{O}(n\sqrt{m}\sqrt{m+n}(T^{(e+1)})^{-1/3})$$
$$\|u_t - \tilde{u}_t\|_2 \le O(z_{\max}\sqrt{mn} r_\theta^{(e+1)}) = \tilde{O}(n\sqrt{m}\sqrt{m+n}(T^{(e+1)})^{-1/3})$$

this is done. □

Lemma 24. *When $\mathcal{E}_{safe}$ is true, under conditions in Theorem 3,*

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

Proof. This is by $|x_t + \tilde{x}_t| = O(1)$ and the lemma above and Lemma 13. □

E.3 Combine the Bounds Above and Obtain a Proof of Theorem 4

By combining Theorem 6, Lemma 24, Lemma 7, and Lemma 21, we can prove the regret bound for our algorithm. Notice that Theorem 6, Lemma 24, Lemma 7 all condition on $\mathcal{E}_{\text{safe}}$, and $\mathcal{E}_{\text{safe}}$ holds w.p. $1-p$. But Lemma 21 conditions on a different event and that event also holds with probability $1-p$. Putting them together, we have our regret bound holds w.p. $1-2p$.

F More General Benchmark Policy Classes and Nonzero Initial Value $x_0 \neq 0$

In this appendix, we generalize our results to broader benchmark policy classes that include linear dynamical policies considered in [15] and one version of robust model predictive control proposed in [30]. We also design a safe algorithm to handle $x_0 \neq 0$.

F.1 Benchmark Policy Class that Includes Linear Dynamic Policies

In this subsection, we consider a broader benchmark policy class that not only includes linear static policies, $u_t = Kx_t$, but also includes linear dynamic policies, $\hat{u}(z) = \hat{K}(z)\hat{x}(z)$, where $\hat{u}, \hat{x}$ are z-transform of time domain trajectories $\{u_t\}_{t\ge 0}$, $\{x_t\}_{t\ge 0}$ and are defined on the frequency domain, and $\hat{K}$ indicates a linear dynamic controller and is also defined on the frequency domain. Linear dynamic policies are adopted in [15] to tackle constrained LQR.

In the following, we briefly review some basic facts of linear dynamic policies. For more details, we refer the reader to [4]. When applying $\hat{u}(z) = \hat{K}(z)\hat{x}(z)$ to the linear system $x_{t+1} = A_* x_t + B_* u_t + w_t$, we denote the transfer functions from $\hat{w}(z) \to \hat{x}(z)$ and $\hat{w}(z) \to \hat{u}(z)$ as $\hat{\Psi}_x$ and $\hat{\Psi}_u$, which are given by

$$\hat{\Psi}_x(z) = (zI - A - B\hat{K})^{-1}, \qquad \hat{\Psi}_u(z) = \hat{K}(zI - A - B\hat{K})^{-1}.$$

By taking inverse z-transform on the transfer functions $\hat{\Psi}_x$ and $\hat{\Psi}_u$, we obtain impulse response functions, denoted by $\mathbf{\Psi} = \{\Psi_x[t]\}_{t\ge 1}$, $\mathbf{\Psi}_u = \{\Psi_u[t]\}_{t\ge 1}$. The states and actions can be represented by the impulse response functions by the following:

$$x_t = \sum_{k=1}^{+\infty} \Psi_x[t-k] w_{t-k}, \qquad u_t = \sum_{k=1}^{+\infty} \Psi_u[t-k] w_{t-k},$$

where $w_t = 0$ if $t < 0$. Notice that the impulse response function $\mathbf{\Psi}_u$ provides another way to represent the linear dynamic controller $\hat{u}(z) = \hat{K}(z)\hat{x}(z)$. This representation is closely related with our DAP controller. Therefore, we will use this representation to define our new benchmark policies. Since $\mathbf{\Psi}$ contains an infinite number of matrices, it is usually called infinite impulse response, to be contrasted with finite impulse response, where only a finite truncation, e.g. $\{\Psi_u[t]\}_{t=1}^H$, is considered. Our DAP controller can be viewed as a finite impulse response function.

To ensure stability, certain function space is usually imposed on the linear dynamic controllers. Such function space can be defined by transfer functions or impulse response functions. For example, the following function space for transfer functions is commonly adopted in the literature [13, 12, 15, 4].

$$\mathcal{RH}_\infty = \{\hat{M}(z) = \sum_{k=1}^{+\infty} M[k]z^{-k} : \|M[k]\|_2 \leq 2\kappa^2(1-\gamma)^{k-1},\ k \geq 1\}$$

The corresponding function space for impulse responses are defined below.

$$\mathcal{M}_\infty = \{\mathbf{M} = \{M[k]\}_{k\geq 1} : \|M[k]\|_\infty \leq 2\sqrt{n}\kappa^2(1-\gamma)^{k-1},\ k \geq 1\}.$$

Here we consider $\|\cdot\|_\infty$ norm to be consistent with our definition of $\mathcal{M}_H$, and the additional factor $\sqrt{n}$ is introduced to make sure $\mathcal{M}_\infty$ is large enough to contain all $\mathbf{M} = \{M[k]\}_{k\geq 1}$ characterized in $\mathcal{RH}_\infty$, i.e. $\|M[k]\|_2 \leq 2\kappa^2(1-\gamma)^{k-1}$.

A broader benchmark policy class Φ_1 that contains linear dynamic controllers. We define a broader benchmark policy class Φ_1 that contains not only linear static controllers but also linear dynamic controllers below:

$$\Phi_1 = \{\mathbf{M} = \{M[k]\}_{k\geq 1} \in \mathcal{M}_\infty : x_t \in \mathbb{X},\ u_t \in \mathbb{U},\ \forall \{w_k \in \mathbb{W}\}_{k\geq 0}\} \tag{22}$$

Roughly, policy class Φ_1 contains all linear dynamic controllers inside $\mathcal{M}_\infty$ that guarantees constraint satisfaction/safety. Even when considering this benchmark class, we can still provide a $\tilde{O}(T^{2/3})$ regret bound.

Corollary 3 (Regret bound with benchmark class Φ_1.)**.** *Under the conditions in Theorem 4, our Algorithm 1 satisfies*

$$\sum_{t=0}^{T-1} l(x_t, u_t) - T \min_{\mathbf{M}\in\Phi_1} J(\mathbf{M}) \leq \tilde{O}(T^{2/3}),$$

where x_t, u_t are generated by Algorithm 1.

The proof is basically the same with the proof of Theorem 4. The only difference is on the bound Part iv. By defining a truncated version of $\mathbf{M}^*$, a similar bound on Part iv can be obtained using similar proof techniques.

F.2 Regret Analysis with RMPC in [30] as the Benchmark

F.2.1 A brief review of RMPC in [30]

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 [40] for a review). In this appendix, we will focus on a tube-based RMPC defined in [30]. The RMPC method in [30] 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 [30]), 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 6. *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 $V_f(x) = x^\top Px$ satisfying certain conditions (see [30] for more details).

RMPC scheme in [30]. Now, we are ready to define the tube-based RMPC proposed in [30]. At each stage t, consider a planning window $t+k|t$ for $0 \leq k \leq W$, RMPC in [30] solves the following optimization:

$$\begin{aligned} \min_{x_{t|t}, u_{t+k|t}} \quad & \sum_{k=0}^{W-1} l(x_{t+k|t}, u_{t+k|t}) + V_f(x_{t+W|t}) \\ \text{s.t.} \quad & x_{t+k+1|t} = A_* \bar{x}_{t+k|t} + B_* u_{t+k|t}, \quad k \geq 0 \\ & x_{t|t} \in x_t \oplus \Xi \\ & x_{t+k|t} \in \mathbb{X} \ominus \Xi, \forall 0 \leq k \leq W-1 \\ & u_{t+k|t} \in \mathbb{U} \ominus \mathbb{K}\Xi, \forall 0 \leq k \leq W-1 \\ & x_{t+W|t} \in X_f \subseteq \mathbb{X} \ominus \Xi \end{aligned} \tag{RMPC [30]}$$

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 [6], 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

$$X_N = \{x_0 : \text{(RMPC [30]) is feasible when } x_t = x_0\}.$$

The RMPC scheme in [30] 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 [30] enjoys robust exponential stability.

Theorem 8 (Theorem 1 in [30])**.** *The set Ξ is robustly exponentially stable for the closed-loop system with* (RMPC [30]) *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 8 suggests that (RMPC [30]) 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.

F.2.2 Infinite-horizon Cost of RMPC in [30] and Regret Compared with (RMPC [30])

In the following, we will consider a broader benchmark policy class that includes (RMPC [30]) reviewed above and still provide a $\tilde{O}(T^{2/3})$ regret bound. This is possible because we establish a connection from the infinite-horizon averaged cost of (RMPC [30], which employs nonlinear policies, to the infinite-horizon averaged cost of linear static controllers. This connection is built upon the robust exponential stability property.

Theorem 9 (Connection between RMPC in [30] and linear control's infinite-horizon costs)**.** *Consider (RMPC [30]) defined above with $\mathbb{K}$ satisfying the requirements in [30]. For any $x_0 \in X_N$, the infinite-horizon averaged cost of (RMPC [30]) equals the infinite-horizon averaged cost of $\mathbb{K}$, i.e.*

$$J(\text{(RMPC [30])}) = J(\mathbb{K}),$$

The proof is deferred to Appendix F.2.4.

Based on Theorem 9, we can show that our Algorithm 1 achieves $\tilde{O}(T^{2/3})$ regret even when compared with (RMPC [30]).

Corollary 4. *Under the conditions in Theorem 4, for any (RMPC [30]) with admissible parameters required by [30], we have*

$$\sum_{t=0}^{T-1} l(x_t, u_t) - TJ((\text{RMPC [30]})) \leq \tilde{O}(T^{2/3}),$$

where x_t, u_t are generated by Algorithm 1.

Remark 3. *Since our proof relies on the robust exponential stability property of RMPC in [30], 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., [48, 35]. However, [48] allows constraint violation during the learning process and allows restarts when policies are updated, and [35] 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.*

F.2.3 Handling non-zero initial value $x_0 \neq 0$

In the main body of this paper, we assume $x_0 = 0$ for simplicity and focus on optimizing the performance around the origin. For non-zero initial values, there is a rich literature on how to safely drive a nonzero x_0 to a neighborhood around the origin, e.g. RMPC [40, 30]. Therefore, it is a natural idea to combine our algorithm and RMPC to handle non-zero initial values. That is, to apply RMPC at first to steer a nonzero x_0 to a neighborhood around the origin and then switch to our Algorithm 1 to optimize the performance around the origin. Though we only reviewed (RMPC [30]) when the model θ_* is known, the method (RMPC [30]) can be easily extended to handle model uncertainties Θ, for example, by (i) replacing θ_* in the constraints as $\hat{\theta}$, (ii) enlarging the disturbance-invariant set Ξ to a robust disturbance-invariant set that allows for model uncertainties in Θ, (iii) requiring $\mathbb{K}$ to be strictly safe for all possible models in Θ, (iv) reducing X_f such that it is safe to implement $u_t = -\mathbb{K}x_t$ for all possible models in Θ starting from X_f. One can still establish robust exponential stability of (RMPC [30]) under model uncertainties. Based on the robust exponential stability, it can be shown that the state x_t enters X_f in finite steps. After entering X_f, we can switch to linear static controller $u_t = -\mathbb{K}x_t$ without violating any constraints (this is a property of RMPC). Then, we can safely switch from $\mathbb{K}$ to $\mathbf{M}_{\dagger}^{(0)}$ by the safe transition method described in Appendix D.3. It can be shown that the number of stages before implementing Algorithm 1 is finite so it will not affect our regret bound's dependence on T.

F.2.4 Proof of Theorem 9

To prove Theorem 9, 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 [6].

Proposition 2 (Corollary 2 and Theorem 4 and Section 4.4 in [6])**.** *Consider (CLQR) with p.d. quadratic costs and polytopic constraints below:*

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

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 can show $\pi_{CLQR}(x)$ is Lispchitz continous with Lipschitz factor $L_{CLQR} = \max_i \|K_i\|_2$.

Lemma 25. *$\pi_{CLQR}(x)$ is Lispchitz continous with Lipschitz factor $L_{CLQR} = \max_i \|K_i\|_2$.*

The proof is deferred to Appendix G.

Next, we will use the exponential convergence results of RMPC in [30].

Proposition 3 (See the proof of Theorem 1 in [30])**.** *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*

$$\|x^*_{t|t}(x_t)\|_2 \le c_1 \rho^t \|x^*_{0|0}(x_0)\|_2.$$

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

Lemma 26. *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*

$$\|u^*_{t|t}(x_t)\|_2 \le c_2 \rho^t,$$

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

Proof. First of all, we point out that for the (RMPC [30]) 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 [30]). 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,

$$\|u^*_{t|t}(x^*_{t|t})\|_2 = \|u^*_{t|t}(x^*_{t|t}) - u^*_{t|t}(0)\|_2 \le L_{rmpc}\|x^*_{t|t}\|_2 \le L_{rmpc} c_1 \rho^t \|x^*_{0|0}(x_0)\|_2 \le c_2 \rho^t$$

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

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

Lemma 27. *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*

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

Now, we are ready for the proof of Theorem 9.

Proof of Theorem 9. The closed-loop system of (RMPC [30]) 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

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

where $A_{\mathbb{K}} = A_* - B_* \mathbb{K}$, and $\upsilon_t = B_*(\mathbb{K} x^*_{t|t}(x_t) + u^*_{t|t}(x_t))$. Notice that by Proposition 3 and Lemma 26, we have

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

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

Therefore,

$$\begin{aligned}\|e_t\|_2 &= \|\upsilon_{t-1} + A_{\mathbb{K}}\upsilon_{t-2} + A_{\mathbb{K}}^{t-1}\upsilon_0\|_2 \\ &\leq c_3\rho^{t-1} + \kappa^2(1-\gamma)c_3\rho^{t-2} + \dots \\ &\leq c_3\kappa^2 t\max(\rho, 1-\gamma)^{t-1} = c_4 t\rho_0^{t-1}\end{aligned}$$

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

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

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,

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

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 Technical proofs

This appendix includes the proofs of the technical lemmas used in this paper.

G.1 Proof of Lemma 4

The proof relies on the following two lemmas.

Lemma 28 (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.

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

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

$\square$

Lemma 29 (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 30. *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 31. *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 $k \geq 1$,

$$
\begin{aligned}
\|\hat{A}^k - A^k\|_2 &= \|\sum_{i=0}^{k-1} \hat{A}^{k-i-1} (\hat{A} - A) A^i\|_2 \\
&\leq \sum_{i=0}^{k-1} \|\hat{A}^{k-i-1}\|_2 \|\hat{A} - A\| \|A^i\|_2 \\
&\leq \sum_{i=0}^{k-1} \kappa(1-\gamma)^{k-i-1} \epsilon \kappa (1-\gamma)^i \\
&= k\kappa^2 r (1-\gamma)^{k-1} \\
\|\hat{A}^k \hat{B} - A^k B\|_2 &\leq \|\hat{A}^k \hat{B} - A^k \hat{B}\|_2 + \|A^k \hat{B} - \hat{A}^k \hat{B}\|_2 \\
&\leq k\kappa^2 \kappa_B r (1-\gamma)^{k-1} \mathbb{1}_{(k \geq 1)} + \kappa(1-\gamma)^k r
\end{aligned}
$$

$\square$

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.

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

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

□

G.2 Proof of Lemma 8

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

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

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

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 10

Case 1: during the transition from $\mathbf{M}_*^{(e-1)}$ to $\mathbf{M}_\dagger^{(e)}$. For any $e \geq 1$, during the transition from $\mathbf{M}_*^{(e-1)}$ to $\mathbf{M}_\dagger^{(e)}$. When $T^{(e)} \leq t \leq T^{(e)} + W_1 - 1$, $\mathbf{M}_t \in \Omega^{(e-1)}$, thus, $g_j^u(\mathbf{M}_t) \leq d_{u,j} - 0 = d_{u,j}$. By (4) and $\eta_t = 0$, we have

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

for all j, so $u_t \in \mathbb{U}$. When $T^{(e)} + W_1 \leq t \leq T^{(e)} + W_1 + W_2 - 1$, $\mathbf{M}_t \in \Omega_\dagger^{(e)}$, so $g_j^u(\mathbf{M}_t) \leq d_{u,j} - \epsilon_{\eta,u}(\bar{\eta}^{(e)}) \leq d_{u,j}$. By (4), and $\eta_t = 0$,

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

for all j, so $u_t \in \mathbb{U}$.

Case 2: during CCE with safe exploration. When $t_1 \leq t \leq t_1 + T_D^{(e)} - 1$, $\mathbf{M}_t \Omega_\dagger^{(e)}$, so $g_j^u(\mathbf{M}_t) \leq d_{u,j} - \epsilon_{\eta,u}(\bar{\eta}^{(e)}) \leq d_{u,j}$. By (4), and $\|\eta_t\|_\infty \leq \bar{\eta}^{(e)}$,

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

for all j, so $u_t \in \mathbb{U}$.

Case 3: during the transition from $\mathbf{M}_\dagger^{(e)}$ to $\mathbf{M}_*^{(e)}$. When $t_1 + T_D^{(e)} \le t \le t_1 + T_D^{(e)} + W'_{s_1} - 1$, $\mathbf{M}_t \in \Omega_\dagger^{(e)}$, thus, $g_j^u(\mathbf{M}_t) \le d_{u,j} - \epsilon_{\eta,u}(\bar{\eta}^{(e)}) < d_{u,j}$. By (4), and $\eta_t = 0$, we have

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

for all j, so $u_t \in \mathbb{U}$. When $t_1 + T_D^{(e)} + W'_{s_1} \le t \le t_1 + T_D^{(e)} + W'_{s_1} + W'_{s_2} - 1$, $\mathbf{M}_t \in \Omega^{(e)}$, so $g_j^u(\mathbf{M}_t) \le d_{u,j}$. By (4), and $\eta_t = 0$,

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

for all j, so $u_t \in \mathbb{U}$.

Case 4: full exploitation. When $t_2 \le t \le T^{(e+1)} - 1$, $\mathbf{M}_t \in \Omega^{(e)}$, so $g_j^u(\mathbf{M}_t) \le d_{u,j}$. By (4) and $\eta_t = 0$,

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

for all j, so $u_t \in \mathbb{U}$.

G.4 Proof of Lemma 11

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

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

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

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

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

G.5 Proof of Lemma 15

Proof. For notational simplicity, we omit $\mathbf{M}$ in this proof. We denote $\tilde{x}(\theta)$ and $\tilde{x}(\hat{\theta})$ as approximate states. Notice that,

$$f(\theta) = \mathbb{E}\, l(\tilde{x}(\theta), \tilde{u}(\theta)),\ f(\hat{\theta}) = \mathbb{E}\, l(\tilde{x}(\hat{\theta}), \tilde{u}(\hat{\theta}))$$

When $\mathbf{M} \in \mathcal{M}_H$, $H \ge \log(2\kappa)/\log((1-\gamma)^{-1})$, by Lemma 24 in technical report.pdf, we have $\|\tilde{x}(\theta)\|_2 \le O(\sqrt{mn})$, and $\tilde{u}(\theta) \in \mathbb{U}$.

Next,

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

where the last inequality uses Lemma 15 in technical report.pdf.

Notice that $u(\hat{\theta}) = u(\theta)$.

Now,

$$\begin{aligned}|f(\mathbf{M};\theta) - f(\mathbf{M};\hat{\theta})| &= |\,\mathbb{E}(l(\tilde{x}(\theta), \tilde{u}(\theta)) - l(\tilde{x}(\hat{\theta}), \tilde{u}(\hat{\theta})))| \\ &\le \mathbb{E}\,|l(\tilde{x}(\theta), \tilde{u}(\theta)) - l(\tilde{x}(\hat{\theta}), \tilde{u}(\hat{\theta}))| \\ &\le \mathbb{E}(\tilde{x}(\theta) + \tilde{x}(\hat{\theta}))^\top Q(\tilde{x}(\theta) - \tilde{x}(\hat{\theta})) + (\tilde{u}(\theta) + \tilde{u}(\hat{\theta}))^\top R(\tilde{u}(\theta) - \tilde{u}(\hat{\theta})) \\ &\le \mathbb{E}\, O(\sqrt{mn}\sqrt{mn}r_\theta) \le O(mnr_\theta)\end{aligned}$$

□

G.6 Proof of Lemma 16

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

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

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

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

$$\|\tilde{x} - \tilde{x}'\|_2 \le \sum_{k=1}^{2H}\|(\Phi_k^x(\mathbf{M}) - \Phi_k^x(\mathbf{M}'))w_{t-k}\|_2$$

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

$$
\|\tilde{u} - \tilde{u}'\|_2 \sum_{k=1}^{H} \|M[k] - M'[k]\|_2 \sqrt{n} w_{\max} \le O(\sqrt{n}\sqrt{H}) \|\mathbf{M} - \mathbf{M}'\|_F
$$

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, $\|\tilde{x}\|_2 \le \sqrt{mn}$. Further, even when we don't have additional conditions on $\mathbf{M}$, we still have $\|\tilde{u}\|_2 \le O(\sqrt{mn})$. Therefore, for $\mathbf{M}, \mathbf{M}' \in \mathcal{M}_{out,H}$,

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

Therefore,

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

□

G.7 Proof of Lemma 18

Proof. Notice that Ω_1 and Ω_3 satisfies the conditions in Proposition 2 in [24]. Therefore,

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

Notice that

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

therefore, $\|\Delta_1 - \Delta_3\|_\infty \le \|\Delta_1 - \Delta_2\|_\infty$. Further, $\{i : (\Delta_3)_i > (\Delta_1)_i\} = \{i : (\Delta_2)_i > (\Delta_1)_i\} \subseteq \{i : (\Delta_1)_i \ne (\Delta_2)_i\}$. So $\min_{\{i:(\Delta_3)_i > (\Delta_1)_i\}} (h - \Delta_1 - C x_F)_i \ge \min_{\{i:(\Delta_1)_i \ne (\Delta_2)_i\}} (h - \Delta_1 - C x_F)_i \ge \min_{\{i:(\Delta_1)_i \ne (\Delta_2)_i\}} (h - \Delta_3 - C x_F)_i$. Therefore,

$$
|\min_{\Omega_1} f(x) - \min_{\Omega_3} f(x)| \le \frac{L d_{\Omega_0} \|\Delta_1 - \Delta_3\|_\infty}{\min_{\{i:(\Delta_1)_i > (\Delta_3)_i\}} (h - \Delta_1 - C x_F)_i} \le \frac{L d_{\Omega_0} \|\Delta_1 - \Delta_2\|_\infty}{\min_{\{i:(\Delta_1)_i \ne (\Delta_2)_i\}} (h - \Delta_3 - C x_F)_i}
$$

Similarly,

$$
|\min_{\Omega_2} f(x) - \min_{\Omega_3} f(x)| \le \frac{L d_{\Omega_0} \|\Delta_2 - \Delta_3\|_\infty}{\min_{\{i:(\Delta_2)_i > (\Delta_3)_i\}} (h - \Delta_2 - C x_F)_i} \le \frac{L d_{\Omega_0} \|\Delta_1 - \Delta_2\|_\infty}{\min_{\{i:(\Delta_1)_i \ne (\Delta_2)_i\}} (h - \Delta_3 - C x_F)_i}
$$

which completes the bound. □

G.8 Proof of Lemma 21

Lemma 32. *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: $\mathbf{M}_*^{(e)} \in \mathcal{F}(\hat{\theta}^{(e+1)}) = \mathcal{F}(\{z_k, x_{k+1}\}_{k=t_1^{(e)}}^{t_1^{(e)}+T_D^{(e)}-1}) = \mathcal{F}(w_0, \ldots, w_{t_1^{(e)}+T_D^{(e)}-1}, \eta_0, \ldots, \eta_{t_1^{(e)}+T_D^{(e)}-1}) = \mathcal{F}^m_{t_1^{(e)}+T_D^{(e)}}$. By $\tilde{W}_1^{(e)} \geq H^{(e)}$, we have $t_1^{(e)} + T_D^{(e)} + H^{(e)} \leq t_2^{(e)}$, and since $\mathcal{F}_t^m \subseteq \mathcal{F}_t$, we have the last claim. $\square$

Lemma 33. *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 $w_{t-2H^{(e)}} \perp\!\!\!\perp \mathcal{F}_{t_2^{(e)}-H^{(e)}}$. $\square$

Lemma 34. *In our Algorithm 1, when* $t \in \mathcal{T}_2^{(e)}$, *we have*

$$\mathbb{E}\, l(\tilde{x}_t, \tilde{u}_t) \mid \mathcal{F}_{t_2^{(e)}-H^{(e)}} = f(\mathbf{M}_*^{(e)}; \theta_*)$$

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 $\tilde{x}_t, \tilde{u}_t$ and $f(\mathbf{M}; \theta_*)$, we have the result. $\square$

Definition 7 (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 5. $\{X_t\}_{t\geq 0}$ *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. Let $\alpha = \sqrt{2t\sigma^2 \log(2/\delta)}$, then we are done. $\square$

Lemma 35. *Define* $q_t = l(\tilde{x}_t, \tilde{u}_t) - f(\mathbf{M}_*^{(e)}; \theta_*)$. *Then,* $|q_t| \leq O(mn)$ *w.p.1.*

Proof. By Lemma 24 in technical report, we have $|l(\tilde{x}_t, \tilde{u}_t)| = O(mn)$. Since $f(\mathbf{M}_*^{(e)}; \theta_*) = \mathbb{E}\, l(\tilde{x}_t, \tilde{u}_t) \mid \mathcal{F}_{t_2^{(e)}-H^{(e)}}$, we have $|f(\mathbf{M}_*^{(e)}; \theta_*)| = O(mn)$. This completes the proof. $\square$

We will define many important concepts!

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

$$\mathcal{T}_{2,h}^{(e)} = \{t \in \mathcal{T}_2^{(e)} : t \equiv h \mod (2H^{(e)})\} =: \{t_h^{(e)} + 2H^{(e)}, \ldots, t_h^{(e)} + 2H^{(e)}k_h^{(e)}\} \tag{23}$$

Lemma 36. $t_h^{(e)} \geq t_2^{(e)} - H^{(e)}$ *and* $k_h^{(e)} \leq T^{(e+1)}/(2H^{(e)})$

Proof. $t_h^{(e)} + 2H^{(e)} \geq t_2^{(e)} + H^{(e)}$, so the first inequality holds.

$2H^{(e)}k_h^{(e)} \leq t_h^{(e)} + 2H^{(e)}k_h^{(e)} \leq T^{(e+1)}$, so the second inequality holds.

$\square$

Define

$$\tilde{q}_{h,j}^{(e)} = q_{t_h^{(e)}+j(2H^{(e)})}, \quad \forall 1 \le j \le k_h^{(e)} \tag{24}$$

Define

$$S_{h,j}^{(e)} = \sum_{s=1}^{j} \tilde{q}_{h,s}^{(e)}, \quad \forall 0 \le j \le k_h^{(e)} \tag{25}$$

we define $\sum_{s=1}^{0} a_s = 0$.

Define

$$\mathcal{F}_{h,j}^{(e)} = \mathcal{F}_{t_h^{(e)}+j(2H^{(e)})}, \quad \forall 0 \le j \le k_h^{(e)} \tag{26}$$

By Lemma 36, $\mathcal{F}_{h,0}^{(e)} = \mathcal{F}_{t_h^{(e)}} \supseteq \mathcal{F}_{t_2^{(e)}-H^{(e)}}$.

Lemma 37. *$S_{h,j}^{(e)}$ 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 $t \in \mathcal{T}_2^{(e)}$, $w_{t-1}, \ldots, w_{t-2H^{(e)}} \in \mathcal{F}_t$. 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 38. *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 37, we can apply Corollary 5, and obtain the bound, where we used $\log(2/\delta) = \tilde{O}(1)$. □

Lemma 39. *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 36 and Cauchy Schwartz.

Now,

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

□

Now, we can prove Lemma 21. By Lemma 39, w.p $1-p$,

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

for all e. Then, by Lemma 13, we have the bound.

G.9 Proof of Lemma 25

Proof. For any two points $x_1, x_2 \in X_N$, consider the line segment $r(s) = x_1 + s(x_2 - x_1)$ for $s \in [0,1]$. Note that $r(0) = x_1$ and $r(1) = x_2$. This line segment goes through a finite number of regions. Denote the points on this line segment that are on the boundary of at two regions as $r(s_1), \ldots, r(s_H)$, for $0 \le s_1 < \cdots < s(H) \le 0$. For each i, we call the two regions that $r_(s_i)$ belongs to as region i and $i+1$.

$$\begin{aligned}
\|\pi_{CLQR}(x_1) - \pi_{CLQR}(x_1)\|_2 &= \|\sum_{i=1}^{H+1} \pi_{CLQR}(r(s_i)) - \pi_{CLQR}(r(s_{i-1}))\|_2 \\
&= \|\sum_{i=1}^{H+1} K_i(r(s_i) - r(s_{i-1}))\|_2 \\
&\le \sum_{i=1}^{H+1} \|K_i\|_2 \|(r(s_i) - r(s_{i-1}))\|_2 \\
&\le \max_i \|K_i\|_2 \sum_{i=1}^{H+1} (s_i - s_{i-1})\|x_2 - x_1\|_2 \\
&= \max_{0 \le i \le N_{clqr}} \|K_i\|_2 \|x_2 - x_1\|_2
\end{aligned}$$

□

Potential Negative Societal Impacts

In this work, we develop a learning-based safe control algorithm, which ensures that the generated control policies satisfy the constraints even under model uncertainties and disturbances. Most practical systems, such as autonomous vehicles and robotics, have to satisfy certain constraints on the states and actions. Thus our algorithm can potentially be very beneficial for plenty of safety-critical applications. However, note that our algorithm relies on a set of technical assumptions mentioned in the paper. These assumptions may not directly hold for all practical applications. Hence, if one uses our algorithm in practice, one has to carefully verify the assumptions or be more conservative than the schemes designed in this paper, otherwise our safety guarantees may not hold.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 1, 2, 4, 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 1
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2, 5
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 5 and Supplementary
3. If you ran experiments: It is a theoretical paper. We do not run experiments.
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

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

[2] 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), 2019.

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

[4] James Anderson, John C Doyle, Steven H Low, and Nikolai Matni. System level synthesis. *Annual Reviews in Control*, 47:364–393, 2019.

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

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

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

[8] 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.

[9] 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.

[10] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in neural information processing systems*, pages 8092–8101, 2018.

[11] 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.

[12] 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.

[13] 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, 2019.

[14] 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, 2019.

[15] 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, 2019.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] Wilbur Langson, Ioannis Chryssochoos, SV Raković, and David Q Mayne. Robust model predictive control using tubes. *Automatica*, 40(1):125–133, 2004.

[23] 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.

[24] Yingying Li, Subhro Das, and Na Li. Online optimal control with affine constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8527–8537. AAAI Press, 2021.

[25] 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.

[26] D Limon, I Alvarado, Teodoro Alamo, and EF Camacho. On the design of robust tube-based mpc for tracking. *IFAC Proceedings Volumes*, 41(2):15333–15338, 2008.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[34] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

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

[36] Marko Nonhoff and Matthias A Müller. Data-driven online convex optimization for control of dynamical systems. *arXiv preprint arXiv:2103.09127*, 2021.

[37] 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.

[38] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with Thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

[39] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*, 2019.

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

[41] Harold E Roland and Brian Moriarty. *System safety engineering and management*. John Wiley & Sons, 1990.

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

[43] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550:354–359, 2017.

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

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence. AAAI Press, online*, 2021.

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