Safe Adaptive Learning for Linear Quadratic Regulators with Constraints

Yingying Li 1  Tianpeng Zhang 2 Subhro Das 3 Jeff Shamma 1 Na Li 2

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

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

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

\[
D_x x_t \leq d_x, \quad D_u u_t \leq d_u.
\]

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

The constraints on LQR bring great difficulties even when the model is known. Unlike unconstrained LQR, which enjoys closed-form optimal policies (Lewis et al., 2012), there is no computationally efficient method to solve the optimal policy for constrained LQR (Rawlings & Mayne, 2009). Thus, most literature sacrifices optimality for computation efficiency by designing policies with certain structures, e.g. linear policies (Dean et al., 2019b; Li et al., 2021), piecewise-affine policies in robust model predictive control (RMPC) (Bemporad & Morari, 1999; Rawlings & Mayne, 2009), etc. Therefore, when the model is unknown, a reasonable goal is to learn and achieve what can be obtained with perfect model information. In this paper, we adopt the optimal safe linear policy as our benchmark/target and leave the discussions on RMPC as future work.

The current literature on learning an optimal safe linear policy adopts an offline/non-adaptive learning approach, which does not improve the policies until the learning terminates (Dean et al., 2019b). To improve the control performance during learning, adaptive/online learning-based control algorithms should be designed. However, though adaptive learning for unconstrained LQR can be designed by direct conversions from offline algorithms (see e.g., (Simchowitz & Foster, 2020; Mania et al., 2019; Dean et al., 2018)), it is much more challenging for the constrained case because direct conversions may cause infeasibility and/or constraint violation for single-trajectory adaptive learning as noted in (Dean et al., 2019b).

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

Theoretically, for safety, we guarantee feasibility and con-

1University of Illinois at Urbana-Champaign 2Harvard University 3MIT-IBM Watson AI Lab, IBM Research. Correspondence to: Yingying Li <yl101@illinois.edu>.

1Efficient algorithms exist for some special cases, e.g. when \(w_t = 0\), the optimal controller is piecewise-affine and can be computed as in (Bemporad et al., 2002).
Safe Adaptive Learning for Linear Quadratic Regulators with Constraints

Constrained LQR by model predictive control (MPC). There is an interesting recent paper (Dogan et al., 2021) considering an unknown model and time-varying cost functions. However, there are limited results on hard constraint satisfaction but lack non-asymptotic optimality bounds. In contrast, there are some recent results on non-asymptotic optimality guarantees compared with the known-model case.

Oldewurtel et al., 2008). With model uncertainties, robust stochastic MPC methods for soft constraints (Mesbah, 2016; Limon et al., 2010; Rawlings & Mayne, 2009), and its variants are popular methods for constrained control, (Dean et al., 2018; Mania et al., 2019; Cohen et al., 2019; Simchowitz & Foster, 2020), and this approach is shown to be optimal for the unconstrained LQR (Mania et al., 2019; Simchowitz & Foster, 2020). Further, similar to (Agarwal et al., 2019a;b; Plevrakis & Hazan, 2020), we adopt the disturbance-action policies to approximate linear policies.

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

Related work. Constrained LQR with linear policies is studied in (Dean et al., 2019b; Li et al., 2021). Dean et al. (2019b) consider an unknown model and propose an offline learning method with sample complexity guarantees. In contrast, (Li et al., 2021) study online constrained LQR with a known model and time-varying cost functions. However, it remains open how to design online learning algorithms to compute safe linear policies under model uncertainties.

Constrained LQR by model predictive control (MPC). MPC and its variants are popular methods for constrained control, e.g. RMPC designed for hard constraints (Mayne et al., 2005; Limon et al., 2010; Rawlings & Mayne, 2009), and stochastic MPC methods for soft constraints (Mesbah, 2016; Oldewurtel et al., 2008). With model uncertainties, robust adaptive MPC (RAMPC) algorithms are proposed to learn the model and update the policies (Zhang & Shi, 2020; Bujarbaruah et al., 2019; Köhler et al., 2019; Lu et al., 2019). Most RAMPC algorithms guarantee recursive feasibility and constraint satisfaction but lack non-asymptotic performance guarantees compared with the known-model case. There is an interesting recent paper (Dogan et al., 2021) that provides a regret bound for learning-based MPC, but its benchmark policy is conservative since it is robustly safe for all uncertain models instead of just for the true model. In contrast, there are some recent results on non-asymptotic regret bounds by sacrificing feasibility and/or constraint satisfaction, e.g., (Wabersich & Zeilinger, 2020) establish a regret bound for an adaptive MPC algorithm that requires restarting the system to some safe feasible state, (Muthirayan et al., 2020) provides a regret bound for an adaptive algorithm without considering state constraints.

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

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

Model estimation for nonlinear systems. There are model estimation bounds for general nonlinear systems (Foster et al., 2020; Sattar & Oymak, 2020), but our estimation error bound leverages the special structure of our problem: nonlinear policies. This is to handle the potential nonlinearity in our designed controllers when the model errors are non-negligible. Our error bound extends the existing results for linear policies in (Dean et al., 2019a;b) and can be useful on its own.

2. Problem formulation

We consider the following constrained LQR problem,

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

where \(l(x,u) = x^T Q x + u^T R u\), \(Q\) and \(R\) are positive definite matrices, \(x_t \in \mathbb{R}^n\) is the state with a given initial state \(x_0\), \(u_t \in \mathbb{R}^m\) is the action, and \(w_t\) is the disturbance. The parameters \(D_x, d_x\), and \(D_u, d_u\) determine the constraint sets of the state and action respectively, where \(d_x \in \mathbb{R}^k, d_u \in \mathbb{R}^k\). Further, the constraint sets on the state and action are assumed to be bounded with non-negligible. Our error bound extends the existing results for linear policies in (Dean et al., 2019a;b) and can be useful on its own.
As an example of applications, we study how to safely maintain the performance of our adaptive learning algorithm by comparing it with the optimal safe linear policy and piecewise affine policies obtained by RMPC (Mayne et al., 2005; Rawlings & Mayne, 2009). In this paper, we focus on the optimal safe linear policy as our learning goal and briefly discuss RMPC in Section 6. We aim to achieve our learning goal by designing safe adaptive learning-based control. Further, we consider single-trajectory learning, which is more challenging since the system cannot be restarted to ensure feasibility and constraint satisfaction.

For simplicity, we assume $x_0 = 0$. Our results can be generalized to $x_0$ in a small neighborhood around 0.\footnote{The appendix will discuss more on nonzero $x_0$. Here, we discuss the implication of small $x_0$. Remember that state 0 represents a desirable system equilibrium. With $x_0$ close to 0, we study how to safely control the performance around the equilibrium instead of safely steering a distant state back to the equilibrium. As an example of applications, we study how to safely maintain a drone around a target in the air despite wind disturbances with minimum battery consumption, instead of flying the drone to the target from a distance. In practice, one can first apply algorithms such as Mayne et al. (2005) to drive the system to around 0, then apply our algorithm to achieve optimality and safety around 0.\cite{Mayne2005}}

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

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

**Definition 2.1.** For $\kappa \geq 1$, $\gamma \in [0, 1)$, a matrix $A$ is called $(\kappa, \gamma)$-stable if $\|A^T\|_2 \leq (1 - \gamma)^\frac{1}{\kappa}$, $\forall t \geq 0$.

Consider the following benchmark policy set:

$$\mathcal{K} = \{ K : (A_s - B_s K) \text{ is } (\kappa, \gamma)\text{-stable}, \|K\|_2 \leq \kappa, D_xx^K_t \leq d_x, D_u u^K_t \leq d_u, \forall t, \forall \{w_k \in \mathcal{W}\} \}_{k \geq 0}.$$ 

where $x^K_t, u^K_t$ are generated by policy $u_t = -K^*x_t$.

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

$$\text{Regret} = \sum_{t=0}^{T-1} l(x^A_t, u^A_t) - T \min_{K \in \mathcal{K}} J(K)$$

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

Next, we provide and discuss the assumptions.

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

**Assumption 2.2.** There is a known model uncertainty set $\Theta_{\text{ini}} = \{ \theta : \|\theta - \theta_\star\|_F \leq r_{\text{ini}} \}$ for some $0 < r_{\text{ini}} < +\infty$ such that (i) $\theta_\star \in \Theta_{\text{ini}}$, and (ii) there exist $\kappa \geq 1$, $\gamma \in [0, 1)$, and $K_{\text{stab}}$ such that for any $(A, B) \in \Theta_{\text{ini}}$, $A - BK_{\text{stab}}$ is $(\kappa, \gamma)$-stable.

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

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

**Assumption 2.3.** There exists $K_F \in \mathcal{K}$ and $\epsilon_{F,w} > 0, \epsilon_{F,u} > 0$ such that $D_xx^K_F \leq d_x - \epsilon_{F,w}1_{k_x}$ and $D_u u^K_F \leq d_u - \epsilon_{F,u}1_{k_u}$ for all $t \geq 0$ under all $w_k \in \mathcal{W}$.

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

**Definition 2.4 (Anti-concentration).** A random vector $w \in \mathbb{R}^n$ satisfies $(s, p)$-anti-concentration for some $s > 0, p \in (0, 1)$ if $P(\lambda^T w \geq s) \leq p$ for any $\|\lambda\|_2 = 1$.

**Assumption 2.5.** $w_t \in \mathcal{W}$ is i.i.d., $\sigma_{sub}^2$-sub-Gaussian, zero mean, and $(s_w, p_w)$-anti-concentration.$^3$

$^2$The symmetry of $\Theta_{\text{ini}}$ is assumed for simplicity and not restrictive. We only need $\Theta_{\text{ini}}$ to be small and contain $\theta_\star$.

$^3$By $\mathcal{W} = \{w : \|w\|_\infty \leq w_{\text{max}}\}$, we have $\sigma_{sub} \leq \sqrt{n}w_{\text{max}}$. 

$\sigma_{sub}$ is the sub-Gaussian scale parameter of $w$.
Remark 2.6 (On bounded disturbances). Here, we assume bounded disturbances because we aim to achieve constraint satisfaction despite any disturbances, which is generally impossible for unbounded disturbances. Nevertheless, we can allow Gaussian disturbances by considering chance constraints as discussed in Oldewurtel et al. (2008).

2.1. Preliminaries

2.1.1 Approximation of optimal safe linear polices with disturbance-action policies. This section reviews a computation-efficient method for approximating the optimal safe linear policy when the model \( \theta_s \) is known (Li et al., 2021). The method is based on the disturbance-action policy (DAP) defined below.

\[
u_t = -K_{\text{stab}}x_t + \sum_{k=1}^{H} M[k]u_{t-k}.
\]

where \( M = \{M[1], \ldots, M[H]\} \) denote the policy parameters and \( H \geq 1 \) denotes the policy’s memory length.

Roughly, Li et al. (2021) show that the optimal safe linear policy can be approximated by the optimal safe DAP for large \( H \), and the optimal safe DAP can be solved efficiently by a quadratic program (QP). More details are as follows.\(^6\)

Firstly, Proposition 2.7 shows that the state and action can be approximated by affine functions of DAP parameters \( M \).

Proposition 2.7 ((Agarwal et al., 2019a)). Under a time-invariant DAP \( M \), the state \( x_t \) and action \( u_t \) can be approximated by affine functions on \( M \):

\[
\tilde{x}_t(M; \theta_s) = \sum_{k=1}^{2H} \Phi_k \tilde{x}_t(M; \theta_s) u_{t-k},
\]

where \( \tilde{x}_t(M; \theta_s) = \sum_{k=1}^{H} \Phi_k \tilde{x}_t(M; \theta_s) u_{t-k} \), and \( A_{k,\text{stab}} \Phi_k(M; \theta_s) = A_{k,\text{stab}} \Phi_k(M; \theta_s) \) for \( 1 \leq k \leq H \).\(^4\)

Secondly, we review the polytopic safe policy set (SPS) in Li et al. (2021) by imposing constraints on the approximate states and actions and by tightening the constraints to allow for approximation errors. Specifically, define constraint functions on \( \tilde{x}_t, \tilde{u}_t \), i.e.,

\[
g^0_t(M; \theta_s) = \sup_{u \in \Omega(M)} D^T_t \tilde{x}_t(M; \theta_s) = \sum_{k=1}^{2H} \| D^T_t \tilde{x}_t(M; \theta_s) \|_2 \max \leq 1 \leq k_t \) and \( g^0_t(M; \theta_s) = \sup_{u \in \Omega(M)} D^T_t \tilde{u}_t(M; \theta_s) = \sum_{k=1}^{2H} \| D^T_t \tilde{u}_t(M; \theta_s) \|_2 \max \leq 1 \leq k_u \). The SPS in Li et al. (2021) is defined as follows.

\[
\Omega(\theta_s, \epsilon_x, \epsilon_u) = \{ M \in \mathcal{M}_H : g^0_t(M; \theta_s) \leq d_{x,t} - \epsilon_x, \epsilon_u, 1 \leq i \leq k_x, g^0_t(M; \theta_s) \leq d_{u,i} - \epsilon_u, 1 \leq j \leq k_u \},
\]

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

Finally, we review the QP reformulation for the optimal safe DAP \( M^* \) when the model \( \theta_s \) is available.

\[
M^* = \arg \min_{M \in \Omega(\theta_s, \epsilon_x, \epsilon_u)} f(M; \theta_s) = E[(\tilde{x}_t(M; \theta_s), \tilde{u}_t(M; \theta_s)]
\]

Notice that \( f(M; \theta_s) \) is a quadratic convex function of \( M \).

Further, Li et al. (2021) showed that the optimal safe DAP \( M^* \) approximates the optimal safe linear policy \( K^* \) with error \( J(K^*) - J(K^*) \leq O((1-\gamma)^H) \).

2.1.2 Safe slowly varying policies. Notice that the SPS in (3) only considers a time-invariant policy \( M \). Nevertheless, Li et al. (2021) show that with additional constraint-tightening terms \( \epsilon^*_x(D_M), \epsilon^*_u(D_M) \) to allow for small policy variation \( \Delta_M \), the SPS can also be used to guarantee the safety of slowly varying policy sequences, which is called a slow-variation trick.

Lemma 2.8 (Slow-variation trick (Li et al., 2021)). Consider a slowly varying DAP sequence \( \{M_t\}_{t \geq 0} \) with \( |M_t - M_{t-1}| \leq \Delta_M, \) where \( \Delta_M \) is called the policy variation budget. \( \{M_t\}_{t \geq 0} \) is safe to implement if \( M_t \in \Omega(\theta_s, \epsilon_x, \epsilon_u) \) for all \( t \geq 0 \), where \( \epsilon_x \geq \epsilon^*_x(D_M), \epsilon_u \geq \epsilon^*_u(D_M) \), and \( \epsilon^*_x(D_M), \epsilon^*_u(D_M) = O(\sqrt{\gamma}D_M) \).

3. Safe Adaptive Control Algorithm

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

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

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

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

\[
    u_t = -K_{stab}x_t + \sum_{k=1}^{H} M[k] \tilde{w}_{t-k} + \eta_t,
\]

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

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

\[
    \Omega(\theta; \tilde{\eta}, \epsilon_{\text{rob}}, \epsilon_{\text{unc}}) = \mathcal{O}(r + \tilde{\eta}) + \mathcal{O}(\mathcal{D}_\eta, \Theta)
\]

Compared with SPS (3) with a known model \( \theta_{\text{true}} \), RSPS approximates the true model by \( \theta \) and adds additional constraint-tightening terms \( \epsilon_{\text{unc}}^{\theta}(r, \tilde{\eta}) \) and \( \epsilon_{\text{unc}}^{\theta}(r, \tilde{\eta}) \) to allow for additional model uncertainties in \( \Theta \) and excitation noises with excitation level \( \tilde{\eta} \). Besides, RSPS (6) also includes constraint-tightening terms \( \epsilon_{\text{rob}}^{\theta}(\mathcal{D}_\eta, \Theta) \) to allow for small policy variation \( \Delta_M \) as discussed in Section 2.1.2. Formulas of the additional constraint-tightening terms are provided below. The proof is by perturbation analysis.

**Lemma 3.1 (RSPS).** Let \( \epsilon_{\text{unc}}^{\theta}(r, \tilde{\eta}) = O(r + \tilde{\eta}) \) and \( \epsilon_{\text{unc}}^{\theta}(r, \tilde{\eta}) = O(r + \tilde{\eta}) \). Then, set \( \Omega(\theta; \tilde{\eta}, \epsilon_{\text{rob}}, \epsilon_{\text{unc}}) \) is RSPS despite uncertainties \( \Theta \) and \( \tilde{\eta} \) and policy variation \( \Delta_M \).

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

\[
    \min_{M \in \Omega(\hat{\theta}, \epsilon_{\text{rob}}, \epsilon_{\text{unc}})} f(M; \hat{\theta})
\]

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

To address this, we design Algorithm 2 to ensure safe policy updates at the start of each phase. The high-level idea of Algorithm 2 is based on the slow-variation trick reviewed in

**Algorithm 1: Safe Adaptive Control**

**Input:** \( \Theta_{\text{ini}}, \mathcal{D}_\eta, K_{\text{stab}}, T^e, H^e, \hat{\eta}, \Delta_M, T^D, \forall \epsilon \).

1. **Initialize:** \( \theta^0 = \Theta_{\text{ini}}, r^0 = r_{\text{ini}} \). Define \( w_t = \tilde{w}_t = 0 \). For \( t < 0, \theta^0 = \). Define \( w_t = \tilde{w}_t = 0 \) for \( t < 0, \theta^0 = \).

2. **for** Episode \( e = 0, 1, 2, \ldots \) **do**

   // **Phase 1:** exploration & exploitation

   3. \( (\mathcal{M}_t^e, \Omega_t^e) \) \( \rightarrow \) RobustCE(\( \theta^e, H^e, \hat{\eta}, \Delta_M^e \)).

   4. If \( e > 0 \), run Algo. 2 to safely update the policy to \( \mathcal{M}_t^e \) with inputs \( (\mathcal{M}^{-1}, \Omega^{-1}, \theta^e, \hat{\eta}, \Delta_M^e) \), \( (\mathcal{M}^{-1}, \Omega^{-1}, \theta^e, \hat{\eta}, \Delta_M^e) \), \( T^e \) and output \( t^e_1 \).

   5. **for** \( t = t^e_1, \ldots, t^e_1 + T^D - 1 \) **do**

           Implement ApproxDAP(\( \mathcal{M}_t^e, \theta^e, \hat{\eta} \)).

   // **Model update by least square estimation**

   7. Estimate \( \theta^{e+1} \) by LSE with projection on \( \Theta_{\text{ini}} \):

   \[
   \theta^{e+1} = \arg\min_{\theta} \sum_{k=1}^{t+1} \|x_{k+1} - Ax_k - Bu_k\|^2_\epsilon.
   \]

   // **Phase 2:** pure exploitation (\( t = 0 \))

   8. \( (\mathcal{M}_t^\epsilon, \Omega_t^\epsilon) \) \( \rightarrow \) RobustCE(\( \theta^{e+1}, H^e, \hat{\eta}, \Delta_M^e \)).

   9. Run Algo. 2 to safely update the policy to \( \mathcal{M}_t^\epsilon \) with output \( t^\epsilon_2 \), inputs \( (\mathcal{M}_t^\epsilon, \Omega_t^\epsilon, \hat{\eta}, \Delta_M^e, t^\epsilon_2 + T^D) \).

   10. **for** \( t = t^\epsilon_2, \ldots, T^{(e+1)} - 1 \) **do**

           Implement ApproxDAP(\( \mathcal{M}_t^\epsilon, \theta^{e+1}, \hat{\eta} \)).

// **Algorithm 2: SafeTransit**

**Input:** \( (\mathcal{M}, \Omega, \Theta, \eta, \Delta_M), (\mathcal{M}', \Omega', \Theta', \hat{\eta}', \Delta_M'), t_0 \).

1. Set \( \eta_{\text{min}} = \min(\tilde{\eta}, \tilde{\eta}') \), \( \theta_{\text{min}} = \theta^{(r) r_{(r)} + \theta^{(r)_r}} \).

2. Set an intermediate policy as \( \mathcal{M}_{\text{mid}} \in \Omega \cap \Omega' \).

   // **Step 1:** slowly move from \( \mathcal{M} \) to \( \mathcal{M}_{\text{mid}} \)

   3. Define \( W_1 = \max(\|\mathcal{M} - \mathcal{M}_{\text{mid}}\|_F, \|H'\|_F) \).

   4. **for** \( t = t_0, t_0 + 1 \) **do**

           Set \( M_t = M_{t-1} + \frac{1}{W_1} (\mathcal{M}_{\text{mid}} - M) \).

   6. Run ApproxDAP(\( M_t, \eta_{\text{min}}, \theta_{\text{mid}} \)).

   // **Step 2:** slowly move from \( \mathcal{M}_{\text{mid}} \) to \( \mathcal{M}' \)

   7. Define \( W_2 = \frac{\|\mathcal{M}_{\text{mid}} - \mathcal{M}'\|_F}{\|\Delta_M\|_F} \).

   8. **for** \( t = t_0 + 1, \ldots, t_0 + W_1 + W_2 - 1 \) **do**

           Set \( M_t = M_{t-1} + \frac{1}{W_2} (M' - M_{\text{mid}}) \).

   9. Run ApproxDAP(\( M_t, \hat{\eta}', \theta' \)).

**Output:** Termination stage \( t_1 = t_0 + W_1 + W_2 \).

**Subroutine ApproxDAP(\( M, \theta, \hat{\eta} \)):**

1. Implement (5) with \( \eta_t \sim \mathcal{D}_\eta \).

2. Observe \( v_{t+1} \) and record \( \tilde{w}_t = \mathbb{T}(v_{t+1} = Ax_t - \tilde{B}u_t) \).

Safe Adaptive Learning for Linear Quadratic Regulators with Constraints
Subroutine $\text{RobustCE}(\Theta, H, \bar{\eta}, \Delta_M)$:
- Construct the robustly safe policy set: $\Omega = \Omega(\bar{\theta}, R_{\text{rob}}, \epsilon_{R_{\text{rob}}})$ for $(R_{\text{rob}}, \epsilon_{R_{\text{rob}}})$ defined in (6).
- Compute the optimal policy $M$ to (7).
- return policy $M$ and robustly safe policy set $\Omega$.

Section 2.1.2. That is, we construct a policy path connecting the old policy to the new policy such that this policy path is contained in some robustly safe policy set with an additional constraint tightening term to allow slow policy variation, then by slowly varying the policies along this path, we are able to safely transit to the new policy.

Next, we discuss the construction of this policy path,\(^ 7\) which is illustrated in Figure 1d. We follow the notations in Algorithm 2, i.e., the old policy is $M$ in an old RSPS $\Omega$ and the new policy is $M'$ in $\Omega'$. Notice that the straight line from $M$ to $M'$ does not satisfy the requirements of the slow variation trick because some parts of the line are outside both RSPSs. To address this, Algorithm 2 introduces an intermediate policy $M_{\text{mid}}$ in $\Omega \cap \Omega'$, and slowly moves the policy from the old one $M$ to the intermediate one $M_{\text{mid}}$ (Step 1), then slowly moves from $M_{\text{mid}}$ to the new policy $M'$ (Step 2). In this way, all the path is included in at least one of the robustly safe policy sets, which allows safe transition from the old policy to the new policy. The choice of $M_{\text{mid}}$ is not unique. In practice, we recommend selecting $M_{\text{mid}}$ with a shorter path length for quicker policy transition. $M_{\text{mid}}$ can be computed efficiently since the set $\Omega \cap \Omega'$ is a polytope. The existence of $M_{\text{mid}}$ can be guaranteed if the first $\text{RobustCE}$ program in Algorithm 1 (Phase 1 of episode 0) is strictly feasible. This is usually called recursive feasibility and will be formally proved in Theorem 4.3.

Remark 3.2 (More discussions on $M_{\text{mid}}$). If RSPSs are monotone, e.g., if $\Omega \subseteq \Omega'$, then we can let $M_{\text{mid}} = M$ and the path constructed by Algorithm 2 reduces to the straight line from $M$ to $M'$. Hence, a non-trivial $M_{\text{mid}}$ is only relevant for non-monotone RSPS, which can be caused by the non-monotone model uncertainty sets generated by LSE (even though the error bound $r$ of LSE decreases with more data, the change in the point estimator $\bar{\theta}$ may cause $\Omega' \not\subseteq \Omega$). Though one can enforce decreasing uncertainty sets by taking joints over all the history uncertainty sets, this approach leads to an increasing number of constraints when determining the RSPS in $\text{RobustCE}$, thus demanding high computation for large episode $e$.

Remark 3.3 (Single trajectory and computation comparison). Though Alg. 1 is implemented by episodes, it is still a single trajectory since no system starts are needed when new episodes start. Compared with the projected-gradient-descent algorithm in Li et al. (2021), our algorithm only solves constrained QP once in a while, but Li et al. (2021) solve constrained QP for projection at every stage. In this sense, our algorithm reduces the computational burden.

Remark 3.4 (Model Estimation). As for the model updates, we can use all the history data in practice, though we only use part of history in $\text{ModelEst}$ for simpler analysis. $\text{ModelEst}$ also projects the estimated model onto $\Theta_{\text{ini}}$ to ensure bounded estimation.

Remark 3.5 (Safe algorithm comparison). Some constrained control methods construct safe state sets and safe action sets for the current state, e.g., control barrier functions (Ames et al., 2016), reachability-set-based methods (Akametalu et al., 2014), regulation maps (Kellett & Teel, 2004), etc. In contrast, this paper constructs safe policy sets in the space of policy parameters $M$. This is possible because our policy structure (linear on history disturbances) allows a transformation from linear constraints on the states and actions to polytopic constraints on the policy parameters.

4. Theoretical Analysis

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

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

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

**Theorem 4.1** (Estimation error bound). Consider actions $u_t = \pi_t(x_t, \{w_{t-k}, \eta_{t-k}\}_{k=0}^{\infty}) + \eta_t$, where $\|\eta_t\|_\infty \leq \bar{\eta}$ is generated as discussed after (5) and policies $\pi_t(\cdot)$ ensure bounded states and actions, i.e., $\|(x_t^1, u_t^1)\|_2 \leq b_z$ for all $t \geq 0$. Let $\bar{\theta}_T = \min_{A,B} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|_2^2$ denote the model estimation. For any $0 < \delta < 1/3$, for $T \geq O((\log(1/\delta) + (m+n) \log(b_\eta/\bar{\eta})))$, with probability (w.p.) $1 - 3\delta$, we have $\|\bar{\theta}_T - \theta_*\|_2 \leq O(\sqrt{n+m} \sqrt{\log(b_\eta/\bar{\eta})}/\sqrt{T \bar{\eta}})$. Theorem 4.1 holds for both linear and nonlinear policies as long as the induced states and actions are bounded, which can be guaranteed by the stability of the policies. Further, the error bound in Theorem 4.1 is $O(\sqrt{n+m} \sqrt{\log(b_\eta/\bar{\eta})}/\sqrt{T \bar{\eta}})$, which coincides with the error bound for linear policies in terms of $T, \bar{\eta}, n, m$ in (Dean et al., 2018; 2019b).

Based on Theorem 4.1, we obtain a formula for the estimation error bound $r^e$ in Line 7 of Algorithm 1.

**Corollary 4.2** (Formula of $r^e$). Suppose $H^0 \geq \log(2\kappa)/\log((1 - \gamma)^{-1})$, $T^{e+1} \geq t^1 \geq T_D$ and $T_D$ satisfies the condition on $T$ in Theorem 4.1. For any $0 < p < 1$ and $e \geq 1$, with probability at least $1 - \frac{k}{T^p}$, we have $\|\bar{\theta}^e_\pi - \theta_*\|_F \leq r^e$, where

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

Corollary 4.2 considers the $\| \cdot \|_F$ norm because Algorithm 1 projects matrix $\bar{\theta}_T$ onto $\Theta^e_{\text{ini}}$ and the $\| \cdot \|_F$ norm is more convenient to analyze and implement when matrix projections are involved. Due to the change of norms, the error bound has an additional $\sqrt{n}$ factor.

4.2. Feasibility and Constraint Satisfaction

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

**Theorem 4.3** (Feasibility). Algorithms 1 and 2 output feasible policies for all $t$ under the following conditions.

(i) (Strict initial feasibility) There exists $\epsilon_0 > 0$ such that $\Omega(\bar{\theta}_T, \epsilon^0, \epsilon^0, \epsilon^0_\pi) \neq \emptyset$, where $\epsilon^0, \epsilon^0_\pi$ are defined by (7) with initial parameters $r^0, \bar{\theta}_0^e, H^0, \Delta_M^0$.

(ii) (Monotone parameters) $\bar{\eta}, H^e, T_D^e, \Delta_M^e$ are selected s.t. $(H^e)^{-1}$, $\sqrt{H^e} \Delta_M^e, \bar{\eta}$, $r^e$ are all non-increasing with $e$, and $r^e \leq c_1 \min(\bar{\eta}, \epsilon^0, \epsilon^0_\pi)$, where $r^e$ is defined in (8), $c_1$ is defined in Lemma 3.1.

Further, under Assumption 2.3, condition (i) is satisfied if (ii) $e^0_\pi + \epsilon_0 \leq \epsilon_{F,x} - \epsilon_{\text{unc}}(r_{\text{ini}})$, $\epsilon^0_\pi \leq \epsilon_{F,u} - \epsilon_{P}$, where $\epsilon_P = O(\sqrt{\min(1 - \gamma)}/T^p)$.

Condition (i) requires the initial policy set $\Omega^0_{\text{ini}}$ to contain a policy that strictly satisfies the constraints on $g^e(\bar{\theta}; \bar{\theta}_0^e)$. Condition (ii) requires monotonic parameters in later phases, where the non-increasing estimation error $r^e$ requires an increasing number of exploration stages $T^e_D$. Conditions (i) and (ii) together establish the recursive feasibility: if our algorithm is (strictly) feasible at the initial stage, then the algorithm is feasible in the future under proper parameters.

The condition (ii) in Theorem 4.3 guarantees strict initial feasibility, which is based on the $e_\pi$-strictly safe policy $u_t = -K_F x_t$ in Assumption 2.3. The term $e_{\pi}$ captures the difference between the policy $K_F$ and its approximate DAP. Consider (iii) requires large enough $H^0$, $T_D^0$ and small enough $\bar{\eta}, \Delta_M^0$. Further, consider (iii) implicitly requires a small enough initial uncertainty radius: $e^0_{\text{unc}}(r_{\text{ini}}, 0) \leq \epsilon_{F,x}$. If the initial uncertainty set is too large but a safe policy is available, one can first explore the system with the safe policy to reduce the model uncertainty until the strict initial feasibility is obtained and then apply our algorithm.

**Theorem 4.4** (Constraint Satisfaction). Under the conditions in Theorem 4.3 and Corollary 4.2, when $T^{e+1} \geq t^2$, we have $u_t \in U$ for all $t \geq 0$ w.p.$1$ and $x_t \in X$ for all $t \geq 0$ w.p. $1 - p$, where $p$ is chosen in Corollary 4.2.

The control constraint satisfaction is always ensured by the projection onto $\mathbb{W}$ in (5). Besides, we can show that the state constraints are satisfied if the true model is inside the confidence sets $\Theta^e$ for all $e \geq 0$, whose probability is at least $1 - p$ by Corollary 4.2.

4.3. Regret Bound

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

**Theorem 4.5** (Regret bound). Suppose $e_{\Theta}(r_{\text{ini}}) \leq \epsilon_{F,x}/4$. For any $0 < p < 1/2$, with parameters $T^1 \geq \bar{\Omega}((\sqrt{n+m})^3)$, $T^e + 1 = 2T^e$, $T_D^e = (T^e + 1 - T^e)/2$, $\Delta_M^e = O(\sqrt{\log(\log(T^e + 1)/\sqrt{n+m})}/\sqrt{n+m})$, $\bar{\eta}^e = O(\min(\epsilon_{\pi}^e, \epsilon_{\pi}^0))$, and $H^e \geq O(\log(\log(T^e + 1)/\sqrt{n+m})))$, Algorithm 1 is feasible and satisfies $\{u_t \in U\}_{t \geq 0}$ w.p.$1$ and $\{x_t \in X\}_{t \geq 0}$ w.p.$(1 - p)$. Further, w.p.$1 - 2p$, we have Regret $\leq O((n^2 m^2 + n^{1.5} m^{1.5}) \sqrt{nm} + k_x + k_u T^{2/3})$.

On $r_{\text{ini}}$. As discussed after Theorem 4.3, the initial feasibility requires a small enough $r_{\text{ini}}$, otherwise more exploration is needed before applying our algorithm. For technical simplicity, Theorem 4.5 assumes $e_{\Theta}(r_{\text{ini}}) \leq \epsilon_{F,x}/4$ and establishes other conditions accordingly.
On parameters. Theorem 4.5 provides choices of parameters that ensure feasibility, safety, and the regret bound. Here, we choose exponentially increasing episode lengths $T^\ast$, and explore for $(T^e)^{2/3}$ stages at each episode with a small enough constant excitation level $\tilde{\eta}$. Compared with the first-explore-then-exploit algorithms, episodic updates allow early improvements in system performance and adaptation to possible changes in the environment. Finally, we select large enough memory lengths $H^e \geq O(\log(T^e))$ and small enough variation budgets $\Delta_M^e \leq O((T^e+1)^{-1/3})$.

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

Discussions on the pure exploitation phase. Algorithm 1 includes a pure exploitation phase with no excitation noise at each episode, which is not present in the unconstrained algorithms (Dean et al., 2018; Mania et al., 2019; Simchowitz & Foster, 2020). This phase is motivated by our regret analysis: the algorithm can still work without this phase but will generate a worse regret bound. Specifically, consider a robust CE policy (7) with respect to $\tilde{\eta}$ and uncertainty radius $r$, the regret of this policy per stage can be roughly bounded by $\tilde{O}(\tilde{\eta} + r)$ in the supplementary (we omit $\Delta_M$ here for simplicity). With no pure exploitation phases, the regret in episode $c$ can be roughly bounded by $\tilde{O}(T^c(\tilde{\eta}^c + r^c))$, where $r^c = \tilde{O}(\frac{1}{\sqrt{T^c(\tilde{\eta}^c + r^c)}})$ by Corollary 4.2 (hiding $n, m$). Therefore, the total regret can be bounded by $\sum_c(\frac{1}{\sqrt{T^c(\tilde{\eta}^c + r^c)}} + \tilde{\eta}^c)T^c \approx \sum_c(\frac{1}{\sqrt{T^c\tilde{\eta}^c}} + \tilde{\eta}^c)T^c$, which is minimized at $\frac{1}{\sqrt{T^c\tilde{\eta}^c}} = \tilde{\eta}^c = (T^c)^{-1/4}$, leading to a worse regret bound $\tilde{O}(T^{3/4})$. Lastly, with a constant $\tilde{\eta}^c$, our algorithm suffers slightly larger stage regret during the Phase 1 (exploration & exploitation) compared with diminishing $\tilde{\eta}^c = (T^c)^{-1/4} \to 0$, but the performance still improves by refining the models and reducing $\Delta_M$.

5. Numerical Experiments

In this section, we numerically test our safe adaptive control algorithm on a quadrotor vertical flight problem. Specifically, we consider the affine dynamical model for vertical flight with linear air drag force as in Luukkonen (2011): $\ddot{z} = \nu/m - I^a \dot{\dot{z}}/m + d$, where $z$ is the quadrotor’s altitude, $\nu$ is the motor thrust, $m = 1$ kg is the mass, $g = 9.8\text{m/s}^2$ is the gravitational acceleration, $I^a = 0.25\text{kg/s}$ is the drag coefficient of the air resistance, $-I^a \dot{z}$ models air drag force, and $d$ represents other system disturbances. The unknown system parameters are the mass due to the unknown load and the air drag coefficient. The constraints on the altitude are $0 \leq z \leq 1.7\text{m}$ and the constraints on the thrust are $0 \leq \nu \leq 12\text{N}$. The control task is to safely maintain the quadrotor around a target altitude $z^{\text{ref}} = 0.7\text{m}$. The corresponding velocity and control input at this equilibrium point is $\dot{z}^{\text{ref}} = 0, \nu^{\text{ref}} = g$. We consider a quadratic cost function to measure the deviation from the desirable equilibrium point: $0.1(z - z^{\text{ref}})^2 + 0.1(\dot{z} - \dot{z}^{\text{ref}})^2 + 0.1(\nu - \nu^{\text{ref}})^2$. Consider unknown mass and air drag coefficient. The initial estimation are $0.83\text{kg} \leq m \leq 2\text{kg}$, and $0.1\text{kg/s} \leq I^a \leq 0.33\text{kg/s}$. We generate the system disturbances i.i.d. from distribution Bern$[-0.2, 0.2]$ with $w_{\text{max}} = 0.2$. We select $K_{\text{stab}} = (2/3, 1)$ to robustly stabilize the system. Let time discretization be $\Delta_e = 1\text{s}$.

Figure 2a shows that the model estimation error decreases as $\tilde{\eta}$ increases, which is consistent with Corollary 4.2. Figure 2b shows that the average regret first decreases then increases as $\tilde{\eta}$ increases. This is because for small $\tilde{\eta}$, the regret benefits from faster estimation error reduction, but for large $\tilde{\eta}$, the regret suffers from larger constraint-tightening terms due to large excitation noises. This shows the tradeoff between exploration and exploitation under safety constraints. Figure 2c,d show that our algorithm ensure safety by satisfying constraint satisfaction. Further, as learning continues, the ranges decrease due to better model estimation.

6. Discussions and Future Work

Comparing with RMPC in (Mayne et al., 2005). (Mayne et al., 2005) propose an RMPC method for constrained
LQR with a known model. Though this method generates piecewise-affine policies, we are able to show in the appendix that its infinite-horizon averaged cost can be characterized by $J(K)$, where $K$ is a pre-fixed safe linear policy in RMPC (Mayne et al., 2005). Thus, our regret remains the same after adding this RMPC to the benchmark.

**Future work.** There are many interesting future directions, e.g., (i) improving practical performance by reducing the constraint tightenings, (ii) regret analysis compared with other RMPC algorithms, (iii) algorithm design and analysis for large initial states, (iv) fundamental regret lower bounds, (v) considering nonlinear policy classes, (vi) safe adaptive control for nonlinear systems, etc.

**References**


**Appendices**

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

- Appendix A provides necessary lemmas that will be used throughout the appendices and defines the constraint-tightening factors promised in Lemma 3.1.
- Appendix B focuses on the estimation error and provides proofs for Theorem 4.1 and Corollary 4.2.
- Appendix C studies feasibility and provides a proof for Theorem 4.3.
- Appendix D focuses on the constraint satisfaction and provides a proof for Theorem 4.4, which also includes a proof for Lemma 3.1.
- Appendix E analyzes the regret and proves Theorem 4.5.
- Appendix F provides additional discussions on RMPC and non-zero $x_0$.
- Appendix G provides proofs to the technical lemmas used in the appendices A-E.

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

**A. Preparations: State Approximation and Constraint-tightening Terms**

This section provides results that will be useful throughout the rest of the appendices. Specifically, the first subsection provides a state approximation lemma and a constraint-decomposition corollary for the approximate DAP, and the second subsection defines and discusses the constraint-tightening terms in Lemma 3.1 and (7) based on the upper bounds of the constraint decomposition terms.

**A.1. State Approximation and Constraint Decompositions**

We consider a more general form of approximate DAP below than that in Section 2.1, i.e., an approximate DAP with time-varying policy matrices $M_t$, time-varying excitation levels $\bar{\theta}_t$, and time-varying model estimations $\theta_t$.

$$u_t = -K_{\text{stat}}x_t + \sum_{t=1}^{H_1} M_t[k] \bar{w}_{t-k} + \eta_t, \quad \bar{w}_t = \Pi_W (x_{t+1} - \bar{\theta}_t z_t), \quad \|\eta_t\|_\infty \leq \bar{\eta}_t, \quad t \geq 0, \quad (9)$$

where $M_t \in \mathcal{M}_{H_1}$ and $\{ H_1 \}_{t \geq 0}$ is non-decreasing.

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

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

$$x_t = A_s^{H_1} x_{t-H_1} + \sum_{k=2}^{H_1} \sum_{i=1}^{H_1} A_s^{-1} B_s M_{t+k-i} [k-i] \bar{w}_{t-k+i} \mathbb{I}(1 \leq k-i \leq H_{\text{est}}) + \sum_{t=1}^{H_1} A_s^{-1} w_{t-i} + \sum_{i=1}^{H_1} A_s^{-1} B_s \eta_{t-i}.$$
The lemma above is a straightforward extension from Proposition 2.7 reviewed in Section 2.1 for the case with perfect model information, thus the proof is omitted.

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

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

\[
\tilde{g}_i^x(M_{t-H_t:t-1}; \theta) = \sup_{\tilde{w}_k \in \mathcal{W}} D_{x,i}^T \tilde{\Phi}_k^x(M_{t-H_t:t-1}; \theta) \tilde{w}_t-k = \sum_{k=1}^{2H_t} \|D_{x,i}^T \tilde{\Phi}_k^x(M_{t-H_t:t-1}; \theta)\|_w w_{\max},
\]

where \(1 \leq i \leq k_x\) and we define \(M_t = M_0\) for \(t \leq 0\) for notational simplicity. Notice that when \(M_t = M\) and \(H_t = H\) (the time-invariant case), the definitions of \(\tilde{\Phi}_k^x(M_{t-H_t:t}; \theta)\) and \(\tilde{g}_i^x(M_{t-H_t:t-1}; \theta)\) above reduce to the definitions of \(\Phi_k^x(M; \theta)\) and \(g_i^x(M; \theta)\) respectively in Section 2.1.

The formula and constraints for \(u_t\) are similar.

**Remark A.2** (Constraint satisfaction of time-varying DAP). Notice that the above discussions show that \(x_t\) and \(u_t\) are heavily influenced by not only \(M_{t-1}\) but also earlier history policies. Therefore, by only restricting \(M_{t-1}\), it is not enough to ensure constraint satisfaction for time-varying DAPs.

Based on Lemma A.1 and the notations defined above, we can obtain the following corollary on the decompositions of the state constraints \(D_{x,i}^T x_t\) and action constraints \(D_{u,i}^T u_t\). The decompositions are crucial when defining our constraint-tightening terms and developing the constraint satisfaction guarantees.

For simplicity, we consider \(K_{\text{stab}} = 0\) from now on. The proof ideas for \(K_{\text{stab}} \neq 0\) are the same.

**Corollary A.3** (Constraint decomposition). When implementing the time-varying approximate DAP (9) to the system \(x_{t+1} = A_s x_t + B_s u_{t} + w_t\), for each \(1 \leq i \leq k_x\) and \(1 \leq j \leq k_u\), we have the following decompositions:

\[
D_{x,i}^T x_t \leq \underbrace{g_i^x(M_t; \hat{\theta}_t^x)}_{\text{estimated state constraint function}} + \sum_{k=1}^{H_t} D_{x,i}^T A^{k-1} (w_{t-k} - \hat{w}_{t-k}) + \underbrace{\sum_{i=1}^{H_t} D_{x,i}^T A^{i-1} (A_s x_i - \hat{x}_{t-i})}_{\text{model estimation errors}} + \underbrace{D_{x,i}^T A^{H_t} x_t - \hat{x}_t}_{\text{history truncation errors}} + \underbrace{(g_x^t(M_t; \theta_s) - g_x^t(M_t; \hat{\theta}_t^x))}_{\text{policy variation errors}},
\]

\[
D_{u,j}^T u_t \leq \underbrace{\theta_j^u(M_t)}_{\text{action constraint function}} + \underbrace{D_{u,j}^T \eta_t}_{\text{excitation error on the action}} + \underbrace{D_{u,j}^T \hat{\eta}_t}_{\text{excitation errors on the state}}.
\]

where \(\hat{\theta}_t^x\) is an estimated model used to approximate the state constraint function, and we allow \(\hat{\theta}_t^x \neq \hat{\theta}_t\) for generality.

**Proof.** The proof is by the definitions of \(g_i^x, g_j^u, \tilde{g}_i^x\), and Lemma A.1. For the action constraints, by the definition (9), we have

\[
D_{u,j}^T u_t = \sum_{i=1}^{H_t} D_{u,j}^T M_t[k] \hat{w}_{t-k} + D_{u,j}^T \hat{\eta}_t \leq \sum_{i=1}^{H_t} \|D_{u,j}^T M_t[k]\|_w w_{\max} + D_{u,j}^T \hat{\eta}_t = g_j^u(M_t) + D_{u,j}^T \hat{\eta}_t
\]

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

\[\square\]

**A.2. The Constraint-tightening Terms**

Define \(\epsilon_H(H) = \epsilon_x^r, \epsilon_{\text{rob}}^u = \epsilon_{\theta}(r) + \epsilon_{\eta, \alpha}(\bar{\eta})\). Notice that when \(K_{\text{stab}} = 0\), \(\epsilon_x^u = 0\) and \(\epsilon_{\text{rob}}^u\) only depends on \(\bar{\eta}\). Define \(\epsilon_{\text{rob}}^u = \epsilon_{\eta, \alpha}(\bar{\eta})\).

This subsection provides the definitions of the constraint-tightening terms introduced and their explanations by showing that
The proof of Lemma A.4 is based on the perturbation analysis and deferred to Appendix G.1. When proving Lemma A.4, each constraint-tightening term serves as an upper bound on an error term in the constraint decomposition in Corollary A.3.

**Definition and explanation of $\epsilon_\theta(r)$.** The next lemma formally shows that the constraint-tightening term $\epsilon_\theta(r)$ is an upper bound on the model estimation errors in the state constraint decomposition in Corollary 4.2, where $r$ is the model estimation error bound.

**Lemma A.4 (Definition of $\epsilon_\theta(r)$).** Consider implementing the time-varying approximate DAP (9) to the system $x_{t+1} = A_x x_t + B_x u_t + w_t$. For a fixed $t$, suppose $\hat{\theta}_{t-k}, \hat{\theta}_{t} \in \Theta_{\text{init}}$, $\|\theta_{t-k} - \theta\|_F \leq r$, and $\|\theta_t - \theta\|_F \leq r$ 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

$$
\begin{align*}
g^r_t(M_i; \theta_*) - g^r_t(M_i; \hat{\theta}_t^0) \leq \epsilon_\theta(r),
\sum_{k=1}^{H_t} D^T_{x,i} A_{s}^{k-1}(w_{t-k} - \hat{w}_{t-k}) \leq \epsilon_w(r), \\
(g^r_t(M_i; \hat{\theta}_t) - g^r_t(M_i; \hat{\theta}_t^0)) + \sum_{k=1}^{H_t} D^T_{x,i} A_{s}^{k-1}(w_{t-k} - \hat{w}_{t-k}) \leq \epsilon_\theta(r)
\end{align*}
$$

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

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

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

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

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

**Definition and explanation of $\epsilon_{\eta,x}(\bar{\eta})$ and $\epsilon_{\eta,u}(\bar{\eta})$.** The next lemma formally shows that the terms $\epsilon_{\eta,x}(\bar{\eta})$ and $\epsilon_{\eta,u}(\bar{\eta})$ bounds the excitation errors on the state and action constraint decompositions in Corollary A.3.

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

$$
\begin{align*}
\sum_{i=1}^{H_t} D^T_{x,i} A_{s}^{i-1} B^* \eta_{t-i} \leq \epsilon_{\eta,x}(\bar{\eta}),
\sum_{i=1}^{H_t} D^T_{x,i} \eta_t \leq \epsilon_{\eta,u}(\bar{\eta}),
\end{align*}
$$

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

**Proof.** The proof is provided below.

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

$$
\|D_u \eta_t\|_\infty \leq \|D_u\|_\infty \|\eta_t\|_\infty \leq \|D_u\|_\infty \bar{\eta}
$$

\qed
Theorem B.2 This section provides a proof for Theorem 4.1 and a proof for Corollary 4.2. When proving Corollary 4.2, we also establish a random process Block Martingale Small-Ball (BMSB) condition for $\Gamma$. By general policies, we allow time-varying policies, nonlinear policies, policies that depend on all the history, and shown to satisfy the BMSB condition. Our contribution is to show that even for general policies, BMSB still holds. BMSB are included below for completeness. In the literature (Dean et al., 2018; 2019b), only linear policies are considered a block matingale small-ball (BMSB) condition (Simchowitz et al., 2018). The general error bound and the definition of $\epsilon_H(\Gamma)$ has been introduced in (Li et al., 2021) for the known-model case to bound the history truncation errors in the state constraint decomposition. Here, we slightly improve its dependence on the problem dimensions and will be provided in Appendix G.2.

Lemma A.8 (Definition of $\epsilon_H$). For any $x_{t-H} \in \mathbb{X}$, we have

$$D_xA^H_{x_{t-H}} \leq \epsilon_H(\Gamma) = \|D_x\|_\infty \kappa \max_t (1 - \gamma)^{H_t} = O((1 - \gamma)^{H_t}).$$

Proof.

$$\|D_xA^H_{x_{t-H}}\| \leq \|D_x\|_\infty \|A^H_{x_{t-H}}\| \leq \|D_x\|_\infty \|A^H_{x_{t-H}}\|_2 \leq \|D_x\|_\infty \kappa (1 - \gamma)^{H_t}.\max_t.$$

Lemma A.8 (Definition of $\epsilon_v(\Delta_M, H)$). Under the conditions in Lemma A.1, suppose $\Delta_M \geq \max_{1 \leq k \leq H_t} \|M_y - M_{k-1}\|_F$, then we have

$$\left(\tilde{g}_t^\top(M_{t-1}; x_{t-1}) - g_t^\top(M_t; x_{t-1})\right) \leq \epsilon_v(\Delta_M, H_t)$$

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

B. Estimation Error Bounds

This section provides a proof for Theorem 4.1 and a proof for Corollary 4.2. When proving Corollary 4.2, we also establishes a.s. upper bounds on the state and action trajectories of our algorithm.

B.1. Proof of Theorem 4.1

Our proof of Theorem 4.1 relies on a recently developed least square estimation error bound for general time series satisfying a block martingale small-ball (BMSB) condition (Simchowitz et al., 2018). The general error bound and the definition of BMSB are included below for completeness. In the literature (Dean et al., 2018; 2019b), only linear policies are considered and shown to satisfy the BMSB condition. Our contribution is to show that even for general policies, BMSB still holds as long as the corresponding states and actions are bounded (which is usually the case if certain stability properties are satisfied). By general policies, we allow time-varying policies, nonlinear policies, policies that depend on all the history, etc., (i.e. we consider $\pi_t(x, \{w_t, \eta_t\}_{k=0}^{t-1} + \eta_t$). More rigorous discussions are provided below.

Definition B.1 (Block Martingale Small-Ball (BMSB) (Definition 2.1 (Simchowitz et al., 2018))). Let $\{X_t\}_{t \geq 1}$ be an $\{F_t\}_{t \geq 1}$-adapted random process taking values in $\mathbb{R}^d$. We say that it satisfies the $(k, \Gamma_{sb}, p)$-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 $\frac{1}{T} \sum_{t=1}^T \mathbb{P}(|X_{t+1}X_{t+1}^\top| \geq \sqrt{\lambda^\top \Gamma_{sb} \lambda} \mid F_j) \geq \gamma$ almost surely.

Theorem B.2 (Theorem 2.4 in (Simchowitz et al., 2018)). Fix $\epsilon \in (0, 1), \delta \in (0, 1/3), T \geq 1$, and $0 < \Gamma_{sb} < \Gamma$. Consider a random process $\{X_t, Y_t\}_{t \geq 1} \in (\mathbb{R}^d \times \mathbb{R}^n)^T$ and a filtration $\{F_t\}_{t \geq 1}$. Suppose the following conditions hold,

1. $Y_t = \theta_t X_t + \eta_t$, where $\eta_t \mid F_t$ is $\sigma_{sb}^2$-sub-Gaussian and mean zero,
2. $\{X_t\}_{t \geq 1}$ is an $\{F_t\}_{t \geq 1}$-adapted random process satisfying the $(k, \Gamma_{sb}, p)$-block martingale small-ball (BMSB) condition,
3. $\mathbb{P}(|\sum_{t=1}^T X_tX_t^\top| \leq \gamma T) \leq \delta$. 


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

\[
T \geq \frac{10k}{p^2} \left( \log\left( \frac{1}{\delta} \right) + 2d \log(10/p) + \log \det(\hat{\Gamma}^{-1}_{ab}) \right),
\]

we have

\[
\| \hat{\theta} - \theta \|_2 \leq \frac{90\sigma_{\text{sub}}}{p} \sqrt{\frac{n + d \log(10/p) + \log \det(\hat{\Gamma}^{-1}_{ab}) + \log(1/\delta)}{Tv_{\text{min}}(\Gamma_{ab})}}
\]

with probability at least \( 1 - 3\delta \).

Next, we present a proof for our Theorem 4.1 by verifying the conditions in Theorem B.2 for general nonlinear policies.

**Proof of Theorem 4.1.** Condition 1 is straightforward: \( x_{t+1} = \theta^*_z z_t + w_t \), and \( w_t \mid F_t = w_t \) which is mean 0 and \( \sigma_{\text{sub}}^2 \)-sub-Gaussian by Assumption 2.5. Condition 3 is also straightforward. Notice that \( \nu_{\max}(z_t z_t^\top) \leq \| z_t \|_2^2 \leq b_2^2 \).

Therefore, we can define \( \Gamma = b_2^2 I_{n+m} \), and then \( P(\sum_{t=1}^{\infty} z_t z_t^\top \not\leq TT) = 0 \leq \delta \).

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

**Lemma B.3 (Verification of BMSB condition).** Define filtration \( F_t = \{w_0, \ldots, w_{t-1}, \eta_0, \ldots, \eta_t\} \). Under the conditions in Theorem 4.1,

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

where \( p_z = \min(p_w, p_\eta) \), \( s_z = \min(s_w/4, \sqrt{3} s_\eta, \frac{s_x s_w}{b_{zu}^2} \eta) \).

**Proof of Lemma B.3.** Define filtration \( F_{t}^{m} = F(w_0, \ldots, w_{t-1}, \eta_0, \ldots, \eta_{t-1}) \). Notice that the policy in Theorem 4.1 can be written as \( u_t = \pi_t(F_t^m) + \eta_t \). Note that \( z_t \in F_t \) is by definition. Next,

\[
z_{t+1} \mid F_t = \begin{bmatrix} x_{t+1} \\ u_{t+1} \end{bmatrix} \mid F_t = \begin{bmatrix} \theta^*_z z_t + w_t \mid F_t \\ \pi_{t+1}(F_t^m) + \eta_{t+1} \mid F_t \end{bmatrix},
\]

where \( F_{t+1} = F(w_0, \ldots, w_t, \eta_0, \ldots, \eta_t) \).

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

Consider any \( \lambda = (\lambda^T_1, \lambda^T_2)^\top \in \mathbb{R}^{m+n} \), 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 \( \lambda^T_1 \theta^*_z z_t \geq 0 \), (ii) when \( \| \lambda_2 \|_2 \leq 1/k_0 \) and \( \lambda^T_1 \theta^*_z z_t < 0 \), (iii) when \( \| \lambda_2 \|_2 > 1/k_0 \). We will show in all three cases,

\[
P(\| \lambda^T z_{t+1} \| \geq s_z) \geq p_z
\]

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

**Case 1:** when \( \| \lambda_2 \|_2 \leq 1/k_0 \) and \( \lambda^T_1 \theta^*_z z_t \geq 0 \)

\[
\lambda^T_1 w_t \leq \lambda^T_1 (w_t + \theta^*_z z_t) \leq |\lambda^T_1 (w_t + \theta^*_z z_t)|
\]

\[
= |\lambda^T z_{t+1} - \lambda^T_2 u_{t+1}| \leq |\lambda^T z_{t+1}| + |\lambda^T_2 u_{t+1}| \leq |\lambda^T z_{t+1}| + \| \lambda_2 \|_2 b_u
\]

\[
\leq |\lambda^T z_{t+1}| + b_u/k_0 \leq |\lambda^T z_{t+1}| + s_w/4
\]

where the last inequality uses \( k_0 \geq 4b_u/s_w \).
Further, notice that \( k_0 \geq 2/\sqrt{3} \), so \( \|\lambda_2\|^2 \leq 1/k_0^2 \leq 3/4 \), thus, \( \|\lambda_1\|^2 \geq 1/4 \), which means \( \|\lambda_1\| \geq 1/2 \). Therefore,

\[
P(\lambda_1^T w_t \geq s_w/2) = P\left(\frac{\lambda_1^T w_t}{\|\lambda_1\|_2} \geq \frac{s_w}{2\|\lambda_1\|_2} \right) \geq P\left(\frac{\lambda_1^T w_t}{\|\lambda_1\|_2} \geq s_w \right) = p_w
\]

Then,

\[
P(|\lambda^T z_{t+1}| \geq s_z) \geq P(|\lambda^T z_{t+1}| \geq s_w/4) = P(|\lambda^T z_{t+1}| + s_w/4 \geq s_w/2)
\]

\[
\geq P(\lambda_1^T w_t \geq s_w/2) \geq p_w
\]

which completes case 1.

**Case 2:** when \( \|\lambda_2\| \leq 1/k_0 \) and \( \lambda_1^T \theta_\tau z_t < 0 \).

\[
\lambda_1^T w_t \geq \lambda_1^T (w_t + \theta_\tau z_t) \geq -|\lambda_1^T (w_t + \theta_\tau z_t)|
\]

\[
= -|\lambda_1^T z_{t+1} - \lambda_2^T u_{t+1}| \geq -|\lambda_1^T z_{t+1}| - |\lambda_2^T u_{t+1}| \geq -|\lambda_1^T z_{t+1}| - \|\lambda_2\|_2 b_u
\]

\[
\geq -|\lambda_1^T z_{t+1}| - b_u/k_0 \geq -|\lambda_1^T z_{t+1}| - s_w/4
\]

where the last inequality uses \( k_0 \geq 4b_u/s_w \).

Further, notice that \( k_0 \geq 2/\sqrt{3} \), so \( \|\lambda_2\|^2 \leq 1/k_0^2 \leq 3/4 \), thus, \( \|\lambda_1\|^2 \geq 1/4 \), which means \( \|\lambda_1\| \geq 1/2 \). Therefore,

\[
P(\lambda_1^T w_t \leq -s_w/2) = P\left(\frac{\lambda_1^T w_t}{\|\lambda_1\|_2} \leq -\frac{s_w}{2\|\lambda_1\|_2} \right) \geq P\left(\frac{\lambda_1^T w_t}{\|\lambda_1\|_2} \leq -s_w \right) = P\left(-\frac{\lambda_1^T w_t}{\|\lambda_1\|_2} \geq s_w \right) = p_w
\]

by \( s_w/(2\|\lambda_1\|_2) \leq s_w \), and thus \(-s_w/(2\|\lambda_1\|_2) \geq -s_w \), and Assumption 2.5.

Consequently,

\[
P(|\lambda^T z_{t+1}| \geq s_z) \geq P(|\lambda^T z_{t+1}| \geq s_w/4) = P(-|\lambda^T z_{t+1}| - s_w/4 \leq -s_w/2)
\]

\[
\geq P(\lambda_1^T w_t \leq -s_w/2) \geq p_w
\]

which completes case 2.

**Case 3:** when \( \|\lambda_2\| \geq 1/k_0 \). Define \( v = \tilde{\eta}s_\eta/k_0 = \min(\sqrt{3}\tilde{\eta}s_\eta/2, s_w\tilde{\eta}s_\eta/(4b_u)) \). Define

\[
\Omega_1^\lambda = \{w_t \in \mathbb{R}^n \mid \lambda_1^T (w_t + \theta_\tau z_t) + \lambda_2^T (\pi_{t+1}(F^m_{t+1})) \geq 0\}
\]

\[
\Omega_2^\lambda = \{w_t \in \mathbb{R}^n \mid \lambda_1^T (w_t + \theta_\tau z_t) + \lambda_2^T (\pi_{t+1}(F^m_{t+1})) < 0\}
\]

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

\[
P(|\lambda^T z_{t+1}| \geq s_z) \geq P(|\lambda^T z_{t+1}| \geq v) = P(\lambda^T z_{t+1} \geq v) + P(\lambda^T z_{t+1} \leq -v)
\]

\[
\geq P(\lambda_1^T z_{t+1} \geq v, w_t \in \Omega_1^\lambda) + P(\lambda_1^T z_{t+1} \leq -v, w_t \in \Omega_2^\lambda)
\]

\[
\geq P(\lambda_2^T \eta_{t+1} \geq v, w_t \in \Omega_1^\lambda) + P(\lambda_2^T \eta_{t+1} \leq -v, w_t \in \Omega_2^\lambda)
\]

\[
= P(\lambda_2^T \eta_{t+1} \geq v)P(w_t \in \Omega_1^\lambda) + P(\lambda_2^T \eta_{t+1} \leq -v)P(w_t \in \Omega_2^\lambda)
\]

\[
\geq p_\eta
\]

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

\[
P(\lambda_2^T \eta_{t+1} \geq v) = P(\lambda_2^T \eta_{t+1}/\|\lambda_2\|_2 \geq v/\|\lambda_2\|_2)
\]

\[
= P(\lambda_2^T \eta_{t+1}/\|\lambda_2\|_2 \geq v/(\|\lambda_2\|_2\tilde{\eta}))
\]

\[
\geq P(\lambda_2^T \eta_{t+1}/\|\lambda_2\|_2 \geq k_0 v/\tilde{\eta})
\]

\[
= P(\lambda_2^T \tilde{\eta}_{t+1}/\|\lambda_2\|_2 \geq s_\eta) \geq p_\eta
\]
Then, \( \mathbb{P}(\lambda_2^T \eta_{t+1} \leq -v) = \mathbb{P}(-\lambda_2^T \eta_{t+1} \geq v) \geq p_\eta \)

This completes the proof of Case 3.

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

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

B.2. Proof of Corollary 4.2

Corollary 4.2 follows directly from Theorem 4.1. We only need to verify the boundedness of the states and actions. In the following, we will show that \( u_t \in \mathbb{U} \) for all \( t \) and \( \|x_t\|_2 \leq O(\sqrt{mn}) \) for all \( t \). Notice that though we can further show a much smaller bound \( \|x_t\|_2 \leq x_{\text{max}} \) with probability \((1 - p)\) in Theorem 4.4, Theorem 4.1 requires an almost sure bound and thus we provide a larger bound \( \|x_t\|_2 \leq O(\sqrt{mn}) \) here.

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

**Lemma B.4** (Action constraint satisfaction). When applying Algorithm 1, \( u_t \in \mathbb{U} \) for all \( t \) and for any \( w_k \in \mathbb{W} \).

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

\[
D_{u,j}^T u_t = D_{u,j}^T \sum_{k=1}^{H_{\text{r}}-1} M_t[k] \hat{w}_{t-k} + D_{u,j}^T \eta_t \leq \sum_{k=1}^{H_{\text{r}}-1} \|D_{u,j}^T M_t[k]\|_1 w_{\text{max}} + \|D_{u,\infty}\| \|\eta_t\|_\infty = g_{u,j}^0(M_t) + \|D_{u,\infty}\| \|\eta_t\|_\infty
\]

Our goal is to show that \( g_{u,j}^0(M_t) + \|D_{u,\infty}\| \|\eta_t\|_\infty \leq d_{u,j} \) for all \( j \) and for all \( t \geq 0 \). This is straightforward when \( t^*_j \leq t \leq t^*_j + T_{\text{b}} - 1 \) and \( t^*_j \leq t \leq T_{\text{c}} + 1 - 1 \). For example, when \( t^*_j \leq t \leq t^*_j + T_{\text{b}} - 1 \), we have \( M_t = M_t^j \) and \( \|\eta_t\|_\infty \leq \tilde{\eta}^j \), which leads to \( g_{u,j}^0(M_t^j) + \|D_{u,\infty}\| \|\eta_t\|_\infty = g_{u,j}^0(M_t^j) + c_3\|\eta_t\|_\infty \leq d_{u,j} \) by RobustCE and Lemma A.6.

Similar results can be shown for \( t^*_j \leq t \leq T_{\text{c}} + 1 - 1 \).

Next, we focus on the safe policy transition stages. It suffices to show that \( u_t = \sum_{k=1}^{H_{\text{r}}-1} M_t[k] \hat{w}_{t-k} + \eta_t \) in all stages of Algorithm 2. In the following, we will adopt the notations in Algorithm 2. In Step 1 of Algorithm 2, we have \( \|\eta_t\|_\infty \leq \eta_{\text{min}} \leq \bar{\eta} \) and \( M_t \in \Omega \) by the convexity of \( \Omega \). Therefore, we have \( g_{u,j}^0(M_t) + \|D_{u,\infty}\| \|\eta_t\|_\infty = g_{u,j}^0(M_t) + c_3\|\eta_t\|_\infty \leq d_{u,j} \), where we used the definition of \( \Omega \) in RobustCE. In Step 2, we have \( \|\eta_t\|_\infty \leq \tilde{\eta}^j \) and \( M_t \in \Omega^j \) by the convexity of \( \Omega^j \). Therefore, we have \( g_{u,j}^0(M_t) + \|D_{u,\infty}\| \|\eta_t\|_\infty = g_{u,j}^0(M_t) + c_3\|\eta_t\|_\infty \leq d_{u,j} \), where we used the definition of \( \Omega^j \) by RobustCE with input \( \tilde{\eta}^j \).

**Lemma B.5** (Almost sure upper bound on \( x_t \)). Consider DAP policy \( u_t = \sum_{k=1}^{H_0} M_t[k] \hat{w}_{t-k} + \eta_t \), where \( M_t \in \mathcal{M}_{H_t} \), \( \{H_t\}_{t=0}^\infty \) is non-decreasing, and \( \|\eta_t\|_\infty \leq \eta_{\text{max}} \). Suppose \( H_0 \geq \log(2k_0) / \log((1 - \gamma)^{-1}) \) and \( \eta_{\text{max}} \leq w_{\text{max}} / \kappa \). Let \( \{x_t, u_t\}_{t=0}^\infty \) denote the trajectory generated by this policy on the system with parameter \( \theta^* \) and disturbance \( w_t \). Then, there exists \( b_z = 4\sqrt{n\kappa w_{\text{max}} / \gamma} + 4\sqrt{mn\kappa^3} \kappa B w_{\text{max}} / \gamma = O(\sqrt{mn}) \) such that

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

This lemma is a natural extension of Lemma 2 in (Li et al., 2021) and the proof is deferred to Appendix G.3.

**Proof of Corollary 4.2.** By letting \( \delta^e = \frac{p}{\delta^e} \) for \( e \geq 1 \), we have that \( \|\hat{\theta}^e - \theta^*\|_2 \leq O\left(\sqrt{m + n} \frac{\log(\sqrt{mn})}{\sqrt{\delta^e}}\right) \) w.p. \( 1 - p / (2e^2) \). Notice that \( \|\hat{\theta}^e - \theta^*\|_F \leq \sqrt{n} \|\hat{\theta}^e - \theta^*\|_2 \), which completes the proof.
C. Feasibility

This appendix provides a proof for Theorem 4.3. We will first establish the recursive feasibility and then prove the initial feasibility. For notational simplicity, we define \( \Omega_0 := \Omega(\bar{\theta}^0, e^0_x + e_0, e_0^u). \)

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

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

\[
g_{ij}^a(M) \leq d_{u,j} - \epsilon_{n, u}(\bar{\eta}^0), \quad \forall 1 \leq j \leq k_u.
\]

Since \( \bar{\eta}^0 \geq \bar{\eta}^e \) by condition (ii) of Theorem 4.3, we have \( \epsilon_{n, u}(\bar{\eta}^0) \geq \epsilon_{n, u}(\bar{\eta}^e) \), so \( M \) satisfies the action constraints in \( \Omega_{\ast} \):

\[
g_{ij}^a(M) \leq d_{u,j} - \epsilon_{n, u}(\bar{\eta}^e), \quad \forall 1 \leq j \leq k_u.
\]

Next, we consider the state constraints. Notice that \( \hat{\theta}^e \in \Theta_{n,k} \) by ModelEst, so \( \|\hat{\theta}^e - \hat{\theta}_0\| \leq r^0 \) for \( \epsilon \geq 1 \). By Lemma A.4, for any \( M \in \Omega_0 \), we have

\[
g_{ij}^r(M; \hat{\theta}^e) \leq g_{ij}^r(M; \hat{\theta}_0^e) + \epsilon_{\theta}(r^0)
\]

Further, since \( r^e \leq r \leq \eta_0 / (c_1 \sqrt{mn}) \) by condition (ii) of Theorem 4.3, we have \( \epsilon_{\theta}(r^e) \leq \epsilon_{\theta}(r) \leq \epsilon_{\theta}(r^0) \), so

\[
g_{ij}^r(M; \hat{\theta}^e) \leq d_{x,i} - \epsilon_{n,x}(\bar{\eta}^0) - \epsilon_{H}(e_i) - \epsilon_{v}(\Delta_{M}^0, H^0) - \epsilon_0 + \epsilon_{\theta}(r^0)
\]

by condition (ii) of Theorem 4.3. So \( M \in \Omega_{\ast} \) for \( \epsilon \geq 1 \). Similarly, we can show \( M \in \Omega_{\ast} \) for \( \epsilon \geq 0 \). This completes the recursive feasibility.

**Proof of initial feasibility.** By Lemma 4 and Corollary 2 in (Li et al., 2021), we can construct \( M_F \) with length \( H^0 \) based on \( K_F \) in Assumption 2.3 such that

\[
g_{ij}^r(M_F; \theta_0) \leq d_{x,i} - \epsilon_{F,x} + \epsilon_{P}(H^0)
\]

\[
g_{ij}^u(M_F) \leq d_{u,j} - \epsilon_{F,u} + \epsilon_{P}(H^0).
\]

where \( \epsilon_{F} \) corresponds to \( \epsilon_1 + \epsilon_3 \) in (Li et al., 2021).\(^8\)

Therefore, by Lemma A.4, we have

\[
g_{ij}^r(M_F; \hat{\theta}^0) \leq d_{x,i} - \epsilon_{F,x} + \epsilon_{P}(H^0) + \epsilon_{\theta}(r^0) \quad g_{ij}^u(M_F) \leq d_{u,j} - \epsilon_{F,u} + \epsilon_{P}(H^0).
\]

Therefore, \( M_F \in \Omega_0 \) if Condition iii in Theorem 4.3 holds.

D. Constraint Satisfaction

This section provides a proof for the constraint satisfaction guarantee in Theorem 4.4. Notice that the control constraint satisfaction has already been established in Lemma B.4. Hence, we will focus on state constraint satisfaction in this appendix.

\(^8\)Notice that \( \epsilon_1 + \epsilon_3 = O(n \sqrt{mn} \gamma(1 - \gamma)^H) \) in (Li et al., 2021), but we improve the bound to \( \epsilon_F = O(\sqrt{mn} \gamma(1 - \gamma)^H) \). Specifically, \( \epsilon_3 = O(\sqrt{mn} \gamma(1 - \gamma)^H) \) remains unchanged, but we can show \( \epsilon_1(H) = O(\sqrt{mn} \gamma(1 - \gamma)^H) \). This is because the proof of Lemma 1 in (Li et al., 2021) shows that \( \epsilon_1 = O(b_0 \gamma(1 - \gamma)^H) \), where \( \|x_t\|_2 \leq b_0 \) a.s. In Lemma B.5, we show \( b_0 = O(\sqrt{mn}) \) in this paper, so we have \( \epsilon_1(H) = O(\sqrt{mn} \gamma(1 - \gamma)^H) \).
D.1. A General State Constraint Satisfaction Lemma

This subsection provides a general state constraint satisfaction lemma for time-varying approximate DAPs, which includes Lemma 3.1 as a special case.

Lemma D.1 (General Constraint Satisfaction Lemma). Consider the time-varying approximate DAPs in (9), where \( M_t \in \mathcal{M}_H \) for non-decreasing \( \{H_t\}_{t \geq 0} \), \( \theta_t \in \Theta_{m_t} \). Define

\[
\epsilon_{H,t} = (1 - \gamma)^{H_t} \cdot \|D_x\|_{\infty} \kappa x_{\text{max}}
\]

\[
\epsilon_{\theta, t} = c_1 \max_{0 \leq k \leq H_t} \|\hat{\theta}_{t-k} - \theta_*\|_F
\]

\[
\epsilon_{\eta, x, t} = c_2 \sqrt{m} \max_{1 \leq k \leq H_t} \bar{\eta}_{t-k},
\]

where \( c_1, c_2 \) are defined in Lemma A.4 and Lemma A.6, and we let \( M_t = M_0, \bar{\eta}_t = 0, H_t = H_0, \hat{\theta}_t = \theta_0, \omega_t = \bar{\omega}_t = x_t = 0 \), for \( t \leq -1 \).

For any \( t \geq 0 \), if \( x_s \in \mathcal{X}, u_s \in \mathcal{U} \) for all \( s \leq t - 1 \) and

\[
g^{\gamma}_t(M_t; \hat{\theta}_t) \leq d_{s,i} - \epsilon_{H,t} - \epsilon_{\eta, x, t} - \epsilon_{\theta, t} - \epsilon_{\omega, t}, \quad \forall 1 \leq i \leq k_x,
\]

then \( x_t \in \mathcal{X} \).

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

Proof. Consider stage \( t \geq 0 \). By Lemma A.1, for any \( 1 \leq i \leq k_x \), we have

\[
D_{x,i}^T x_t = D_{x,i}^T \Phi_k(M_{t-H_t-1}; \theta_*) \hat{\theta}_t - k_x + \sum_{k=1}^{H_t} D_{x,i}^T A_{k,i}^{-1} B_* M_{t-k} \hat{\omega}_{t-k} + \sum_{k=1}^{H_t} D_{x,i}^T A_{k,i}^{-1} B_* \eta_{t-k}
\]

\[
\leq \|D_x\|_{\infty} \kappa (1 - \gamma)^{H_t} x_{\text{max}} + g^{\gamma}_t(M_{t-H_t-1}; \theta_*) + \|D_x\|_{\infty} \kappa / \gamma \max_{1 \leq k \leq H_t} \|\hat{\omega}_{t-k} - \theta_*\|_2 \bar{z}_{\text{max}}
\]

\[
+ \|D_x\|_{\infty} \kappa \sqrt{m} \max_{1 \leq k \leq H_t} \|\hat{\eta}_{t-k}\|
\]

\[
\leq \epsilon_{H,t} + \epsilon_{\omega,t} + \epsilon_{\omega,t} - \|D_x\|_{\infty} \kappa / \gamma \max_{1 \leq k \leq H_t} \|\hat{\theta}_{t-k} - \theta_*\|_2 \bar{z}_{\text{max}} + \epsilon_{\eta, x, t}
\]

\[
\leq \epsilon_{H,t} + \epsilon_{\omega,t} + \epsilon_{\omega,t} + \|D_x\|_{\infty} \kappa / \gamma \max_{1 \leq k \leq H_t} \|\hat{\theta}_{t-k} - \theta_*\|_2 \bar{z}_{\text{max}} + \epsilon_{\eta, x, t}
\]
We discuss three possible cases based on the value of $T$

\textbf{Case 3: when $T \geq t_e + t_1 - 1$.} In this case, $M_t \in \Omega^{e-1}$, so

\[
g^e_t(M_t; \hat{\theta}^e) \leq d_{x,i} - \epsilon_H(H^{e-1}) - \epsilon_v(\Delta^{e-1}_M) - \epsilon_\theta(r^e)
\]

Notice that $H_t = H^{e-1}$, so $\epsilon_{H,t} = \epsilon_H(H^{e-1})$. Further, by our algorithm design, $\Delta_{M,t} \leq \Delta^{e-1}_M$. Since $\bar{W}_t \geq H^{e-1}$ and $\eta_k = 0$ for $t_{n}^{e} - 1 + T_{D}^{e} - 1 \leq k \leq t$, we have $\epsilon_{\eta,x,t} = \max_{1 \leq k \leq t} e_{\eta,t} e_{\eta,x,t} \epsilon_{\eta} = c_2 \sqrt{\lambda_{t}} 0. $ Next, since $r^e \leq r^e - 1$ by Condition 2 of Theorem 4.3 for $e \geq 1$, $\epsilon_{\theta,t} \leq \epsilon_\theta(r^e)$. So we satisfy (12).

\textbf{Case 2: when $T^e + W_1^e \geq t \leq t_1^e + T_D^e + \bar{W}_1^e - 1$.} We have $M_t \in \Omega^{e-1}_1$. So

\[
g^e_t(M_t; \hat{\theta}^e) \leq d_{x,i} - \epsilon_H(H^e) - \epsilon_v(\Delta^e_M) - \epsilon_\theta(r^e)
\]

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

\textbf{Case 3: when $t_1^e + T_D^e + \bar{W}_1^e \geq t \leq T^{e+1} - 1$.} We have $M_t \in \Omega^e$. So

\[
g^e_t(M_t; \hat{\theta}^{e+1}) \leq d_{x,i} - \epsilon_H(H^e) - \epsilon_v(\Delta^e_M) - \epsilon_\theta(r^{e+1})
\]

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

In conclusion, we satisfy (12) for all $t \geq 0$. By Lemma D.1 and Lemma B.5, we can show state constraint satisfaction under $E_{safe}$.

\section{E. Regret Analysis}

In this section, we provide a proof for Theorem 4.5. Specifically, we first prove the regret bound and then verify the conditions for feasibility and constraint satisfaction. Before the formal proof, we note that the statement in Theorem 4 has the following typos.

\[
\leq \epsilon_{H,t} + \hat{g}^e_t(M_t; \hat{\theta}_t) + \epsilon_{\theta,t} + \epsilon_{v,t} + \epsilon_{\eta,x,t}
\]

\[
\leq d_{x,i}
\]

where we used Lemma A.5, Lemma A.4, $x_s \in X$, $u_s \in U$ for all $s \leq t - 1$, and (12). The last inequality guarantees $x_t \in X$. Therefore, the proof can be completed by induction.
E.1. Proof of the Regret Bound

Our proof of the regret bound relies on decomposing the regret into several parts and bounding each part. First, we decompose the \( T \) stages into two parts and decompose the regret accordingly. For \( e \geq 0 \), define

\[
T_1^e = \{ t^e \leq t \leq t_2(e) + H^e - 1 \}, \quad T_2^e = \{ t_2(e) + H^e \leq t \leq T^{e+1} - 1 \}.
\]

Then, decompose the regret by the stage decomposition below:

\[
\text{Regret} = \sum_{t=0}^{T-1} (l(x_t, u_t) - J^*) = \sum_{e=0}^{N-1} \sum_{t \in T_1^e} (l(x_t, u_t) - J^*) + \sum_{e=0}^{N-1} \sum_{t \in T_2^e} (l(x_t, u_t) - J^*)
\]

(14)

The first term can be bounded straightforwardly by the fact that the single-stage regret is bounded and the total number of stages in \( T_1^e \) for all \( e \) can be bounded by \( O(T^{2/3}) \).

**Lemma E.1 (Regret Bound of the First Term).** When the event \( \mathcal{E}_{\text{safe}} \) defined in (13) happens, under the conditions in Theorem 4.5, we have

\[
\sum_{e=0}^{N-1} \sum_{t \in T_1^e} (l(x_t, u_t) - J^*) \leq O(T^{2/3})
\]

**Proof.** When \( \mathcal{E}_{\text{safe}} \) is true, by Theorem 4.4, we have \( x_t \in \mathcal{X} \) and \( u_t \in \mathcal{U} \), thus \( \|x_t\|_2 \leq x_{\text{max}} \) and \( \|u_t\|_2 \leq u_{\text{max}} \) and \( l(x_t, u_t) - J^* \leq \|Q\|_2 x_{\text{max}}^2 + \|R\|_2 u_{\text{max}}^2 = O(1) \).

Next, we bound the number of stages in \( T_1^e \). Under the conditions in Theorem 4.5, the number of the stages in \( T_1^e \) is \( T_1^e + H^e \) plus the safe policy transition stages in Phase 1 and Phase 2. Since \( \mathcal{M}_{H^e} \) is a bounded set, the number of stages in SafeTransit between any two policies in \( \mathcal{M}_{H^e} \) can be bounded by \( O(\max(1/\Delta^e_M, H^e)) = O(\sqrt{\log T^{2/3}}) \), where we used \( H^e = O(\log(T^{e+1})) \) and \( \Delta^e_M = O(e_{\text{safe}} H^e) \). Further, by \( T^{e+1} = 2T^e \), \( T_D^e = (T^{e+1} - T^e)^{2/3} \), we have \( T_D^e = O((T^{e+1})^{2/3}) \). Consequently, the total number of stages in \( T_1^e \) can be bounded by \( O((T^{e+1})^{2/3}) \) (notice that \( T^{e+1})^{1/3} \geq \sqrt{\log T} \) by our condition of \( T^1 \) in Theorem 4.5).

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

Lemma E.2 is a technical fact that will be used throughout our regret proof.

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

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

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

The second term in (14) is more complicated to bound, so we further decompose it into four parts as follows.
where we introduced auxiliary states \( \hat{x}_t \) and actions \( \hat{u}_t \) defined as

\[
\hat{x}_t = \sum_{k=1}^{2H^{*}} \Phi^*_k(M^*; \theta_t) w_{t-k}, \quad \hat{u}_t = \sum_{k=1}^{H^{*}} M^*[k] w_{t-k},
\]

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

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

### E.1.1. Bound on Part iii

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

Specifically, we establish the following general perturbation bound. This bound is not only useful for our bound on Part iii but also helps the discussions after Theorem 4.5 on the reasons for including the pure exploitation phases.

**Lemma E.3** (Perturbation analysis for CCE). Consider a fixed memory length \( H \geq \log(2\kappa)/\log((1 - \gamma)^{-1}) \) and \( \theta_1, \theta_2 \in \Theta_{ini} \). Consider two CCE programs \( M_1 = \arg\min_{M \in \Omega_{t^*}} f(M; \theta_1) \) and \( M_2 = \arg\min_{M \in \Omega_{t^*}} f(M; \theta_2) \). Suppose there exists \( \epsilon_g > 0 \) such that \( \Omega(\theta_1, \epsilon_{x1}, \epsilon_{u1}, \epsilon_g) \cap \Omega(\theta_2, \epsilon_{x2}, \epsilon_{u2}, \epsilon_g) \) is non-empty. Then, we have

\[
f(M_1, \theta_2) - f(M_2, \theta_2) \leq O(mn r + (\sqrt{mn} + \sqrt{k_x + k_u}) m H \max(|\epsilon_{x1} - \epsilon_{x2}|, |\epsilon_{u1} - \epsilon_{u2}|))/\epsilon_g\]

where \( ||\theta_1 - \theta_2||_F \leq r \).

**Proof.** Notice that both the objective functions and the constraints are different in the two CCE programs above, so we introduce an auxiliary policy \( M_3 = \arg\min_{M \in \Omega_{t^*}} f(M; \theta_2) \) to discuss the perturbation bounds on cost differences and constraint differences separately. We will first bound \( f(M_1, \theta_2) - f(M_3, \theta_2) \) and then bound \( f(M_3, \theta_2) - f(M_2, \theta_2) \) below.

**Perturbation on the cost functions.**

\[
f(M_1, \theta_2) - f(M_3, \theta_2) = f(M_1, \theta_2) - f(M_1, \theta_1) + f(M_1, \theta_1) - f(M_3, \theta_1) + f(M_3, \theta_1) - f(M_3, \theta_2)
\]

\[
\leq f(M_1, \theta_2) - f(M_1, \theta_1) + f(M_3, \theta_1) - f(M_3, \theta_2)
\]

\[
\leq O(mn ||\theta_1 - \theta_2||_F)
\]

where the first inequality is because \( M_1 \) and \( M_3 \) are in the same set and \( M_1 \) minimizes the cost \( f(M, \theta_1) \) in this set, and the second inequality is because of the following perturbation lemma on the cost functions.

**Lemma E.4** (Perturbation bound on \( f \) with respect to \( \theta \)). For any \( H \geq \log(2\kappa)/\log((1 - \gamma)^{-1}) \), \( M \in \mathcal{M}_H \), any \( \theta, \hat{\theta} \in \Theta_{ini} \), we have

\[
||f(M; \theta) - f(M; \hat{\theta})|| \leq O(mn r)
\]

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

**Proof.** We let \( \hat{x}_t(\theta) \) and \( \hat{x}_t(\hat{\theta}) \) denote the approximate states defined by Proposition 2.7, and we omit \( M \) in this proof for notational simplicity. Notice that

\[
||\hat{x}(\theta) - \hat{x}(\hat{\theta})||_2 = ||\sum_{k=1}^{2H} (\Phi^*_k(\theta) - \Phi^*_k(\hat{\theta})) w_{t-k}||_2 \leq \sum_{k=1}^{2H} ||(\Phi^*_k(\theta) - \Phi^*_k(\hat{\theta})) w_{t-k}||_2
\]
vise versa, so we need to generalize Proposition 2 to this setting as follows.

Lemma E.6 on absolute values (see Lemma 10 in (Li et al., 2021) for more details). Therefore, we can apply Lemma E.5 to obtain the bound.

Now, we are ready to bound \( f(M_3, \theta_2) - f(M_2, \theta_2) \) is established based on Proposition 2 and Lemma 9 in (Li et al., 2021). The major difference is that (Li et al., 2021) only considers changes in the right-hand-side of the constraint inequalities but in our setting, the left-hand-side of the constraint inequalities also change. To handle this difference, we introduce two auxiliary sets \( \Omega_1 \) and \( \Omega_2 \) with the same left-hand-sides in the constraint inequalities such that \( \Omega_1 = \Omega(\theta_1, \epsilon_1, \epsilon_{u1}) \) and \( \Omega_2 = \Omega(\theta_2, \epsilon_2, \epsilon_{u2}) \), which is achieved by adding inactive inequality constraints. Specifically, define

\[
\Omega_1 = \Omega(\theta_1, \epsilon_1, \epsilon_{u1}) \cap \Omega(\theta_2, \epsilon_1 - \epsilon_2(r), \epsilon_{u1}), \quad \Omega_2 = \Omega(\theta_2, \epsilon_2, \epsilon_{u2}) \cap \Omega(\theta_1, \epsilon_2 - \epsilon_2(r), \epsilon_{u2}).
\]

Notice that the constraints in \( \Omega(\theta_2, \epsilon_1 - \epsilon_2(r), \epsilon_{u1}) \) and \( \Omega(\theta_1, \epsilon_2 - \epsilon_2(r), \epsilon_{u2}) \) are inactive due to Lemma A.4. Further, notice that \( \Omega_1, \Omega_2 \) are both polytopes.

Another difference between our setting and that in Proposition 2 of (Li et al., 2021) is that \( \Omega_1 \) may not be a subset of \( \Omega_2 \) and vice versa, so we need to generalize Proposition 2 to this setting as follows.

Lemma E.5 (Extension from Proposition 2 (Li et al., 2021)). Consider two polytopes: \( \Omega_1 = \{ x : Cx \leq h - \Delta_1 \} \), \( \Omega_2 = \{ x : Cx \leq h - \Delta_2 \} \), where \( \Delta_1, \Delta_2 \) are two vectors. Define \( \Delta_0 = \min(\Delta_1, \Delta_2) \) elementwise. Define \( \Delta_1 = \max(\Delta_1, \Delta_2) \) elementwise. Suppose the l2-diameter of \( \Omega_0 \) is \( d_{\Omega_0} \), \( f(x) \) is L-Lipschitz continuous, and there exists \( x_F \in \Omega_3 \), then we have

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

The proof is deferred to Appendix G.5.

Now, we are ready to bound \( f(M_3, \theta_2) - f(M_2, \theta_2) \). By our definitions and discussions above, we have \( f(M_3, \theta_2) = \min_{M \in \Omega_1} f(M; \theta_2) \) and \( f(M_2, \theta_2) = \min_{M \in \Omega_2} f(M; \theta_2) \). Further, notice that \( \Omega_1, \Omega_2 \) are both polytopes and can be written in the form of \( \{ \tilde{P} : C\tilde{P} \leq h - \Delta_1 \} \), \( \{ \tilde{P} : C\tilde{P} \leq h - \Delta_2 \} \) by introducing auxiliary variables to represent the absolute values (see Lemma 10 in (Li et al., 2021) for more details). Therefore, we can apply Lemma E.5 to obtain the bound on \( f(M_3, \theta_2) - f(M_2, \theta_2) \) by bounding the corresponding constants \( d_{\Omega_0}, L, \| \Delta_1 - \Delta_2 \|_\infty \), and \( \min_{i \in \{i : (\Delta_1) \neq (\Delta_2)\}} (h - \Delta_3 - Cx_F)_i \). Same as the proof of Lemma 9, we can show the l2-diameter of \( \Omega_0 \) is \( d_{\Omega_0} = O(\sqrt{m} + \sqrt{E} + \sqrt{G}) \) and \( \| \Delta_1 - \Delta_2 \|_\infty = \max(\| \epsilon_1 - \epsilon_2 \|, \| \epsilon_{u1} - \epsilon_{u2} \|) \). Further, the Lipschitz factor \( L \) can be obtained from the gradient bound \( L = G_f = O(\sqrt{n^2mH}) \) as provided below, whose proof is provided in Appendix G.4.

Lemma E.6 (Gradient bound of \( f(M; \theta) \)). For any \( H \geq 1, M \in M_H, \theta \in \Theta_{in} \), we have \( \| \nabla f(M; \theta) \|_F \leq G_f = O(\sqrt{n^2mH}) \).
Next, since \( \Omega(\theta_1, \epsilon_1 + \epsilon_g, \epsilon_u + \epsilon_g) \cap \Omega(\theta_2, \epsilon_x, \epsilon_g, \epsilon_u + \epsilon_g) \) is non-empty, there exists \( \tilde{P}_F \in \Omega_3 \) such that 
\[
\min_{(h: \Delta_3, \neq (\Delta_2), h) \geq \epsilon_0}, \text{Lastly, by applying the constants above to Lemma E.5, we can show}
\[
f(M_3, \theta_2) - f(M_2, \theta_2) \leq \hat{O}( (\sqrt{\bar{m}n} + \sqrt{k_x + k_u}) \sqrt{mn} ) \max(|\epsilon_x|, |\epsilon_u|)/\epsilon_g.
\]
The proof of Lemma E.3 is completed by combining the bounds on \( f(M_1, \theta_2) - f(M_3, \theta_2) \) and \( f(M_3, \theta_2) - f(M_2, \theta_2) \).

Based on Lemma E.3, we can show the following bound on Part iii when \( \epsilon_\text{safe} \) is true.

\[
\text{Part iii} = \sum_{\epsilon = 0}^{N-1} \sum_{t \in T_2^*} (f(M^\epsilon; \theta_*^t) - f(M_{H^\epsilon}; \theta_*^t)) \leq \hat{O}(m^2 + n^{2.5}m^{1.5}) \sqrt{mn + k_x + k_u T^{2/3}}
\]

The proof for (15) is provided below. For each \( \epsilon \geq 0 \), notice that \( M^\epsilon = \arg\min_{M \in \Omega(\theta^+, \epsilon_0, 0)} f(M; \theta^+ + \epsilon) \), where we define \( \tilde{e}_x^\epsilon = \epsilon_0(e^{+1} + \epsilon)(H* + \epsilon)(\Delta_3^*, H^*) \); and \( M_{H^\epsilon} = \arg\min_{M \in \Omega(\theta^*, \epsilon_\mu(H^*), 0)} f(M; \theta^*) \). Therefore, we can apply Lemma E.3. First verify the conditions of Lemma E.3. Notice that \( \theta^{+1}, \theta^* \in \Theta_{\min} \) is large enough. Further, we have \( \epsilon_g = \min(\epsilon_1, \epsilon_2)/4 > 0 \) such that \( \Omega(\theta^{+1}, \tilde{e}_x^\epsilon + \epsilon_g, \epsilon_g) \cap \Omega(\theta^*, \epsilon_\mu(H^*) + \epsilon_g, \epsilon_g) \) is not empty due to the feasibility conditions in Theorem 4.3. Therefore, when \( \epsilon_\text{safe} \) is true, by Lemma E.3, under our choices of parameters in Theorem 4.5, we have

\[
f(M^\epsilon; \theta_*) - f(M_{H^\epsilon}; \theta_*) \leq \hat{O}(mn^{\epsilon^{+1}} + (\sqrt{\bar{m}n} + \sqrt{k_x + k_u}) n\sqrt{m} H^*(\epsilon_0^{+1} + \epsilon_\epsilon(h_0, H^*))
\]

\[
\leq \hat{O}(mn(n\sqrt{m} + m\sqrt{n}) (\epsilon^{+1} + \epsilon_\epsilon(h_0, H^*)))
\]

\[
\leq \hat{O}(n^{2.5}m^{1.5}) \sqrt{mn + k_x + k_u T^{2/3}}
\]

Consequently, by Lemma E.2, we prove the bound (15).

### E.1.2. Bound on Part i

When \( \epsilon_\text{safe} \) is true, we are able to show

\[
\text{Part i} = \sum_{\epsilon} \sum_{t \in T_2^*} l(x_t, u_t) - l(\hat{x}_t, \hat{u}_t) \leq \hat{O}(n\sqrt{m} \sqrt{m + n} T^{2/3}).
\]

The proof is provided below.

Firstly, under \( \epsilon_\text{safe} \), we have \( x_t \in X \) and \( u_t \in U \), so \( \|x_t\|_2 \leq O(1), \|u_t\|_2 \leq O(1) \). Further, by the definitions of \( \hat{x}_t, \hat{u}_t \), and the proof of Theorem 4.4, we can also verify that \( \hat{x}_t \in X \) and \( \hat{u}_t \in U \) based on the proof of Lemma B.4 and Lemma D.1. Therefore, we have \( \|\hat{x}_t\|_2 \leq O(1), \|\hat{u}_t\|_2 \leq O(1) \)

Next, by Lemma A.1, for \( t \in T_2^* \), we can write \( x_t, \hat{x}_t, u_t, \hat{u}_t \) as

\[
x_t = A^H_{x} x_{t-H_x} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} A^{i-1}_{x} B_s M_{t-i}[k-i] u_{t-k} \chi_{1 \leq i \leq H_{t-i}} + \sum_{i=1}^{H_t} A^{i-1}_{x} u_{t-i}
\]

\[
\hat{x}_t = A^H_{x} x_{t-H_x} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} A^{i-1}_{x} B_s M_{t-i}[k-i] u_{t-k} \chi_{1 \leq i \leq H_{t-i}} + \sum_{i=1}^{H_t} A^{i-1}_{x} u_{t-i}
\]

\[
u_t = \sum_{i=1}^{H_t} M_t[k] u_{t-k}, \quad \hat{u}_t = \sum_{i=1}^{H_t} M_t[k] u_{t-k}.
\]

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

\[
\|x_t - \hat{x}_t\|_2 \leq O((1 - \gamma) H^* + \sqrt{mn} \epsilon^{+1}) = \hat{O}(mn \sqrt{m + n}(T^{+1})^{-1/3})
\]

\[
\|u_t - \hat{u}_t\|_2 \leq O(\sqrt{mn} \epsilon^{+1}) = \hat{O}(nm \sqrt{m + n}(T^{+1})^{-1/3})
\]
Consequently, by applying Lemma E.2 and the quadratic structure of \( l(x, u) \), we can bound the Part i by
\[
\sum_e \sum_{t \in T_2^e} (l(x_t, u_t) - l(\hat{x}_t, \hat{u}_t)) \leq \sum_e T^{e+1} \tilde{O}(nm \sqrt{m + n(T^{e+1})^{-1/3}}) \leq \tilde{O}(nm \sqrt{m + nT^{2/3}}).
\]

**E.1.3. Bound on Part II**

**Lemma E.7** (Bound on Part ii). With probability \( 1 - p \), Part ii \( \leq \tilde{O}(mn \sqrt{T}) \).

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

**E.1.4. Bound on Part iv**

In the following, we will show that
\[
\text{Part iv} = \sum_{e=0}^{N-1} \sum_{t \in T_2^e} (f(M^*_e; \theta_s) - J^* - J^*) = \tilde{O}(n \sqrt{m} \sqrt{mn + k_c \sqrt{m}})
\]  

(17)

The proof is provided below.

Remember that \( J^* \) is generated by the optimal safe linear policy \( K^* \). By Lemma 4 and Corollary 2 in (Li et al., 2021), for a memory length \( H^e \), we can define \( M^*(K^*) \in \Omega(\theta_s, -\epsilon^+_p, 0) \), where \( \epsilon^+_p = \sqrt{mn(1 - \gamma)}H^e \) corresponds to \( \epsilon_1 + \epsilon_3 \) in (Li et al., 2021) with \( H = H^e \).\(^9\) Further, by Lemma 6 in (Li et al., 2021), we have
\[
f(M^*_e; \theta_s) - J^* = \lim_{T+\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} f(M^*_e; \theta_s) - \mathbb{E}(l(x_t^*, u_t^*)) \leq O(n^2 m(H^e)^2 (1 - \gamma)^{H^e})
\]

In addition, we have
\[
f(M^*_e; \theta_s) - f(M^*_e; \theta_s) \leq f(M^*_e; \theta_s) - \min_{M^* \in \Omega(\theta_s, -\epsilon^+_p, 0)} f(M; \theta_s) 
\leq \tilde{O}(n \sqrt{m} \sqrt{mn + k_x + k_u(\epsilon^+_p + \epsilon_H(H^e)))} 
\leq \tilde{O}(n \sqrt{m} \sqrt{mn + k_c \sqrt{mn(1 - \gamma)^{H^e}}})
\]

By combining the bounds above and by choosing \( H^e \geq \log(T^{e+1}) / \log((1 - \gamma)^{-1}) \), we have
\[
f(M^*_e; \theta_s) - J^* = f(M^*_e; \theta_s) - f(M^*_e; \theta_s) + (M^*_e; \theta_s) - J^* 
\leq \tilde{O}(n \sqrt{m} \sqrt{mn + k_c \sqrt{mn(T^{e+1})}),
\]

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

**Completing the Proof of the Regret Bound.**

By combining (16), (15), (17), and Lemma E.7, we obtain our regret bound in Theorem 4.5. Notice that (16), (15), (17) all condition on \( E_{safe} \), and \( E_{safe} \) holds w.p. \( 1 - p \). But Lemma E.7 conditions on a different event and that event also holds with probability \( 1 - p \). Putting them together, we have that our regret bound holds w.p. \( 1 - 2p \).

**E.2. Condition Verification for Feasibility and Constraint Satisfaction**

In this subsection, we briefly show that there exist parameters characterized by Theorem 4.5 that satisfy the conditions for feasibility and constraint satisfaction in Theorem 4.3 and Theorem 4.4, which include: \( T^e_D \) satisfying the condition in Theorem 4.1 (Corollary 4.2’s condition), condition (iii), condition (ii) of Theorem 4.3, and \( T^{e+1} \geq t^e_0 \).

\(^9\)As discussed in footnote 8, our \( \epsilon^+_p \) has smaller dependence on \( n, m, H \) compared with (Li et al., 2021).
Firstly, in Corollary 4.2, we need $T^e_\epsilon \geq O(\log(2\epsilon^2/p) + (m + n) \log(1/\eta))$ for $\epsilon \geq 0$. By our choices, we have $T^e_\epsilon = (T^{(\min(\epsilon,1))}/3$ and $T^{e+1} = 2T^e$, so $T^e_\epsilon$ increases exponentially. Therefore, $T^e_\epsilon \geq O(\log(2\epsilon^2/p) + (m + n) \log(1/\eta))$ can be guaranteed if $T^e_\epsilon \geq O((m + n) \log(1/\eta))$) for some sufficiently large constant factor, which requires $T^1 \geq O((m + n)^{3/2})$.

Secondly, for condition (ii), we set $\epsilon_0 = \epsilon_{F,\epsilon}/4$ and let $\epsilon_P + \epsilon_H(H^0) \leq \epsilon_{F,\epsilon}/12$, $\epsilon_{\eta,\epsilon} \leq \epsilon_{F,\epsilon}/12$, $\epsilon_\epsilon(O_{-}, H^0) \leq \epsilon_{F,\epsilon}/12$, and $\epsilon_P \leq \epsilon_{F,\epsilon}/4$, $\epsilon_{\eta,\epsilon} \leq \epsilon_{F,\epsilon}/4$. These conditions require $H^0 \geq O(\log(\sqrt{mn}/\min(\epsilon_F)), \eta^0 = O(\min(\epsilon_F^0, \epsilon_P^0))$, and $\Delta^e_H = O(\sqrt{mn}H^0/(T^{e+1})^{1/3})$.

Thirdly, for the condition (ii) of Theorem 4.4, the monotonicity for $H^e$, $\sqrt{mn} \Delta^e_M$ are satisfied and $\eta^e$ is a constant, so its monotonicity condition is also satisfied. With exponentially increasing $T^e_\epsilon$, the decreasing $r^e$ is also satisfied. We only need to verify that $r^e \leq r_m$. This requires $T^1 \geq \tilde{O}((\sqrt{mn} + n\sqrt{\eta}))$.

Lastly, for $T^{e+1} \geq T^e_2$, notice that Phase 1 only takes $(T^{e+1} - T^e)^{2/3}$ stages, and the safe transitions only takes $\tilde{O}((T^{e+1})^{1/3})$ stages, so $T^{e+1} \geq T^e_2$ for all $\epsilon$ for large enough initial $T^1$.

F. More Discussions

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

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

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

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

**Definition F.1.** $\Xi$ is called a disturbance-invariant set for $x_{t+1} = A x_t - B \xi x_t + w_t$ if $x_{t+1} \in \Xi$ for any $x_t \in \Xi, w_t \in \mathcal{W}$.

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 $\mathcal{X}_f$ such that for any $x_0 \in \mathcal{X}_f$, implementing the controller $u_t = -K_t x_t$ is safe, as well as a terminal cost function $V_f(x) = x^T P x$ satisfying certain conditions (see Mayne et al., 2005 for more details).

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

$$
\min_{x_{t|t+k} \in \mathcal{X}} \sum_{k=0}^{W-1} l(x_{t+k|t}, u_{t+k|t}) + V_f(x_{t+W|t})
$$

s.t. $x_{t+k+1|t} = A x_{t+k|t} + B u_{t+k|t}$, $k \geq 0$

$x_{t|t} \in x_t \oplus \Xi$

$x_{t+k|t} \in \mathcal{X} \oplus \Xi, \forall 0 \leq k \leq W - 1$

$u_{t+k|t} \in \mathcal{U} \oplus \mathcal{K} \Xi, \forall 0 \leq k \leq W - 1$

$x_{t+W|t} \in \mathcal{X} f \subseteq \mathcal{X} \oplus \Xi

(RMPC (Mayne et al., 2005))
Then, implement control:

\[ u_t = -K(x_t - x_t^*) + u_t^*. \]

Notice that \( x_t^*, u_t^* \) are functions of \( x_t \). Further, by (Bemporad et al., 2002), \( u_t \) is a piece-wise affine (PWA) function of the state \( x_t \) when \( \Xi \) is a polytope. Define the set of feasible initial values as

\[ X_N = \{ x_0 : (\text{RMPC (Mayne et al., 2005)}) \text{ is feasible when } x_1 = x_0 \}. \]

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

**Theorem F.2** (Theorem 1 in (Mayne et al., 2005)). The set \( \Xi \) is robustly exponentially stable for the closed-loop system with (RMPC (Mayne et al., 2005)) for \( w_k \in \mathbb{W} \) with an attraction region \( X_N \), i.e., there exists \( c > 0, \gamma_1 \in (0, 1) \), such that for any \( x_0 \in X_N \), for any \( w_k \in \mathbb{W} \),

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

Theorem F2 suggests that (RMPC (Mayne et al., 2005)) can quickly reduce the distance between \( x_t \) and \( \Xi \), i.e. it can drive a large initial state \( x_0 \neq 0 \) quickly to a neighborhood around \( \Xi \), which is also a neighborhood around the origin.

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

**Theorem F.3** (Connection between RMPC in (Mayne et al., 2005) and linear control’s infinite-horizon costs). Consider (RMPC (Mayne et al., 2005)) defined above with \( K \) satisfying the requirements in (Mayne et al., 2005). For any \( x_0 \in X_N \), the infinite-horizon averaged cost of RMPC in (Mayne et al., 2005) equals the infinite-horizon averaged cost of \( K \), i.e.

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

The proof is deferred to the end of this appendix.

Notice that \( K \) is a pre-fixed safe linear policy, so by Theorem F.3, we have \( J(K^*) \leq J(\text{RMPC in (Mayne et al., 2005)}) \), where \( K^* \) is our regret benchmark, i.e., the optimal safe linear policy. This suggests that RMPC in (Mayne et al., 2005) achieves similar or worse performance than the optimal safe linear policy in the long run. Since our adaptive control algorithm enjoys a sublinear regret compared to the optimal safe linear policy, Theorem F.3 suggests that our algorithm achieves the same regret bound even if we include RMPC in (Mayne et al., 2005) to the benchmark policy set. Further, if \( K \neq K^* \), our adaptive algorithm can even achieve better performance than RMPC in (Mayne et al., 2005) at around the equilibrium point 0.

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

**Remark F.4.** The policy set (3) includes safe DAPs starting from \( x_0 = 0 \). With a non-zero \( x_0 \), additional linear constraints should be imposed to ensure state constraint satisfaction for \( t < H \) because Proposition 2.7 only holds for \( t \geq H \) with a nonzero \( x_0 \). Due to the additional constraints, only small \( x_0 \) can ensure the existence of a safe DAP linear policy. For larger \( x_0 \), other types of policies should be considered to ensure safety, e.g., (Mayne et al., 2005). For too large \( x_0 \), it is possible that no policy can ensure safety. See e.g., (Rawlings & Mayne, 2009), for more discussions.

**Remark F.5.** Since our proof relies on the robust exponential stability property of RMPC in (Mayne et al., 2005), for other RMPC schemes without this property, we still cannot include them to our benchmark policy class and generate a sublinear regret. We leave the regret analysis compared with other RMPC schemes without robust exponential stability as future work. Further, we note that there are a few papers on the regret analysis with RMPC as the benchmark, e.g., (Wabersich & Zeilinger, 2018; Muthirayan et al., 2020). However, (Wabersich & Zeilinger, 2018) allows constraint violation during

$^{10}$Though RMPC in (Mayne et al., 2005) requires a known model, there are standard approaches to extend RMPC to handle model uncertainties, e.g., (Köhler et al., 2019; Lu et al., 2019).
the learning process and allows restarts when policies are updated, and (Muthirayan et al., 2020) does not consider state constraints and the proposed algorithm involves an intractable oracle. In conclusion, the regret analysis with RMPC as the benchmark is largely under-explored and is an important direction for future research.

**Proof of Theorem F.3.** To prove Theorem F.3, we introduce some necessary results from the existing literature and some lemmas based on these existing results.

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

**Proposition F.6** (Corollary 2 and Theorem 4 and Section 4.4 in (Bemporad et al., 2002)). Consider (CLQR) with p.d. quadratic costs and polytopic constraints below:

\[
\min_{u_{t+k|t}} \sum_{k=0}^{W-1} l(x_{t+k|t}, u_{t+k|t}) + x_{t+W|t}^TPx_{t+W|t}
\]

s.t.

- \(x_{t+k+1|t} = A_s x_{t+k|t} + B_s u_{t+k|t}, \quad k \geq 0\)
- \(D_x x_{t+k|t} \leq d_x, \quad \forall 0 \leq k \leq W - 1\)
- \(D_u u_{t+k|t} \leq d_u, \quad \forall 0 \leq k \leq W - 1\)
- \(D_{term} x_{t+W|t} \leq d_{term}\)
- \(x_t = x\)

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,

\[\pi_{CLQR}(x) = K_ix + b_i, \quad G_ix \leq h_i, \quad i = 0, 1, \ldots, N_{clqr}\]

Further, the number of different gain matrices can bounded by a constant \(N_{clqr} - gain\) that only depends on the dimensionality of the problem.

Based on Proposition F.6, we have that \(\pi_{CLQR}(x)\) is Lipschitz continuous with Lipschitz factor \(L_{CLQR} = \max_i \|K_i\|_2\) since \(\pi_{CLQR}(x)\) is continuous and piecewise-affine with respect to \(x\).

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

**Proposition F.7** (See the proof of Theorem 1 in (Mayne et al., 2005)). There exists \(c_1 > 0\) and \(\rho \in (0, 1)\) such that for any \(x_0 \in X_N\), and for any admissible disturbances \(w_k\), we have

\[\|x^*_t|_t(x_t)\|_2 \leq c_1 \rho^t \|x^*|_0(x_0)\|_2\]

Based on this, we can also show the exponential decay of \(u^*_t|_t(x_t)\).

**Lemma F.8.** 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 continuous with a finite factor denoted as \(L_{rmpec}\) on a convex feasible set. Further, we have \(\|u^*_t|_t(x_t)\|_2 \leq c_2 \rho^t\), where \(c_2 = L_{rmpec} c_1 x_{max}\).

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

Further, notice that \(u^*_t|_t(0) = 0\). Therefore,

\[\|u^*_t|_t(x^*_t|_t)\|_2 = \|u^*_t|_t(x^*_t|_t) - u^*_t|_t(0)\|_2 \leq L_{rmpec} \|x^*_t|_2 \leq \|L_{rmpec} c_1 \rho^t \|x^*|_0(x_0)\|_2 \leq c_2 \rho^t\]

where \(c_2 = L_{rmpec} c_1 x_{max}\).  

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

**Lemma F.9.** Consider $y^+ = A_k y + w$, where $y_0 = x_0 \in X$ and $p = -K y$. Since $K$ is $(\kappa, \gamma)$ strongly convex, both $y$ and $p$ are bounded by

\[
\|y_t\|_2 \leq \|u\|_2 \kappa^2 / \gamma + \kappa^2 x_{\text{max}} = y_{\text{max}}, \quad \|p_t\|_2 \leq \|u\|_2 \kappa^3 / \gamma + \kappa^2 x_{\text{max}} = p_{\text{max}}.
\]

Now, we are ready for the proof of Theorem F.3.

**Proof of Theorem F.3.** The closed-loop system of (RMPC (Mayne et al., 2005)) is

\[
x_{t+1} = A_s x_t + B_s x_{RMP}^1(x_t) + w_t = A_s x_t - B_s K x_t + B_s (K x_{\text{a}1}(x_t) + u_{\text{a}1}(x_t)) + w_t.
\]

Consider a possibly unsafe system:

\[
y_{t+1} = A_s y_t + B_s p_t + w_t, \quad p_t = -K y_t
\]

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

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

\[
e_{t+1} = A_k e_t + v_t
\]

where $A_k = A_s - B_s K$, and $v_t = B_s (K x_{\text{a}1}(x_t) + u_{\text{a}1}(x_t))$. Notice that by Proposition F.7 and Lemma F.8, we have

\[
\|v_t\|_2 \leq \|B_s\|_2 (\kappa c_1 x_{\text{max}} + c_2), \quad \text{where } c_3 = \|B_s\|_2 (\kappa c_1 x_{\text{max}} + c_2).
\]

Therefore,

\[
\|e_t\|_2 = \|v_{t-1} + A_k v_{t-2} + A_k^{t-1} v_0\|_2 \\
\leq c_3 \rho^{t-1} + \kappa^2 (1 - \gamma) c_3 \rho^{t-2} + \ldots \\
\leq c_3 \kappa^2 \max(\rho, 1 - \gamma)^{t-1} = c_4 \rho^{t-1}
\]

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

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

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

Therefore, the stage cost difference is

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

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

Therefore,

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

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

□
G. Additional Proofs

G.1. Proof of Lemma A.4

The proof relies on the following two lemmas.

**Lemma G.1** (Definition of $\epsilon_\varnothing$). Under the conditions in Lemma A.6,

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

*Proof.*

\[
\|D_x \sum_{k=1}^{H_t} A_{*}^{k-1}(w_{t-k} - \hat{w}_{t-k})\|_\infty \leq \|D_x\|_\infty \sum_{k=1}^{H_t} \|A_{*}^{k-1}(w_{t-k} - \hat{w}_{t-k})\|_\infty \\
\leq \|D_x\|_\infty \sum_{k=1}^{H_t} \|A_{*}^{k-1}(w_{t-k} - \hat{w}_{t-k})\|_2 \\
\leq \|D_x\|_\infty \sum_{k=1}^{H_t} \kappa(1 - \gamma)^{k-1} r z_{max} \\
\leq \|D_x\|_\infty \kappa/\gamma z_{max} r = \epsilon_\varnothing(r)
\]

**Lemma G.2** (Definition of $\epsilon_{\hat{\theta}}$). For any $M \in \mathcal{M}$, any $\hat{\theta}, \theta \in \Theta^0$ such that $\|\hat{\theta} - \theta\|_F \leq r$, we have

\[
|g^x_i(M; \hat{\theta}) - g^x_i(M; \theta)| \leq \epsilon_{\hat{\theta}}(r)
\]

where $\epsilon_{\hat{\theta}}(r) = c_{\hat{\theta}} r \sqrt{m n}$.

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

\[
|g^x_i(M; \hat{\theta}) - g^x_i(M; \theta)| = \left| \sum_{k=1}^{2H} \|D_{x,i}^T \Phi^x_k(M; \hat{\theta})\|_1 - \|D_{x,i}^T \Phi^x_k(M; \theta)\|_1 w_{max} \right| \\
\leq \sum_{k=1}^{2H} \|D_{x,i}^T \Phi^x_k(M; \hat{\theta})\|_1 - \|D_{x,i}^T \Phi^x_k(M; \theta)\|_1 w_{max} \\
\leq \sum_{k=1}^{2H} \|D_{x,i}^T \Phi^x_k(M; \hat{\theta}) - D_{x,i}^T \Phi^x_k(M; \theta)\|_1 w_{max} \\
\leq \sum_{k=1}^{2H} \|D_x\|_\infty \|\Phi^x_k(M; \hat{\theta}) - \Phi^x_k(M; \theta)\|_\infty w_{max}
\]

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

**Lemma G.3.** 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 G.4.** For any $k \geq 0$, any $\hat{\theta}, \theta \in \Theta^0$ such that $\|\hat{\theta} - \theta\|_F \leq r$, we have

\[
\|A^k - \hat{A}^k\|_2 \leq k\kappa^2(1 - \gamma)^{k-1} r \mathbb{1}_{(k \geq 1)}
\]
Safe Adaptive Learning for Linear Quadratic Regulators with Constraints

\[ \| A^k B - \hat{A}^k \hat{B} \|_2 \leq \kappa n^2 \kappa_B (1 - \gamma)^{k-1} r \mathbb{I}_{(k \geq 1)} + \kappa (1 - \gamma)^k r \]

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

\[
\| \hat{A}^k - A^k \|_2 = \| \sum_{i=0}^{k-1} \hat{A}^{k-i-1}(\hat{A} - A)A^i \|_2 \\
\leq \| \sum_{i=0}^{k-1} \hat{A}^{k-i-1} \|_2 \| \hat{A} - A \| \| A^i \|_2 \\
\leq \| \sum_{i=0}^{k-1} \kappa (1 - \gamma)^{k-i-1} \kappa (1 - \gamma)^i \|_2 \\
= \kappa n^2 \kappa_B r (1 - \gamma)^{k-1} \\
\| \hat{A}^k \hat{B} - A^k \hat{B} \|_2 \leq \| \hat{A}^k \hat{B} - A^k \hat{B} \|_2 + \| A^k \hat{B} - \hat{A}^k \hat{B} \|_2 \\
\leq \kappa n^2 \kappa_B r (1 - \gamma)^{k-1} \mathbb{I}_{(k \geq 1)} + \kappa (1 - \gamma)^k r
\]

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

\[
\| \Phi_k^r (M; \hat{\theta}) - \Phi_k^r (M; \theta) \|_\infty \\
= \| \hat{A}^{k-1} \mathbb{I}_{(k \leq H)} + \sum_{i=1}^{H} \hat{A}^{i-1} \hat{B} \hat{M}_{i-1} \mathbb{I}_{(k=1)} - A^{k-1} \mathbb{I}_{(k \leq H)} - \sum_{i=1}^{H} A^{i-1} B M_{i-1} \mathbb{I}_{(k=1)} \|_\infty \\
\leq \| \hat{A}^{k-1} - A^{k-1} \|_{\infty} \mathbb{I}_{(k \leq H)} + \sum_{i=1}^{H} \| (\hat{A}^{i-1} \hat{B} - A^{i-1} B) M_{i-1} \mathbb{I}_{(k=1)} \|_\infty \mathbb{I}_{(1 \leq k = i \leq H)} \\
\leq \sqrt{n} \| \hat{A}^{k-1} - A^{k-1} \|_2 \mathbb{I}_{(k \leq H)} + \sqrt{n} \sum_{i=1}^{H} \| \hat{A}^{i-1} \hat{B} - A^{i-1} B \|_2 2\sqrt{n} \kappa (1 - \gamma)^{k-i-1} \mathbb{I}_{(1 \leq k = i \leq H)}
\]

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

\[
\sum_{k=1}^{2H} \sqrt{n} \| \hat{A}^{k-1} - A^{k-1} \|_2 \mathbb{I}_{(k \leq H)} \leq \sum_{k=1}^{2H} \sqrt{n} (k-1) \kappa^2 (1 - \gamma)^{k-2} r \mathbb{I}_{(2 \leq k \leq H)} \leq \sqrt{n} \kappa^2 r / \gamma^2 \\
\sum_{k=1}^{2H} \sqrt{n} \sum_{i=1}^{H} \| \hat{A}^{i-1} \hat{B} - A^{i-1} B \|_2 2\sqrt{n} \kappa (1 - \gamma)^{k-i-1} \mathbb{I}_{(1 \leq k = i \leq H)} \\
\leq \sum_{k=1}^{2H} \sqrt{n} \sum_{i=1}^{H} (i-1) \kappa_B (1 - \gamma)^{i-2} r \mathbb{I}_{(i \geq 2)} 2\sqrt{n} \kappa^3 (1 - \gamma)^{k-i-1} \mathbb{I}_{(1 \leq k = i \leq H)} \\
+ \sum_{k=1}^{2H} \sqrt{n} \sum_{i=1}^{H} \kappa (1 - \gamma)^{i-1} r 2\sqrt{n} \kappa (1 - \gamma)^{k-i-1} \mathbb{I}_{(1 \leq k = i \leq H)} \\
= 2 \sqrt{n} \kappa^4 \kappa_B r \sum_{i=1}^{H} \sum_{j=1}^{H} (i-1) (1 - \gamma)^{i-2} (1 - \gamma)^{j-1} + 2 \sqrt{n} \kappa^3 r \sum_{i=1}^{H} \sum_{j=1}^{H} (1 - \gamma)^{i-1} (1 - \gamma)^{j-1} \\
= 2 \sqrt{n} \kappa^4 \kappa_B r / \gamma^3 + 2 \sqrt{n} \kappa^3 r / \gamma^2
\]
G.2. Proof of Lemma A.8

For notational simplicity, we omit the subscript $t$ in $H_t$ in this proof. Remember that $g^x_t(M_{t-H,t-1}; \theta) = \sum_{s=1}^{2H} \|D_x \Phi^x_s(M_{t-H,t-1}; \theta)\|_1 w_{\text{max}}$.

$$\left| \tilde{g}^x_t(M_{t-H,t-1}; \theta) - g^x_t(M; \theta) \right| = \left| \sum_{k=1}^{2H} \|D_x \Phi^x_k(M_{t-H,t-1}; \theta)\|_1 - \|D_x \Phi^x_k(M_t; \theta)\|_1 \right| w_{\text{max}}$$

$$\leq \sum_{k=1}^{2H} \|D_x \Phi^x_k(M_{t-H,t-1}; \theta)\|_1 - \|D_x \Phi^x_k(M_t; \theta)\|_1 w_{\text{max}}$$

$$\leq \sum_{k=1}^{2H} \|D_x \Phi^x_k(M_{t-H,t-1}; \theta) - \Phi^x_k(M_t; \theta)\|_1 w_{\text{max}}$$

$$\leq \sum_{k=1}^{2H} \|D_x\|_\infty \|\Phi^x_k(M_{t-H,t-1}; \theta) - \Phi^x_k(M_t; \theta)\|_\infty w_{\text{max}}$$

$$\leq \sum_{k=1}^{2H} \|D_x\|_\infty \sum_{i=1}^{H} \|A^{i-1}B(M_{t-i}[k-i] - M_t[k-i])\|_\infty \mathbb{1}_{(1 \leq k-i \leq H)} w_{\text{max}}$$

$$\leq \sum_{k=1}^{2H} \|D_x\|_\infty \sum_{i=1}^{H} \|A^{i-1}B\|_\infty \|M_{t-i}[k-i] - M_t[k-i]\|_\infty \mathbb{1}_{(1 \leq k-i \leq H)} w_{\text{max}}$$

$$= \|D_x\|_\infty \sqrt{m} w_{\text{max}} \sum_{k=1}^{H} \sum_{i=1}^{H} (1-\gamma)^{i-1} \kappa B \|M_{t-i}[k-i] - M_t[k-i]\|_\infty$$

$$\leq \|D_x\|_\infty \sqrt{m} w_{\text{max}} \kappa B \sqrt{nH} \sum_{i=1}^{H} (1-\gamma)^{i-1} \|B\|_F$$

$$\leq \|D_x\|_\infty \sqrt{mnH} w_{\text{max}} \kappa B \sum_{i=1}^{H} (1-\gamma)^{i-1} \Delta_M$$

$$\leq \|D_x\|_\infty \sqrt{mnH} w_{\text{max}} \kappa B / \gamma^2 \Delta_M$$

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

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

G.3. Proof of Lemma B.5

For notational simplicity, we define $y_t = \sum_{i=1}^{H_t} A_i^{i-1} w_{t-i} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} A_i^{i-1} B_s M_{t-i}[k-i] \hat{\omega}_{t-k} \mathbb{1}_{1 \leq k-i \leq H_t} + \sum_{i=1}^{H_t} A_i^{i-1} B_s \eta_{t-i}$. Since $A_s$ is $(\kappa, \gamma)$-stable, we have

$$\|y_t\|_2 \leq \sum_{i=1}^{H_t} \|A_i^{i-1}\|_2 \|w_{t-i}\|_2 + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} \|A_i^{i-1} B_s M_{t-i}[k-i] \hat{\omega}_{t-k} \mathbb{1}_{1 \leq k-i \leq H_t} + \sum_{i=1}^{H_t} \|A_i^{i-1} B_s \eta_{t-i}\|_2$$

$$\leq \sum_{i=1}^{H_t} \kappa (1-\gamma)^{i-1} \sqrt{m} w_{\text{max}} + \sum_{k=2}^{2H_t} \sum_{i=1}^{H_t} \|A_i^{i-1} B_s\|_2 \|M_{t-i}[k-i] \hat{\omega}_{t-k} \mathbb{1}_{1 \leq k-i \leq H_t} + \sum_{i=1}^{H_t} \|A_i^{i-1} B_s\|_2 \|\eta_{t-i}\|_2$$
where the third inequality uses \( \theta \).

Further, even though we make \( M \) in this proof for simplicity of notations.

\[ \text{G.4. Proof of Lemma E.6} \]

**Proof.** We omit \( \theta \) in this proof for simplicity of notations.

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

\[
\| \nabla f(M; \theta) \|_F = \sup_{\Delta M \neq 0, M + \Delta M \in \mathcal{M}_{out, H}} \frac{\langle \nabla f(M; \theta), \Delta M \rangle}{\| \Delta M \|_F} \leq \sup_{\Delta M \neq 0, M + \Delta M \in \mathcal{M}_{out, H}} \frac{f(M + \Delta M) - f(M)}{\| \Delta M \|_F}
\]

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

\[
\| \tilde{x} - \tilde{x}' \|_2 \leq \sum_{k=1}^{2H} \| (\Phi_k^x(M) - \Phi_k^x(M'))w_{t-k} \|_2 \\
\leq \sum_{k=1}^{2H} \| \sum_{i=1}^{H} A^{-1} B(M[k-i] - M'[k-i])I_{1 \leq k-i \leq H}w_{t-k} \|_2 \\
\leq \sum_{j=1}^{H} O(\sqrt{n})\| M[j] - M'[j] \|_2 \\
\leq \sum_{j=1}^{H} O(\sqrt{n})\| M[j] - M'[j] \|_F \\
\leq O(\sqrt{n}H)\| M - M' \|_F \\
\| \hat{u} - \hat{u}' \|_2 \sum_{k=1}^{H} \| M[k] - M'[k] \|_2 \sqrt{n}w_{max} \leq O(\sqrt{n}H)\| M - M' \|_F
\]

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

\[
|f(M) - f(M')| \leq O(\sqrt{mn}n^{\frac{1}{2}})\left\|M - M'\right\|_F
\]

Therefore,

\[
\|\nabla f(M; \theta)\|_F \leq \sup_{\Delta M \neq 0, M + \Delta M \in M_{out,H}} \frac{f(M + \Delta M) - f(M)}{\|\Delta M\|_F}
\]

\[
\leq \sup_{\Delta M \neq 0, M + \Delta M \in M_{out,H}} \frac{O(\sqrt{mn}n^{\frac{1}{2}})\|\Delta M\|_F}{\|\Delta M\|_F} \leq O(n\sqrt{mn})
\]

\( \square \)

### G.5. Proof of Lemma E.5

**Proof.** Notice that \( \Omega_1 \) and \( \Omega_3 \) satisfies the conditions in Proposition 2 in (Li et al., 2021). Therefore,

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

Notice that

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

therefore, \( \|\Delta_1 - \Delta_3\|_\infty \leq \|\Delta_1 - \Delta_2\|_\infty \). Further, \( \{i:(\Delta_3)_i > (\Delta_1)_i\} = \{i:(\Delta_2)_i > (\Delta_1)_i\} \subseteq \{i:(\Delta_1)_i \neq (\Delta_2)_i\} \).

So \( \min_{\{i:(\Delta_3)_i > (\Delta_1)_i\}}(h - \Delta_1 - Cx_F)_i \geq \min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}}(h - \Delta_1 - Cx_F)_i \geq \min_{\{i:(\Delta_1)_i \neq (\Delta_2)_i\}}(h - \Delta_3 - Cx_F)_i \).

Therefore,

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

Similarly,

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

which completes the bound. \( \square \)

### G.6. Proof of Lemma E.7

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

**Lemma G.5.** In our Algorithm 1, \( M^e \in \mathcal{F}(w_0, \ldots, w_{T_1^e + T_2^e - 1}, \eta_0, \ldots, \eta_{T_1^e + T_2^e - 1}) = \mathcal{F}_{t_1^e + T_2^e} \subseteq \mathcal{F}_{t_2^e - H^e} \).

**Proof.** By definition, we have the following fact: \( M^e \in \mathcal{F}(\hat{\theta}^{e+1}) = \mathcal{F}(\{z_k, x_{k+1}^{t_1^e + T_2^e - 1}\}_{k=t_1^e}^{t_1^e + T_2^e - 1}) = \mathcal{F}(w_0, \ldots, w_{t_1^e + T_2^e - 1}, \eta_0, \ldots, \eta_{t_1^e + T_2^e - 1}) = \mathcal{F}_{t_1^e + T_2^e} \). By \( \tilde{W}_{t_1}^e \geq H^e \), we have \( t_1^e + T_2^e + H^e \leq t_2^e \), and since \( \mathcal{F}_{t_1^e} \subseteq \mathcal{F}_t \), we have the last claim. \( \square \)

**Lemma G.6.** When \( t \in \mathcal{T}_2^e \) we have \( w_{t-2}^{e+1} \perp \mathcal{F}_{t_2^e - H^e} \).

**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 \( \mathcal{T}_t \), we have \( w_{t-2}^{e+1} \perp \mathcal{F}_{t_2^e - H^e} \). \( \square \)

**Lemma G.7.** In our Algorithm 1, when \( t \in \mathcal{T}_2^e \), we have \( \mathbb{E}[l(\hat{x}_t, \hat{u}_t) \mid \mathcal{F}_{t_2^e - H^e}] = f(M^e; \theta_e) \).
Proof. By our lemmas above, $M^e \in F_{t^2 + H^e}$, but $w_{t-2H^e} \downarrow F_{t^2 + H^e}$. Then, by our definition of $\hat{x}_t, \hat{u}_t$ and $f(M; \theta_s)$, we have the result.

Definition G.8 (Martingale). $\{X_t\}_{t \geq 0}$ is a martingale wrt $\{F_t\}_{t \geq 0}$ if (i) $E \left| X_t \right| < +\infty$, (ii) $X_t \in F_t$, (iii) $E(X_{t+1} \mid F_t) = X_t$ for $t \geq 0$.

Proposition G.9 (Azuma-Hoeffding Inequality). $\{X_t\}_{t \geq 0}$ is a martingale with respect to $\{F_t\}_{t \geq 0}$. If (i) $X_0 = 0$, (ii) $|X_t - X_{t-1}| \leq \sigma$ for any $t \geq 1$, then, for any $\alpha > 0$, any $t \geq 0$,

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

Corollary G.10. $\{X_t\}_{t \geq 0}$ is a martingale wrt $\{F_t\}_{t \geq 0}$. If (i) $X_0 = 0$, (ii) $|X_t - X_{t-1}| \leq \sigma$ for any $t \geq 1$, then, for any $\delta \in (0, 1)$,

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

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

Proof. The proof is by letting $\alpha = \sqrt{2\sigma^2\log(2/\delta)}$ in Proposition G.9.

Lemma G.11. Define $q_t = l(\hat{x}_t, \hat{u}_t) - f(M^e; \theta_s)$. Then, $|q_t| \leq (mn)$ w.p.1.

Proof. We can show that $|\|\hat{x}_t\|_2 \leq O(\sqrt{mn})$ a.s. and $\hat{u}_t \in U$ a.s. by the proofs of Lemmas B.4 and B.5. Therefore, we have $|l(\hat{x}_t, \hat{u}_t)| = O(mn)$. Since $f(M^e; \theta_s) = E[l(\hat{x}_t, \hat{u}_t) \mid F_{t^2 + H^e}]$, we have $|f(M^e; \theta_s)| = O(mn)$. This completes the proof.

Notations and definitions. Define, for $0 \leq h \leq 2H^e - 1$,

$$T^2_{k, h} = \{t \in T^r : t \equiv h \mod (2H^e) \} = \{t^e_h \pm 2H^e, \ldots, t^e_h + 2H^e k^e_h\}$$

Lemma G.12. $t^e_h \geq t^e_2 - H^e$ and $k^e_h \leq T^{e+1}/(2H^e)$

Proof. Notice that $t^e_h \pm 2H^e \geq t^e_2 + H^e$, so the first inequality holds. Besides, notice that $2H^e k^e_h \leq t^e_h + 2H^e k^e_h \leq T^{e+1}/(2H^e)$, so the second inequality holds.

Define

$$\tilde{q}^e_{h,j} = q_{t^e_{h,j} - 2H^e} \quad \forall 1 \leq j \leq k^e_h$$

$$S^e_{h,j} = \sum_{s=1}^j \tilde{q}^e_{h,s} \quad \forall 0 \leq j \leq k^e_h,$$

$$F^e_{h,j} = F^e_{t^e_{h,j} - 2H^e} \quad \forall 0 \leq j \leq k^e_h,$$

where we define $\sum_{s=1}^0 a_s = 0$. By Lemma G.12, we have $F^e_{h,0} = F^e_{t^e_h - 2H^e}$.

Lemma G.13. $S^e_{h,j}$ is a martingale wrt $F^e_{h,j}$ for $j \geq 0$. Further, $S^e_{h,0} = 0$, $|S^e_{h,j+1} - S^e_{h,j}| \leq O(mn)$.

Proof. Since $|q_t| \leq O(mn)$, $E |S^e_{h,j}| \leq O(Tmn) < +\infty$. Notice that, for $t \in T^r_{2h}, w_{t-1}, \ldots, w_{t-2H^e} \in F_t$, and $M^e \in F_t$, so $t^e_{h,j} \in F^e_{h,j}$. Next, $E[S^e_{h,j+1} \mid F^e_{h,j}] = S^e_{h,j} + E[q^e_{h,j+1} \mid F^e_{h,j}] = S^e_{h,j}$. So this is done. The rest is by definition, and $q_t$'s bound.

Lemma G.14. Consider our choice of $H^e$ in Theorem 4.4. Let $\delta = \frac{p}{2^{2^{2^{2^2 + H^e}}}}$, w.p. $1 - \delta$, we have $|S^e_{h,k^e_h}| \leq \tilde{O} \left( \sqrt{k^e_h mn} \right)$.

Proof. By Lemma G.13, we can apply Corollary G.10, and obtain the bound, where we used $\log(2/\delta) = \tilde{O}(1)$.
Lemma G.15. Consider our choice of $H^e$ in Theorem 4.4. For any $e$, w.p. $1 - 2H^e\delta$, where $\delta = \frac{p}{2 \sum_{\tau \leq T^e} H^e}$,

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

Proof. Define event

$$\mathcal{E}^e_h = \{ |S^e_{h,k_h^e}| \leq \tilde{O}(\sqrt{kn^e}) \}$$

When $\cap_h \mathcal{E}^e_h$ holds,

$$| \sum_{t \in \mathcal{T}^e} q_t | = | \sum_{h=0}^{2H^e-1} S^e_{h,k_h^e} | \tilde{O}(mn) \sqrt{\sum_k k_h^e \sqrt{2H^e}} \leq \tilde{O}(mnT^{e+1})$$

where we used Lemma G.12 and Cauchy Schwartz.

Then, we have

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

Now, we can prove Lemma E.7. By Lemma G.15, w.p. $1 - p$, we have $| \sum_{h=0}^{2H^e-1} S^e_{h,k_h^e} | \leq \tilde{O} \left( \sqrt{T^{e+1}mn} \right)$ for all $e$. Then, by Lemma E.2, we completed the proof.