Online Policy Optimization in Unknown Nonlinear Systems

Yiheng Lin YIHENGL@CALTECH.EDU
California Institute of Technology

James A. Preiss JAPREISS@CALTECH.EDU
California Institute of Technology

Fengze Xie FXXIE@CALTECH.EDU
California Institute of Technology

Emile Anand EMILEA@ANDREW.CMU.EDU
Carnegie Mellon University

Soon-Jo Chung SJCHUNG@CALTECH.EDU
California Institute of Technology

Yisong Yue YYUE@CALTECH.EDU
California Institute of Technology

Adam Wierman ADAMW@CALTECH.EDU
California Institute of Technology

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study online policy optimization in nonlinear time-varying systems where the true dynamical models are unknown to the controller. This problem is challenging because, unlike in linear systems, the controller cannot obtain globally accurate estimations of the ground-truth dynamics using local exploration. We propose a meta-framework that combines a general online policy optimization algorithm (ALG) with a general online estimator of the dynamical system's model parameters (EST). We show that if the hypothetical joint dynamics induced by ALG with known parameters satisfies several desired properties, the joint dynamics under inexact parameters from EST will be robust to errors. Importantly, the final regret only depends on EST's predictions on the visited trajectory, which relaxes a bottleneck on identifying the true parameters globally. To demonstrate our framework, we develop a computationally efficient variant of Gradient-based Adaptive Policy Selection, called Memoryless GAPS (M-GAPS), and use it to instantiate ALG. Combining M-GAPS with online gradient descent to instantiate EST yields (to our knowledge) the first local regret bound for online policy optimization in nonlinear time-varying systems with unknown dynamics.

Keywords: Online policy optimization, online control, regret analysis.

1. Introduction

We consider a class of discrete-time policy optimization problems with unknown time-varying nonlinear dynamics (called non-autonomous systems in nonlinear control theory; Slotine et al. (1991)). Our setting specifies a particular functional form of the dynamics and the parameterized policy class that is broad enough to capture applications from drone control to robotic manipulation (Dawson et al., 2023; O'Connell et al., 2022; Kadirkamanathan and Fabri, 1995; Shi et al., 2020). Our goal is to optimize the control policy online to minimize the total cost even if the online agent cannot obtain a globally accurate model of the true dynamics.

Online policy optimization and the broader field of learning-based control have received significant attention over the last several years due to their ability to leverage data and adapt to time-varying dynamical systems (Fazel et al., 2018; Lin et al., 2023; Arous et al., 2021; Mokhtari et al., 2016; Zhao et al., 2022; Zhou et al., 2023; Hazan and Seshadhri, 2007, 2009; Gradu et al., 2023; Baby and Wang, 2021). Online policy optimization faces two major challenges in practice. The first comes from the unknown dynamical model, which increases the difficulty of deciding the right direction for policy improvement. The second comes from the generality of dynamics and policy classes, which requires flexible algorithms. Early works establishing finite-sample regret/complexity guarantees in this field considered the linear-quadratic regulator (LQR) (Fazel et al., 2018) and linear time-invariant dynamics with adversarial disturbances (Agarwal et al., 2019a; Cohen et al., 2019; Chen and Hazan, 2021; Simchowitz et al., 2018; Li et al., 2019), where the dynamics are known, linear, and time-invariant. Since then, much progress has been made to address challenges arising from considering either the unknown dynamical models or the general nonlinear dynamics (Muthirayan et al., 2022; Minasyan et al., 2021; Yu et al., 2023b; Dogan et al., 2021), but addressing the two challenges simultaneously is still open.

One line of work focuses on online policy optimization under increasingly general classes of dynamical systems and policies, but under the assumption that the true dynamical models are known (Chen et al., 2023; Zhou et al., 2023; Agarwal et al., 2019b). For example, Lin et al. (2023) propose an algorithm with provable regret guarantees that can be applied to nonlinear time-varying dynamical systems and general policy classes. However, assuming exact knowledge of the dynamical systems can be particularly restrictive in many applications when the system is nonlinear and time-varying. Even if the online agent has oracle access to a good dynamical model estimator, such as one with bounded error, it is unclear whether the model estimation errors will accumulate in the policy update.

Another line of work about online policy optimization focuses on learning the unknown dynamical models but is generally restricted to linear systems with specific policy classes (Qu et al., 2021; Minasyan et al., 2021). A common approach in the literature is random local exploration, where the controller either sets the control input to be a random perturbation or a randomly added perturbation to obtain sufficiently accurate estimations of the true dynamical model with high probability (Dean et al., 2018; Lale et al., 2022). However, extrapolating from local data to global models is only valid in general for linear dynamics.

The adaptive control literature also studies a similar problem (Annaswamy and Fradkov, 2021; Slotine et al., 1991; Ioannou and Sun, 2012; Wise et al., 2006). Typically, a set of linear parameters over a set of basis functions (Shi et al., 2020; Kadirkamanathan and Fabri, 1995; O'Connell et al., 2022) is dynamically adjusted online to compensate for the effect of time-varying disturbances, in order to stabilize the system and improve its trajectory tracking performance. Li et al. (2023) recently proposed an adaptive stabilizing algorithm for unknown linear systems without any system identification. Shi et al. (2021) and O'Connell et al. (2022) proposed a meta-online adaptive control algorithm to address the time-varying prediction errors; however, their methods do not optimize the gains in the control policy. Another related work from control theory is online robust control, where the goal is to ensure stability (Yu et al., 2023b; Li et al., 2023) or staying in safety sets (Ho et al., 2021), subject to model uncertainty.

The motivating insight we take from adaptive control is that the controller does not need to learn the true dynamical model to stabilize a system. The controller only needs to focus on "fitting" the actual trajectory it visited rather than "actively exploring" with the purpose of identifying the true

model parameter. This idea of "lazy learning" is shared by some works on online robust control (Boffi et al., 2021; Ho et al., 2021; Yu et al., 2023a), which maintains a set of possible dynamical models that are consistent with past observations. In this work, we are interested in developing a general approach for online policy optimization that can address the challenges of dealing with both unknown dynamical models and general nonlinear dynamics/policy classes.

Contributions: We make three main contributions. First, we develop a meta-framework that combines an online policy optimization algorithm (ALG) with an online parameter estimator (EST), where ALG focuses on optimizing the policy parameters while EST focuses on estimating the unknown component in the dynamics. We specify a set of properties that, if satisfied, implies that our meta-framework can mimic the behavior of applying ALG with known dynamical models up to an error that depends on EST's estimations. This setup enables us to reason about how to use existing results in online policy optimization and online regression (for model learning) as subroutines.

Second, we provide a theoretical analysis of our meta-framework, establishing conditions under which we can derive regret guarantees. We study the behavior of our meta-framework in two steps: The first step (Section 3.1) focuses on the behaviors of ALG and treats the estimations of EST as external inputs. We specify a set of properties that make the joint dynamics of applying ALG to the original system robust to the errors injected by using EST instead of the true dynamical models. The second step (Section 3.2) formulates the task of EST as an online regression problem, where the states visited by ALG are treated as adversarial inputs that can adapt to the history, i.e., the adversary is non-oblivious. Compared to a standard online regression problem of minimizing the errors, EST faces the additional challenge of minimizing the errors of the model's partial derivatives with respect to the state. We address this challenge by showing a reduction from the regret of estimating these partial derivatives to the regret with a standard regression loss when the original dynamics contain a certain level of randomness.

Third, we provide a concrete instantiation of our meta-framework on matched-disturbance dynamics. For ALG, we develop Memoryless Gradient-based Adaptive Policy Selection (M-GAPS), which extends the GAPS algorithm for online policy optimization (Lin et al., 2023) to utilize only $O(1)$ computational/memory complexity per time step and may be of independent interest. For EST, we utilize standard online regression. Combining these components, we obtain a bound on the local regret, an online analog of the stationary point conditions in nonconvex optimization. To our knowledge, this is the first local regret bound for online policy optimization in nonlinear time-varying systems with unknown dynamics.

2. Problem Setting

We consider online policy optimization in a discrete-time dynamical system that varies over time with dynamics $x_{t+1} = g_t\big(x_t, u_t, f_t(x_t, a_t^*)\big) + w_t$, where $x_t \in \mathbb{R}^n$ denotes the system state, $u_t \in \mathbb{R}^m$ denotes the control input, and g_t is the dynamical function. Here, $f_t(x_t, a_t^*) \in \mathbb{R}^k$ is a nonlinear residual term of which the online agent can make (noisy) observations. It has a known function form f_t and an unknown parameter $a_t^* \in \mathcal{A} \subseteq \mathbb{R}^p$. The disturbance term $w_t \in \mathcal{W} \subseteq \mathbb{R}^n$ does not depend on the states or the control inputs.

To control this system, the online agent adopts a time-varying control policy π_t that is parameterized by a policy parameter $\theta_t \in \Theta \subseteq \mathbb{R}^d$ and depends on the current value of the nonlinear residual. Specifically, the online agent picks the control input from the policy class $u_t = \pi_t\big(x_t, \theta_t, f_t(x_t, \hat{a}_t)\big)$. Here, function $f_t(\cdot, \hat{a}_t)$ reflects the online agent's current estimation of the ground-truth nonlinear

residual function $f_t(\cdot, a_t^*)$ at time step t. Intuitively, we assume the policy class π_t cares about predicting the nonlinear residual $f_t(x_t, a_t^*)$ rather than the unknown model parameter a_t^*. The objective of the online agent is to minimize the total cost $\sum_{t=0}^{T-1} c_t$ incurred over a finite horizon, where the stage cost at time step t is given by $c_t = h_t(x_t, u_t, \theta_t)$[1].

We provide a simple nonlinear control example that can be captured by our online policy optimization framework to help the readers understand the concepts we discussed.

Example 1 *Consider the control problem with a scalar discrete-time nonlinear dynamical system:*

$$x_{t+1} = x_t + \Delta(u_t + f_t(x_t, a_t^*) + w_t), \text{ where } f_t(x_t, a_t^*) = \phi(x_t) \cdot a_t^*, \tag{1}$$

where $\Delta > 0$ is the discretization step size. The nonlinear residual takes the form $\phi(x_t) \cdot a_t^*$, where $\phi : \mathbb{R} \to \mathbb{R}^p$ is a (nonlinear) feature map and a_t^* is the unknown model parameter. To control this system, the online agent with an estimated model parameter $\hat{a}_t$ can adopt the policy class:

$$u_t = -k_t x_t - f_t(x_t, \hat{a}_t), \text{ where } f_t(x_t, \hat{a}_t) = \phi(x_t) \cdot \hat{a}_t, \text{ and } k_t = \theta_t \in \mathbb{R}. \tag{2}$$

Here, the goal of the second term $-f_t(x_t, \hat{a}_t)$ is to cancel out the true nonlinear residual $f_t(x_t, a_t^*)$. In an ideal case where the online agent has access to the true model parameter a_t^*, policy (2) achieves the effect of removing the nonlinear residual and directly doing feedback control, resulting in the closed-loop dynamics $x_{t+1} = x_t + \Delta(-k_t x_t + w_t)$. In this case, the problem reduces to finding the optimal policy parameters (gains) $\{\theta_t\}$ in a known time-varying dynamical system.

2.1. Performance Metrics

In the literature of online optimization, *regret* is a common performance metric that directly compares the total cost $\sum_{t=0}^{T-1} c_t$ incurred by the online policy optimization algorithm against the optimal total cost one can achieve in hindsight. Before introducing the variants of regret we study, we first introduce the concept of the *surrogate cost*. The present formulation extends the similarly named concept of Lin et al. (2023) to the setting where the true dynamical models are unknown.

Definition 1 (Surrogate Cost) *The surrogate cost function is $F_t(\theta) := h_t(\tilde{x}_t^*(\theta), \tilde{u}_t^*(\theta), \theta)$, where $\tilde{x}_t^*(\theta)$ and $\tilde{u}_t^*(\theta)$ are the system state and control input at time t if the agent applies the control input $\tilde{u}_\tau^*(\theta) := \pi_\tau(\tilde{x}_\tau^*(\theta), \theta, f_\tau(x_\tau, a_\tau^*))$ at all previous time steps $\tau = 0, 1, \ldots, t$.*

Intuitively, the surrogate cost $F_t(\theta)$ evaluates how good a policy parameter θ is at an intermediate time step t by eliminating the interference of inexact estimations $\hat{a}_{0:t}$ of model parameters and any other policy parameters $\theta_{0:t-1}$ in the history that may be different with θ. This concept is useful for defining different regret metrics. For example, the *static regret* is a widely-used metric in the literature of online policy optimization (Cesa-Bianchi and Lugosi (2006); Hazan and Seshadhri (2009)) that compares the total cost of an online agent with the best static policy parameter in hindsight can be written as $R^S(T) := \sum_{t=0}^{T-1} c_t - \min_{\theta^* \in \Theta} \sum_{t=0}^{T-1} F_t(\theta^*)$.

However, as noted in previous works (Lin et al., 2023; Hazan et al., 2017), regret metrics that directly compare the cost difference (like $R^S(T)$) are not always suitable for nonconvex cost functions because gradient-based online optimization algorithms may easily get stuck in local minima even when the cost functions are time-invariant. Thus, the metric of *local regret* is used. For a sequence of policy parameters $\theta_{0:T-1}$, the local regret is defined as $R_\eta^L(T) := \sum_{t=0}^{T-1} \|\nabla_{\eta,\Theta} F_t(\theta_t)\|^2$.

1. We include θ_t in the stage cost for generality, allowing e.g. regularization. $c_t = h_t(x_t, u_t)$ is a special case.

Here, the projected gradient $\nabla_{\eta,\Theta} F_t(\theta_t)$ (parameterized by η) is a surrogate of the original gradient $\nabla F_t(\theta_t)$ that also considers the constraint set Θ. Specifically, an update step with the projected gradient is equivalent to projecting the output of the original gradient descent step back onto Θ, i.e., for any $\theta \in \Theta$, $\theta - \eta \nabla_{\eta,\Theta} F_t(\theta) = \Pi_\Theta(\theta - \eta \nabla F_t(\theta))$ [2]. This notion of local regret is first introduced by Hazan et al. (2017), and we provide a formal definition in Definition 14 in Appendix B. Intuitively, the local regret measures how well the policy parameter sequence $\theta_{0:T-1}$ tracks the changing stationary points of the surrogate cost functions $F_{0:T-1}$ (when $\Theta = \mathbb{R}^d$). In the time-invariant setting, sublinear local regret implies convergence to a stationary point.

Although local regret is useful for measuring the performance of an online policy optimization algorithm under nonconvex surrogate costs, a limitation of applying it alone to our setting with unknown dynamical models is that the surrogate cost F_t is defined in terms of ALG's behavior with known true dynamics. To address this limitation, in addition to bounding the local regret of the policy parameters $\theta_{0:T-1}$, we also bound the distance between the actual trajectory of the online agent and the trajectory it would achieve with the same policy parameters $\theta_{0:T-1}$ and exact knowledge of true model parameters $a^*_{0:T-1}$.

3. Main Results

Our approach is outlined in Algorithm 1, where two modules ALG and EST work together to update the policy and estimated model parameter at each time step (see Figure 1 for an illustration). ALG and EST are responsible for optimizing the policy parameters $\theta_{0:T-1}$ and learning the unknown model parameters $a^*_{0:T-1}$ of the nonlinear residual terms respectively:

- **ALG:** At time step t, ALG receives the current state x_t, policy parameter θ_t, and the known part of the time-varying system π_t, g_t, h_t, f_t. It also receives the current estimation $\hat{a}_t$ of the unknown model parameter a^*_t. Then, ALG outputs the new policy parameter θ_{t+1}. Note that we allow ALG to leverage/memorize history by maintaining an internal state y_t.
- **EST:** At time step t, EST receives the current state x_t and a (noisy) observation $\tilde{f}_t$ of the unknown component $f_t(x_t, a^*_t)$. Then, EST outputs the new estimation $\hat{a}_{t+1}$. Like ALG, we allow EST to keep internal state/memory (e.g., to memorize historical input data). We require EST to minimize the *trajectory-dependent model mismatches*:

$$\text{Zeroth-order model mismatch: } \varepsilon_t(x_t, \hat{a}_t, a^*_t) \coloneqq \|f_t(x_t, \hat{a}_t) - f_t(x_t, a^*_t)\|, \tag{3a}$$

$$\text{First-order model mismatch: } \varepsilon'_t(x_t, \hat{a}_t, a^*_t) \coloneqq \|\nabla_x f_t(x_t, \hat{a}_t) - \nabla_x f_t(x_t, a^*_t)\|_F. \tag{3b}$$

We adopt the shorthand $\varepsilon_t = \varepsilon_t(x_t, \hat{a}_t, a^*_t)$ and $\varepsilon'_t = \varepsilon'_t(x_t, \hat{a}_t, a^*_t)$ when the context is clear.

The key idea in analyzing our meta-framework (Algorithm 1) is to characterize how the inexact model estimations generated by EST affect the behavior ALG. We start by considering the "ideal" dynamics of applying ALG with exact model parameters $a^*_{0:T-1}$, which we denote as ALG*. We then state the key insight of our analysis in the informal lemma below, which connects the performance of the meta-framework to 1) the performance of ALG*, and 2) the model mismatches.

Lemma 2 (Informal) *Suppose ALG* satisfies the desired properties in Section 3.1. Then, the meta-framework (Algorithm 1) generates the same policy parameters as ALG* with perturbation ζ_t on the update of θ_{t+1} (see Figure 2). Further, $\sum_{t=0}^{T-1} \|\zeta_t\| = O\left(\sum_{t=0}^{T-1} \varepsilon_t + \sum_{t=0}^{T-1} \varepsilon'_t\right)$.*

2. Π_Θ is the Euclidean projection to Θ.

Algorithm 1: Meta-Framework

Require: ALG and EST
Require: Knowing functions $\{\pi_t, g_t, h_t, f_t\}$ at each time step t
Initialize: State x_0; Policy parameter θ_0; Model parameter estimation $\hat{a}_0$.
for $t = 0, 1, \ldots, T-1$ **do**
  Decide control input $u_t = \pi_t(x_t, \theta_t, f_t(x_t, \hat{a}_t))$.
  Incur stage cost $h_t(x_t, u_t, \theta_t)$.
  $\theta_{t+1} \leftarrow$ ALG.update$(x_t, \theta_t, \pi_t, g_t, h_t, f_t, \hat{a}_t)$. */* ALG can have internal memory. */*
  System evolves to $x_{t+1} = g_t(x_t, u_t, f_t(x_t, a_t^*)) + w_t$.
  Receive a (noisy) observation $\tilde{f}_t$ of $f_t(x_t, a_t^*)$.
  $\hat{a}_{t+1} \leftarrow$ EST.update$(x_t, \tilde{f}_t, \hat{a}_t)$. */* EST can have internal memory. */*
end

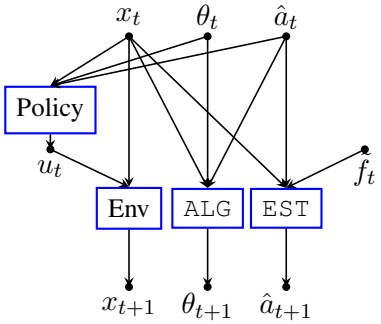

Figure 1: The meta-framework.

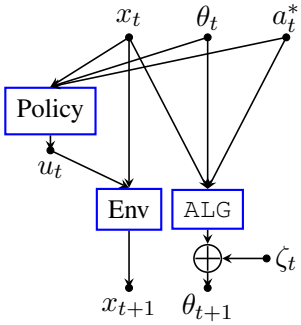

Figure 2: Theoretical comparand: ALG* with perturbations.

The formal statement of Lemma 2 can be found in Theorem 6.

The rest of this section is organized as following: In Section 3.1, we specify the properties of ALG* that enable the meta-framework to be robust against inexact model parameters in the policy parameter update. Then, in Section 3.2, we formulate EST's task of learning $f_t(x_t, a_t^*)$ as an online optimization problem, where we view the state x_t as picked by an adaptive adversary. We also discuss how this problem reduces to existing results on online optimization.

3.1. Online Policy Optimization

In this section, we take a perspective that views the updates performed by ALG as part of a joint dynamics formed together with the original dynamical system. Compared to the common approach of analyzing ALG separately from the dynamical system to which it applies, our dynamical view enables us to compare the differences of applying ALG under different external inputs (i.e. different $\hat{a}_t$ estimates) more efficiently.

We consider the class of online policy optimization algorithms whose joint dynamics with the original system can be written in the following form: When the model parameter a_t is given as the

input to ALG at time step t, the joint dynamics can be written as

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \\ \theta_{t+1} \end{pmatrix} = q_t(x_t, y_t, \theta_t, a_t) = \begin{pmatrix} q_t^x(x_t, y_t, \theta_t, a_t) \\ q_t^y(x_t, y_t, \theta_t, a_t) \\ q_t^\theta(x_t, y_t, \theta_t, a_t) \end{pmatrix}, \text{ for } x_t \in \mathbb{R}^n, y_t \in \mathbb{R}^p, \theta_t \in \Theta \subset \mathbb{R}^d. \quad (4)$$

Here, $y_t \in \mathbb{R}^p$ is an auxiliary state that ALG can use to store something besides the system state x_t and the policy parameter θ_t to help it perform the update. For example, y_t can be a finite memory buffer that stores information from the past. It can also be the integral of past tracking error in an integral controller. Thus, we introduce y_t to allow broader classes of online policy optimization algorithms, and we will provide a concrete example of y_t later in Section 4.1.

The goal of formulating joint dynamics (4) is to compare the behaviors of the meta-framework and ALG* with perturbations on policy parameter updates. Specifically, recall that $\hat{a}_{0:T-1}$ denotes the estimated model parameters of EST. The actual trajectory of the meta-framework is

$$\text{Meta-framework: } (x_{t+1}, y_{t+1}, \theta_{t+1})^\top = q_t(x_t, y_t, \theta_t, \hat{a}_t). \quad (5)$$

We compare it with the joint dynamics of ALG* (see Figure 2). Recall that ALG* denotes the scenario when ALG has access to exact model parameters $a^*_{0:T-1}$, and we consider a perturbed variant:

$$\text{ALG}^* \text{ with perturbations: } (x_{t+1}, y_{t+1}, \theta_{t+1})^\top = q_t(x_t, y_t, \theta_t, a_t^*) + (0, 0, \zeta_t)^\top. \quad (6)$$

Here, ζ_t is an additive perturbation on the update equation of policy parameter θ_{t+1}. To understand (6) intuitively, it is helpful to draw connections with the process of using a gradient-based optimizer to update θ_t iteratively, where $\zeta_t \equiv 0$ corresponds to the case when exact gradients are available. In contrast, nonzero perturbations correspond to the more practical case when the optimizer can only use biased gradient estimations, which still perform well in general.

Note that the estimated model parameters $\hat{a}_{0:T-1}$ generated by EST may also depend on the state x_t and other parts of the dynamical system. Thus, a natural question is whether we should also incorporate the update rule of EST into the joint dynamical system in (5), where we include $\hat{a}_t$ as another element of the joint state. However, we still choose to model $\hat{a}_t$ as an external input in (5) and handle the update of $\hat{a}_t$ separately in Section 3.2. This is because our approach requires comparing the actual joint dynamics with (6). Since a_t^* is an external input decided by the environment in (6), keeping the joint state space identical in (5) makes the comparison easier. Further, a strength of our proof framework based on the joint dynamics is that we can show the actual trajectory (5) will stay close to (6). However, we know that the estimated model parameter sequence $\{\hat{a}_t\}$ will not converge to the true sequence $\{a_t^*\}$ in general, even if EST has low regret.

We require three important properties of the joint dynamics induced by ALG. The first property is about the Lipschitzness with respect to the prediction errors ε_t and ε_t'.

Property 3 *[Lipschitzness] For any $x_t, y_t, \theta_t, \hat{a}_t$ that satisfies $\|x_t\| \le R_x$, $\|y_t\| \le R_y$, $\theta_t \in \Theta$, and $\hat{a}_t \in \mathcal{A}$, the following Lipschitzness conditions hold:*

$$\begin{aligned}
\|q_t^x(x_t, y_t, \theta_t, a_t^*) - q_t^x(x_t, y_t, \theta_t, \hat{a}_t)\| &\le \alpha_x \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_x \varepsilon_t'(x_t, \hat{a}_t, a_t^*), \\
\|q_t^y(x_t, y_t, \theta_t, a_t^*) - q_t^y(x_t, y_t, \theta_t, \hat{a}_t)\| &\le \alpha_y \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_y \varepsilon_t'(x_t, \hat{a}_t, a_t^*), \\
\left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\le \alpha_\theta \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_\theta \varepsilon_t'(x_t, \hat{a}_t, a_t^*).
\end{aligned}$$

Further, $q_t^\theta(x, y, \theta, a_t^*)$ is $(L_{\theta,x}, L_{\theta,y})$-Lipschitz in (x, y).

Intuitively, Property 3 says that the error brought by an inexact estimation $\hat{a}_t$ only "distorts" the ideal joint dynamics of $\texttt{ALG}^*$ in the form of zeroth-order and first-order prediction errors. Therefore, to bound the error injected into the joint dynamics at every step, $\texttt{EST}$ only needs to minimize ε_t and ε'_t on the actual state trajectory $x_{0:T-1}$ that the online agent visits. Note that this property can be viewed as a standard assumption about Lipschitzness if $\texttt{ALG}$ is a gradient-based algorithm. This is because all terms that involve the unknown model parameter will take the form $f_t(x_t, \hat{a}_t)$ and $\nabla_x f_t(x_t, \hat{a}_t)$ in the joint dynamics.

The second property is about contraction stability of x_t and y_t under exact predictions $a^*_{0:T-1}$. As we show in Theorem 6, this property guarantees that the dynamical updates of states x_t and y_t in the joint dynamics are robust to the model mismatches $\{\varepsilon_t, \varepsilon'_t\}_{0:T-1}$.

Property 4 *[Contraction Stability] For any sequence $\theta_{0:T-1}$ that satisfies the slowly time-varying constraint $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$, the partial dynamical system*

$$x_{t+1} = q^x_t(x_t, y_t, \theta_t, a^*_t), \qquad y_{t+1} = q^y_t(x_t, y_t, \theta_t, a^*_t) \tag{7}$$

satisfies that $\|x_t\| \le R^*_x < R_x$ and $\|y_t\| \le R^*_y < R_y$ always hold if the system starts from $(x_\tau, y_\tau) = (0, 0)$. Further, there exists a function $\gamma : \mathbb{Z}_{\ge 0} \to \mathbb{R}_{\ge 0}$ that satisfies $\sum_{t=0}^{\infty} \gamma(t) \le C$, such that from any initial states $(x_\tau, y_\tau), (x'_\tau, y'_\tau)$ where $\|x_\tau\|, \|x'_\tau\| \le R_x$ and $\|y_\tau\|, \|y'_\tau\| \le R_y$, the trajectory satisfies $\left\|(x_{\tau+t}, y_{\tau+t}) - (x'_{\tau+t}, y'_{\tau+t})\right\| \le \gamma(t) \cdot \|(x_\tau, y_\tau) - (x'_\tau, y'_\tau)\|$.

Note that Property 4 is different with the contraction assumption of Lin et al. (2023) because it also considers the internal state y_t of $\texttt{ALG}$ besides the system state x_t. The requirement that $\sum_{t=0}^{\infty} \gamma(t) \le C$ is also weaker than the exponential decay rate in Lin et al. (2023).

Intuitively, Property 4 guarantees that if (x_t, y_t) is perturbed at one step t (e.g., due to an update under the inexact model parameter $\hat{a}_t$), the resulting trajectory $\{(x_\tau, y_\tau)\}_{\tau \ge t}$ still converges quickly towards the trajectory that $\texttt{ALG}$ would achieve under exact predictions without this perturbation. Property 4 can be viewed as an extension of the time-varying stability and contractive perturbation property in Lin et al. (2023) to include state y_t maintained by $\texttt{ALG}$. This is required in our framework because y_t can be affected by the estimation errors and is involved in the dynamics of updating θ_t.

The third property we need is the robustness of the update rule that $\texttt{ALG}^*$ uses to update the policy parameter θ_t. Specifically, it requires the regret guarantee achieved by $\texttt{ALG}^*$ to be robust against a certain level of adversarial disturbances $\{\zeta_t\}$ on the update dynamics.

Property 5 *[Robustness] Consider the joint dynamics in (6). When $\|\zeta_t\| \le \bar{\zeta}$ holds for all t, the resulting $\{\theta_t\}$ satisfies the slowly-time-varying constraint $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ for all time t. Further, $\texttt{ALG}^*$ with perturbations (6) can achieve a regret guarantee $R(T, \sum_{t=0}^{T-1} \|\zeta_t\|)$ that depends on the total magnitude of the perturbation sequence $\zeta_{0:T-1}$.*

To understand Property 5, we can think about online gradient descent (OGD) in online optimization problems without state or dynamics. It is known that this approach is robust to (biased) disturbances on the gradient estimation, and the total amount of added disturbances will affect the final regret bound (see, for example, Theorem 27).

Now, we present our main results about the stability of applying $\texttt{ALG}$ with inexact model parameters and the regret bound in Theorem 6. Additionally, Theorem 6 bounds the distances between the actual trajectory and the trajectory achieved by applying the same policy parameter sequence with the exact model parameter sequence.

Theorem 6 *Suppose Properties 3, 4, and 5 hold. Let $\xi = \{x_t, y_t, \theta_t\}_{0:T-1}$ be the trajectory of the meta-framework (Algorithm 1). If the prediction errors $\{\varepsilon_t, \varepsilon'_t\}_{0:T-1}$ are uniformly bounded such that the following inequalities hold for all time step t: $\alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t \le \bar{\zeta}/2$, and*

$$(\alpha_x + \alpha_y)\varepsilon_t + (\beta_x + \beta_y)\varepsilon'_t \le \min\left\{ \frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C} \right\},$$

then the trajectory ξ satisfies $\|x_t\| \le R_x$, $\|y_t\| \le R_y$, and $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ for all time steps t. Further, define $\tilde{\xi} := \{\tilde{x}_t, \tilde{y}_t, \theta_t\}_{0:T-1}$, where $\{\tilde{x}_t, \tilde{y}_t\}_{0:T-1}$ are obtained by implementing the policy parameters $\theta_{0:T-1}$ with exact model parameters $a^*_{0:T-1}$, i.e., the trajectory of partial joint dynamics (7). The trajectory $\tilde{\xi}$ achieves the regret $R(T, \sum_{t=0}^{T-1}\|\zeta_t\|)$, with $\sum_{t=0}^{T-1}\|\zeta_t\|$ upper bounded by

$$\Big(\alpha_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\alpha_x + \alpha_y)\Big)\sum_{t=0}^{T-1}\varepsilon_t + \Big(\beta_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\beta_x + \beta_y)\Big)\sum_{t=0}^{T-1}\varepsilon'_t.$$

The total distances between the states on the trajectories ξ and $\tilde{\xi}$ satisfies that

$$\sum_{t=1}^{T}\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| \le C\left((\alpha_x + \alpha_y)\sum_{t=0}^{T-1}\varepsilon_t + (\beta_x + \beta_y)\sum_{t=0}^{T-1}\varepsilon'_t\right).$$

We defer the proof of Theorem 6 to Appendix D. Intuitively, Theorem 6 states that when the model mismatches $\{\varepsilon_t, \varepsilon'_t\}_{0:T-1}$ are uniformly bounded, the actual trajectory ξ of applying ALG with inexact model parameters $\hat{a}_{0:T-1}$ will be uniformly bounded. Further, if the actual parameter sequence of $\theta_{0:T-1}$ is applied with exact model parameters $a^*_{0:T-1}$, the resulting trajectory $\tilde{\xi}$ achieves a regret guarantee that depends on the total magnitude of the model mismatches. It is worth noticing that the regret in Theorem 6 can be any regret that depends on the trajectory $\tilde{\xi}$. And as we discussed in Section 2.1, we evaluate the regret on trajectory $\tilde{\xi}$ rather than ξ because the metrics like the local regret are designed for evaluating the actual policy parameters $\theta_{0:T-1}$ rather than the whole trajectory ξ. The distances between ξ and $\tilde{\xi}$ are bounded in the last inequality in Theorem 6.

To show Theorem 6, the key idea is to fit the trajectory $\tilde{\xi}$ into the dynamical equation (6), where we design ζ_t to compensate the difference between the update rules $q_t(\tilde{x}_t, \tilde{y}_t, \theta_t, a^*_t)$ and $q_t(x_t, y_t, \theta_t, \hat{a}_t)$. To leverage Property 5, we show the perturbations $\zeta_{0:T-1}$ we constructed are uniformly bounded by $\bar{\zeta}$. We bound ζ_t and the distances between ξ and $\tilde{\xi}$ by induction. The induction is important because the magnitude of ζ_t depends on the distance between $\{x_t, y_t\}_{0:T-1}$ and $\{\tilde{x}_t, \tilde{y}_t\}_{0:T-1}$ in the past time steps. On the other hand, to bound the distance between $\{x_t, y_t\}$ and $\{\tilde{x}_t, \tilde{y}_t\}$, we need to leverage the contraction property in Property 19, which relies on $\|\zeta_t\| \le \bar{\zeta}$ so that $\theta_{0:T-1}$ is slowly time-varying. Lastly, we conclude the proof with the bounds on the distance between ξ and $\tilde{\xi}$ as well as the norm of ζ_t that depend on the model mismatches $\{\varepsilon_t, \varepsilon'_t\}_{0:T-1}$.

3.2. Online Parameter Estimation

The second part of our meta-framework focuses on predicting the unknown model parameter based on possibly noisy observations of the true nonlinear residual $f_t(x_t, a^*_t)$. A critical difference with prior works on system identification or model-based learning (e.g., Dean et al. (2020)) is that we only seek to optimize the zeroth-order and first-order model mismatches $\{\varepsilon_t, \varepsilon'_t\}$ (defined in (3)) on the actual trajectory that the online agent experiences. It is worth noticing that, although learning the

ground-truth model parameter a_t^* is impossible for a general nonlinear residual, minimizing the sum of zeroth-order model mismatches incurred on the actual trajectory can be formulated as a classic online regression problem, which we discuss below:

Online regression problem: At the beginning, the environment commits a sequence of error functions $e_t : \mathbb{R}^n \times \mathcal{A} \to \mathbb{R}, t = 0, \dots, T-1$, which are defined as $e_t(x, a) \coloneqq f_t(x, a_t^*) - f_t(x, a)$ for $t = 0, \dots, T-1$. The relationship between the error function e_t and the model mismatches $\{\varepsilon_t, \varepsilon_t'\}$ is $\varepsilon_t = \|e_t(x_t, \hat{a}_t)\|$, and $\varepsilon_t' = \|\nabla_x e_t(x_t, \hat{a}_t)\|$. At each time step t, the online parameter estimator $\texttt{EST}$ predicts $\hat{a}_t = \texttt{EST}(x_{0:t-1}, \hat{a}_{0:t-1}) \in \mathcal{A}$, which means the estimation $\hat{a}_t$ can be a general function of the historical states and estimations. Then, the environment reveals $x_t \in B_n(0, R_x)$ that can depend on the history $x_{0:t-1}$ and $\hat{a}_{0:t-1}$. We define the stage loss of $\texttt{EST}$ as $\ell_t = \|e_t(x_t, \hat{a}_t)\|^2$, which is equal to the squared ℓ_2-norm of the model mismatch $e_t(x_t, \hat{a}_t)$.

Under different sets of assumptions on the error functions and the sequence of true model parameters $\{a_t^*\}$, existing online algorithms can achieve regret guarantees. We consider a general form of expected regret bound: $\mathbb{E}\left[\sum_{t=1}^T \ell_t\right] \le R_0^\ell(T)$, where the expectation is taken over the randomness of implementing $\texttt{EST}$ and generating x_t. While different assumptions and designs of $\texttt{EST}$ can achieve different bounds on $R_0^\ell(T)$, an example we provide in Section 4 shows that a simple gradient estimator can achieve sublinear $R_0^\ell(T)$ when the nonlinear residual can be decomposed as $f_t(x, a) = \phi(x) \cdot a$ under the path length constraint $\sum_{t=1}^{T-1} \left\|a_{t+1}^* - a_t^*\right\| \le C$ (see Section 4.2).

While most prior works focus on minimizing the magnitude of the zeroth-order model mismatch $e_t(x_t, \hat{a}_t)$, we also need to bound the first-order model mismatch $\nabla_x e_t(x_t, \hat{a}_t)$ because it contributes to the regret bound in Theorem 6 (recall that $\|\nabla_x e_t(x_t, \hat{a}_t)\|_F = \varepsilon_t'$). Our main result in this section is about an automatic reduction from the regret bound $R_0^\ell(T)$ to a bound on the expected sum of the squared gradients $\mathbb{E}\left[\sum_{t=1}^T \|\nabla_x e_t(x_t, \hat{a}_t)\|_F^2\right]$.

Remark 7 *Besides the online policy optimization problem for control, the regret bound that concerns $\|\nabla_x e_t(x_t, \hat{a}_t)\|$ can be of independent interest for the problem of online regression because it characterizes how sensitive the regression loss is to any perturbations on the input sequence $x_{0:T-1}$ under the same estimations $\hat{a}_{0:T-1}$. Intuitively, if gradients of the error functions are small, the estimations $\hat{a}_{0:T-1}$ will be robust to small perturbations on the input sequence.*

To enable a reduction from the regret bound $R_0^\ell(T)$ to the gradient error bound, we employ Property 8 about the dynamical system that generates the state x_t. Specifically, we require there to be at least a small level of randomness when choosing x_t. Recall that $\hat{a}_{t+1}$ is decided based on the history $x_{0:t}$ and $\hat{a}_{0:t}$. We define the filtrations $\mathcal{F}_t \coloneqq \sigma(x_{1:t}, \hat{a}_{1:t})$ and $\mathcal{F}_t' \coloneqq \sigma(x_{1:t}, \hat{a}_{1:t+1})$, which satisfy $\mathcal{F}_t \subseteq \mathcal{F}_t' \subseteq \mathcal{F}_{t+1}$.

Property 8 *There is a certain level of random disturbances when generating each state x_t, i.e., for some $\bar{\epsilon} > 0$ and $\underline{\sigma} > 0$, one can find a σ-algebra $\mathcal{G}_t$ such that $\mathcal{F}_t' \subseteq \mathcal{G}_t \subseteq \mathcal{F}_{t+1}$ and $\|x_{t+1} - \mathbb{E}[x_{t+1} \mid \mathcal{G}_t]\| \le \bar{\epsilon}$, $Cov(x_{t+1} \mid \mathcal{G}_t) \succeq \underline{\sigma} I$.*

Note that $\bar{\epsilon}$ in Property 8 is not an upper bound on the noise w_t. To see this, consider an example where the state $x_t \in \mathbb{R}$ and noise w_t are sampled independently from $\mathrm{Uniform}([-1, 1])$. Although the noise magnitude $|w_t|$ can go up to 1, we can let Property 8 hold for arbitrary $\bar{\epsilon} > 0$ by designing the σ-algebra $\mathcal{G}_t$: Let I_t be the indicator of which of the $s \coloneqq \lceil 2/\bar{\epsilon} \rceil$ intervals $[-1, -1 + \bar{\epsilon}), [-1 + \bar{\epsilon}, -1 + 2\bar{\epsilon}), \dots, [-1 + (s-1)\bar{\epsilon}, 1]$ that w_t belongs to. If we define $\mathcal{G}_t \coloneqq \mathcal{F}_t' \vee \sigma(I_t)$,

then $|x_{t+1} - \mathbb{E}[x_{t+1} \mid \mathcal{G}_t]| \le \bar{\epsilon}$ holds. In contrast, if w_t is sampled uniformly from the discrete set $\{-1, 1\}$, one cannot find $\mathcal{G}_t$ to make $\bar{\epsilon} < 1$. This is consistent with the intuition that one cannot learn the gradients well if the states only take values from a discrete set despite the noise having positive variances. Further, the randomness enforced by Property 8 will "force" EST to also minimize the gradient of the error functions. To see this, suppose an input state x_t is given by $\bar{x}_t + v_t$, where $\bar{x}_t$ is the mean and v_t is a random disturbance. When v_t is sufficiently small, we know that $e_t(x_t, \hat{a}_t) \approx e_t(\bar{x}_t, \hat{a}_t) + \nabla_x e_t(\bar{x}_t, \hat{a}_t) \cdot v_t$ by Taylor's expansion. Since we can pick v_t randomly in different directions, we know the zeroth-order loss $\mathbb{E}[e_t(x_t, \hat{a}_t)^2]$ cannot converge to zero unless the magnitude of the gradient $\nabla_x e_t(\bar{x}_t, \hat{a}_t)$ converges to zero. We follow this intuition to show the reduction from the regret bound $R_0^\ell(T)$ to the total gradient error in Theorem 9.

Theorem 9 *Suppose each dimension $i \in [k]$ of the prediction error function satisfies*

$$\|\nabla_x e_t(x, a)_i\| \le \beta_e, \text{ and } \left\|\nabla_x^2 e_t(x, a)_i\right\| \le \gamma_e, \text{ for any } x \in B(0, R_x) \text{ and } a \in \mathcal{A}.$$

Suppose Property 8 holds with $\bar{\epsilon} \le \min\{\frac{1}{4}, \frac{1}{2\gamma_e}, \frac{1}{4\beta_e\gamma_e}\}$ and $\underline{\sigma} > 0$. If EST achieves the zeroth-order regret $\mathbb{E}\left[\sum_{t=1}^T \ell_t\right] \le R_0^\ell(T) \le \bar{\epsilon}^3 T$, the expected total squared gradient loss satisfies that

$$\mathbb{E}\left[\sum_{t=1}^T \|\nabla_x e_t(x_t, \hat{a}_t)\|_F^2\right] \le \frac{2k}{\underline{\sigma}}(1 + \gamma_e + \beta_e\gamma_e)\bar{\epsilon}^3 T + 2k\gamma_e^2\bar{\epsilon}^2 T.$$

Recall that k is the dimension of the unknown component $f_t(x_t, a_t^*)$. We defer the proof of Theorem 9 to Appendix E. We provide the following corollary to clarify this result in the special case when $R_0^\ell(T)$ is $O(\sqrt{T})$. For example, a gradient estimator can achieve this regret bound if ℓ_t is convex in a (see Section 4.2).

Corollary 10 *Under the same assumptions as Theorem 9, if EST achieves $R_0^\ell(T) = O(\sqrt{T})$ and Property 8 holds with $\bar{\epsilon} = \theta(T^{1/6})$ and $\underline{\sigma} = \Omega(\bar{\epsilon}^2)$, then the expected total squared gradient loss is bounded by $\mathbb{E}\left[\sum_{t=1}^T \|\nabla_x e_t(x_t, \hat{a}_t)\|_F^2\right] = O(kT^{5/6})$.*

In summary, with the help of Theorem 9, we reduce the problem of bounding the total squared first-order prediction errors $\sum_{t=0}^{T-1}(\varepsilon_t')^2$ to the standard online optimization problem. By substituting the bounds on $\varepsilon_{0:T-1}$ and $\varepsilon'_{0:T-1}$ into Theorem 6 in Section 3.1, one can derive the local regret bound for the actual joint dynamics and bound the distance between trajectories ξ and $\tilde{\xi}$.

4. Application: Matched Disturbance

In this section, we consider an instantiation of our setting to demonstrate the effectiveness of our meta-framework. Specifically, we study the matched-disturbance dynamics (Ferguson et al., 2020; Sinha et al., 2022; Garofalo et al., 2012), where the controller can choose a control input to "cancel out" the nonlinear residual term $f_t(x_t, a_t^*)$ when the exact model parameter a_t^* is available. The dynamics have the form

$$x_{t+1} = g_t(x_t, u_t, f_t(x_t, a_t^*)) + w_t = \phi_t(x_t, u_t + f_t(x_t, a_t^*)) + w_t. \tag{8}$$

To control a matched-disturbance system, a natural policy class is to first cancel out the nonlinear residual with $-f_t(x_t, \hat{a}_t)$ and then apply an actuation term $\psi_t(x_t, \theta_t)$ to achieve the optimal costs. This policy class can be expressed as

$$u_t = \pi_t(x_t, \theta_t, f_t(x_t, \hat{a}_t)) = -f_t(x_t, \hat{a}_t) + \psi_t(x_t, \theta_t). \tag{9}$$

Example 1 is in matched disturbance form. More generally, a common case is the joint-space dynamics of robotic manipulators (Siciliano et al., 2008) when the system has actuators for every joint, as detailed in Example 2. The form in Example 2 also applies to tilted-rotor rotorcraft (Rajappa et al., 2015; Zheng et al., 2020), which can move in six degrees of freedom independently.

Example 2 *Consider a robot arm with $J \in \mathbb{N}$ joints. Let $q \in \mathbb{R}^J$ denote the joint angles and $\dot{q}$ their angular velocities. The state is $x = (q, \dot{q})$; the input $u \in \mathbb{R}^J$ is per-joint motor torques. The dynamics are $M(q)\ddot{q} + f_e(q, \dot{q}) = u$, where $M(q) \succ 0$ encapsulates mass/inertia and f_e encapsulates known forces like gravity. Any further effect (e.g., friction) that we wish to learn must enter as an unmodeled torque due to Newton's laws, i.e., $M(q)\ddot{q} + f_e(q, \dot{q}) = u + f(x, a)$, where a is the target of* EST*. A forward Euler time discretization will satisfy (8). An example of ψ in the policy (9) is linear feedback to track a target joint trajectory $(x_t^d)_{t=1}^T$, giving $\psi_t(x_t, \theta_t) = M(q_t)\theta_t(x_t^d - x_t) + f_e(q_t, \dot{q}_t)$, where the gain matrix $\theta_t \in \mathbb{R}^{J \times 2J}$ is under control of* ALG*. The closed-loop dynamics will be a stable linear system, thereby satisfying Property 4.*

To establish Properties 3-5 and 8 for the meta-framework, we need assumptions (Assumptions 18-22) on the dynamics, policy classes, and costs, which we discuss in detail in Appendix H. Note that the matched-disturbance dynamics/policy class we consider can recover the setting (Lin et al., 2023) as a special case when f_t and w_t are always zero (so there is no need to estimate a_t^*). We recover the same regret bound as Lin et al. (2023) in that special case (see Lemma 11).

4.1. Online Policy Selection: M-GAPS

This section introduces a general online policy optimization algorithm, Memoryless Gradient-based Adaptive Policy Selection (M-GAPS, Algorithm 2), which can serves as ALG in our meta-framework. M-GAPS use $\hat{a}_t$ to estimate how the current state x_t and policy parameter θ_t would affect the next state x_{t+1} and the current cost. The estimations are characterized by

$$\hat{g}_{t+1|t}(x_t, \theta_t) \coloneqq g_t(x_t, \pi_t(x_t, \theta_t, f_t(x_t, \hat{a}_t)), f_t(x_t, \hat{a}_t)), \text{ and} \tag{10a}$$

$$\hat{h}_{t|t}(x_t, \theta_t) \coloneqq h_t(x_t, \pi_t(x_t, \theta_t, f_t(x_t, \hat{a}_t)), \theta_t) \tag{10b}$$

Although M-GAPS can be applied to any online policy optimization problems that fit into the setting we discussed in Section 2, we focus on its application to the matched-disturbance dynamics and policy class for theoretical analysis. We verify that the joint dynamics of M-GAPS satisfy the required properties of our meta-framework to derive a concrete regret bound in Appendix C.

The design of M-GAPS takes inspiration from the Gradient-based Adaptive Policy Selection (GAPS) algorithm (Lin et al., 2023), but it significantly improves computational efficiency and generality. Specifically, in the setting where the online agent has exact knowledge of the time-varying dynamical models, M-GAPS can achieve the same regret guarantees as GAPS, while the memory/computational complexities are reduced from $O(\log T)$ to $O(1)$. To understand why this improvement is possible, note that the core problem of GAPS is to design an efficient estimation

Algorithm 2: Memoryless Gradient-based Adaptive Policy Selection (M-GAPS, for ALG)

Parameters: Learning rate η, initial parameter θ_0.
Initialize: Policy parameter θ_0; Internal state $y_0 = \mathbf{0}$.
for $t = 0, 1, \ldots, T-1$ **do**
  Take inputs $x_t, g_t, \pi_t, h_t, f_t$ and $\hat{a}_t$. */* Inputs given when meta-framework calls ALG.update. */*
  Use $\hat{a}_t$ to obtain $\hat{g}_{t+1|t}$ and $\hat{h}_{t|t}$. */* $\hat{g}_{t+1|t}$ and $\hat{h}_{t|t}$ are defined in (10). */*
  Update $y_{t+1} \leftarrow \left.\frac{\partial \hat{g}_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial \hat{g}_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}$. */* Update partial derivatives accumulator. */*
  Let $G_t \leftarrow \left.\frac{\partial \hat{h}_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}$.
  Update and output $\theta_{t+1} \leftarrow \prod_{\Theta}(\theta_t - \eta G_t)$. */* Π_Θ is the Euclidean projection to Θ. */*
end

of $\nabla F_t(\theta_t)$. GAPS does two steps of approximations: first replacing the imaginary trajectory with the actual trajectory and then doing a bounded-memory truncation. In comparison, M-GAPS only keeps the first approximation step of GAPS but greatly simplifies the computation by introducing the auxiliary internal state y_t that accumulates past partial derivatives. Intuitively, the estimation of M-GAPS is even closer to $\nabla F_t(\theta_t)$ than GAPS, so it can achieve the same regret guarantees as GAPS. A more detailed comparison between M-GAPS and GAPS can be found in Appendix J.

A key step of our proof shows that, when exact model parameters $a^*_{0:T-1}$ are available, M-GAPS is robust against perturbations on policy parameter updates as required by Property 5 in Section 3.1.

Lemma 11 *Under Assumptions 18 and 19, Property 5 holds when $\eta \le \bar{\eta}$ for some positive constant $\bar{\eta}$ and $R^L_\eta(T, \sum_{t=0}^{T-1}\|\zeta_t\|) = O\Big(\frac{1}{\eta}(1 + V_{sys} + V_w) + \eta T + \eta^3 T + \frac{1}{\eta}\sum_{t=1}^{T-1}\|\zeta_t\|\Big)$, where the variation intensities are defined as $V_w = \sum_{t=1}^{T-1}\|w_t - w_{t-1}\|$ and*

$$V_{sys} = \sum_{t=1}^{T-1}\Big(\sup_{x\in\mathcal{X},u\in\mathcal{U}}\|\phi_t(x,u) - \phi_{t-1}(x,u)\| + \sup_{x\in\mathcal{X},\theta\in\Theta}\|\psi_t(x,\theta) - \psi_{t-1}(x,\theta)\| + \sup_{x\in\mathcal{X},u\in\mathcal{U},\theta\in\Theta}|h_t(x,u,\theta) - h_{t-1}(x,u,\theta)|\Big).$$

The formal statement and proof of Lemma 11 can be found in Appendix F. Note that in the special case of the nonconvex local regret result in Lin et al. (2023), we have $\Theta = \mathbb{R}^d$, $V_w = 0$, and $\sum_{t=1}^{T-1}\|\zeta_t\| = 0$. The local regret bound $R^L_\eta(T, 0)$ of M-GAPS given by Lemma 11 matches the local regret bound of GAPS in Lin et al. (2023) because the projected gradients are identical with the original gradients when $\Theta = \mathbb{R}^d$.

4.2. Online Parameter Estimation: Gradient Estimator

In the application of matched-disturbance dynamics, we assume the online parameter estimator EST can make a noisy observation $\tilde{f}_t$ of the true nonlinear residual $f_t(x_t, a^*_t)$ after it decides $\hat{a}_t$ at each time step t. Recall that the error function is defined as $e_t(x, a) := f_t(x, a) - f_t(x, a^*_t)$ and the estimation loss at time step t as $\ell_t := \|e_t(x_t, \hat{a}_t)\|^2$. We instantiate EST with the gradient estimator

Algorithm 3: Gradient Estimator (for EST)

Parameters: Learning rate ι; **Initialize:** Model parameter estimation $\hat{a}_0$.

for $t = 0, 1, \ldots, T-1$ **do**

Take inputs x_t, $\tilde{f}_t$, and $\hat{a}_t$. */* Inputs given when meta-framework calls EST.update. */*

Incur loss $\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) \coloneqq \|f_t(x_t, \hat{a}_t) - \tilde{f}_t\|^2$.

Update and output $\hat{a}_{t+1} \leftarrow \prod_{\mathcal{A}}\Big(\hat{a}_t - \iota \cdot \partial \ell_t / \partial a_t|_{x_t, \hat{a}_t, \tilde{f}_t}\Big)$.

end

(Algorithm 3), where $\tilde{f}_t$ is a (noisy) observation of $f_t(x_t, a_t^*)$ provided by the environment. It performs online gradient descent on an approximate estimation loss function constructed from $\tilde{f}_t$.

Using Theorems 6 and 9, we show a local regret guarantee of our meta-framework in Theorem 12 and test it numerically in the setting of Example 1. Due to space limit, we defer the proof of Theorem 12 to Appendix H and the simulation results to Appendix A.

Theorem 12 *Under Assumptions 18-22, if we use M-GAPS for ALG and Gradient Estimator for EST, the trajectory $\xi = \{x_t, y_t, \theta_t\}$ achieves an expected local regret of*

$$R_\eta^L(T) = O\Big(\eta^{-1}(1 + V_{sys} + \bar{\epsilon} \cdot T) + \eta T + (\sqrt{m\bar{\epsilon}} + m\bar{\epsilon}) \cdot T\Big),$$

where V_{sys} is the total variation of the system and $\bar{\epsilon}$ is the magnitude of the random disturbance w_t (see Appendix C for detailed definitions). Under the same definition of $\tilde{\xi}$ as Theorem 6, the expected total distance between ξ and $\tilde{\xi}$ is bounded by $\sum_{t=0}^{T-1}(\|x_t - \tilde{x}_t\| + \|y_t - \tilde{y}_t\|) = O\big(T^{3/4} + \sqrt{m\bar{\epsilon}} \cdot T\big)$.

5. Conclusion

In this work, we propose a meta-framework that combines an online policy optimization algorithm ALG with an online parameter estimator EST to address the challenge of unknown time-varying dynamics. We specify a set of properties that, if satisfied, imply that ALG can act as if EST is providing the true dynamics models, while EST only needs to minimize the model mismatches on the actual trajectory visited by ALG. To demonstrate our framework, we propose an efficient candidate for ALG called M-GAPS (Algorithm 2). When our meta-framework is instantiated with M-GAPS (as ALG) and the gradient estimator (as EST), it achieves the first local regret bound (Theorem 12) for online policy optimization in a class of nonlinear time-varying systems with unknown dynamics.

Our meta-framework also motivates interesting future work: The structural properties of ALG and EST that we identify can serve as guidelines for the design of improved online policy optimization and dynamics estimations methods. For example, tools from online nonconvex optimization and adaptive control could be leveraged to handle more general classes of dynamics.

Acknowledgements

This work was supported by NSF Grants CNS-2146814, CPS-2136197, CNS-2106403, NGSDI-2105648, CCF-1918865, DARPA, a Gift from Latitude AI, and the Caltech Bren Professorship. The research of Yiheng Lin was additionally supported by Amazon AI4Science Fellowship and PIMCO Graduate Fellowship in Data Science. We would like to thank Chuxin Cheng for the inspiring discussions during the development of this research work.

References

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In International Conference on Machine Learning, pages 111–119. PMLR, 2019a. URL https://proceedings.mlr.press/v97/agarwal19c.html.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In Advances in Neural Information Processing Systems, volume 32, 2019b. URL https://dl.acm.org/doi/pdf/10.5555/3454287.3455200.

Anuradha M. Annaswamy and Alexander L. Fradkov. A historical perspective of adaptive control and learning. Annu. Reviews in Control, 52:18–41, October 2021. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2021.10.014. URL https://www.sciencedirect.com/science/article/pii/S1367578821000894.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. J. Mach. Learn. Res., 22:106–1, 2021. URL https://jmlr.csail.mit.edu/papers/volume22/20-1288/20-1288.pdf.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In Conference on Learning Theory, pages 359–409. PMLR, 2021. URL https://proceedings.mlr.press/v134/baby21a/baby21a.pdf.

Nicholas M. Boffi, Stephen Tu, and Jean-Jacques E. Slotine. Regret bounds for adaptive nonlinear control. In Ali Jadbabaie, John Lygeros, George J. Pappas, Pablo Anbsp;Parrilo, Benjamin Recht, Claire J. Tomlin, and Melanie N. Zeilinger, editors, Proceedings of the 3rd Conference on Learning for Dynamics and Control, volume 144 of Proceedings of Machine Learning Research, pages 471–483. PMLR, 07 – 08 June 2021. URL https://proceedings.mlr.press/v144/boffi21a.html.

Nicolo Cesa-Bianchi and Gabor Lugosi. Prediction, Learning, and Games. Cambridge University Press, 2006. URL https://www.cambridge.org/core/books/prediction-learning-and-games/A05C9F6ABC752FAB8954C885D0065C8F.

Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. In Conference on Learning Theory, pages 1114–1143. PMLR, 2021. URL https://proceedings.mlr.press/v134/chen21c.html.

Xinyi Chen, Edgar Minasyan, Jason D Lee, and Elad Hazan. Regret guarantees for online deep control. Proceedings of Machine Learning Research vol 211, 1:34, 2023. URL https://proceedings.mlr.press/v211/chen23b/chen23b.pdf.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{t}$ regret. In International Conference on Machine Learning, pages 1300–1309. PMLR, 2019. URL https://proceedings.mlr.press/v97/cohen19b.html.

Charles Dawson, Sicun Gao, and Chuchu Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. IEEE Transactions on Robotics, pages 1–19, 2023. doi: 10.1109/TRO.2022.3232542. URL https://dl.acm.org/doi/abs/10.1109/TRO.2022.3232542.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In Advances in Neural Information Processing Systems, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/0ae3f79a30234b6c45a6f7d298ba1310-Paper.pdf.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. Foundations of Computational Mathematics, 20(4):633–679, 2020. URL https://link.springer.com/article/10.1007/s10208-019-09426-y.

Ilgin Dogan, Zuo-Jun Max Shen, and Anil Aswani. Regret analysis of learning-based mpc with partially-unknown cost function. arXiv preprint arXiv:2108.02307, 2021. URL https://arxiv.org/abs/2108.02307.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In International conference on machine learning, pages 1467–1476. PMLR, 2018. URL https://proceedings.mlr.press/v80/fazel18a/fazel18a.pdf.

Joel Ferguson, Alejandro Donaire, Romeo Ortega, and Richard H. Middleton. Matched disturbance rejection for a class of nonlinear systems. IEEE Transactions on Automatic Control, 65(4):1710–1715, 2020. doi: 10.1109/TAC.2019.2933398. URL https://ieeexplore.ieee.org/document/8788577.

Gianluca Garofalo, Christian Ott, and Alin Albu-Schäffer. Walking control of fully actuated robots based on the bipedal slip model. In 2012 IEEE International Conference on Robotics and Automation, pages 1456–1463, 2012. doi: 10.1109/ICRA.2012.6225272. URL https://ieeexplore.ieee.org/document/6225272.

Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. In Nikolai Matni, Manfred Morari, and George J. Pappas, editors, Proceedings of The 5th Annual Learning for Dynamics and Control Conference, volume 211 of Proceedings of Machine Learning Research, pages 560–572. PMLR, 15–16 Jun 2023. URL https://proceedings.mlr.press/v211/gradu23a.html.

Elad Hazan. Introduction to online convex optimization. Foundations and Trends® in Optimization, 2(3-4):157–325, 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL http://dx.doi.org/10.1561/2400000013.

Elad Hazan and C. Seshadhri. Efficient learning algorithms for changing environments. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, page 393–400, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553425. URL https://doi.org/10.1145/1553374.1553425.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. In Electronic Colloquium on Computational Complexity (ECCC), volume 14, 2007. URL https://www.cs.princeton.edu/techreports/2007/798.pdf.

Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In International Conference on Machine Learning, pages 1433–1441. PMLR, 2017. URL https://proceedings.mlr.press/v70/hazan17a/hazan17a.pdf.

Dimitar Ho, Hoang Le, John Doyle, and Yisong Yue. Online robust control of nonlinear systems with large uncertainty. In International Conference on Artificial Intelligence and Statistics, pages 3475–3483. PMLR, 2021. URL https://proceedings.mlr.press/v130/ho21a/ho21a.pdf.

P.A. Ioannou and J. Sun. Robust Adaptive Control. Dover Books on Electrical Engineering Series. Dover Publications, Incorporated, 2012. ISBN 9780486498171. URL https://books.google.com/books?id=pXWFY_vbg1MC.

Visakan Kadirkamanathan and Simon Fabri. Stable nonlinear adaptive control with growing radial basis function networks. 5th IFAC Symp. on Adaptive Systems in Control and Signal Processing, 28(13):245–250, June 1995. ISSN 1474-6670. doi: https://doi.org/10.1016/S1474-6670(17)45357-2. URL https://www.sciencedirect.com/science/article/pii/S1474667017453572.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pages 5354–5390. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/lale22a.html.

Yingying Li, Xin Chen, and Na Li. Online optimal control with linear dynamics and predictions: Algorithms and regret analysis. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://openreview.net/pdf?id=HkgmTHBlLH.

Yingying Li, James A Preiss, Na Li, Yiheng Lin, Adam Wierman, and Jeff S Shamma. Online switching control with stability and regret guarantees. In Learning for Dynamics and Control Conference, pages 1138–1151. PMLR, 2023. URL https://proceedings.mlr.press/v211/li23a/li23a.pdf.

Yiheng Lin, James A. Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations. In Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=hDajsofjRM.

Edgar Minasyan, Paula Gradu, Max Simchowitz, and Elad Hazan. Online control of unknown time-varying dynamical systems. Advances in Neural Information Processing Systems, 34:15934–15945, 2021. URL https://proceedings.neurips.cc/paper/2021/file/856b503e276cc491e7e6e0ac1b9f4b17-Paper.pdf.

Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In 2016 IEEE 55th Conference on Decision and Control (CDC), pages 7195–7201. IEEE, 2016. URL https://dl.acm.org/doi/10.1109/CDC.2016.7799379.

Deepan Muthirayan, Ruijie Du, Yanning Shen, and Pramod P. Khargonekar. Adaptive control of unknown time varying dynamical systems with regret guarantees. CoRR, abs/2210.11684, 2022. URL https://arxiv.org/abs/2210.11684.

Michael O'Connell, Guanya Shi, Xichen Shi, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural-fly enables rapid learning for agile flight in strong winds. Science Robotics, 7(66):eabm6597, 2022. URL https://www.science.org/doi/10.1126/scirobotics.abm6597.

Guannan Qu, Chenkai Yu, Steven Low, and Adam Wierman. Exploiting linear models for model-free nonlinear control: A provably convergent policy gradient approach. In 2021 60th IEEE Conference on Decision and Control (CDC), pages 6539–6546. IEEE, 2021. URL https://ieeexplore.ieee.org/document/9683735.

Sujit Rajappa, Markus Ryll, Heinrich H. Bülthoff, and Antonio Franchi. Modeling, control and design optimization for a fully-actuated hexarotor aerial vehicle with tilted propellers. In IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015, pages 4006–4013. IEEE, 2015. doi: 10.1109/ICRA.2015.7139759. URL https://doi.org/10.1109/ICRA.2015.7139759.

Rolf Schneider. Convex Bodies: the Brunn–Minkowski Theory. Cambridge University Press, 2014. URL https://books.google.com/books/about/Convex_Bodies.html?id=2QhT8UCKx2kC.

Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, and Yisong Yue. Meta-adaptive nonlinear control: Theory and algorithms. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=nm3sOq42Gmx.

Xichen Shi, Patrick Spieler, Ellande Tang, Elena-Sorina Lupu, Phillip Tokumaru, and Soon-Jo Chung. Adaptive Nonlinear Control of Fixed-Wing VTOL with Airflow Vector Sensing. In IEEE Int. Conf. on Robotics and Automation (ICRA), pages 5321–5327, May 2020. ISBN 978-1-72817-395-5. doi: 10.1109/ICRA40945.2020.9197344. URL https://ieeexplore.ieee.org/document/9197344/.

Bruno Siciliano, Lorenzo Sciavicco, Luigi Villani, and Giuseppe Oriolo. Robotics: Modelling, Planning and Control. Springer, 2008. ISBN 1846286417. URL https://link.springer.com/book/10.1007/978-1-84628-642-1.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In Conference On Learning Theory, pages 439–473. PMLR, 2018. URL https://proceedings.mlr.press/v75/simchowitz18a/simchowitz18a.pdf.

Rohan Sinha, James Harrison, Spencer M. Richards, and Marco Pavone. Adaptive robust model predictive control with matched and unmatched uncertainty. In 2022 American Control Conference (ACC), pages 906–913, 2022. doi: 10.23919/ACC53348.2022.9867457. URL https://ieeexplore.ieee.org/document/9867457.

Jean-Jacques E Slotine, Weiping Li, et al. Applied nonlinear control, volume 199. Prentice hall Englewood Cliffs, NJ, 1991. URL https://lewisgroup.uta.edu/ee5323/notes/Slotine%20and%20Li%20applied%20nonlinear%20control-%20bad%20quality.pdf.

K.A. Wise, E. Lavretsky, and N. Hovakimyan. Adaptive control of flight: theory, applications, and open problems. In 2006 American Control Conference, pages 6 pp.–, 2006. doi: 10.1109/ACC.2006.1657677. URL https://ieeexplore.ieee.org/document/1657677/.

Jing Yu, Varun Gupta, and Adam Wierman. Online adversarial stabilization of unknown linear time-varying systems. In 2023 62nd IEEE Conference on Decision and Control (CDC), pages 8320–8327. IEEE, 2023a. URL https://dl.acm.org/doi/10.1145/3579452.

Jing Yu, Dimitar Ho, and Adam Wierman. Online adversarial stabilization of unknown networked systems. Proc. ACM Meas. Anal. Comput. Syst., 7(1), mar 2023b. doi: 10.1145/3579452. URL https://doi.org/10.1145/3579452.

Peng Zhao, Yu-Xiang Wang, and Zhi-Hua Zhou. Non-stationary online learning with memory and non-stochastic control. In International Conference on Artificial Intelligence and Statistics, pages 2101–2133. PMLR, 2022. URL https://www.jmlr.org/papers/volume24/22-0218/22-0218.pdf.

Peter Zheng, Xinkai Tan, Basaran Bahadir Kocer, Erdeng Yang, and Mirko Kovac. Tiltdrone: A fully-actuated tilting quadrotor platform. IEEE Robotics and Automation Letters, 5(4):6845–6852, 2020. doi: 10.1109/LRA.2020.3010460. URL https://ieeexplore.ieee.org/document/9144371.

Hongyu Zhou, Zirui Xu, and Vasileios Tzoumas. Efficient online learning with memory via Frank-Wolfe optimization: Algorithms with bounded dynamic regret and applications to control. CoRR, abs/2301.00497, 2023. URL https://arxiv.org/abs/2301.00497.

Outline of the appendices.

- In Appendix A, we provide the simulation results for a nonlinear control application (Example 1).
- In Appendix B, we provide a notation table and list important definitions used in the proofs.
- In Appendix C, we present the details about the application of matched-disturbance dynamics.
- In Appendix D, we prove Theorem 6 on the joint dynamics of ALG in our meta-framework.
- In Appendix E, we prove Theorem 9 on the regret of EST in our meta-framework.
- In Appendix F, we show M-GAPS satisfies the properties for ALG in the application of matched-disturbance dynamics.
- In Appendix G, we show the gradient estimator satisfies the properties for EST in the application of matched-disturbance dynamics.
- In Appendix H, we prove Theorem 12 about instantiating our meta-framework with M-GAPS (for ALG) and the gradient estimator (for EST) in the application of matched-disturbance dynamics.
- In Appendix I, we show online gradient descent with inexact updates can achieve local regret bounds in online nonconvex optimization, which is used in the proof of M-GAPS.
- In Appendix J, we make a detailed comparison between our M-GAPS algorithm with the GAPS algorithm proposed by Lin et al. (2023). We summarize some results from Lin et al. (2023) that are useful for us to analyze M-GAPS.

Appendix A. Simulation Results

In this section, we show our simulation results in the setting of Example 1, where the dynamics model, control policy and the update law of the online parameter estimator EST are given by

$$x_{t+1} = x_t + \Delta(u_t + \phi(x_t)a_t^*) + w_t \tag{11a}$$
$$u_t = -\theta_t x_t - \phi(x_t)\hat{a}_t \tag{11b}$$
$$\hat{a}_{t+1} = \hat{a}_t - (\phi(x_t)\hat{a}_t - \phi(x_t)a_t^*)\Delta P\phi(x_t), \tag{11c}$$

where $\phi(x) = \begin{bmatrix}1 & \sin(x) & \sin(2x)\end{bmatrix}$ is a basis function, $\Delta = 0.01$ is the time interval between steps and P is a constant gain. The disturbances $w_{0:T-1}$ are generated by Ornstein-Uhlenbeck random walk. The Ornstein-Uhlenbeck random walk updates the random disturbance w_t following the dynamics $w_{t+1} = \gamma w_t + \delta_t$, where parameter $\gamma \in [0, 1)$ is a constant decay factor and δ_t is Gaussian-distributed with zero mean and variance σ^2. In the simulation, we set $\gamma = 0.95$ and $\sigma = 0.1$. We consider a time-varying dynamical system where the true model parameter a_t^* changes at the end of each period of 40000 time steps. When a period ends, we resample a_t^* from a uniform distribution over the closed interval $[-0.25, 0.25]^3$. Figure 3 contains a plot of time-varying a_t^*. The stage cost function is given by $h_t(x_t, u_t) = x_t^2 + \beta u_t^2$, where we set $\beta = 0.1$.

We test our meta-framework by 1) instantiating ALG using M-GAPS with the learning rate $\eta = 1 \times 10^{-4}$, and 2) instantiating EST with the gradient estimator (Algorithm 3), whose update is given by (11c). For EST, we make a minor modification to the prediction loss function by changing

Table 1: Important notations in this paper.

Notation	Meaning
$t_1 : t_2$	The integer sequence $\{t_1, \dots, t_2\}$;
$a_{t_1:t_2}$	A sequence of variables $\{a_t\}_{t=t_1,\dots,t_2}$;
$\|\cdot\|$	ℓ_2 (Euclidean) norm;
$\|\cdot\|_F$	Frobenius norm;
$\|\cdot\|_P$	Norm induced by matrix P;
$\mathbb{Z}_{\geq 0}$	The set of non-negative integers;
$\mathbb{R}_{\geq 0}$	The set of non-negative reals;
$\sigma(z_{1:t}, z'_{1:t})$	Product sigma-algebra generated by sequences $z_{1:t}$ and $z'_{1:t}$;
x_t	$x_t \in \mathbb{R}^n$ is the system state;
u_t	$u_t \in \mathbb{R}^m$ is the control input;
w_t	$w_t \in \mathcal{W} \subseteq \mathbb{R}^n$ is a disturbance term;
$f_t(x_t, a_t^*)$	f_t is a nonlinear residual term that the agent makes (noisy) observations of;
a_t^*	$a_t^* \in \mathcal{A} \subseteq \mathbb{R}^p$ is the unknown parameter in f_t
$f_t(\cdot, \hat{a}_t)$	An estimation of the true nonlinear residual function $f_t(\cdot, a_t^*)$;
$q_t(x_t, y_t, \theta_t, a_t)$	The joint dynamics of the system at time t;
$\Pi_\Theta(y)$	Euclidean projection of y to set Θ;

it from the ℓ_2-squared prediction error $\|\phi(x_t)\hat{a}_t - \phi(x_t)a_t^*\|^2$ to $\|\phi(x_t)\hat{a}_t - \phi(x_t)a_t^*\|_P^2$, where $\|\cdot\|_P$ denotes the norm induced by matrix P. Note that the update law of the gradient estimator in this setting g is identical with a classic adaptive controller (Slotine et al., 1991, §8.7.3). We compare our algorithm with a common benchmark that uses the adaptive controller to learn $\hat{a}_t$ while the policy parameter (the feedback gain) θ_a is fixed:

$$x_{t+1} = x_t + \Delta(u_t + \phi(x_t)a_t^*) + w_t \tag{12a}$$

$$u_t = -\theta_a x_t - \phi(x_t)\hat{a}_t \tag{12b}$$

$$\hat{a}_{t+1} = \hat{a}_t - (\phi(x_t)\hat{a}_t - \phi(x_t)a_t^*)\Delta P\phi(x_t). \tag{12c}$$

The result is shown in Figure 3-7. Similar to traditional adaptive controllers, adaptive parameters $\hat{a}_t$ do not necessarily converge to the real value a_t^*. Still, our model and tracking errors converge to a small error ball. Our algorithm optimizes the cost function with control input and improves the accumulative costs significantly compared with the baseline adaptive controller (12).

Appendix B. Notations and Definitions

We provide a notation table (Table 1) that summarizes the important notations in this paper.

A key concept that we explore in this paper is how to compare the actual trajectory of our meta-framework with an "ideal" trajectory that the agent could achieve with exact knowledge of the true model parameters $a^*_{0:T-1}$, we introduce the important notations of multi-step dynamics/cost that characterize how the system would evolve under a sequence of policy parameters $\theta_{0:T-1}$ when $a^*_{0:T-1}$ is known. The concepts of multi-step dynamics/cost are first introduced in Lin et al. (2023), which studies online policy selection with known dynamical systems. In this work, we replace all estimated $\hat{a}_t$ in the policy classes with true a_t^* to reproduce the same definition as Lin et al. (2023).

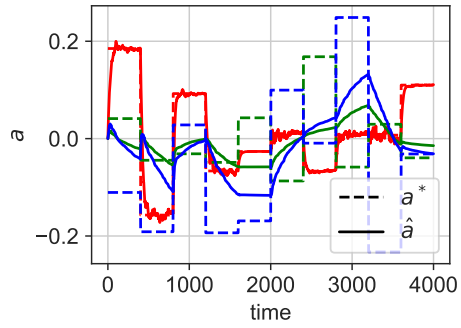

Figure 3: a^* and $\hat{a}$

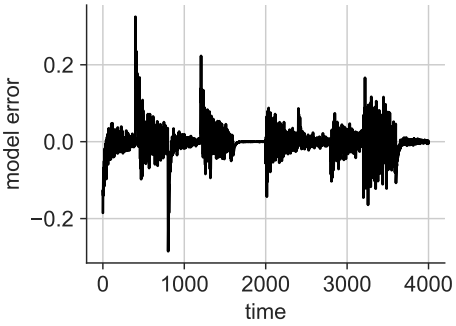

Figure 5: Model error $\phi(x)\hat{a} - \phi(x)a^*$

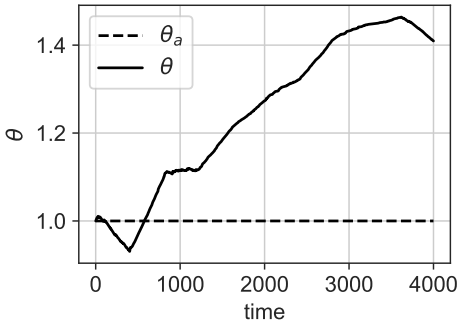

Figure 4: θ_a and $\hat{\theta}$

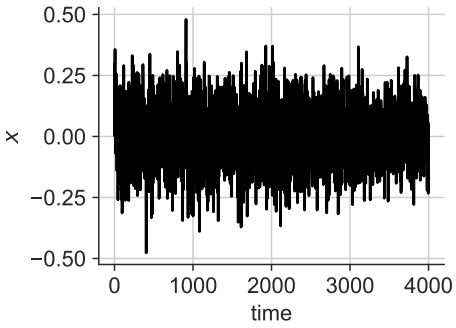

Figure 6: Tracking error x

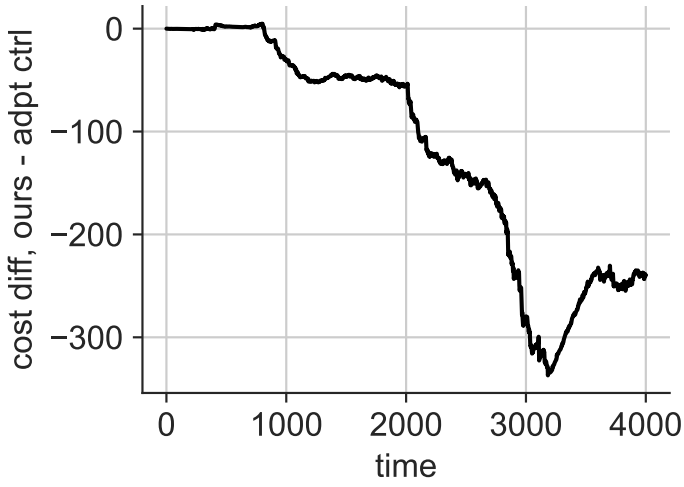

Figure 7: Regret difference between M-GAPS and adaptive control

Definition 13 (Multi-Step Dynamics and Cost) *The multi-step dynamics $g^*_{t|\tau}$ between two time steps $\tau \le t$ specifies the state x_t as a function of the previous state x_τ and previous policy parameters $\theta_{\tau:t-1}$ under exact predictions $\{a^*_t\}$. It is defined recursively, with the base case $g^*_{\tau|\tau}(x_\tau) \coloneqq x_\tau$ and the recursive case*

$$g^*_{t+1|\tau}(x_\tau, \theta_{\tau:t}) = g_t(z_t, \pi_t(z_t, \theta_t, f_t(z_t, a^*_t)), f_t(z_t, a^*_t)) + w_t, \ \forall\, t \ge \tau,$$

in which $z_t \coloneqq g^*_{t|\tau}(x_\tau, \theta_{\tau:t-1})$.[3] *The multi-step cost* $h^*_{t|\tau}$ *specifies the cost* c_t *as function of* x_τ *and* $\theta_{\tau:t}$. *It is defined as*

$$h^*_{t|\tau}(x_\tau, \theta_{\tau:t}) \coloneqq h_t(z_t, \pi_t(z_t, \theta_t, f_t(z_t, a^*_t)), \theta_t).$$

It is worth emphasizing that, in our work, the concepts of multi-step dynamics/cost are only used for the theoretical analysis, because their definitions involve true model parameters that are unknown to the online agent. When doing online policy optimization, the online agent may use the estimations $\hat{g}_{t+1|t}$ and $\hat{h}_{t|t}$ (see (10)) as the estimations of $g^*_{t+1|t}$ and $h^*_{t|t}$ respectively. Note that this is different than the case when true dynamics are known (Lin et al., 2023), where the online agent can directly construct multi-step dynamics/cost ($g^*_{t+1|t}$ and $h^*_{t|t}$) or compute the exact Jacobian matrices.

Another important definition that we require is the *projected gradient*, which is introduced in Hazan et al. (2017) to accommodate the challenge of analyzing gradient-based online policy approaches on a constrained set.

Definition 14 (Projected gradient) *Let* $F : \Theta \to \mathbb{R}$ *be a differentiable function on a closed convex set* $\Theta \subseteq \mathbb{R}^d$. *For* $\eta > 0$, *the* (Θ, η)*-projected gradient of* F *is defined as*

$$\nabla_{\Theta,\eta} F(\theta) \coloneqq \frac{1}{\eta}(\theta - \Pi_\Theta(\theta - \eta \nabla F(\theta))).$$

When Θ is equal to the whole Euclidean space $\mathbb{R}^d$ (unconstrained), the projected gradient in Definition 14 will be identical with the normal gradient $\nabla F(\theta)$. This concept of projected gradient is used to define the local regret in Section 2.1, Theorem 12, and Appendix I that studies online gradient descent for online nonconvex optimization with constraints.

Appendix C. Matched-Disturbance Dynamics

In this section, we discuss the detailed assumptions on the application of matched-disturbance dynamics and how they enable us to apply the theory for our meta-framework.

We first focus on the `ALG` part that is instantiated with M-GAPS (Algorithm 2). Note that in the setting of this application, the definition of multi-step dynamics (Definition 13) can be simplified to

$$g^*_{t+1|\tau}(x_\tau, \theta_{\tau:t}) = \phi_t(z_t, \psi_t(z_t, \theta_t)) + w_t, \text{ where } z_t \coloneqq g^*_{t|\tau}(x_\tau, \theta_{\tau:t-1}).$$

We see that $g^*_{t+1|\tau}$ has exactly the same form as the multi-step dynamics studied in Lin et al. (2023). Since $g^*_{t+1|\tau}$ corresponds to the case when the nonlinear residual has been canceled out with the exact model parameter a^*_t, it is reasonable to make the same assumptions as Lin et al. (2023) about the stability and contraction properties of $g^*_{t+1|\tau}$. These assumptions rely on the key definitions of time-varying contractive perturbation and time-varying stability.

Definition 15 *We denote the set of policy parameter sequences with* ϵ_θ*-constrained step size by*

$$S_{\epsilon_\theta}(t_1 : t_2) \coloneqq \{\theta_{t_1:t_2} \in \Theta^{t_2-t_1+1} \mid \|\theta_{\tau+1} - \theta_\tau\| \le \epsilon_\theta, \forall \tau \in [t_1 : t_2 - 1]\}.$$

3. z_t is an auxiliary variable to denote the state at t under initial state x_τ and parameters $\theta_{\tau:t}$.

Definition 16 (ϵ_θ-Time-varying Contractive Perturbation) *For $\epsilon_\theta \geq 0$, the ϵ_θ-time-varying contractive perturbation property holds for $R_C > 0, C > 0$, and $\rho \in (0,1)$ if, for any $\theta_{\tau:t-1} \in S_{\epsilon_\theta}(\tau : t-1)$, the following inequality holds for arbitrary $x_\tau, x'_\tau \in B_n(0, R_C)$ and time steps $\tau \leq t$:*

$$\left\| g^*_{t|\tau}(x_\tau, \theta_{\tau:t-1}) - g^*_{t|\tau}(x'_\tau, \theta_{\tau:t-1}) \right\| \leq C\rho^{t-\tau} \left\| x_\tau - x'_\tau \right\|$$

Definition 17 (ϵ_θ-Time-varying Stability) *For $\epsilon_\theta \geq 0$, the ϵ_θ-time-varying stability property holds for $R_S > 0$ if, for any $\theta_{\tau:t-1} \in S_{\epsilon_\theta}(\tau : t-1)$, $\left\| g^*_{t|\tau}(0, \theta_{\tau:t-1}) \right\| \leq R_S$ holds for any $t \geq \tau$.*

With the definitions of time-varying contractive perturbation and time-varying stability, we state our key assumptions below: Assumption 18 is about the Lipschitzness/smoothness properties of the dynamics, policy, nonlinear residual, and the cost functions.

Assumption 18 *The dynamics $\phi_{0:T-1}$, policies $\psi_{0:T-1}$, residuals $f_{0:T-1}$, and costs $h_{0:T-1}$ are differentiable at every time step and satisfy that, for any convex compact sets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{U} \subseteq \mathcal{R}^m$, one can find Lipschitzness/smoothness constants (can depend on $\mathcal{X}$ and $\mathcal{U}$) such that:*

1. *$\phi_t(x,u)$ is $(L_{\phi,x}, L_{\phi,u})$-Lipschitz and $(\ell_{\phi,x}, \ell_{\phi,u})$-smooth in (x,u) on $\mathcal{X} \times \mathcal{U}$.*
2. *$\psi_t(x,\theta)$ is $(L_{\psi,x}, L_{\psi,\theta})$-Lipschitz and $(\ell_{\psi,x}, \ell_{\psi,\theta})$-smooth in (x,θ) on $\mathcal{X} \times \Theta$.*
3. *$f_t(x,a)$ is $(L_{f,x}, L_{f,a})$-Lipschitz and $(\ell_{f,x}, \ell_{f,a})$-smooth in (x,a) on $\mathcal{X} \times \mathcal{A}$.*
4. *$h_t(x,u,\theta)$ is $(L_{h,x}, L_{h,u}, L_{h,\theta})$-Lipschitz and $(\ell_{h,x}, \ell_{h,u}, \ell_{h,\theta})$-smooth in (x,u,θ) on $\mathcal{X} \times \mathcal{U} \times \Theta$.*

Compared with Assumption 2.1 in Lin et al. (2023), our Assumption 18 additionally assumes the Lipschitzness and smoothness of the nonlinear residual function f_t, which is part of our dynamics and policy classes. The second assumption (Assumption 19) is on the contractive perturbation and the stability of the multi-step dynamics $g^*_{t|\tau}$.

Assumption 19 *Let $\mathcal{G}$ denote the set of all possible sequences $\{\phi_t, f_t, w_t, \psi_t\}_{t \in \mathcal{T}}$ the environment may provide. For a fixed $\epsilon_\theta \in \mathbb{R}_{\geq 0}$, the ϵ_θ-time-varying contractive perturbation holds with $(R_C, \bar{C}, \rho)$ for any sequence in $\mathcal{G}$. The ϵ_θ-time-varying stability holds with $R_S < R_C$ for any sequence in $\mathcal{G}$. We assume that the initial state satisfies $\|x_0\| < (R_C - R_S)/\bar{C}$. Further, we assume that if $\{\phi, f, w, \psi\}$ is the dynamics/residual/disturbance/policy at an intermediate time step of a sequence in $\mathcal{G}$, then the time-invariant sequence $\{\phi, f, w, \psi\}_{\times T}$ is also in $\mathcal{G}$.*[4]

Compared with Assumption 2.2 in Lin et al. (2023), our Assumption 19 also includes the disturbance w_t as a part of the system configuration. This is because for every time t, $g^*_{t+1|t}$ is formed by ϕ_t, π_t, and w_t together. While w_t can also be viewed as a part of the dynamics ϕ_t, we choose to represent it separately because we will leverage the randomness of w_t to bound the first-order model mismatches of `EST`. Like Lin et al. (2023), in Assumption 19, we assume there exists a positive real number $\bar{R}_C$ such that $R_C > \bar{R}_C > R_S + \bar{C}\|x_0\|$. Here, we introduce the real constant $\bar{R}_C$ because R_C can be $+\infty$ when time-varying contractive perturbation (Definition 16) holds globally. Similarly, to leverage the Lipschitzness/smoothness property, we require $\mathcal{X} \supseteq B(0, R_x)$ where $R_x \geq \bar{C}\bar{R}_C + R_S$ and $\mathcal{U} = \{-f(x,a) + \pi(x,\theta) \mid x \in \mathcal{X}, \theta \in \Theta, a \in \mathcal{A}, \pi, f \in \mathcal{G}\}$. Since the coefficients in Assumption 18 depend on $\mathcal{X}$ and $\mathcal{U}$, we will set $\mathcal{X} = B_n(0, R_x)$ and $R_x = \bar{C}\bar{R}_C + R_S$

4. For $\{\phi, f, w, \psi\}_{\times T}$ to be in $\mathcal{G}$, it must satisfy other assumptions about contractive perturbation and stability that we impose on $\mathcal{G}$ but does not need to occur in real problem instances. This assumption can be made without the loss of generality for time-invariant dynamics and policy classes.

by default when presenting these constants. We also set $\mathcal{Y} = B_p(0, R_y)$ with $R_y = \frac{\bar{C} L_{\phi,u} L_{\psi,\theta}}{\rho(1-\rho)}$, so that the internal state y_t will stay in $\mathcal{Y}$.

It is straightforward to verify that the joint dynamics of M-GAPS satisfy the three properties required by the meta-framework. We state this result in Lemma 20.

Lemma 20 *Under Assumptions 18 and 19, M-GAPS (Algorithm 2) satisfy Properties 3, 4, and 5 when applied to dynamics (8) and policy class (9).*

We present the specific constants and the formal proof of Lemma 20 in Appendix F.

For the part of `EST` that is instantiated with the gradient estimator (Algorithm 3), we first introduce an assumption about the magnitude of the nonlinear residual to guarantee that (several) bad estimations of the unknown model parameters will not destabilize the system or violate the constraints of the contractive perturbation property.

Assumption 21 *The set of all possible model parameters $\mathcal{A}$ is a convex compact subset of $\mathbb{R}^p$. For any fixed $x \in B_n(0, R_x)$, $f_t(x, \cdot) : \mathcal{A} \to \mathbb{R}^m$ is an affine function whose gradient is uniformly bounded, i.e., for some positive constant D'_f, $\|\nabla_x f_t(x, a)\| \le D'_f$ hold for all $a \in \mathcal{A}$. It also satisfies that for any $a, a' \in \mathcal{A}$,*

$$\big\|f_t(x,a) - f_t(x,a')\big\| \le C_f,\ \big\|\nabla_x f_t(x,a) - \nabla_x f_t(x,a')\big\|_F \le \beta \le C'_f,\ \text{and}$$
$$\big\|\nabla_x^2 f_t(x,a)_i - \nabla_x^2 f_t(x,a')_i\big\|_F \le \gamma,\ \text{for any dimension } i \in [1:m].$$

hold with some positive constants β, γ, and the upper bounds C_f and C'_f are given by

$$C_f = \min\left\{\frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C\alpha_x}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C\alpha_x}, \frac{\bar{\zeta}}{2\alpha_\theta}\right\},$$
$$C'_f = \min\left\{\frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C\beta_x}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C\beta_x}, \frac{\bar{\zeta}}{2\beta_\theta}\right\}.$$

The expressions of $\alpha_x, \beta_x, \alpha_\theta, \beta_\theta, L_{\theta,x}, L_{\theta,y}, R_x^*, R_y^*$, and $\bar{\zeta}$ are given in Appendix F.

Note that we need Assumption 21 to bound the model mismatches uniformly because even if an online parameter estimator performs well in the long term (e.g., achieving a regret bound on the total model mismatches), it may incur a large error at a single time step that can potentially destabilize the system especially when a_t^* changes abruptly. Our simulation (Appendix A) provides a good illustration of this intuition: The model mismatches may increase dramatically right after the system switches to a different true model parameter a_t^*; Then, the error converges back to near zero as the gradient estimator learns the model (see Figure 5). Addressing this challenge with other assumptions like slowly time-varying a_t^* is an interesting future direction.

The second assumption we need is about the randomness provided by the environment:

Assumption 22 *The total path length of the true model parameters satisfies*

$$1 + \sum_{t=1}^{T-1} \big\|a_t^* - a_{t-1}^*\big\| \le C_p$$

for some positive constant C_p. At every time step t, the noisy observation $\tilde{f}_t$ satisfies that

$$\left\|\tilde{f}_t - f_t(x_t, a_t^*)\right\| \le e_f, \text{ and } \mathbb{E}[\tilde{f}_t \mid \mathcal{F}_t] = f_t(x_t, a_t^*).$$

Further, the random disturbance w_t in the dynamical system (8) satisfies that

$$\|w_t\| \le \bar{\epsilon}, \ \mathbb{E}[w_t \mid \mathcal{F}'_t] = 0, \text{ and } Cov(w_t \mid \mathcal{F}'_t) \succeq c\bar{\epsilon}^2 I.$$

Here, $\bar{\epsilon}$ satisfies that

$$(C_f + e_f)\left(2D'_f\sqrt{3C_p/T}\right)^{\frac{1}{3}} \le \bar{\epsilon} \le \min\{\frac{1}{4}, \frac{1}{2\gamma}, \frac{1}{4\beta\gamma}\},$$

where β and γ are defined in Assumption 21.

Intuitively, Assumption 22 put requirements on both the lower and upper bounds of the level of randomness in the system. The lower bound $\bar{\epsilon} = \Omega(T^{-1/6})$ is required due to the condition $R_0^\ell(T) \le \bar{\epsilon}^3 T$ in Theorem 9. This guarantees that the zeroth-order regret is sufficiently small to be used for bounding the first-order gradients in Taylor's expansion, which are multiplied by $\bar{\epsilon}^2$ when we take the square. The upper bound of $\bar{\epsilon}$ is required to ignore the higher-order terms in Taylor's expansion. With these assumptions, we show the following guarantee on the total model mismatches achieved by the gradient estimator (Algorithm 3).

Lemma 23 *Under Assumptions 21 and 22, the total squared zeroth-order model mismatches of the gradient estimator can be bounded by $\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon_t^2\right] \le 2\sqrt{3}(C_f + e_f)^3 D'_f\sqrt{C_p T}$.*

We defer the proof of Lemma 23 to Appendix G. Note that our meta-framework only requires us to bound the total squared zeroth-order model mismatch incurred by an instantiation of EST. Under Assumptions 21 and 22, we can apply Theorem 9 in the meta-framework to bound the total squared first-order model mismatch of the gradient estimator by $\mathbb{E}\left[\sum_{t=0}^{T-1}(\varepsilon'_t)^2\right] = O(m\bar{\epsilon}T)$. Recall that m is the dimension of the unknown component, which is identical with the control input in this application.

In Lemmas 20 and 23, we have shown that M-GAPS and the gradient estimator satisfy all the required properties for ALG and EST respectively in our meta-framework. Therefore, we can obtain the local regret guarantees for instantiating our meta-framework with M-GAPS and the gradient estimator in the application with matched-disturbance dynamics (Theorem 12).

Appendix D. Proof of Theorem 6

To simplify the notation, we define

$$\bar{\varepsilon} := \min\left\{\frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C}\right\}.$$

By the assumption, we know that the following inequality holds for all time step t:

$$(\alpha_x + \alpha_y)\varepsilon_t + (\beta_x + \beta_y)\varepsilon'_t \le \bar{\varepsilon}. \tag{13}$$

Now we show that $\|x_t\| \le R_x$, $\|y_t\| \le R_y$, and $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ by induction. These inequalities hold for time step 0. Suppose they hold for all time steps $\tau \le t$. Then, for time step $t+1$, by Theorem 4, we see that

$$\|\tilde{x}_{t+1}\| \le R_x^*, \text{ and } \|\tilde{y}_{t+1}\| \le R_y^*. \tag{14}$$

By Property 4 about the contraction of states x_t and y_t under policy parameters $\theta_{0:t}$, we see that

$$\begin{aligned}
&\|(x_{t+1}, y_{t+1}) - (\tilde{x}_{t+1}, \tilde{y}_{t+1})\| \\
&\le \sum_{\tau=0}^{t} \left\| q^{(x,y)*}_{t+1|t+1-\tau}(x_{t+1-\tau}, y_{t+1-\tau}, \theta_{t+1-\tau:t}) - q^{(x,y)*}_{t+1|t-\tau}(x_{t-\tau}, y_{t-\tau}, \theta_{t-\tau:t}) \right\| && \text{(15a)} \\
&\le \sum_{\tau=0}^{t} \gamma(\tau) \left\| (x_{t+1-\tau}, y_{t+1-\tau}) - q^{(x,y)*}_{t+1-\tau|t-\tau}(x_{t-\tau}, y_{t-\tau}, \theta_{t-\tau}) \right\| && \text{(15b)} \\
&\le \sum_{\tau=0}^{t} \gamma(\tau) \big((\alpha_x + \alpha_y)\varepsilon_{t-\tau} + (\beta_x + \beta_y)\varepsilon'_{t-\tau} \big) && \text{(15c)} \\
&\le \sum_{\tau=0}^{t} \gamma(\tau)\bar{\varepsilon} \le C\bar{\varepsilon} && \text{(15d)} \\
&\le \min\{R_x - R_x^*, R_y - R_y^*\}, && \text{(15e)}
\end{aligned}$$

where we use the triangle inequality in (15a); we use the contractive perturbation property in Property 4 in (15b); we use the induction assumption and Theorem 3 in (15c); we use (13) in (15d) and the definition of $\bar{\varepsilon}$ in (15e). By (14) and (15), we see that

$$\begin{aligned}
&\|x_{t+1}\| \le \|\tilde{x}_{t+1}\| + \|\tilde{x}_{t+1} - x_{t+1}\| \le R_x, \text{ and} \\
&\|y_{t+1}\| \le \|\tilde{y}_{t+1}\| + \|\tilde{y}_{t+1} - y_{t+1}\| \le R_y.
\end{aligned} \tag{16}$$

Note that we can construct the disturbance sequence $\{\zeta_t\}$ in Theorem 5 such that the dynamics

$$\begin{pmatrix} \tilde{x}_{t+1} \\ \tilde{y}_{t+1} \\ \theta_{t+1} \end{pmatrix} = q_t(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*) + \begin{pmatrix} 0 \\ 0 \\ \zeta_t \end{pmatrix}$$

produce the same policy parameter sequence $\{\theta_t\}$ as the dynamics

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \\ \theta_{t+1} \end{pmatrix} = q_t(x_t, y_t, \theta_t, \hat{a}_t).$$

Therefore, under this construction, we see that

$$\begin{aligned}
\|\zeta_t\| &\le \left\| \theta_{t+1} - q_t^\theta(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*) \right\| \\
&= \left\| q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) - q_t^\theta(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*) \right\| \\
&\le \left\| q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) - q_t^\theta(x_t, y_t, \theta_t, a_t^*) \right\| + \left\| q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*) \right\| && \text{(17a)}
\end{aligned}$$

$$\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + L_{\theta,x}\|x_t - \tilde{x}_t\| + L_{\theta,y}\|y_t - \tilde{y}_t\| \tag{17b}$$
$$\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + \sqrt{2}(L_{\theta,x} + L_{\theta,y})\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| \tag{17c}$$
$$\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + \sqrt{2}C\bar{\varepsilon}(L_{\theta,x} + L_{\theta,y}) \leq \bar{\zeta}, \tag{17d}$$

where we use the triangle inequality in (17a); we use Theorem 3 in (17b); we use the inequality

$$\|x_t - \tilde{x}_t\| + \|y_t - \tilde{y}_t\| \leq \sqrt{2}\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\|$$

in (17c) and (15d) in (17d). Thus, by Theorem 5, we see that $\|\theta_{t+1} - \theta_t\| \leq \epsilon_\theta$. Therefore, we have shown that

$$\|x_t\| \leq R_x, \|y_t\| \leq R_y, \text{ and } \|\theta_t - \theta_{t-1}\| \leq \epsilon_\theta$$

hold for all time step t by induction.

By (17c) and (15c), we also see that

$$\begin{aligned}\|\zeta_t\| &\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + \sqrt{2}(L_{\theta,x} + L_{\theta,y})\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| \\ &\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + \sqrt{2}(L_{\theta,x} + L_{\theta,y}) \sum_{\tau=0}^{t-1} \gamma(\tau)\big((\alpha_x + \alpha_y)\varepsilon_{t-1-\tau} + (\beta_x + \beta_y)\varepsilon'_{t-1-\tau}\big). \end{aligned} \tag{18}$$

Summing (18) over $t = 0, 1, \ldots, T-1$ gives that

$$\begin{aligned}&\sum_{t=0}^{T-1}\|\zeta_t\| \\ &\leq \Big(\alpha_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\alpha_x + \alpha_y)\Big)\sum_{t=0}^{T-1}\varepsilon_t + \Big(\beta_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\beta_x + \beta_y)\Big)\sum_{t=0}^{T-1}\varepsilon'_t.\end{aligned}$$

Summing (15c) over $t = 0, 1, \ldots, T-1$ gives that

$$\sum_{t=1}^{T}\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| \leq C\left((\alpha_x + \alpha_y)\sum_{t=0}^{T-1}\varepsilon_t + (\beta_x + \beta_y)\sum_{t=0}^{T-1}\varepsilon'_t\right).$$

Appendix E. Proof of Theorem 9

To simplify the notation, we let $\check{x}_{t+1} \coloneqq \mathbb{E}[x_{t+1} \mid \mathcal{G}_t]$ and let $\iota_{t+1} \coloneqq x_{t+1} - \check{x}_{t+1}$.

We first focus on one dimension i of the model mismatch. By Taylor's expansion, we see that

$$e_t(x_t, \hat{a}_t)_i = e_t(\check{x}_t, \hat{a}_t)_i + \nabla_x e_t(\check{x}_t, \hat{a}_t)_i \iota_t + \frac{1}{2}\iota_t^\top \nabla_x^2 e_t(\check{x}_t, \hat{a}_t)_i \iota_t. \tag{19}$$

Note that we have

$$\mathbb{E}[e_t(\check{x}_t, \hat{a}_t)_i \nabla_x e_t(\check{x}_t, \hat{a}_t)_i \iota_t \mid \mathcal{G}_{t-1}] = e_t(\check{x}_t, \hat{a}_t)_i \nabla_x e_t(\check{x}_t, \hat{a}_t)_i \mathbb{E}[\iota_t \mid \mathcal{G}_{t-1}] = 0. \tag{20}$$

Thus, we see that

$$\mathbb{E}\big[e_t(x_t, \hat{a}_t)_i^2 \mid \mathcal{G}_{t-1}\big] \geq e_t(\check{x}_t, \hat{a}_t)_i^2 + \nabla_x e_t(\check{x}_t, \hat{a}_t)_i^\top \mathrm{Cov}(\iota_t \mid \mathcal{G}_{t-1}) \nabla_x e_t(\check{x}_t, \hat{a}_t)_i$$

$$- \bar{\epsilon}^2\gamma_e|e_t(\check{x}_t,\hat{a}_t)_i| - \bar{\epsilon}^3\beta_e\gamma_e \tag{21a}$$
$$\geq e_t(\check{x}_t,\hat{a}_t)_i^2 + \underline{\sigma}\|\nabla_x e_t(\check{x}_t,\hat{a}_t)_i\|^2 - \bar{\epsilon}^2\gamma_e|e_t(\check{x}_t,\hat{a}_t)_i| - \bar{\epsilon}^3\beta_e\gamma_e. \tag{21b}$$

Summing over $t = 1,\dots,T$ and taking expectation on both sides gives that

$$R_0^\ell(T) \geq \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right] + \underline{\sigma}\mathbb{E}\left[\sum_{t=1}^T\|\nabla_x e_t(\check{x}_t,\hat{a}_t)\|_i^2\right] - \bar{\epsilon}^2\gamma_e\mathbb{E}\left[\sum_{t=1}^T|e_t(\check{x}_t,\hat{a}_t)_i|\right] - \bar{\epsilon}^3\beta_e\gamma_e T. \tag{22}$$

Now we show that $\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right] \leq \bar{\epsilon}^2 T$. For the sake of contradiction, suppose

$$\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right] > \bar{\epsilon}^2 T.$$

By (22), we see that

$$R_0^\ell(T) \geq \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right] - \bar{\epsilon}^2\gamma_e\mathbb{E}\left[\sum_{t=1}^T|e_t(\check{x}_t,\hat{a}_t)_i|\right] - \bar{\epsilon}^3\beta_e\gamma_e T$$
$$\geq \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right] - \bar{\epsilon}^2\gamma_e\sqrt{\mathbb{E}\left[\left(\sum_{t=1}^T|e_t(\check{x}_t,\hat{a}_t)_i|\right)^2\right]} - \bar{\epsilon}^3\beta_e\gamma_e T \tag{23a}$$
$$\geq \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right] - \bar{\epsilon}^2\gamma_e\sqrt{T\cdot\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right]} - \bar{\epsilon}^3\beta_e\gamma_e T \tag{23b}$$
$$= \sqrt{\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right]}\cdot\left(\sqrt{\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right]} - \bar{\epsilon}^2\gamma_e\sqrt{T}\right) - \bar{\epsilon}^3\beta_e\gamma_e T$$
$$> \frac{1}{4}\bar{\epsilon}^2 T, \tag{23c}$$

where we use Jensen's inequality in (23a); we use Cauchy-Schwarz inequality in (23b); we use the assumptions that $\bar{\epsilon}\gamma_e \leq \frac{1}{2}$ and $\bar{\epsilon}\beta_e\gamma_e \leq \frac{1}{4}$ in (23c). (23) contradicts with our assumption that $R_0^\ell(T) \leq \bar{\epsilon}^3 T$. Thus, we have shown that $\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right] \leq \bar{\epsilon}^2 T$.

Using the same argument as (23a) and (23b), we see that

$$\mathbb{E}\left[\sum_{t=1}^T|e_t(\check{x}_t,\hat{a}_t)_i|\right] \leq \sqrt{T\cdot\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t,\hat{a}_t)_i^2\right]} \leq \bar{\epsilon}T. \tag{24}$$

Substituting (24) into (22) gives that

$$\underline{\sigma}\mathbb{E}\left[\sum_{t=1}^T\|\nabla_x e_t(\check{x}_t,\hat{a}_t)_i\|^2\right] \leq R_0^\ell(T) + \bar{\epsilon}^2\gamma_e\mathbb{E}\left[\sum_{t=1}^T|e_t(\check{x}_t,\hat{a}_t)_i|\right] + \bar{\epsilon}^3\beta_e\gamma_e T$$

$$\leq (1 + \gamma_e + \beta_e\gamma_e)\bar{\epsilon}^3 T.$$

Therefore, we see that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla_x e_t(x_t, \hat{a}_t)_i\|^2\right] \\
&\leq 2\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla_x e_t(\check{x}_t, \hat{a}_t)\|^2\right] + 2\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla_x e_t(\check{x}_t, \hat{a}_t)_i - \nabla_x e_t(x_t, \hat{a}_t)_i\|^2\right] \\
&\leq \frac{2}{\underline{\sigma}}(1 + \gamma_e + \beta_e\gamma_e)\bar{\epsilon}^3 T + 2\gamma_e^2\bar{\epsilon}^2 T. \qquad (25)
\end{aligned}$$

Summing (25) over dimensions $i \in [1:k]$ finishes the proof of Theorem 9.

Appendix F. Proof of Lemma 20

When applied to the dynamical system (8) and the policy class (9), the joint dynamics induced by applying M-GAPS with exact model parameters $a^*_{0:T-1}$ are given by

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \qquad (26a)$$

$$y_{t+1} = q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g^*_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g^*_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}, \qquad (26b)$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right). \qquad (26c)$$

The joint dynamics induced by applying M-GAPS with inexact parameters $\hat{a}_{0:T-1}$ are given by

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t, \hat{a}_t) = \phi_t(x_t, \psi_t(x_t, \theta_t) + f_t(x_t, a_t^*) - f_t(x_t, \hat{a}_t)) + w_t, \qquad (27a)$$

$$y_{t+1} = q_t^y(x_t, y_t, \theta_t, \hat{a}_t) = \left.\frac{\partial g^*_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g^*_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}, \qquad (27b)$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial \hat{h}_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right), \qquad (27c)$$

where recall that we view $\hat{a}_{0:T-1}$ as external inputs as discussed in Section 3.1.

Since Lemma 20 consists three properties, we show them separately in Lemmas 24-26.

Lemma 24 *Consider the dynamical system*

$$\begin{aligned}
x_{t+1} &= q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \\
y_{t+1} &= q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g^*_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g^*_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t},
\end{aligned}$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right).$$

For any $x_t, y_t, \theta_t, \hat{a}_t$ that satisfies

$$\|x_t\| \le R_x, \|y_t\| \le R_y, \theta_t \in \Theta, \hat{a}_t \in \mathcal{A},$$

the following Lipschitzness conditions hold:

$$\begin{aligned}
\|q_t^x(x_t, y_t, \theta_t, a_t^*) - q_t^x(x_t, y_t, \theta_t, \hat{a}_t)\| &\le \alpha_x \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_x \varepsilon_t'(x_t, \hat{a}_t, a_t^*),\\
\|q_t^y(x_t, y_t, \theta_t, a_t^*) - q_t^y(x_t, y_t, \theta_t, \hat{a}_t)\| &\le \alpha_y \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_y \varepsilon_t'(x_t, \hat{a}_t, a_t^*),\\
\left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\le \alpha_\theta \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_\theta \varepsilon_t'(x_t, \hat{a}_t, a_t^*),
\end{aligned}$$

where

$$\begin{aligned}
&\alpha_x = \ell_{h,u} L_{\psi,\theta},\ \beta_x = \alpha_y = \beta_y = 0,\\
&\alpha_\theta = \eta(R_y(\ell_{h,x} + \ell_{h,u} L_{f,x} + \ell_{h,u} L_{\psi,x}) + \ell_{h,u} L_{\psi,\theta}),\ \beta_\theta = \eta R_y L_{h,u}.
\end{aligned}$$

Further, $q_t^\theta(x, y, \theta, a_t^*)$ is $(L_{\theta,x}, L_{\theta,y})$-Lipschitz in (x, y), where

$$\begin{aligned}
L_{\theta,x} &= \eta R_y((\ell_{h,x} + \ell_{h,u}(L_{f,x} + L_{\psi,x}))(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x}))\\
&\quad + \eta(\ell_{h,x} L_{\psi,\theta} + L_{h,u}\ell_{\psi,x} + \ell_{h,u} L_{\psi,\theta}(L_{f,x} + L_{\psi,x})),\\
L_{\theta,y} &= \eta(L_{h,x} + L_{h,u}(L_{f,x} + L_{\psi,x})).
\end{aligned}$$

The proof of Lemma 24 can be found in Appendix F.1.

Lemma 25 *Suppose the sequence $\theta_{0:T-1}$ is given and it satisfies the constraint that $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ for all time step t. Consider the dynamical system*

$$\begin{aligned}
x_{t+1} &= q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t,\\
y_{t+1} &= q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g_{t+1|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}.
\end{aligned}$$

We have that $\|x_t\| \le R_x^* < R_x, \|y_t\| \le R_y^* < R_y$ always hold if the system starts from $(x_\tau, y_\tau) = (0, 0)$. Here,

$$R_x^* = R_S, \text{ and } R_y^* = \frac{C_{L,g,\theta}}{1-\rho},$$

where recall that ρ is the decay factor defined in Assumption 19. Further, from any initial states $(x_\tau, y_\tau), (x_\tau', y_\tau')$ that satisfy $\|x_\tau\|, \|x_\tau'\| \le R_x$ and $\|y_\tau\|, \|y_\tau'\| \le R_y$, the trajectory satisfies

$$\left\|(x_t, y_t) - (x_t', y_t')\right\| \le \gamma(t-\tau) \cdot \left\|(x_\tau, y_\tau) - (x_\tau', y_\tau')\right\|,$$

where

$$\gamma(\tau) = \left(\bar{C} + C_{\ell,g,(x,x)} R_y + C_{\ell,g,(\theta,x)} \bar{C}\tau\right)\rho^\tau.$$

Note that γ satisfies

$$\sum_{\tau=0}^{\infty} \gamma(\tau) \le C, \text{ where } C = \frac{\bar{C} + C_{\ell,g,(x,x)} R_y}{1-\rho} + \frac{C_{\ell,g,(\theta,x)} \bar{C}}{(1-\rho)^2}.$$

The definitions of the coefficients $C_{L,g,\theta}, C_{\ell,g,(x,x)}, C_{\ell,g,(\theta,x)}$ can be found in Lemma 29 in Appendix J. And the proof of Lemma 25 can be found in Appendix F.2.

Lemma 26 *Consider the dynamical system*

$$\begin{aligned}
x_{t+1} &= q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \\
y_{t+1} &= q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g_{t+1|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}, \\
\theta_{t+1} &= q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right) + \zeta_t. \qquad (28)
\end{aligned}$$

Suppose the learning rate η satisfies $\eta < \min\left\{\frac{(1-\rho)\epsilon_\theta}{C_{L,h,\theta}}, \frac{1-\rho}{2C_{\ell,h,(\theta,\theta)}}\right\}$. When $\|\zeta_t\| \le \bar{\zeta} := \min\{1, \epsilon_\theta - \frac{C_{L,h,\theta}\eta}{1-\rho}\}$ holds for all t, the resulting $\{\theta_t\}$ satisfies the slowly-time-varying constraint $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$. Further, the trajectory $\{\theta_t\}$ achieves the local regret guarantee

$$R_\eta^L(T, \{\|\zeta_t\|\}_{0\le t\le T-1}) \le \frac{2}{\eta}(F_0(\theta_0) + S_0) + \frac{2}{1-\rho}(C_{L,h,\theta}S_1 + C_{\ell,h,(\theta,\theta)}\eta S_2), \text{ where}$$

$$\begin{aligned}
S_0 &:= \frac{2\bar{C}L_h(1 + L_{\psi,x} + L_{f,x})(1 + L_{\phi,u})}{(1-\rho)^2\rho} \cdot (V_{sys} + V_w) \\
&\quad + \frac{2\bar{C}L_h(1 + L_{\psi,x} + L_{f,x})}{1-\rho} \cdot \left(2\bar{C}\bar{R}_C + 2R_S\right), \\
S_1 &:= \left(\frac{1}{\eta} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right)\sum_{t=0}^{T-1}\|\zeta_t\| + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta T, \\
S_2 &:= \left(1 + \frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2 + \hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_2 + \hat{C}_4}{(1-\rho)^3}\right) \cdot \\
&\quad \left[\left(\frac{1}{\eta^2} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right)\sum_{t=0}^{T-1}\|\zeta_t\|^2 + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta^2 T\right].
\end{aligned}$$

Here, the variation intensity is defined as

$$\begin{aligned}
V_{sys} &= \sum_{t=1}^{T-1}\Big(\sup_{x\in\mathcal{X},u\in\mathcal{U}}\|\phi_t(x,u) - \phi_{t-1}(x,u)\| + \sup_{x\in\mathcal{X},\theta\in\Theta}\|\psi_t(x,\theta) - \psi_{t-1}(x,\theta)\| \\
&\quad + \sup_{x\in\mathcal{X},u\in\mathcal{U},\theta\in\Theta}|h_t(x,u,\theta) - h_{t-1}(x,u,\theta)|\Big), \text{ and} \\
V_w &= \sum_{t=1}^{T-1}\|w_t - w_{t-1}\|,
\end{aligned}$$

The bound can be simplified to

$$R_\eta^L(T, \{\|\zeta_t\|\}_{0\le t\le T-1}) = O\left(\frac{1}{\eta}(1 + V_{sys} + V_w) + \eta T + \eta^3 T + \frac{1}{\eta}\sum_{t=0}^{T-1}\|\zeta_t\|\right),$$

where the $O(\cdot)$ notation hides dependence on $\frac{1}{1-\rho}, R_x, R_y, \bar{C}$, and the Lipchitzness/smoothness coefficients defined in Assumption 18.

The definition of the coefficient $C_{L,h,\theta}$ can be found in Corollary 30 in Appendix J. The proof of Lemma 26 can be found in Appendix F.3.

F.1. Proof of Lemma 24

By Assumption 18, we see that

$$\|q_t^x(x_t, y_t, \theta_t, \hat{a}_t) - q_t^x(x_t, y_t, \theta_t, a_t^*)\| \le L_{\phi,u}\|f_t(x_t, a_t^*) - f_t(x_t, \hat{a}_t)\| = L_{\phi,u}\varepsilon_t. \tag{29}$$

We also have that

$$q_t^y(x_t, y_t, \theta_t, \hat{a}_t) = q_t^y(x_t, y_t, \theta_t, a_t^*). \tag{30}$$

Note that

$$h_{t|t}^*(x_t, \theta_t) = h_t(x_t, u_t^1, \theta_t), \text{ where } u_t^1 = -f_t(x_t, a_t^*) + \psi_t(x_t, \theta_t),$$
$$\hat{h}_{t|t}(x_t, \theta_t) = h_t(x_t, u_t^2, \theta_t), \text{ where } u_t^2 = -f_t(x_t, \hat{a}_t) + \psi_t(x_t, \theta_t).$$

Therefore, we see that

$$\begin{aligned}
&\left\| \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \hat{h}_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \right\| \\
\le &\left\| \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t,u_t^1,\theta_t} - \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t,u_t^2,\theta_t} \right\| + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t^1,\theta_t} \cdot \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t,a_t^*} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t^2,\theta_t} \cdot \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t,\hat{a}_t} \right\| \\
&+ \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t^1,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t^2,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t,\theta_t} \right\| && \text{(31a)} \\
\le &\ \ell_{h,x}\varepsilon_t + (\ell_{h,u}L_{f,x}\varepsilon_t + L_{h,u}\varepsilon_t') + \ell_{h,u}L_{\psi,x}\varepsilon_t && \text{(31b)} \\
= &\ (\ell_{h,x} + \ell_{h,u}L_{f,x} + \ell_{h,u}L_{\psi,x})\varepsilon_t + L_{h,u}\varepsilon_t',
\end{aligned}$$

where we use the chain rule and the triangle inequality in (31a); we use Assumption 18 in (31b). Similarly, we also see that

$$\begin{aligned}
\left\| \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t} \right\| &= \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t^1,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t^2,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial \theta_t}\right|_{x_t,\theta_t} \right\| && \text{(32a)} \\
&\le \ell_{h,u}L_{\psi,\theta}\varepsilon_t, && \text{(32b)}
\end{aligned}$$

where we use the chain rule in (32a) and Assumption 18 in (32b).

For q_t^θ, we see that

$$\begin{aligned}
&\left\|q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) - q_t^\theta(x_t, y_t, \theta_t, a_t^*)\right\| \\
&\leq \eta \left\| \left(\left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \hat{h}_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \right) \cdot y_t + \left(\left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t} \right) \right\| && \text{(33a)} \\
&\leq \eta(R_y(\ell_{h,x} + \ell_{h,u}L_{f,x} + \ell_{h,u}L_{\psi,x}) + \ell_{h,u}L_{\psi,\theta})\varepsilon_t + \eta R_y L_{h,u}\varepsilon_t', && \text{(33b)}
\end{aligned}$$

where we use the property that projection onto Θ is contractive in (33a); we use (31) and (32) in (33b). We also see that

$$\begin{aligned}
&\left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x_t', y_t', \theta_t, a_t^*)\right\| \\
&\leq \eta \left\| \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t - \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t',\theta_t} \cdot y_t' + \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t',\theta_t} \right\| && \text{(34a)} \\
&\leq \eta \left\| \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t',\theta_t} \right\| \cdot \|y_t\| + \eta \left\| \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t',\theta_t} \right\| \cdot \|y_t - y_t'\| \\
&\quad + \eta \left\| \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t',\theta_t} \right\|, && \text{(34b)}
\end{aligned}$$

where we use the property that projection onto Θ is contractive in (34a), and apply the triangle inequality in (34b). Note that

$$\begin{aligned}
h_{t|t}^*(x_t, \theta_t) &= h_t(x_t, u_t, \theta_t), \text{ where } u_t = -f_t(x_t, a_t^*) + \psi_t(x_t, \theta_t), \\
h_{t|t}^*(x_t', \theta_t) &= h_t(x_t', u_t', \theta_t), \text{ where } u_t' = -f_t(x_t', a_t^*) + \psi_t(x_t', \theta_t).
\end{aligned}$$

Therefore, we see that

$$\begin{aligned}
&\left\| \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t',\theta_t} \right\| \\
&\leq \left\| \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t,u_t,\theta_t} - \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t',u_t',\theta_t} \right\| + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} \cdot \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t,a_t^*} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \cdot \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t',a_t^*} \right\| \\
&\quad + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t',\theta_t} \right\| && \text{(35a)} \\
&\leq \left\| \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t,u_t,\theta_t} - \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t',u_t',\theta_t} \right\| + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t,a_t^*} \right\| \\
&\quad + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t,a_t^*} - \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t',a_t^*} \right\| + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t,\theta_t} \right\|
\end{aligned}$$

$$+ \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t',\theta_t} \right\| \tag{35b}$$

$$\leq \ell_{h,x}\|x_t - x_t'\| + \ell_{h,u}\|u_t - u_t'\| + L_{f,x}\big(\ell_{h,x}\|x_t - x_t'\| + \ell_{h,u}\|u_t - u_t'\|\big)$$
$$+ L_{h,u}\ell_{f,x}\|x_t - x_t'\| + L_{\psi,x}\big(\ell_{h,x}\|x_t - x_t'\| + \ell_{h,u}\|u_t - u_t'\|\big) + L_{h,u}\ell_{\psi,x}\|x_t - x_t'\| \tag{35c}$$

$$= (\ell_{h,x}(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x}))\|x_t - x_t'\| + \ell_{h,u}(1 + L_{f,x} + L_{\psi,x})\|u_t - u_t'\|,$$

$$\leq ((\ell_{h,x} + \ell_{h,u}(L_{f,x} + L_{\psi,x}))(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x}))\|x_t - x_t'\|, \tag{35d}$$

where we use the triangle inequality in (35a) and (35b); we use Assumption 18 in (35c) and (35d). Similarly, we also see that

$$\left\| \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t',\theta_t} \right\|$$

$$= \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial \theta_t}\right|_{x_t',\theta_t} \right\| \tag{36a}$$

$$\leq \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \left.\frac{\partial \psi_t}{\partial \theta_t}\right|_{x_t,\theta_t} \right\| + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \left.\frac{\partial \psi_t}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \psi_t}{\partial \theta_t}\right|_{x_t',\theta_t} \right\| \tag{36b}$$

$$\leq (\ell_{h,x}L_{\psi,\theta} + L_{h,u}\ell_{\psi,x})\|x_t - x_t'\| + \ell_{h,u}L_{\psi,\theta}\|u_t - u_t'\|, \tag{36c}$$

$$\leq (\ell_{h,x}L_{\psi,\theta} + L_{h,u}\ell_{\psi,x} + \ell_{h,u}L_{\psi,\theta}(L_{f,x} + L_{\psi,x}))\|x_t - x_t'\|, \tag{36d}$$

where we use the chain rule in (36a); we use the triangle inequality in (36b); we use Assumption 18 in (36c). Substituting (35) and (36) into (34) gives that

$$\begin{aligned}
&\left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x_t', y_t', \theta_t, a_t^*)\right\| \\
&\leq \eta R_y((\ell_{h,x} + \ell_{h,u}(L_{f,x} + L_{\psi,x}))(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x}))\|x_t - x_t'\| \\
&\quad + \eta(L_{h,x} + L_{h,u}(L_{f,x} + L_{\psi,x}))\|y_t - y_t'\| \\
&\quad + \eta(\ell_{h,x}L_{\psi,\theta} + L_{h,u}\ell_{\psi,x} + \ell_{h,u}L_{\psi,\theta}(L_{f,x} + L_{\psi,x}))\|x_t - x_t'\| \\
&\leq L_{\theta,x}\|x_t - x_t'\| + L_{\theta,y}\|y_t - y_t'\|.
\end{aligned} \tag{37}$$

F.2. Proof of Lemma 25

Consider two trajectories $\{x_{t_1:t_2}, y_{t_1:t_2}\}$ and $\{x'_{t_1:t_2}, y'_{t_1:t_2}\}$ given by

$$\begin{aligned}
x_{\tau+1} &= \phi_\tau(x_\tau, \psi_t(x_\tau, \theta_\tau)) + w_\tau, \\
y_{\tau+1} &= \left.\frac{\partial g_{\tau+1|\tau}^*}{\partial x_\tau}\right|_{x_\tau,\theta_\tau} \cdot y_\tau + \left.\frac{\partial g_{\tau+1|\tau}^*}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau},
\end{aligned}$$

and

$$x'_{\tau+1} = \phi_\tau(x'_\tau, \psi_t(x'_\tau, \theta_\tau)) + w_\tau,$$

$$y'_{\tau+1} = \left.\frac{\partial g^*_{\tau+1|\tau}}{\partial x_\tau}\right|_{x'_\tau,\theta_\tau} \cdot y'_\tau + \left.\frac{\partial g^*_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_\tau},$$

where $\tau = t_1, t_1+1, \ldots, t_2$. Note that by Assumption 19, we have that $\|x_{t_2}\| \le R_S$ and for any x_{t_1}, x'_{t_1} whose norms are upper bounded by R_C

$$\left\|x_{t_2} - x'_{t_2}\right\| \le \bar{C}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\|. \tag{38}$$

where ρ is the decay factor of the contractive perturbation property defined in Assumption 19. For the y sequence, note that y_{t_2} and y'_{t_2} can be expressed equivalently as

$$y_{t_2} = \left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x_{t_1},\theta_{t_1:t_2-1}} \cdot y_{t_1} + \sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}, \tag{39a}$$

$$y'_{t_2} = \left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}} \cdot y'_{t_1} + \sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_{\tau:t_2-1}}. \tag{39b}$$

By Lemma 29, we see that if $y_{t_1} = 0$, then

$$\|y_{t_2}\| = \left\|\sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}\right\| \le \sum_{\tau=t_1}^{t_2-1}\left\|\left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}\right\| \le \sum_{\tau=t_1}^{t_2-1} C_{L,g,\theta}\rho^{t_2-\tau} = \frac{C_{L,g,\theta}}{1-\rho}. \tag{40}$$

We also see that

$$\begin{aligned}
&\left\|y_{t_2} - y'_{t_2}\right\| \\
&= \left\|\left(\left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x_{t_1},\theta_{t_1:t_2-1}} - \left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}}\right) \cdot y_{t_1}\right\| + \left\|\left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}} \cdot (y_{t_1} - y'_{t_1})\right\| \\
&\quad + \sum_{\tau=t_1}^{t_2-1}\left\|\left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}} - \left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_{\tau:t_2-1}}\right\| && \text{(41a)} \\
&\le C_{\ell,g,(x,x)}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| \cdot R_y + C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\| \\
&\quad + C_{\ell,g,(\theta,x)}\sum_{\tau=t_1}^{t_2-1}\rho^{t_2-\tau}\left\|x_\tau - x'_\tau\right\| && \text{(41b)} \\
&\le C_{\ell,g,(x,x)}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| \cdot R_y + C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\| \\
&\quad + C_{\ell,g,(\theta,x)}\sum_{\tau=t_1}^{t_2-1}\rho^{t_2-\tau} \cdot \bar{C}\rho^{\tau-t_1}\left\|x_{t_1} - x'_{t_1}\right\| && \text{(41c)} \\
&\le \left(C_{\ell,g,(x,x)}R_y + C_{\ell,g,(\theta,x)}\bar{C}(t_2-t_1)\right)\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| + C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\|. && \text{(41d)}
\end{aligned}$$

Therefore, we see that

$$\left\|(x_{t_2}, y_{t_2}) - (x'_{t_2}, y'_{t_2})\right\|$$

$$\leq \left\|x_{t_2} - x'_{t_2}\right\| + \left\|y_{t_2} - y'_{t_2}\right\| \tag{42a}$$
$$\leq \bar{C}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| + \left(C_{\ell,g,(x,x)}R_y + C_{\ell,g,(\theta,x)}\bar{C}(t_2-t_1)\right)\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| + \bar{C}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\| \tag{42b}$$
$$\leq \gamma(t_2-t_1)\left\|(x_{t_1}, y_{t_1}) - (x'_{t_1}, y'_{t_1})\right\|, \tag{42c}$$

where we use the triangle inequality in (42a); we use (38) and (41) and $\bar{C} = C_{L,g,x}$ in (42b); we use the inequality that

$$\left\|x_{t_1} - x'_{t_1}\right\| + \left\|y_{t_1} - y'_{t_1}\right\| \leq \sqrt{2}\left\|(x_{t_1}, y_{t_1}) - (x'_{t_1}, y'_{t_1})\right\|$$

and the definition of $\gamma(\cdot)$ in (42c).

F.3. Proof of Lemma 26

We compare the dynamical system (28) with the Ideal OGD update rule:

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta\nabla F_t(\theta_t)). \tag{43}$$

Note that the update on θ_t that the dynamical system (28) performs can be written equivalently as

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta G_t) + \zeta_t, \tag{44}$$

where

$$G_t := \sum_{\tau=0}^{t} \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}}\bigg|_{x_0,\theta_{0:t}}. \tag{45}$$

By Theorem 32, we know that

$$\|G_t - \nabla F_t(\theta_t)\| \leq \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta + \sum_{\tau=0}^{t-1}\left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5(t-\tau)\right)\rho^{t-\tau}\|\zeta_\tau\|,$$

where the constants $\hat{C}_{0:5}$ are given in Theorem 32. Let θ_{t+1} be the actual next policy parameter (following the update rule (44)). By Lemma 28, we see that

$$\begin{aligned}\|\theta_{t+1} - \Pi_\Theta(\theta_t - \eta\nabla F_t(\theta_t))\| &\leq \eta\|G_t - \nabla F_t(\theta_t)\| + \|\zeta_t\| \\ &\leq \|\zeta_t\| + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta^2 \\ &\quad + \eta\sum_{\tau=0}^{t-1}\left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5(t-\tau)\right)\rho^{t-\tau}\|\zeta_\tau\|.\end{aligned}$$

Then, we can apply Theorem 27 to obtain that

$$\sum_{t=0}^{T-1}\|\nabla_{\Theta,\eta}F_t(\theta_t)\|^2 \leq \frac{1}{\eta(1-\eta\ell_F)}\left(F_0(\theta_0) + \sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1})\right) + \frac{L_F S_1 + \ell_F\eta S_2}{1-\eta\ell_F}, \tag{46}$$

where dist_S is a metric that measures the distance between two surrogate cost functions (see Theorem 27 for definition), and S_1 and S_2 are given by

$$S_1 \coloneqq \left(\frac{1}{\eta} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right)\sum_{t=0}^{T-1}\|\zeta_t\| + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta T,$$

$$S_2 \coloneqq \left(1 + \frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2 + \hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_2 + \hat{C}_4}{(1-\rho)^3}\right)\cdot \left[\left(\frac{1}{\eta^2} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right)\sum_{t=0}^{T-1}\|\zeta_t\|^2 + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta^2 T\right].$$

By applying Lemma F.4 in Lin et al. (2023), we can bound the total variational intensity on the surrogate costs by

$$\sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1}) \le \frac{2\bar{C}L_h(1 + L_{\psi,x} + L_{f,x})(1 + L_{\phi,u})}{(1-\rho)^2\rho}\cdot(V_{sys} + V_w) + \frac{2\bar{C}L_h(1 + L_{\psi,x} + L_{f,x})}{1-\rho}\cdot\left(2\bar{C}\bar{R}_C + 2R_S\right).$$

Substituting the above inequality and $L_F = \frac{C_{L,f,\theta}}{1-\rho}, \ell_F = \frac{C_{\ell,h,(\theta,\theta)}}{1-\rho}$ into (46) finishes the proof.

Appendix G. Proof of Lemma 23

By Assumptions 21 and 22, we see that for any $a \in \mathcal{A}$,

$$\tilde{\ell}_t(x_t, a, \tilde{f}_t) \le \left(f_t(x_t, a) - \tilde{f}_t\right)^2 = \left((f_t(x_t, a) - f_t(x_t, a_t^*)) - (\tilde{f}_t - f_t(x_t, a_t^*))\right)^2 \le (C_f + e_f)^2.$$

We also see that

$$\left\|\nabla_a \tilde{\ell}_t(x_t, a, \tilde{f}_t)\right\| \le 2\|\nabla_a f_t(x_t, a)\| \cdot \left|f_t(x_t, a) - \tilde{f}_t\right| \le 2D_f'(C_f + e_f).$$

By Theorem 10.1 in Hazan (2016), we know that Algorithm 3 with the learning rate $\iota = \frac{C_f + e_f}{D_f'}\cdot\sqrt{\frac{C_p}{T}}$ always achieves the guarantee that

$$\sum_{t=0}^{T-1}\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \sum_{t=0}^{T-1}\tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t) \le R_0^{\ell}(T) \coloneqq 2\sqrt{3}(C_f + e_f)^3 D_f'\sqrt{C_p T}. \tag{47}$$

Let $v_t \coloneqq \tilde{f}_t - f_t(x_t, a_t^*)$. We see that

$$\begin{aligned}&\mathbb{E}\left[\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t) \mid \mathcal{F}_t\right]\\ &= \mathbb{E}\left[\|(f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*)) - v_t\|^2 - \|v_t\|^2 \mid \mathcal{F}_t\right]\end{aligned}$$

$$= \|f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*)\|^2 - 2(f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*))^\top \mathbb{E}[v_t \mid \mathcal{F}_t]$$
$$= \|f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*)\|^2 = \varepsilon_t^2.$$

Therefore, we obtain that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \sum_{t=0}^{T-1} \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t)\right] &= \sum_{t=0}^{T-1} \mathbb{E}\left[\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t)\right] \\ &= \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t) \mid \mathcal{F}_t\right]\right] \\ &= \sum_{t=0}^{T-1} \mathbb{E}[\varepsilon_t^2] = \mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t^2\right].\end{aligned}$$

Combining this with (47) gives that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t^2\right] \le 2\sqrt{3}(C_f + e_f)^3 D_f' \sqrt{C_p T}.$$

Then, we can apply Theorem 9 to conclude that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} (\varepsilon_t')^2\right] \le \frac{2m}{c}(1 + \gamma + \beta\gamma)\bar{\epsilon}T + 2m\gamma^2\bar{\epsilon}^2 T.$$

Appendix H. Proof of Theorem 12

By Lemma 23 and Theorem 9, we know that the expected total prediction errors achieved by the gradient estimator satisfy that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t^2\right] &\le 2\sqrt{3}(C_f + e_f)^3 D_f' \sqrt{C_p T}, \text{ and} \\ \mathbb{E}\left[\sum_{t=0}^{T-1} (\varepsilon_t')^2\right] &\le \frac{2m}{c}(1 + \gamma + \beta\gamma)\bar{\epsilon}T + 2m\gamma^2\bar{\epsilon}^2 T. \end{aligned} \tag{48}$$

By Hölder's inequality, we see that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t\right] &\le \sqrt[4]{12}(C_f + e_f)^{\frac{3}{2}} (D_f')^{\frac{1}{2}} C_p^{\frac{1}{4}} T^{\frac{3}{4}}, \text{ and} \\ \mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t'\right] &\le \sqrt{\frac{2}{c}(1 + \gamma + \beta\gamma) + 2\gamma^2\bar{\epsilon}} \cdot \sqrt{m\bar{\epsilon}} \cdot T. \end{aligned} \tag{49}$$

By Lemma 20, we know that trajectory $\tilde{\xi}$ achieves the local regret

$$R_L(T, \{\|\zeta_t\|\}_{0 \le t \le T-1}) = O\left(\frac{1}{\eta}(1 + V_{\text{sys}} + V_w) + \eta T + \eta^3 T + \frac{1}{\eta}\sum_{t=1}^{T-1} \|\zeta_t\|\right), \tag{50}$$

By Theorem 6 and Lemma 24, we know that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\zeta_t\|\right] &\le \Big(\alpha_\theta + \sqrt{2}C(L_{\theta,x}+L_{\theta,y})(\alpha_x+\alpha_y)\Big)\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon_t\right] \\ &\quad + \Big(\beta_\theta + \sqrt{2}C(L_{\theta,x}+L_{\theta,y})(\beta_x+\beta_y)\Big)\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon'_t\right] \\ &\le C_0\eta\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon_t\right] + R_y L_{h,u}\cdot\eta\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon'_t\right], \end{aligned} \tag{51}$$

where $C = \frac{\bar{C}+C_{\ell,g,(x,x)}R_y}{1-\rho} + \frac{C_{\ell,g,(\theta,x)}\bar{C}}{(1-\rho)^2}$ by Lemma 25 and

$$\begin{aligned}C_0 &= R_y(\ell_{h,x}+\ell_{h,u}L_{f,x}+\ell_{h,u}L_{\psi,x}) + \ell_{h,u}L_{\psi,\theta} \\ &\quad + \sqrt{2}C\ell_{h,u}L_{\psi,\theta}\Bigg(R_y((\ell_{h,x}+\ell_{h,u}(L_{f,x}+L_{\psi,x}))(1+L_{f,x}+L_{\psi,x}) + L_{h,u}(\ell_{f,x}+\ell_{\psi,x})) \\ &\quad + \ell_{h,x}L_{\psi,\theta} + L_{h,u}\ell_{\psi,x} + \ell_{h,u}L_{\psi,\theta}(L_{f,x}+L_{\psi,x}) + L_{h,x} + L_{h,u}(L_{f,x}+L_{\psi,x})\Bigg).\end{aligned}$$

Substituting (51) into (50) and applying (48) and (49) give that

$$R_L(T,\{\|\zeta_t\|\}_{0\le t\le T-1}) = O\left(\frac{1}{\eta}(1+V_{sys}+\bar{\epsilon}\cdot T)+\eta T+(\sqrt{m\bar{\epsilon}}+m\bar{\epsilon})\cdot T\right).$$

Further, by the last statement of Theorem 6, we obtain that

$$\mathbb{E}\left[\sum_{t=0}^{T-1}(\|x_t-\tilde{x}_t\|+\|y_t-\tilde{y}_t\|)\right] = O\Big(T^{3/4}+\sqrt{m\bar{\epsilon}}\cdot T\Big).$$

Appendix I. Local Regret of Online Gradient Descent

Theorem 27 *Consider the parameter sequence $\{\theta_t\}$ that satisfies*

$$\|\theta_{t+1}-(\theta_t-\eta\nabla_{\Theta,\eta}F_t(\theta_t))\| \le \eta\beta_t, \text{ for all } t\ge 0.$$

Suppose at every time t, F_t is ℓ_F-smooth and L_F-Lipschitz in Θ. If the learning rate $\eta \le \frac{1}{\ell_F}$, then the local regret $\sum_{t=0}^{T-1}\|\nabla_{\Theta,\eta}F_t(\theta_t)\|^2$ is upper bounded by

$$\frac{1}{\eta(1-\eta\ell_F)}\left(F_0(\theta_0)+\sum_{t=1}^{T-1}\mathsf{dist}_s(F_t,F_{t-1})\right)+\frac{L_F\sum_{t=0}^{T-1}\beta_t+\ell_F\eta\sum_{t=0}^{T-1}\beta_t^2}{1-\eta\ell_F},$$

where $\mathsf{dist}_s(F,F') := \sup_{\theta\in\Theta}|F(\theta)-F'(\theta)|$.

Next, we state a property of projection onto the compact convex set $\Theta\in\mathbb{R}^d$ in Lemma 28. This is a classic result in convex optimization (see, for example, Theorem 1.2.1 in Schneider (2014)).

Lemma 28 *Let q and q' be arbitrary points in $\mathbb{R}^d$. Let $p = \Pi_\Theta(q)$ and $p' = \Pi_\Theta(q')$. Then, the following inequality holds:*

$$\|p - p'\| \leq \|q - q'\|.$$

Now we come back to the proof of Theorem 27.

Define the quantity

$$\epsilon_t := \frac{1}{\eta}(\theta_{t+1} - (\theta_t - \eta \nabla_{\Theta,\eta} F_t(\theta_t))).$$

We see that

$$\theta_{t+1} - \theta_t = -\eta \nabla_{\Theta,\eta} F_t(\theta_t) + \eta \epsilon_t. \tag{52}$$

By the smoothness of $F_t(\cdot)$, we see that

$$\begin{aligned}
F_t(\theta_{t+1}) &\leq F_t(\theta_t) + \langle \nabla F_t(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\ell_F}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) - \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\|\nabla_{\Theta,\eta} F_t(\theta_t) - \epsilon_t\|^2 \qquad (53a) \\
&= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) \rangle + \frac{\ell_F \eta^2}{2}\|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2 \\
&\quad + \eta \langle \nabla F_t(\theta_t), \epsilon_t \rangle - \ell_F \eta^2 \langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\|\epsilon_t\|^2, \qquad (53b)
\end{aligned}$$

where we use (52) in (53a). Recall that Θ is a closed convex subset of $\mathbb{R}^d$. Since $\theta_t - \eta \nabla_{\Theta,\eta} F_t(\theta_t)$ is the projection of $\theta_t - \eta \nabla F_t(\theta_t)$ onto Θ and $\theta_t \in \Theta$, we have

$$\langle (\theta_t - \eta \nabla F_t(\theta_t)) - (\theta_t - \eta \nabla_{\Theta,\eta} F_t(\theta_t)), \theta_t - (\theta_t - \eta \nabla_{\Theta,\eta} F_t(\theta_t)) \rangle \leq 0.$$

Rearranging terms gives that

$$\langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) \rangle \geq \|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2.$$

Substituting this inequality into (53) gives that

$$\begin{aligned}
F_t(\theta_{t+1}) &\leq F_t(\theta_t) - \eta \|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2 + \frac{\ell_F \eta^2}{2}\|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2 \\
&\quad + \eta \langle \nabla F_t(\theta_t), \epsilon_t \rangle - \ell_F \eta^2 \langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\|\epsilon_t\|^2 \\
&\leq F_t(\theta_t) - \eta(1 - \ell_F \eta)\|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2 + \eta \|\nabla F_t(\theta_t)\| \cdot \|\epsilon_t\| \\
&\quad - \frac{\ell_F \eta^2}{2}\|\nabla_{\Theta,\eta} F_t(\theta_t) + \epsilon_t\|^2 + \ell_F \eta^2 \|\epsilon_t\|^2 \qquad (54a) \\
&\leq F_t(\theta_t) - \eta(1 - \ell_F \eta)\|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2 + \eta L_F \beta_t + \ell_F \eta^2 \beta_t^2, \qquad (54b)
\end{aligned}$$

where we rearrange the terms and use the Cauchy-Schwarz inequality in (54a); In (54b), we use the assumption $\|\epsilon_t\| \leq \beta_t$. Summing (54) over $t = 0, 1, \ldots, T-1$ gives that

$$\eta(1 - \ell_F \eta) \sum_{t=0}^{T-1} \|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2$$

$$\leq \sum_{t=0}^{T-1}(F_t(\theta_t) - F_t(\theta_{t+1})) + \eta L_F \sum_{t=0}^{T-1}\beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1}\beta_t^2$$

$$\leq F_0(\theta_0) + \sum_{t=1}^{T-1}(F_t(\theta_t) - F_{t-1}(\theta_t)) + \sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1}) + \eta L_F \sum_{t=0}^{T-1}\beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1}\beta_t^2 \quad (55a)$$

$$\leq F_0(\theta_0) + \sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1}) + \eta L_F \sum_{t=0}^{T-1}\beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1}\beta_t^2, \quad (55b)$$

where we rearrange the terms and use $F_{T-1}(\theta_T) \geq 0$ in (55a); we use the definition of $\mathsf{dist}_s(\cdot,\cdot)$ in (55b).

Appendix J. Useful Lemmas

In this section, we summarize some useful existing results in Lin et al. (2023) that can help us in the proof of M-GAPS (Algorithm 2). We can build our proof upon some results shown in Lin et al. (2023) because of the similarity between our M-GAPS algorithm and the GAPS algorithm proposed by Lin et al. (2023) when applied to known dynamical systems: Both algorithms are designed to efficiently approximate the gradient $\nabla F_t(\theta_t)$ of the surrogate cost. Note that $\nabla F_t(\theta_t)$ can be expressed as

$$\nabla F_t(\theta_t) = \sum_{\tau=0}^{t} \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}}\bigg|_{x_0,(\theta_t)_{\times(t+1)}}.$$

M-GAPS adopts the approximation G_t that is given by

$$G_t = \sum_{\tau=0}^{t} \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}}\bigg|_{x_0,\theta_{0:t}}, \quad (56)$$

which simplifies $\nabla F_t(\theta_t)$ by replacing the imaginary trajectory achieved by using policy parameter θ_t repeatedly with the actual trajectory. The approximator of GAPS, which we denote as G'_t, takes an additional step to approximate G_t by truncating the summation from time 0 to t to at most B time steps, i.e.,

$$G'_t = \sum_{\tau=\max\{0,t-B\}}^{t} \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}}\bigg|_{x_0,\theta_{0:t}},$$

where B is the buffer length parameter decided by the algorithm. Intuitively, the approximation G_t adopted by M-GAPS is closer to $\nabla F_t(\theta_t)$, which allows us to show the same guarantees for M-GAPS as GAPS when the true dynamics are known.

In this section, we translate some results from Lin et al. (2023) into the settings of matched-disturbance dynamics application discussed in Section 4. Lemma 29 is Lemma D.3 in Lin et al. (2023). We changed the condition $x_\tau, x'_\tau \in B_n(0, R_S + C\|x_0\|)$ to $x_\tau, x'_\tau \in B_n(0, \bar{R}_C)$, where $\bar{R}_C$ can be any positive number that satisfies $\bar{R}_C < R_C$ and $C\bar{R}_C + R_S \leq R_x$. This minor change will not affect the proof provided in Lin et al. (2023).

Lemma 29 (Lipschitzness/Smoothness of the Multi-Step Dynamics) *Suppose Assumptions 18 and 19 hold. Given two time steps $t > \tau$, for any $x_\tau, x'_\tau \in B_n(0, \bar{R}_C)$ and $\theta_\tau, \theta'_\tau \in \Theta$, $\theta_{\tau+1:t-1} \in$*

$S_\varepsilon(\tau+1:t-1)$, if $x'_{\tau+1} \coloneqq g_{\tau+1|\tau}(x'_\tau, \theta'_\tau)$ is also in $B_n(0, \bar{R}_C)$, the multi-step dynamical function $g_{t|\tau}$ satisfies that

$$\left\| \left.\frac{\partial g^*_{t|\tau}}{\partial x_\tau}\right|_{x_\tau,\theta_{\tau:t-1}} \right\| \le C_{L,g,x}\rho^{t-\tau}, \quad \left\| \left.\frac{\partial g^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t-1}} \right\| \le C_{L,g,\theta}\rho^{t-\tau}, \forall \theta_{\tau:t-1} \in S_\varepsilon(\tau:t-1),$$

$$\left\| \left.\frac{\partial g^*_{t|\tau}}{\partial x_\tau}\right|_{x_\tau,\theta_{\tau:t-1}} - \left.\frac{\partial g^*_{t|\tau}}{\partial x_\tau}\right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \le C_{\ell,g,(x,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,g,(x,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|,$$

$$\left\| \left.\frac{\partial g^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t-1}} - \left.\frac{\partial g^*_{t|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \le C_{\ell,g,(\theta,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,g,(\theta,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|,$$

where $C_{L,g,x} = \bar{C}, C_{L,g,\theta} = \frac{\bar{C}L_{\phi,u}L_{\psi,\theta}}{\rho}$, and

$$\begin{aligned}
C_{\ell,g,(x,x)} &= ((1+L_{\psi,x})(\ell_{\phi,x}+\ell_{\phi,u}L_{\psi,x}) + L_{\phi,x}\ell_{\psi,x})C^3\rho^{-1}(1-\rho)^{-1},\\
C_{\ell,g,(x,\theta)} &= ((1+L_{\psi,x})(\ell_{\phi,x}+\ell_{\phi,u}L_{\psi,x}) + L_{\phi,x}\ell_{\psi,x})C^3L_{\phi,u}L_{\psi,\theta}\rho^{-1}(1-\rho)^{-1}\\
&\quad + ((1+L_{\psi,x})\ell_{\phi,u}L_{\psi,\theta} + L_{\phi,u}\ell_{\psi,\theta})C\rho^{-1}(1-\rho)^{-1},\\
C_{\ell,g,(\theta,x)} &= ((1+L_{\psi,x})(\ell_{\phi,x}+\ell_{\phi,u}L_{\psi,x}) + L_{\phi,x}\ell_{\psi,x})(L_{\phi,x}+L_{\phi,u}L_{\psi,x})\cdot\\
&\quad C^3L_{\phi,u}L_{\psi,\theta}\rho^{-2}(1-\rho)^{-1} + C(L_{\psi,\theta}(\ell_{\phi,x}+\ell_{\phi,u}L_{\psi,x}) + L_{\phi,u}\ell_{\psi,x})\rho^{-1},\\
C_{\ell,g,(\theta,\theta)} &= ((1+L_{\psi,x})(\ell_{\phi,x}+\ell_{\phi,u}\cdot L_{\psi,x}) + L_{\phi,x}\cdot\ell_{\psi,x})L^2_{\phi,u}L^2_{\psi,\theta}C^3\rho^{-2}(1-\rho)^{-1}\\
&\quad + \big(L_{\phi,u}\ell_{\psi,\theta} + \ell_{\phi,u}L^2_{\psi,\theta}\big)C\rho^{-1}.
\end{aligned}$$

Corollary 30 is implied by Lemma 29 and corresponds to Corollary D.4 in Lin et al. (2023).

Corollary 30 (Lipschitzness/Smoothness of the Multi-Step Costs) *Under the same assumptions as Lemma 29, the multi-step cost function $h_{t|\tau}$ satisfies that*

$$\left\| \left.\frac{\partial h_{t|\tau}}{\partial x_\tau}\right|_{x_\tau,\theta_{\tau:t}} \right\| \le C_{L,h,x}\rho^{t-\tau}, \left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t}} \right\| \le C_{L,h,\theta}\rho^{t-\tau},$$

$$\left\| \left.\frac{\partial h_{t|\tau}}{\partial x_\tau}\right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \left.\frac{\partial h_{t|\tau}}{\partial x_\tau}\right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t}} \right\| \le C_{\ell,h,(x,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,h,(x,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|,$$

$$\left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t}} \right\| \le C_{\ell,h,(\theta,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,h,(\theta,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|,$$

where $C_{L,h,x} = L_hC(1+L_{\psi,x}), C_{L,h,\theta} = L_h\max\{C_{L,\phi,\theta}(1+L_{\psi,x}), L_{\psi,\theta}\}$, and

$$\begin{aligned}
C_{\ell,h,(x,x)} &= L_h(1+L_{\psi,x})C_{\ell,\phi,(x,x)} + ((\ell_{h,x}+\ell_{h,u}L_{\psi,x})(1+L_{\psi,x}) + L_h\ell_{\psi,x})C^2_{L,\phi,x},\\
C_{\ell,h,(x,\theta)} &= L_h(1+L_{\psi,x})C_{\ell,\phi,(x,\theta)} + ((\ell_{h,x}+\ell_{h,u}L_{\psi,x})(1+L_{\psi,x}) + L_h\ell_{\psi,x})C_{L,\phi,x}C_{L,\phi,\theta},\\
C_{\ell,h,(\theta,x)} &= L_h(1+L_{\psi,x})C_{\ell,\phi,(\theta,x)} + ((\ell_{h,x}+\ell_{h,u}L_{\psi,x})(1+L_{\psi,x}) + L_h\ell_{\psi,x})C_{L,\phi,x}C_{L,\phi,\theta},\\
C_{\ell,h,(\theta,\theta)} &= L_h(1+L_{\psi,x})C_{\ell,\phi,(\theta,\theta)} + ((\ell_{h,x}+\ell_{h,u}L_{\psi,x})(1+L_{\psi,x}) + L_h\ell_{\psi,x})C^2_{L,\phi,\theta}.
\end{aligned}$$

Theorem 31 bounds the distances between the trajectory of M-GAPS with the imaginary trajectory achieved by using θ_t repeatedly from time step 0. It can be shown using a similar approach as Theorem D.5 in Lin et al. (2023), while a difference is that we consider an additional disturbance ζ_t in the update rule of policy parameters. We include the proof of Theorem 31 in Appendix J.1 for completeness.

Theorem 31 *Suppose Assumptions 18 and 19 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of*

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \tag{57a}$$

$$y_{t+1} = q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g_{t+1|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}, \tag{57b}$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right) + \zeta_t. \tag{57c}$$

Suppose η and $\bar{\zeta}$ satisfy the constraint that $\bar{\varepsilon} := \frac{C_{L,h,\theta}\eta}{1-\rho} + \bar{\zeta} \le \varepsilon$. Then, both $\|G_t\|$ and $\|\nabla F_t(\theta_t)\|$ are upper bounded by $\frac{C_{L,h,\theta}}{1-\rho}$, and the following inequalities holds for any two time steps τ, t ($\tau \le t$):

$$\|\theta_t - \theta_\tau\| \le \frac{C_{L,h,\theta}}{1-\rho}\cdot(t-\tau)\eta + \sum_{\tau'=\tau}^{t-1}\|\zeta_{\tau'}\|, \text{ and } \|x_\tau - \hat{x}_\tau(\theta_t)\| \le$$

$$\frac{C_{L,h,\theta}C_{L,\phi,\theta}\rho}{(1-\rho)^2}\left((t-\tau)+\frac{1}{1-\rho}\right)\cdot\eta + \frac{C_{L,\phi,\theta}\rho}{1-\rho}\cdot\left(\sum_{\tau'=\tau}^{t-1}\|\zeta_{\tau'}\| + \sum_{\tau'=0}^{\tau-1}\rho^{\tau-\tau'}\|\zeta_{\tau'}\|\right),$$

where we use the notation $\hat{x}_\tau(\theta) := g_{\tau|0}^*(x_0, \theta_{\times(\tau+1)}), \forall\theta\in\Theta$. Further, we have that

$$|h_t(x_t, u_t, \theta_t) - F_t(\theta_t)| \le \frac{C_{L,h,\theta}C_{L,\phi,\theta}L_h(1+L_{\psi,x}+L_{f,x})\rho}{(1-\rho)^3}\cdot\eta + \frac{C_{L,\phi,\theta}L_h(1+L_{\psi,x}+L_{f,x})\rho}{1-\rho}\cdot\sum_{\tau=0}^{t-1}\rho^{t-\tau}\|\zeta_\tau\|.$$

Recall that we define the gradient approximation G_t for M-GAPS in (56). Using this notation, the update rule of $\theta_{0:T-1}$ in joint dynamics (57) can be simplified as

$$\theta_{t+1} = \Pi_\Theta(\theta_{t+1} - \eta G_t) + \zeta_t.$$

To compare the trajectory of M-GAPS with the trajectory achieved by the online gradient descent trajectory $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta\nabla F_t(\theta_t))$, we bound the difference between G_t and $\nabla F_t(\theta_t)$ in Theorem 32. We provide its proof in Appendix J.2 for completeness.

Theorem 32 (Gradient Bias) *Suppose Assumptions 18 and 19 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of (57). Suppose η and $\bar{\zeta}$ satisfy the constraint that $\bar{\varepsilon} := \frac{C_{L,h,\theta}\eta}{1-\rho} + \bar{\zeta} \le \varepsilon$. Then, the following holds for all $\tau \le t$:*

$$\left\|\left.\frac{\partial h_{t|0}^*}{\partial\theta_\tau}\right|_{x_0,\theta_{0:t}} - \left.\frac{\partial h_{t|0}^*}{\partial\theta_\tau}\right|_{x_0,(\theta_t)_{\times(t+1)}}\right\|$$

$$\leq \left(\hat{C}_0 + \hat{C}_1(t-\tau) + \hat{C}_2(t-\tau)^2\right)\rho^{t-\tau}\cdot\eta + \left(\hat{C}_3\sum_{\tau'=\tau}^{t-1}\|\zeta_{\tau'}\| + \hat{C}_4\sum_{\tau'=\tau}^{t-1}(\tau'-\tau)\|\zeta_{\tau'}\|\right)\cdot\rho^{t-\tau}$$
$$+ \hat{C}_5\sum_{\tau'=0}^{\tau-1}\rho^{t-\tau'}\|\zeta_{\tau'}\|.$$

for

$$\hat{C}_0 = \frac{\rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}}{(1-\rho)^3},\ \hat{C}_1 = \frac{(1-\rho)C_{L,h,\theta}C_{\ell,h,(\theta,x)} + \rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,\theta)}}{(1-\rho)^2},$$
$$\hat{C}_2 = \frac{C_{L,h,\theta}C_{\ell,h,(x,\theta)C_{L,\phi,\theta}}}{1-\rho},\ \hat{C}_3 = \frac{C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}\rho}{1-\rho},$$
$$\hat{C}_4 = C_{\ell,h,(x,\theta)}C_{L,\phi,\theta},\ \hat{C}_5 = \frac{C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}\rho}{1-\rho}.$$

Next,

$$\|G_t - \nabla F_t(\theta_t)\| \leq \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta$$
$$+ \sum_{\tau=0}^{t-1}\left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5(t-\tau)\right)\rho^{t-\tau}\|\zeta_\tau\|.$$

J.1. Proof of Theorem 31

We first use induction to show that for all time step $t \in \mathcal{T}$,

$$\|G_t\| \leq \frac{C_{L,h,\theta}}{1-\rho}, x_t \in B_n(0, R_S + \bar{C}\|x_0\|), u_t \in \mathcal{U}, \text{ and } \|\theta_{t+1} - \theta_t\| \leq \epsilon_\theta, \tag{58}$$

where $\mathcal{U} = \{\psi(x,\theta) - f(x,a) \mid x \in B_n(0,R_x), \theta \in \Theta, a \in \mathcal{A}, (\psi, f) \in \mathcal{G}\}$.

Note that $\|G_0\| \leq C_{L,h,\theta} \leq \frac{C_{L,h,\theta}}{1-\rho}$ by Theorem 30. We also have $x_0 \in B_n(0, R_S + C\|x_0\|)$ and $u_0 \in \mathcal{U}$.

Suppose $\|G_{t-1}\| \leq \frac{C_{L,h,\theta}}{1-\rho}$ for some $t \geq 1$. Then, since $\eta \leq \frac{(1-\rho)\epsilon_\theta}{C_{L,h,\theta}}$ and the projection onto Θ is a contraction, we see that

$$\|\theta_t - \theta_{t-1}\| \leq \|\eta G_{t-1}\| + \|\zeta_t\| \leq \epsilon_\theta.$$

Suppose $\|\theta_\tau - \theta_{\tau-1}\| \leq \epsilon_\theta$ holds for all $\tau \leq t$, i.e., $\theta_{0:t} \in S_{\epsilon_\theta}(0:t)$. By Lemma D.2 in Lin et al. (2023), we see that

$$x_t \in B_n(0, R_S + C\|x_0\|), \text{ and } u_t \in \mathcal{U}.$$

Therefore, by taking norm on both sides of (56), we see that

$$\|G_t\| = \left\|\sum_{\tau=0}^{t}\frac{\partial h_{t|t-\tau}}{\partial\theta_{t-\tau}}\Big|_{x_{t-\tau},\theta_{t-\tau:t}}\right\|$$
$$\leq \sum_{\tau=0}^{t}\left\|\frac{\partial h_{t|t-\tau}}{\partial\theta_{t-\tau}}\Big|_{x_{t-\tau},\theta_{t-\tau:t}}\right\| \tag{59a}$$

$$\leq \sum_{\tau=0}^{t} C_{L,h,\theta}\rho^{\tau} \tag{59b}$$
$$\leq \frac{C_{L,h,\theta}}{1-\rho},$$

where we use the triangle inequality in (59a) and Theorem 30 in (59b). Note that we can apply Theorem 30 because $x_t \in B_n(0, R_S + C\|x_0\|)$. Therefore, we have shown (58) by induction. One can use the same technique as (59) to show $\|\nabla F_t(\theta_t)\| \leq \frac{C_{L,f,\theta}}{1-\rho}$.

Since the projection onto the set Θ is a contraction, we obtain that for any $t > \tau$,

$$\|\theta_t - \theta_\tau\| \leq \frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau)\eta + \sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\|. \tag{60}$$

Now we bound the distance between x_τ and $\hat{x}_\tau(\theta_t)$ for $\tau \leq t$. We see that

$$\begin{aligned}
\|x_\tau - \hat{x}_\tau(\theta_t)\| &= \left\| g^*_{\tau|0}(x_0, \theta_{0:\tau-1}) - g^*_{\tau|0}(x_0, (\theta_t)_{\times\tau}) \right\| \\
&\leq \sum_{\tau'=0}^{\tau-1} \left\| g^*_{\tau|0}(x_0, \theta_{0:\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g^*_{\tau|0}(x_0, \theta_{0:\tau'-1}, (\theta_t)_{\times(\tau-\tau')}) \right\| && \text{(61a)} \\
&\leq \sum_{\tau'=0}^{\tau-1} \left\| g^*_{\tau|\tau'}(x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g^*_{\tau|\tau'}(x_{\tau'}, (\theta_t)_{\times(\tau-\tau')}) \right\| && \text{(61b)} \\
&\leq \sum_{\tau'=0}^{\tau-1} C_{L,g,\theta}\rho^{\tau-\tau'} \|\theta_t - \theta_{\tau'}\| && \text{(61c)} \\
&\leq \frac{C_{L,h,\theta}C_{L,g,\theta}\eta}{1-\rho} \sum_{\tau'=0}^{\tau-1} \left(\frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau')\eta + \sum_{\tau''=\tau'}^{t-1} \|\zeta_{\tau''}\| \right) && \text{(61d)} \\
&\leq \frac{C_{L,h,\theta}C_{L,\phi,\theta}\rho}{(1-\rho)^2} \left((t-\tau) + \frac{1}{1-\rho} \right) \cdot \eta \\
&\quad + \frac{C_{L,\phi,\theta}\rho}{1-\rho} \cdot \left(\sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\| + \sum_{\tau'=0}^{\tau-1} \rho^{\tau-\tau'} \|\zeta_{\tau'}\| \right),
\end{aligned}$$

where we use the triangle inequality in (61a); we use the definition of multi-step dynamics in (61b); we use Theorem 29 in (61c); we use (60) in (61d).

Similarly, since $x_t \in B_n(0, R_S + C\|x_0\|)$ and we also see that $\hat{x}_t(\theta_t) \in B_n(0, R_S + C\|x_0\|)$, we obtain that

$$\begin{aligned}
|h_t(x_t, u_t, \theta_t) - F_t(\theta_t)| &= |h_t(x_t, u_t, \theta_t) - h_t(\hat{x}_t(\theta_t), \hat{u}_t(\theta_t), \theta_t)| \\
&\leq L_h(\|x_t - \hat{x}_t(\theta_t)\| + \|u_t - \hat{u}_t(\theta_t)\|) && \text{(62a)} \\
&\leq L_h(1 + L_{\psi,x} + L_{f,x})\|x_t - \hat{x}_t(\theta_t)\| && \text{(62b)} \\
&\leq \frac{C_{L,h,\theta}C_{L,\phi,\theta}L_h(1 + L_{\psi,x} + L_{f,x})\rho}{(1-\rho)^3} \cdot \eta \\
&\quad + \frac{C_{L,\phi,\theta}L_h(1 + L_{\psi,x} + L_{f,x})\rho}{1-\rho} \cdot \sum_{\tau=0}^{t-1} \rho^{t-\tau}\|\zeta_\tau\|, && \text{(62c)}
\end{aligned}$$

where we use Theorem 18 in (62a) and (62b); we use (61) in (62c).

J.2. Proof of Theorem 32

To simplify the notation, we adopt the shorthand notations $\hat{x}_\tau(\theta) \coloneqq g^*_{\tau|0}(x_0, \theta_{\times\tau})$ and $\hat{u}_\tau(\theta) \coloneqq \pi_\tau(\hat{x}_\tau(\theta), \theta)$ throughout the proof.

We use the triangle inequality to do the decomposition

$$
\begin{aligned}
&\left\| \left.\frac{\partial h^*_{t|0}}{\partial \theta_\tau}\right|_{x_0, \theta_{0:t}} - \left.\frac{\partial h^*_{t|0}}{\partial \theta_\tau}\right|_{x_0, (\theta_t)_{\times(t+1)}} \right\| \\
=& \left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:t}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)}} \right\| \\
\le& \left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_\tau, (\theta_t)_{\times(t-\tau)}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)}} \right\| \\
&+ \sum_{\tau'=\tau+1}^{t-1} \left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:\tau'}, (\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:\tau'-1}, (\theta_t)_{\times(t-\tau'+1)}} \right\|.
\end{aligned} \tag{63}
$$

Note that we can apply Theorem 30 to bound each term in (63). For the first term in (63), since $x_\tau, \hat{x}_\tau(\theta_t), x_{\tau+1} \in B_n(0, \bar{R}_C)$, we see that

$$
\begin{aligned}
&\left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_\tau, (\theta_t)_{\times(t-\tau)}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)}} \right\| \\
\le& \rho^{t-\tau}\big(C_{\ell,h,(\theta,x)}\|x_\tau - \hat{x}_\tau(\theta_t)\| + C_{\ell,h,(\theta,\theta)}\|\theta_t - \theta_\tau\|\big) \qquad (64a) \\
\le& \frac{(1-\rho)C_{L,h,\theta}C_{\ell,h,(\theta,x)} + \rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,\theta)}}{(1-\rho)^2} \cdot (t-\tau)\rho^{t-\tau} \cdot \eta \\
&+ \frac{\rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}}{(1-\rho)^3} \cdot \rho^{t-\tau} \cdot \eta \\
&+ \frac{C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}\rho}{1-\rho} \cdot \left(\sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\| + \sum_{\tau'=0}^{\tau-1} \rho^{\tau-\tau'}\|\zeta_{\tau'}\| \right) \cdot \rho^{t-\tau}, \qquad (64b)
\end{aligned}
$$

where we use Theorem 30 in (64a) and Theorem 31 in (64b).

For any $\tau' \in [\tau+1 : t-1]$, since $x_{\tau'}, x_{\tau'+1} \in B_n(0, \bar{R}_C)$, we see that

$$
\begin{aligned}
&\left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:\tau'}, (\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:\tau'-1}, (\theta_t)_{\times(t-\tau'+1)}} \right\| \\
=& \left\| \left(\left.\frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'}, (\theta_t)_{\times(t-\tau'+1)}} \right) \left.\frac{\partial g^*_{\tau'|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:\tau'-1}} \right\|
\end{aligned}
$$

$$\leq \left\| \left. \frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}} \right|_{x_{\tau'},\theta_{\tau'},(\theta_t)_{\times(t-\tau')}} - \left. \frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}} \right|_{x_{\tau'},(\theta_t)_{\times(t-\tau'+1)}} \right\| \cdot \left\| \left. \frac{\partial g^*_{\tau'|\tau}}{\partial \theta_\tau} \right|_{x_\tau,\theta_{\tau:\tau'-1}} \right\|$$

$$\leq C_{\ell,h,(x,\theta)} \rho^{t-\tau'} \|\theta_t - \theta_{\tau'}\| \cdot C_{L,\phi,\theta} \rho^{\tau'-\tau} \tag{65a}$$

$$\leq C_{\ell,h,(x,\theta)} C_{L,\phi,\theta} \cdot \rho^{t-\tau} \cdot \left(\frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau')\eta + \sum_{\tau''=\tau'}^{t-1} \|\zeta_{\tau''}\| \right), \tag{65b}$$

where we use Theorem 29 and Theorem 30 in (65a); we use Theorem 31 in (65b). Substituting (64) and (65) into (63) finishes the proof of the first inequality.

For the second inequality, recall that G_t and $\nabla F_t(\theta_t)$ are given by

$$G_t := \sum_{\tau=0}^{t} \left. \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}} \right|_{x_0,\theta_{0:t}}, \nabla F_t(\theta_t) = \sum_{\tau=0}^{t} \left. \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}} \right|_{x_0,(\theta_t)_{\times(t+1)}}.$$

Therefore, we see that

$$\begin{aligned}
\|G_t - \nabla F_t(\theta_t)\| &= \left\| \sum_{\tau=0}^{t} \left. \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}} \right|_{x_0,\theta_{0:t}} - \sum_{\tau=0}^{t} \left. \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}} \right|_{x_0,(\theta_t)_{\times(t+1)}} \right\| \\
&\leq \sum_{\tau=0}^{t} \left\| \left. \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}} \right|_{x_0,\theta_{0:t}} - \left. \frac{\partial h^*_{t|0}}{\partial \theta_{t-\tau}} \right|_{x_0,(\theta_t)_{\times(t+1)}} \right\| && \text{(66a)} \\
&\leq \sum_{\tau=0}^{t} \left(\hat{C}_0 + \hat{C}_1 \tau + \hat{C}_2 \tau^2 \right) \rho^\tau \eta && \text{(66b)} \\
&\quad + \sum_{\tau=0}^{t-1} \left(\hat{C}_3 \sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\| + \hat{C}_4 \sum_{\tau'=\tau}^{t-1} (\tau'-\tau)\|\zeta_{\tau'}\| \right) \cdot \rho^{t-\tau} \\
&\quad + \hat{C}_5 \sum_{\tau=0}^{t-1} \sum_{\tau'=0}^{\tau-1} \rho^{t-\tau'} \|\zeta_{\tau'}\| && \text{(66c)} \\
&\leq \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3} \right) \eta \\
&\quad + \sum_{\tau=0}^{t-1} \left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5 (t-\tau) \right) \rho^{t-\tau} \|\zeta_\tau\|,
\end{aligned}$$

where we use the triangle inequality in (66a); we use the first inequality in Theorem 32 that we have shown and Theorem 30 in (66c).