Proceedings of Machine Learning Research vol xxx:1–28, 2021

A Regret Minimization Approach to Iterative Learning Control

Naman Agarwal NAMANAGARWAL@GOOGLE.COM
Google AI Princeton

Elad Hazan EHAZAN@PRINCETON.EDU
Princeton University & Google AI Princeton

Anirudha Majumdar ANU.MAJUMDAR@PRINCETON.EDU
Princeton University & Google AI Princeton

Karan Singh KARANS@PRINCETON.EDU
Princeton University

Abstract

We consider the setting of iterative learning control, or model-based policy learning in the presence of uncertain, time-varying dynamics. In this setting, we propose a new performance metric, *planning regret*, which replaces the standard stochastic uncertainty assumptions with worst case regret. Based on recent advances in non-stochastic control, we design a new iterative algorithm for minimizing planning regret that is more robust to model mismatch and uncertainty. We provide theoretical and empirical evidence that the proposed algorithm outperforms existing methods on several benchmarks.

Keywords: Iterative Learning Control, Planning, Online Learning.

1. Introduction

Consider a robotic system learning to perform a novel task, e.g., a quadrotor learning to fly to a specified goal, a manipulator learning to grasp a new object, or a fixed-wing airplane learning to perform a new maneuver. We are particularly interested in settings where (i) the task requires one to *plan* over a given time horizon, (ii) we have access to an *inaccurate* model of the world (e.g., due to unpredictable external disturbances such as wind gusts or misspecification of physical parameters such as masses, inertias, and friction coefficients), and (iii) the robot is allowed to iteratively refine its control policy via multiple executions (i.e., rollouts) on the real world. Motivated by applications where real-world rollouts are expensive and time-consuming, our goal in this paper is to learn to perform the given task as rapidly as possible. More precisely, given a cost function that specifies the task, our goal is to learn a low-cost control policy using a small number of rollouts.

The problem described above is challenging due to a number of factors. The primary challenge we focus on in this paper is the existence of unmodeled deviations from nominal dynamics, and external disturbances acting on the system. Such disturbances may either be random or potentially even adversarial. In this paper we adopt a *regret minimization* approach coupled with a recent paradigm called non-stochastic control to tackle this problem in generality. Specifically, consider the time-varying dynamical system given by the equation

$$x_{t+1} = f_t(x_t, u_t) + w_t, \tag{1.1}$$

where x_t is the state, u_t is the control, and w_t is the disturbance at time t. One could choose $u_1 \dots u_T$, for horizon length T to minimize a sequence of cost functions $c_1 \dots c_T$. Further, a *closed-*

loop correction policy π_t, that can modify u_t as a function of the observed history thus far, may be added to address disturbances. We denote the cost of a single rollout/episode of such a policy on a horizon of length T trajectory by

$$J(u_{1:T}, \pi_{1:T}|w_{1:T}) = \frac{1}{T}\left(\sum_{t=1}^{T} c_t(x_t, \pi_t(u_{1:t}, w_{1:t-1}))\right),$$

where π is the closed loop correction over trajectory $u_{1:H}$. We define a comparative performance metric, which we call **Planning Regret**. In an episodic setting, for every rollout i, an open-loop sequence $u^i_{1:T}$ is selected along with a non-stationary closed-loop correction policy $\pi^i_{1:T}$. The rollout is performed under the influence of an arbitrary disturbance sequence $w^i_{1:T}$. Planning regret is the difference between the total cost of our actions and that of the best open-loop plan coupled with the best closed-loop policy in a policy class Π for each individual rollout in hindsight. Formally for a total of N rollouts, each of horizon T, it is defined as,

$$\boxed{\textbf{Planning Regret}: \sum_{i=1}^{N} J(u^i_{1:T}, \pi^i_{1:T}|w^i_{1:T}) - \min_{u^\star_{1:T}} \sum_{i=1}^{N} \min_{\pi^\star_i \in \Pi} J(u^\star_{1:T}, \pi^\star_i|w^t_{i,1:T})}$$

The motivation for our performance metric arises from the setting of Iterative Learning Control (ILC), where one assumes access to an imperfect (differentiable) simulator of real-world dynamics as well as access to a limited number of rollouts in the real world. In such a setting the disturbances capture the model-mismatch between the simulator and the real-world. The main novelty in our formulation is the fact that, under vanishing regret, the closed-loop corrections π^i_t are almost *instance-wise optimal* on the specific trajectory, and therefore adapt to the passive controls, dynamics and disturbance for the particular unrolling at time t. Indeed, this deterministic regret is a stronger metric of performance than commonly considered in the planning/learning for control literature.

Our main result is an efficient algorithm that guarantees vanishing average planning regret for non-stationary linear systems and disturbance-action policies. The effectiveness of the algorithm is experimentally demonstrated on two environments to yield substantial improvement over ILC.

Paper structure. We present the relevant definitions including the setting in Section 2. The algorithm and the formal statement of the main result can be found in Section 3. In Section 4 we provide an overview of the algorithm and the proof via the definition of a more general abstract *nested online convex optimization game*. This can be of independent interest. Finally in Section 5, we provide the results and details of the experiments. An extended version of this paper is available online (Agarwal et al., 2020) and contains proofs and other details. References to the Appendix correspond to this version.

1.1. Related Work

The literature on planning and learning in a partially known MDP is vast, and we focus here on the setting with the following characteristics:

1. We consider *model-based* learning, which is suitable for situations in which the learner has some information about the dynamics, i.e. the mapping f_t in Equation (1.1), but not the disturbances w_t. We further assume that we can differentiate through the model. This allows for more efficient algorithms, but is of course not as general as model-free learning.

2. We focus on the task of learning a single optimal trajectory, rather than an optimal policy. This is similar to the Pontryagin optimality (Pontryagin et al., 1962; Ross, 2015) and different from solving for an optimal policy by solving the Bellman equation via Q-learning, policy-gradient methods and other general-purpose algorithms (Sutton and Barto, 2018).

3. We allow for arbitrary disturbance processes, and choose regret as a performance metric. This is a significant deviation from the literature on optimal and robust controls (Zhou et al., 1996; Stengel, 1994), and follows the lead of the recent paradigm of non-stochastic control (Agarwal et al., 2019; Hazan et al., 2020; Simchowitz et al., 2020).

4. We allow for multiple real-world rollouts. This access model is most similar to the iterative learning control (ILC) paradigm (Owens and Hätönen, 2005; Ahn et al., 2007). For comparison, the model-predictive control (MPC) paradigm allows for only one real-world rollout on which performance is measured, and all other learning is permitted via access to a simulator.

We now survey more specific work in similar context to our setting as outlined above.

Planning with inaccurate models. Model predictive control (MPC) (Mayne, 2014) provides a general scheme for planning with inaccurate models. MPC operates by applying model-based planning, (eg. iLQR (Li and Todorov, 2004; Todorov and Li, 2005)), in a receding-horizon manner. MPC can also be extended to robust versions (Bemporad and Morari, 1999; Mayne et al., 2005; Langson et al., 2004) that explicitly reason about the parametric uncertainty or external disturbances in the model. Recently, MPC has also been viewed from the lens of online learning (Wagener et al., 2019), which allows for more robust methods. The setting we consider here is more general than MPC, allowing for iterative policy improvement across *multiple rollouts* on the real world.

Iterative Learning Control (ILC). ILC is a popular approach for tackling the setting considered. ILC operates by iteratively constructing a policy using an inaccurate model, executing this policy on the real world, and refining the policy based on the real-world rollout. ILC can be extended to use real-world rollouts to update the model (see, e.g., Abbeel et al. (2006)). For further details regarding ILC, we refer the reader to the text Moore (2012). Robust versions of ILC have also been developed in the control theory literature (de Roover, 1996), using H-infinity control to capture bounded disturbances or uncertainty in the model. However, robust control more generally, typically account for *worst-case* deviations from the model and can lead to extremely conservative behavior. In contrast, here we leverage the recently-proposed framework of *non-stochastic control* to capture *instance-specific* disturbances. We demonstrate both empirically and theoretically that the resulting algorithm provides significant gains in terms of sample efficiency over the standard ILC approach.

Optimal, Robust and Non-stochastic Control. Classic results (Stengel, 1994; Zhou et al., 1996) in optimal control characterize the optimal policy for linear systems subject to i.i.d. perturbations given explicit knowledge of the system in advance. Recently, (Abbasi-Yadkori and Szepesvári, 2011; Dean et al., 2018; Mania et al., 2019; Cohen et al., 2018) obtained finite rates for unknown systems. Beyond stochastic perturbations, robust control approaches (Zhou and Doyle, 1998) compute the best controller under worst-case noise. Recently, the nonstochastic control techniques (Agarwal et al., 2019; Hazan et al., 2020; Simchowitz et al., 2020; Simchowitz, 2020) have been proposed for linear systems. Unlike optimal and robust control, these regard regret as the measure the performance of a controller, and aim for (near) instance optimality (vs. worst-case optimality) on the realized disturbances.

2. Problem Setting and Definitions

2.1. Notation

The norm $\|\cdot\|$ refers to the ℓ_2 norm for vectors and spectral norm for matrices. For any natural number n, the set $[n]$ refers to the set $\{1, 2 \ldots n\}$. We use the notation $v_{a:b} \triangleq \{v_a \ldots v_b\}$ to denote a sequence of vectors/matrices. Given a set S, we use $v_{a:b} \in S$ to represent element wise inclusion, i.e. $\forall j \in [a, b], v_j \in S$; $\mathrm{Proj}_S(v_{a:b})$ represents the element-wise ℓ_2 projection onto to the set S. $v_{a:b,c:d}$ denotes a sequence of sequences, i.e. $v_{a:b,c:d} = \{v_{a,c:d} \ldots v_{b,c:d}\}$ with $v_{a,c:d} = \{v_{a,c} \ldots v_{a,d}\}$.

2.2. Basic Definitions

A **dynamical system** is specified via a start state $x_0 \in \mathbb{R}^{d_x}$, a time horizon T and a sequence of transition functions $f_{1:T} = \{f_t | f_t : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}^n\}$. The system produces a H-length sequence of states $(x_1, \ldots x_{T+1})$ when subject to an H-length sequence of actions $(u_1 \ldots u_T)$ and a sequence of disturbances $\{w_1, \ldots w_T\}$ as the dynamical equation[1] below dictates.

$$x_{t+1} = f_t(x_t, u_t) + w_t$$

Through the paper the only assumption we make about the disturbances w_t is that it supported on a set of bounded diameter W. We assume full observation of the system, i.e. the states x_t are visible to the controller. We also assume the dynamical system to be **known** beforehand. These assumptions imply that we fully observe the instantiation of the perturbations $w_{1:H}$ during runs of the system.

The actions above may be adaptively chosen based on the observed state sequence, ie. $u_t = \pi_t(x_1, \ldots x_t)$ for some non-stationary policy $\pi_{1:T} = \{\pi_1, \ldots \pi_T\}$. We consider the policy to be deterministic (a restriction made for convenience). Therefore the state-action sequence $\{x_t, u_t\}_{t=1}^T$ defined as $x_{t+1} = f_t(x_t, u_t), u_t = \pi_t(x_1 \ldots x_t)$, thus produced is a sequence determined by $w_{1:T}$, fixing the policy, and the system.

A ***rollout*** *of horizon* $\mathbf{T}$ *on* $f_{1:T}$ refers to an evaluation of the above sequence for T time steps. When the dynamical system will be clear from the context, for the rest of the paper, we drop it from our notation. Given a cost function sequence $\{c_t\} : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ the ***loss*** of executing a policy π on the dynamical system f with a particular process noise sequence given by $w_{1:T}$ is defined as

$$J(\pi_{1:T} | f_{1:T}, w_{1:T}) \triangleq \frac{1}{T}\left[\sum_{\tau=1}^{T} c_t(x_t, u_t)\right].$$

Assumption 1 *We will assume that the cost c_t is a twice differentiable convex function and that the value, gradient and hessian of the cost function c_t is available. Further we assume,*

- ***Lipschitzness:****There exists a constant G such that if $\|x\|, \|u\| \leq D$ for some $D > 0$, then $\|\nabla_x c_t(x, u)\|, \|\nabla_u c_t(x, u)\| \leq GD$.*
- ***Smoothness:*** *There exists a constant β such that for all x, u, $\nabla^2 c_t(x, u) \preceq \beta I$*

When the dynamical system and the noise sequence are clear from the context we suppress this notation for the cost and denote it by $J(\pi_{1:T})$. A particular sub-case which will be of special interest

1. For the sake of simplicity, we do not consider a terminal cost, and consequently drop the last state from the description.

to us is the case of linear dynamical systems (LDS). Formally a (non-stationary) linear dynamical system is described by a sequence of matrices $AB_{1:T} = \{(A_t, B_t) \in \mathbb{R}^{d_x, d_x} \times \mathbb{R}^{d_x, d_u}\}_{t=1}^T$ and the transition function is defined as $x_{t+1} = A_t x_t + B_t u_t$.

Assumption 2 *We will assume that the linear dynamical system $AB_{1:T}$ is (κ, δ)- strongly stable for some $\kappa > 0$ and $\delta \in (0, 1]$, if every t, we have that $\|A_t\| \leq 1 - \delta$, $\|B_t\| \leq \kappa$.*

We note that all the results in the paper can be easily generalized to a weaker notion of strong stability where the linear dynamical system is (κ, δ)- strongly stable if there exists a sequence of matrices $K_{1:T}$, such that for every t, we have that $\|A_t - B_t K_t\| \leq 1 - \delta$, $\|B_t\|, \|K_t\| \leq \kappa$. A system satisfying such an assumption can be easily transformed to a system satisfying Assumption 2 by setting $A_t = A_t - B_t K_t$. This redefinition is equivalent to appending the linear policy K_t on top of the policy being executed. As such the only difference to our analysis such an execution makes is the norm of the played actions which we can be shown to be bounded, leading to a difference only in factors polynomial in the system parameters to our main result. Hence for convenience, we state our results under Assumption 2. The assumption of strong-stability (in a weaker form as allowed by stationary systems) has been popular in recent works on online control (Cohen et al., 2018; Agarwal et al., 2019) and the above notion generalizes it to non-stationary systems.

2.3. Policy Classes

Open-Loop Policies. Given a convex set $\mathcal{U} \in \mathbb{R}^{d_u}$, consider a sequence of control actions, $u_{1:T} \in \mathcal{U}$. We define (by an overload of notation), the open-loop policy $u_{1:T}$ as a policy which plays a time t, the action u_t. The set of all such policies is defined as $\Pi_{\mathcal{U}} \triangleq \mathcal{U}^{\otimes T}$.

Given two policies we define sum of the two denoted by $\pi_1 + \pi_2$ as, the policy for which the action at time t is the sum of the action recommended by policy π_1 and π_2.

Linear Policies. Given a sequence of matrices $K_{1:T} \in \mathbb{R}^{d_u, d_x}$, a *linear policy* [2] denoted (via an overload of notation) by $K_{1:T}$ is a policy that plays action $u_t = K_t x_t$. This linear state-feedback policies are known to be optimal for the LQR problem and for H_∞ control (Zhou et al., 1996).

Disturbance Action Policies. A generalization of the class of linear policies can be obtained via the notion of disturbance-action policies (see Agarwal et al. (2019)) defined as follows. A disturbance action policy $\pi_{M_{1:L}}$ of memory length L is defined by a sequence of matrices $M_{1:L} \triangleq \{M_1 \dots M_L\}$ where each $M_i \in \mathcal{M} \subseteq \{\mathbb{R}^{d_u \times d_x}\}$, with the action at time step t given by

$$\pi_{M_{1:L}} \triangleq \sum_{j=1}^{L} M_j w_{t-j} \tag{2.1}$$

A natural class of matrices from which the above feedback matrices can be picked is given by fixing a number $\gamma > 0$ and picking matrices spectrally bounded by γ, i.e. $\mathcal{M}_\kappa \triangleq \{M | M \in \mathbb{R}^{d_u \times d_x}, \|M\| \leq \gamma\}$. We further overload the notation for a disturbance action policy to incorporate an open-loop control sequence $u_{1:T}$, defined as $\pi_{M_{1:L}}(u_{1:T}) \triangleq u_t + \sum_{j=1}^{L} M_j w_{t-j}$.

2. For notational simplicity, we do not include an affine offset c_t in the definition of our linear policy, this can be included with no change in results across the paper.

2.4. Planning Regret With Disturbance-Action Policies

As discussed, a natural idea to deal with adversarial process disturbance is to plan (potentially oblivious to it), producing a sequence of open loop ($u_{1:H}$) actions and appending an adaptive controller to *correct* for the disturbance online. However the disturbance in practice could have structure across rollouts, which can be leveraged to improve the plan($u_{1:H}$), with the knowledge that we have access to an adaptive controller. To capture this, we define the notion of an online planning game and the associated notion of planning regret below.

Definition 3 (Online Planning) *It is defined as an N round/rollout game between a player and an adversary, with each round defined as follows:*

- *At every round i the player given the knowledge of a new dynamical system $f^i_{1:T} = \{f^i_1 \dots f^i_T\}$, proposes a policy $\pi^i_{1:T} = \{\pi^i_1 \dots \pi^i_T\}$.*
- *The adversary then proposes a noise sequence $w^i_{1:T}$ and a cost sequence $c^i_{1:T}$.*
- *A rollout of policy $\pi^i_{1:T}$ is performed on the system $f^i_{1:T}$ with disturbances $w^i_{1:T}$ and the cost suffered by the player $J_i(\pi^i_{1:T}) \triangleq J(\pi^i_{1:T}|f^i_{1:T}, w^i_{1:T})$.*

The task of the controller is to minimize the cost suffered. We measure the performance of the controller via the following objective, defined as **Planning-Regret**, which measures the performance against the metric of producing the best in hindsight open-loop plan, having been guaranteed the optimal adaptive control policy for every single rollout. The notion of adaptive control policy we use is the disturbance-action policy class defined in (2.1). In Appendix (Section A), we discuss the expressiveness of the disturbance-actions policies. In particular, they generalize linear policies for stationary systems and lend convexity. Formally planning regret is defined as follows:

$$\boxed{\textbf{Planning Regret}: \sum_{i=1}^{N} J_i(\pi^i_{1:T}) - \min_{u_{1:T}} \sum_{i=1}^{N} \left(\min_{M_{1:L}} J_i\left(\pi_{M_{1:L}}(u_{1:T})\right)\right)}.$$

3. Main Algorithm and Result

In this section we present the algorithm **iGPC**(Iterative Gradient Perturbation Controller)(Algorithm 1), we propose to minimize Planning Regret. The algorithm at every iteration given an open-loop policy $u_{1:T}$ performs a rollout overlaying an online DAC adaptive controller (Algorithm 2). Further the base policy $u_{1:T}$ is updated by performing gradient descent(or any other local policy improvement) on u fixing the offsets suggested by GPC. [3] We show the following guarantee on average planning regret for Algorithm 1 for linear dynamical systems,

Theorem 4 *Let $\mathcal{U} \subseteq \mathbb{R}^{d_u}$ be a bounded convex set with diameter U. Consider the online planning game(Definition 3) with linear dynamical systems $\{AB^i_{1:T}\}_{i=1}^N$ satisfying Assumption 2 and cost functions $\{c_{1:T}\}_{i=1}^N$ satisfying Assumption 1. Then we have that Algorithm 1(when executed with appropriate parameters), for any sequence of disturbances $\{w^i_{1:T}\}_{i=1}^N$ with each $\|w^i_t\| \le W$ and any $\gamma \ge 0$, produces a sequence of actions with planning regret bounded as*

$$\frac{1}{N}\left(\sum_{i=1}^{N} J_i(\pi^i_{1:T}) - \min_{u_{1:T}\in\mathcal{U}} \left(\sum_{i=1}^{N} \min_{M_{1:L}\in\mathcal{M}_\gamma} J_i\left(\pi_{M_{1:L}}(u_{1:T})\right)\right)\right) \le \tilde{O}\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right).$$

3. In Appendix Section C, we provide a more general version of the algorithm defined for any base policy class.

where $\mathcal{M}_\gamma = \{M | M \in \mathbb{R}^{d_u, d_x}, \|M\| \leq \gamma\}$.

The $\tilde{O}$ notation above subsumes factors polynomial in system parameters $\kappa, \gamma, \delta^{-1}, U, W, G$ and $\log(T)$. The precise bound along with details regarding the parameters to be supplied to the algorithm are listed in a detailed restatement of the theorem in the Appendix (Section B).

Algorithm 1 iGPC Algorithm

Require: [Online] $f_{1:T}^{1:N}$: Dynamical Systems, $w_{1:T}^{1:N}$: Disturbances , $c_{1:T}^{1:N}$
Parameters: Set : $\mathcal{U}$, η_{out} : Learning Rate

1: Initialize $u_{1:T}^1 \in \mathcal{U}$ arbitrarily.
2: **for** $i = 1 \dots N$ **do**
3: Receive the dynamical system $f_{1:T}^i$ for the next rollout.
4: **Rollout**: Collect trajectory data by rolling out policy $u_{1:T}^i$ with GPC ▷ (Algorithm 2)

$$\{x_{1:T}^i, a_{1:T}^i, w_{1:T}^i, o_{1:T}^i\} = \text{GPCRollout}(f_{1:T}^i, u_{1:T}^i)$$

5: **Update**: Compute update to the policy

$$\pi_{1:T}^{i+1} = \text{Proj}_{\mathcal{U}}\left(u_{1:T}^i - \eta_{\text{out}} \nabla_{u_{1:T}} J(u_{1:T}^i + o_{1:T}^i) | f_{1:T}^i, w_{1:T}^i)\right)$$

6: **end for**

Algorithm 2 GPCRollout

Require: $f_{1:T}$: dynamical system, $u_{1:T}$: input policy, [Online] $w_{1:T}$: disturbances, $c_{1:T}$: costs.
Parameters: L:Window, η_{in}: Learning rate, γ: Feedback bound, S: Lookback

1: Initialize $M_{1,1:L} = \{M_{1,j}\}_{j=1}^L \in \mathcal{M}_\gamma$. Set $w_i = 0$ for any $i \leq 0$.
2: **for** $t = 1 \dots T$ **do**
3: Compute GPC Offset: $o_t = \sum_{r=1}^L M_{t,r} \cdot w_{t-r}$, and play action: $a_t = u_t + o_t$.
4: Suffer Cost: $c_t(x_t, a_t)$ and observe state: x_{t+1}.
5: Compute perturbation: $w_t = x_{t+1} - f_t(x_t, a_t)$.
6: Update $M_{t+1,1:T}$ for the next round as: ▷ GPCLoss defined in Equation 4.1

$$M_{t+1,1:L} = \text{Proj}_{\mathcal{M}_\kappa}\big(M_{t,1:L} - \eta_{\text{in}} \nabla_{M_{1:L}} \text{GPCLoss}(M_{t,1:L}, u_{t-S+1:t}, w_{t-S-L+1:t-1}, f_{t-S+1,t-1}, c_t)\big)$$

7: **end for**
8: **return** $x_{1:T}, a_{1:T}, w_{1:T}, o_{1:T}$.

4. Algorithm and Analysis Overview

In this section we provide an overview of the derivation of the algorithm and the proof for Theorem 4. The formal proof is deferred to Appendix (Section B). We introduce an online learning setting that is the main building block of our algorithm. The setting applies more generally to control/planning and our formulation of planning regret in linear dynamical systems is a specification of this setting.

4.1. Nested OCO and Planning Regret

Setting Consider an online convex optimization(OCO) problem (Hazan, 2016), where the iterations have a nested structure, divided into inner and outer iterations. Fix two convex sets $\mathcal{K}_1$ and $\mathcal{K}_2$. After every one out of N outer iterations, the player chooses a point $x_i \in \mathcal{K}_1$. After that there is a sequence of T inner iterations, where the player chooses $y_t^i \in \mathcal{K}_2$ at every iteration. After this choice, the adversary chooses a convex cost function $f_t^i \in \mathcal{F} \subseteq \mathcal{K}_1 \times \mathcal{K}_2 \to \mathbb{R}$, and the player suffers a cost of $f_t^i(x_i, y_t^i)$. The goal of the player is to minimize Planning Regret, defined as

$$\boxed{\textbf{Planning Regret}: \sum_{i=1}^{N} \frac{1}{T}\left(\sum_{t=1}^{T} f_t^i(x_i, y_t^i)\right) - \min_{x^\star \in \mathcal{K}_1} \sum_{i=1}^{N} \min_{y^\star \in \mathcal{K}_2} \frac{1}{T}\left(\sum_{t=1}^{T} f_t^i(x^\star, y^\star)\right)}$$

To state a general result, we assume we have access to two online learners denoted by $\mathcal{A}_1, \mathcal{A}_2$, that are guaranteed to provide sub-linear regret bounds over *linear* cost functions on the sets $\mathcal{K}_1, \mathcal{K}_2$ respectively in the standard OCO model. We denote the corresponding regrets achieved by $R_N(\mathcal{A}_1), R_T(\mathcal{A}_2)$. A canonical algorithm for online linear optimization (OLO) is online gradient descent (Zinkevich, 2003), which is what we use in the sequel. The theory presented here applies more generally. [4] Algorithm 3 lays out a general algorithm for the Nested-OCO setup.

Algorithm 3 Nested-OCO Algorithm

Require: Algorithms $\mathcal{A}_1, \mathcal{A}_2$. Initialize $x_1 \in \mathcal{K}_1$ arbitrarily.

for $i = 1 \dots N$ **do**
  Initialize $y_0^i \in \mathcal{K}_2$ arbitrarily.
  for $t = 1 \dots T$ **do**
    Define loss function over $\mathcal{K}_2$ as $h_t^i(y) \triangleq \nabla_y f_t^i(x_i, y_t^i) \cdot y$
    Update $y_{t+1} \leftarrow \mathcal{A}_2(h_0^i \dots h_t^i)$
  end for
  Define loss function over $\mathcal{K}_1$ as $g_i(x) \triangleq \sum_{t=1}^{T} \nabla_x f_t^i(x_i, y_t^i) \cdot x$
  Update $x_{s+1} \leftarrow \mathcal{A}_1(g_1, ..., g_i)$
end for

Theorem 5 *Algorithm 3 with sub-algorithms $\mathcal{A}_1, \mathcal{A}_2$ with regrets $R_N(\mathcal{A}_1), R_T(\mathcal{A}_2)$ ensures the following regret guarantee on the average planning regret.*

$$\frac{\text{PlanningRegret}}{N} \leq \frac{R_N(\mathcal{A}_1)}{N} + \frac{R_T(\mathcal{A}_2)}{T}$$

When using Online Gradient Descent as the base algorithm, the average regret scales as $O\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{T}}\right)$.

4. Regret for OLO depends on function bounds, which correspond to gradient bounds here. For clarity we omit these.

Proof [Proof of Theorem 5] Let $x^\star \in \mathcal{K}_1$ be any point and let $y^\star_{1:T} \in \mathcal{K}_2$ be any sequence. We have

$$\frac{\sum_{i=1}^N \sum_{t=1}^T f_t^i(x_i, y_t^i) - f_t^i(x^\star, y_i^\star)}{TN} \leq \frac{\sum_{i=1}^N \sum_{t=1}^T \nabla_x f_t^i(x_i - x^\star) + \sum_{i=1}^N \sum_{t=1}^T \nabla_y f_t^i(y_t^i - y^\star)}{TN}$$
$$= \frac{\sum_{i=1}^N [g_i(x_i) - g_i(x^\star)]}{TN} + \frac{\sum_{i=1}^N \sum_{t=1}^T [h_t^i(y_t) - h_t^i(y_i^\star)]}{TN}$$
$$\leq \frac{R_N(\mathcal{A}_1)}{N} + \frac{\cdot R_T(\mathcal{A}_2)}{T},$$

where the first inequality follows by convexity and the last inequality follows by the regret guarantees and noting that the functions g_i are naturally scaled up by a factor of T. ■

4.2. Proof Sketch

The main idea behind the proof is to reduce to the setting of Theorem 5. In the reduction the x variable corresponds to the open loop controls $u_{1:T} \in \mathcal{U}$ and the variables y_t^i correspond to the closed-loop disturbance-action policy $M^i_{t,1:L} \in \mathcal{M}_\gamma$. The algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ are instantiated as Online Gradient Descent with appropriately chosen learning rates.

We begin the reduction by using the observation in Agarwal et al. (2019) that costs are convex with respect to the variables u, M, for *linear dynamical systems* with convex costs. With convexity, prima-facie the reduction seems immediate, however this is impeded by the counterfactual notion of policy regret which implies that cost at any time is dependent on previous actions. This difference in the reduction is only applicable to the closed loop policies M, the open loop part $u_{1:T}$)(i.e. x), follow according to the reduction and hence direct OGD is applied(Line 6, Algorithm 1).

To resolve the issue of memory we use the techniques introduced in the OCO with memory framework proposed by Anava et al. (2015) and recently employed in the work of Agarwal et al. (2019). We leverage the underlying stability of the dynamical system to ensure that cost at time t approximately depends only a bounded number of previous rounds, say S. We then define a proxy loss denoted by GPCLoss, corresponding to the cost incurred by a stationary closed-loop policy executing for the previous S time steps. Formally, given a dynamical system $f_{1:S}$, perturbations $w_{1:S}$, a cost function c, a non-stationary open-loop policy $u_{1:S}$, GPCLoss is a function of closed-loop transfer $M_{1:L}$ defined as follows. Consider the following iterations with $y_1 = 0$,

$$a_j \triangleq u_j + \sum_{r=1}^L M_r w_{j-r} \qquad y_j \triangleq f_{j-1}(y_{j-1}, a_{j-1}) + w_{j-1} \quad \forall j \in [1, S]$$

$$\text{GPCLoss}(M_{1:L}, u_{1:S}, w_{-L+1:S-1}, f_{1:S-1}, c) \triangleq c(y_S, a_S) \tag{4.1}$$

The algorithm updates by performing a gradient descent step on this loss, i.e. $M^i_{t+1,1:L} = M^i_{t,1:L} - \eta \nabla_M \text{GPCLoss}(M^i_{t,1:L})$. The proof proceeds by showing that the actual cost and its gradient is closely tracked by their proxy GPC Loss counterparts with the difference proportional to the learning rate (Appendix Lemma 10). Choosing the learning rate appropriately then completes the proof.

5. Experiments

Consider the following setup: the agent is scored on the cost incurred on a handful of sequentially executed real world rollouts on a dynamical system $g(x, u)$; all the while, the agent has access to an inccurate simulator $f(x,u) \neq g(x,u)$. In particular, while limited to simply observing its trajectories in the real world g, the agent is permitted to compute the function value and Jacobian of the simulator $f(x,y)$ along arbitrary state-action pairs. The disturbances here are thus the difference between g and f along the state-action pairs visited along any given real world rollout.

We briefly review the methods that we compare to: **ILQR (oracle)** is an *infeasible agent* that executes the Iterative Linear Quadratic Regulator algorithm (Li and Todorov, 2004) directly via Jacobians of the real world dynamics g, indicating a lower bound on the best possible cost. The **ILQG** agent obtains a closed loop policy via the Iterative Linear Quadratic Gaussian algorithm (Todorov and Li, 2005), proposed originally to handle Gaussian noise while planning on non-linear systems, on the simulator dynamics f, and then executes the policy thus obtained – this approach does not *learn* from multiple rollouts. The Iterative Learning Control **ILC** agent (Abbeel et al., 2006) *learns* from past trajectories to refine its actions on the next real world rollout. We provide precise details in the Appendix (Section D). Finally, the **IGPC** agent adapts Algorithm 1 by replacing the policy update step (Line 5) with an **LQR** step.

Quadrotor with Wind The simulator models an underactuated planar quadrotor (6 dimensional state, 2 dimensional control) attempting to fly to $(1,1)$ from origin. The real world dynamics differ from the simulator in the presence of a dispersive force field $(x\hat{\mathbf{i}} + y\hat{\mathbf{j}})$, to accomodate wind. The cost is measured as the distance sqaured from the origin along with a quadratic penalty on the actions.

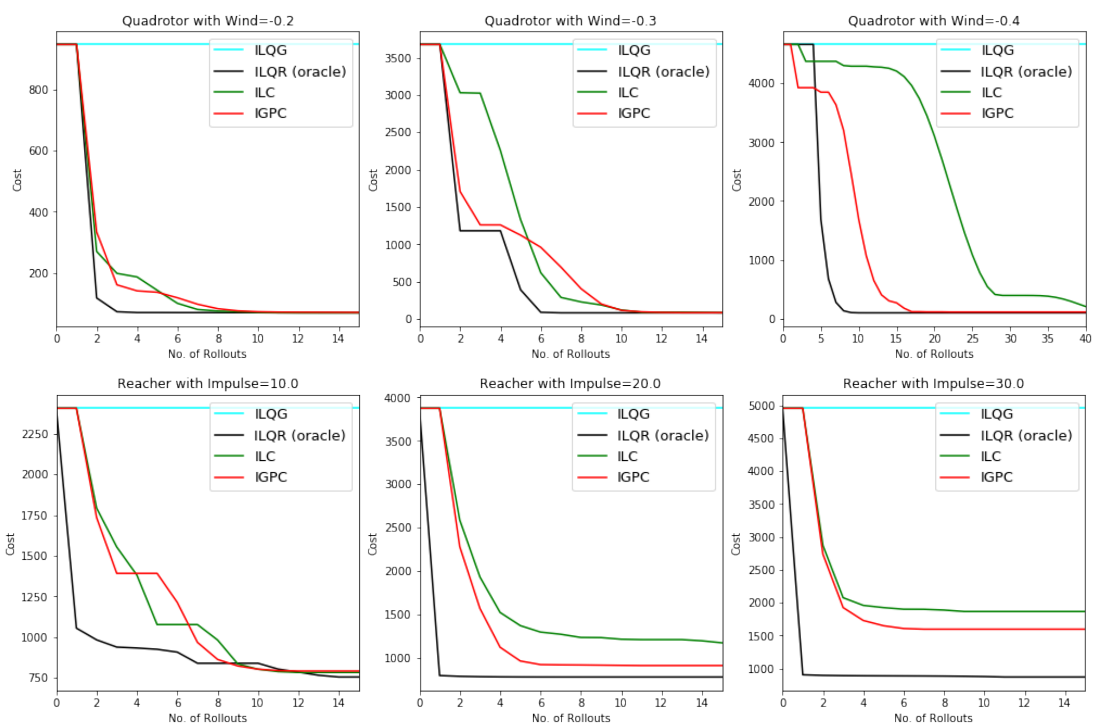

Figure 1: On top is the quadrotor environment for varying magnitudes of wind. Bottom figure captures performance on the reacher environment with varying magnitudes of periodic impulses. **ILQR (oracle)** is an infeasible agent with access to Jacobians on the real world.

Reacher with Impulse The simulator dynamics model a 2-DOF arm (6 dimensional state, 2 dimensional control) attempting to place its end effector at a pre-specified goal. The true dynamics g differs from the simulator in the application of periodic impulses to the center of mass of the arm links. The cost involves a quadratic on the controls and the distance of end effector from the goal.

6. Conclusion

In this work, we cast the task of disturbance-resilient planning into a regret minimization framework. We outline a gradient-based algorithm that refines an open loop plan in conjunction with a near instance-optimal closed loop policy. We provide a theoretical justification for the approach by proving a vanishing average regret bound. We also demonstrate our approach on simulated examples and observe empirical gains compared to the popular iterative learning control (ILC) approach. For a brief discussion of directions for future work, see the full version (Agarwal et al., 2020).

Acknowledgments

Anirudha Majumdar was partially supported by the Office of Naval Research [Award Number: N00014-18- 1-2873]. Elad Hazan was partially supported by NSF award # 1704860 as well as the Google corporation.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Pieter Abbeel, Morgan Quigley, and Andrew Y Ng. Using inaccurate models in reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 1–8. ACM, 2006.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning*, pages 111–119, 2019.

Naman Agarwal, Elad Hazan, Aniruddha Majumdar, and Karan Singh. A regret minimization approach to iterative learning control, 2020. URL https://i-am-karan-singh.github.io/regret-iterative-control.pdf.

Hyo-Sung Ahn, YangQuan Chen, and Kevin L Moore. Iterative learning control: Brief survey and categorization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1099–1121, 2007.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pages 784–792, 2015.

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

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pages 1028–1037, 2018.

Dick de Roover. Synthesis of a robust iterative learning controller using an h/sub/spl infin//approach. In *Proceedings of 35th IEEE Conference on Decision and Control*, volume 3, pages 3044–3049. IEEE, 1996.

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

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, Sham Kakade, and Karan Singh. The nonstochastic control problem. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 408–421, San Diego, California, USA, 08 Feb–11 Feb 2020. PMLR. URL http://proceedings.mlr.press/v117/hazan20a.html.

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

Weiwei Li and Emanuel Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *ICINCO (1)*, pages 222–229, 2004.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

David Q Mayne. Model predictive control: Recent developments and future promise. *Automatica*, 50(12):2967–2986, 2014.

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

Kevin L Moore. *Iterative learning control for deterministic systems*. Springer Science & Business Media, 2012.

David H Owens and Jari Hätönen. Iterative learning control—an optimization paradigm. *Annual reviews in control*, 29(1):57–70, 2005.

Lev S Pontryagin, VG Boltyanskii, RV Gamkrelidze, and EF Mishchenko. The mathematical theory of optimal processes, translated by kn trirogoff. *New York*, 1962.

I Michael Ross. *A primer on Pontryagin's principle in optimal control*. Collegiate publishers, 2015.

Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic. *arXiv preprint arXiv:2006.05910*, 2020.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control, 2020.

Robert F Stengel. *Optimal control and estimation*. Courier Corporation, 1994.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Russ Tedrake. *Underactuated Robotics: Algorithms for Walking, Running, Swimming, Flying, and Manipulation (Course Notes for MIT 6.832)*. 2020.

Emanuel Todorov and Weiwei Li. A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 300–306. IEEE, 2005.

Nolan Wagener, Ching-An Cheng, Jacob Sacks, and Byron Boots. An online learning approach to model predictive control. *arXiv preprint arXiv:1902.08967*, 2019.

Kemin Zhou and John Comstock Doyle. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.

Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice-Hall, Inc., USA, 1996. ISBN 0134565673.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

Appendix A. Comparison of Policy Classes

In this section we make a comparison of various policy classes introduced in the paper.

Linear state-action policies. In classical optimal control with full observation, the cost function is typically assumed to be quadratic in the state and control, i.e.

$$c_t(x, u) = x^\top Qx + u^\top Ru.$$

Under this assumption and infinite horizon time-invariant $(A_i, B_i = A_j, B_j)$ linear dynamical system (LDS), and assuming independent Gaussian disturbances at every time step, the optimal solution can be computed using the Bellman optimality equations (see e.g. Tedrake (2020)). This gives rise to the Discrete time Algebraic Riccati Equation (DARE), whose solution is a linear policy commonly denoted by

$$u_t = Kx_t.$$

The finite-horizon solution is also computable and results in a non-stationary linear policy, where the linear policies converge exponentially fast to the first solution of the Riccati equation. It is thus reasonable to consider the class of all linear policies as a reasonable comparator class. Denote the class of all linear policies as

$$\Pi_L = \{K \in \mathbb{R}^{d_x \times d_u}\}.$$

State of the art: linear dynamical control policies. A generalization of static state-action control policies is that of linear dynamical controllers (LDC). LDC are particularly useful for partially observed LDS and maintain their own internal dynamical system according to the observations in order to recover the hidden state of the system. A formal definition is below.

Definition 6 (Linear Dynamic Controllers) *A linear dynamic controller π is a linear dynamical system $(A_\pi, B_\pi, C_\pi, D_\pi)$ with internal state $s_t \in \mathbb{R}^{d_\pi}$, input $x_t \in \mathbb{R}^{d_x}$ and output $u_t \in \mathbb{R}^{d_u}$ that satisfies*

$$s_{t+1} = A_\pi s_t + B_\pi x_t, \;\; u_t = C_\pi s_t + D_\pi x_t.$$

LDC are state-of-the-art in terms of performance and prevalence in control applications involving LDS, both in the full and partial observation settings. They are known to be theoretically optimal for partially observed LDS with quadratic cost functions and normally distributed noise, but are more widely used. Denote the class of all LDC as

$$\Pi_{LDC} = \{A \in \mathbb{R}^{d_s \times d_s}, B \in \mathbb{R}^{d_s \times d_x}, C \in \mathbb{R}^{d_u \times d_s}, D \in \mathbb{R}^{d_u \times d_x}\}.$$

Disturbance-Action Controllers (DAC) As we have defined earlier, we consider an even more general class of policies, i.e. that of disturbance-action control. For linear time invariant systems, this policy class is more general than that of LDC and linear controllers, in the sense that for every LDS there exists a DAC which outputs exactly the same controls on the same system and sequence of noises. With a finite and fixed H, an approximate version of this statement is true. The precise approximation statement and formal proof can be found in Agarwal et al. (2019). A similar statement can be made for LDC as well.

However we note that all of the above statements hold only in linear time invariant case. In the time varying case, these generalizations are not necessarily true, however note that we are using disturbance action feedback control only as an adaptive control policy to correct against noise, and it is added upon an open-loop plan.

Appendix B. Main Theorem and Proof

We provide the following restatement of Theorem 4 with details regarding the parameters and the dependence on the system parameters. To state the results concisely, we assume that all the appropriate assumed constants, i.e. $\kappa, \gamma, G, \beta, U, W$ are greater than 1. This is done to upper bound the sum of two constants by twice their product. All the results hold by replacing any of these constants by the max of the constant and 1.

Theorem 7 *Let $\mathcal{U} \subseteq \mathbb{R}^{d_u}$ be a bounded convex set with diameter U. Consider the online planning game(Definition 3) with linear dynamical systems $\{AB^i_{1:T}\}_{i=1}^N$ satisfying Assumption 2 and cost functions $\{c_{1:T}\}_{i=1}^N$ satisfying Assumption 1. Then we have that Algorithm 1(when executed with appropriate parameters), for any sequence of disturbances $\{w^i_{1:T}\}_{i=1}^N$ with each $\|w^i_t\| \leq W$ and any $\gamma \geq 0$, produces a sequence of actions with planning regret bounded as*

$$\frac{1}{N}\left(\sum_{i=1}^N J_i(\pi^i_{1:T}) - \min_{u_{1:T}\in\mathcal{U}}\left(\sum_{i=1}^N \min_{M_{1:L}\in\mathcal{M}_\gamma} J_i\left(\pi_{M_{1:L}}(u_{1:T})\right)\right)\right) \leq \left(\frac{c_{\text{in}}\log^2(T)}{\sqrt{T}} + \frac{c_{\text{out}}}{\sqrt{N}}\right).$$

where $\mathcal{M}_\gamma = \{M | M \in \mathbb{R}^{d_u, d_x}, \|M\| \leq \gamma\}$and $c_{\text{in}}, c_{\text{out}}$ are constants depending on system parameters as follows

$$c_{\text{out}} = \tilde{O}\left(GU(U+\gamma LW)\kappa^2\delta^{-2}\right)$$
$$c_{\text{in}} = \tilde{O}\left(\sqrt{\gamma^3\kappa^4\delta^{-3}\beta G^2 L^5 W^3 (U+\gamma LW)^2}\right).$$

Here $\tilde{O}$ subsumes constant factors and factors poly-logarithimic in the the arguments of $\tilde{O}$. To achieve the above bound, Algorithm 1 is to be executed with parameters, learning rate $\eta_{\text{out}} = \frac{U}{G\kappa\delta^{-2}(\kappa U + \kappa\gamma LW + W)\sqrt{N}}$, with the inner execution of Algorithm 2 is performed with parameters $\eta_{\text{in}} = \frac{\gamma^2 L^2}{\sqrt{12\gamma\kappa^4\delta^{-5}\beta G^2 L^3 W^3 (U+\gamma LW)^2}}$ and $S = \delta^{-1}\log(\eta_{\text{in}})$.

B.1. Requisite Definitions

Before proving the theorem we set up some useful definitions. Fix a linear dynamical system $AB_{1:T}$ and a disturbance sequence $w_{1:T}$. For any sequence $u_{1:T} \in \mathcal{U}$ and $M_{1:T,1:L} \in \mathcal{M}_\gamma$, we define T functions $x_{1:T}(\cdot|AB_{1:T}, w_{1:T}), a_{1:T}(\cdot|AB_{1:T}, w_{1:T})$, denoting the action played and the state visited at time t upon execution of the policies together. Herein we drop $AB_{1:T}, w_{1:T}$ from the notation when clear from the context. Formally, consider the following definitions for all t,

$$a_t(u_{1:T}, M_{1:T,1:L}) \triangleq u_t + \sum_{r=1}^L M_{t,r} w_{t-r} \tag{B.1}$$

$$x_1(u_{1:T}, M_{1:T,1:L}) \triangleq 0 \qquad x_{t+1}(u_{1:T}, M_{1:T,1:L}) \triangleq A_t x_t(u_{1:T}, M_{1:T,1:L}) + B_t a_t + w_t \tag{B.2}$$

Given a sequence of cost functions $c_{1:T}(x,u) : \mathbb{R}^{d_x \times d_u} \to \mathbb{R}$, satisfying Assumption 1, define via an overload of notation, the cost functions c_t as a function of $u_{1:T}, M_{1:T,1:L}$ as follows

$$\forall t \in [1:T], \qquad c_t(u_{1:t}, M_{1:T,1:L}) = c_t(x_t(u_{1:t}, M_{1:T,1:L}), a_t(u_{1:T}, M_{1:T,1:L})) \tag{B.3}$$

Naturally, according to our definition of the total cost J of the rollout we get that

$$J(u_{1:T}, M_{1:T,1:L}) = \frac{1}{T}\sum_{t=1}^{T} c_t(u_{1:t}, M_{1:T,1:L})$$

Next, we expand upon the recursive definition of $x_t(\cdot,\cdot)$ via the following operators,

Definition 8 *Given a linear dynamical system $AB_{1:T}$, define the following transfer matrices*

$$\forall j \in [T], \forall k \in [j+1, T] \quad T_{j\to k} \in \mathbb{R}^{d_x \times d_u} \qquad T_{j\to k} \triangleq \begin{cases} I & \text{if } k = j+1 \\ \left(\Pi_{t=j+2}^{k} A_t\right) & \text{otherwise} \end{cases}$$

Additionally given a disturbance sequence $w_{1:T}$, define the following linear operator over matrix sequences $M_{1:T,1:L}$

$$\forall j \in [T], \forall k \in [j+1, T] \quad \psi^M_{j\to k} : [\mathbb{R}^{d_u \times d_x}]^{T\times L} \to \mathbb{R}^{d_x}$$

$$\psi^M_{j\to k}(M_{1:T,1:L}) = \sum_{t=j}^{k-1}\left(T_{t\to k} B_t \left(\sum_{r=1}^{L} M_{t,r} w_{k-r}\right)\right)$$

It can be observed via unrolling the recursion and the definitions above that

$$x_t(u_{1:T}, M_{1:T,1:L}) = \sum_{j=1}^{t-1} T_{j\to t}(B_j u_j + w_j) + \psi^M_{1\to t}(M_{1:T,1:L}). \tag{B.4}$$

Since x_t, a_t are linear functions of $u_{1:T}, M_{1:T,1:L}$, therefore we have that $c_t(u_{1:T}, M_{1:T,1:L})$ is a convex function of its arguments. The next lemma further shows that the gradient of the total cost with respect to the argument $u_{1:T}$ is bounded, as stated in the following lemma.

Lemma 9 *Given a linear system $AB_{1:T}$ satisfying Assumption 2, a bounded disturbance sequence $w_{1:T}$ and a cost sequence c_t satisfying Assumption 1, then for any $\gamma \geq 0, \mathcal{U}$, let $u_{1:T} \in \mathcal{U}, M_{1:T,1:L} \in \mathcal{M}_\gamma$ be two sequences, then we have that*

$$\left\| \nabla_{u_j}\left(\sum_{t=1}^{T} c_t(u_{1:T}, M_{1:T,1:L})\right)\right\| \leq 2G\kappa\delta^{-2}(\kappa U + \kappa\gamma L W + W)$$

We provide the proof of the lemma further in the section. Using the lemma we are now ready to prove Theorem 4.

Proof [Proof of Theorem 4] Lets fix a particular rollout i. Let $AB^i_{1:T}$ be the dynamical system and $w^i_{1:T}$ be the disturbance supplied. Further $u^i_{1:T}$ be the open loop control sequence played at round i and $M^i_{1:T,1:L}$ be the disturbance feedback sequence played by the GPC subroutine. By definition we have that the state achieved

$$x^i_t = x_t(u^i_{1:T}, M^i_{1:T,1:L}) \qquad a^i_t = a_t(u^i_{1:T}, M^i_{1:T,1:L})$$

We have for convenience dropped the system and disturbance from our notation. The total cost at round i incurred by the algorithm by definition is

$$J = \sum_{i=1}^{N} \frac{1}{T} \left(\sum_{t=1}^{T} c_t^i(u_{1:T}^i, M_{1:T,1:L}^i) \right)$$

Fix the sequence of comparators to be $\overset{*}{u}_{1:T}, \{\overset{*}{M}{}_{1:L}^i\}_{i=1}^N$. The comparator cost by definition then is

$$\overset{*}{J} = \sum_{i=1}^{N} \frac{1}{T} \left(\sum_{t=1}^{T} c_t^i(\overset{*}{u}_{1:T}, \mathcal{T}_T \overset{*}{M}{}_{1:L}^i) \right),$$

where given a sequence $v_{a:b}$, we define the tiling operator $\mathcal{T}_k$, which creates a nested sequence of outer length k by tiling with copies of the sequence $v_{a:b}$, i.e. $\mathcal{T}_k v_{a:b} = [v_{a:b}, v_{a:b} \dots v_{a:b}]$. We therefore have the following calculation for the regret which follows from the convexity of the cost function c_t with respect to u, M as established before,

$$\begin{aligned}
&\sum_{i=1}^{N} \sum_{t=1}^{T} \left(c_t^i(u_{1:T}^i, M_{1:T,1:L}^i) - c_t^i(\overset{*}{u}_{1:T}, \mathcal{T}_T \overset{*}{M}{}_{1:L}^i) \right) \\
&\leq \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\nabla_u c_t^i(u_{1:T}^i, M_{1:T,1:L}^i)(u_{1:T}^i - \overset{*}{u}_{1:T}) + \nabla_M c_t^i(u_{1:T}^i, M_{1:T,1:L}^i)(M_{1:T,1:L}^i - \overset{*}{M}{}_{1:L}^i) \right) \\
&= \underbrace{\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\nabla_u c_t^i(u_{1:T}^i, M_{1:T,1:L}^i)(u_{1:T}^i - \overset{*}{u}_{1:T}) \right)}_{\text{Outer Regret}} + \underbrace{\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\nabla_M c_t^i(u_{1:T}^i, M_{1:T,1:L}^i)(M_{1:T,1:L}^i - \overset{*}{M}{}_{1:L}^i) \right)}_{\text{Inner Regret}}
\end{aligned}$$

We analyze the both the terms above separately. We begin by analyzing the first term.

Outer Regret: Consider the following calculation

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\nabla_u c_t^i(u_{1:T}^i, M_{1:T,1:L}^i)(u_{1:T}^i - \overset{*}{u}_{1:T}) \right) = \sum_{j=1}^{T} \sum_{i=1}^{N} \underbrace{\nabla_{u_j} \left(\sum_{t=1}^{T} c_t^i(u_{1:T}^i, M_{1:T,1:L}^i) \right)}_{\triangleq g_{ij}^u} (u_j^i - \overset{*}{u}_j).$$

Note that by definition of the algorithm, we have that for all i, j

$$u_j^{i+1} = \mathrm{Proj}_{\mathcal{U}}(u_j^i - \eta_{\mathrm{out}} g_{ij}^u),$$

which via the pythagorean inequality implies that

$$\|u_j^{i+1} - \overset{*}{u}_j\|^2 \leq \|u_j^i - \eta_{\mathrm{out}} g_{ij}^u - \overset{*}{u}_j\|^2$$

Combining the above equations we immediately get that

$$\begin{aligned}
\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\nabla_u c_t^i(u_{1:T}^i, M_{1:T,1:L}^i)(u_{1:T}^i - \overset{*}{u}_{1:T}) \right) &\leq \sum_{j=1}^{T} \sum_{i=1}^{N} \frac{1}{2} \left(\eta_{\mathrm{out}} \|g_{ij}^u\|^2 + \frac{(u_j^i - \overset{*}{u}_j)^2 - (u_j^{i+1} - \overset{*}{u}_j)^2}{\eta_{\mathrm{out}}} \right) \\
&\leq \sum_{j=1}^{T} \frac{1}{2} \left(\eta_{\mathrm{out}} \left(\sum_{i=1}^{N} \|g_{ji}^u\|^2 \right) + \frac{(u_j^1 - \overset{*}{u}_j)^2}{\eta_{\mathrm{out}}} \right) \\
&\leq 2UG\kappa\delta^{-2}(\kappa U + \kappa\gamma LW + W)T\sqrt{N} \qquad \text{(B.5)}
\end{aligned}$$

where the last inequality follows using Lemma 9 and choice of η_{out}.

Inner Regret: Next we analyze the second Inner Regret term. Before doing so we recommend the reader to re-familiarize with the notations defined in Definition 8 and Equations B.1,B.2,B.3. We will also need the following further definitions again for a fixed rollout. Therefore given a dynamical system $AB_{1:T}$, a disturbance sequence $w_{1:T}$, and an open loop sequence $u_{1:T}$ define the notion of surrogate state at time t which is parameterized by a lookback window S and is a function of an input sequence $M_{1:L} \in \mathbb{R}^{d_u \times d_x}$. Intuitively it corresponds to the state achieved by executing the stationary policy $M_{1:L}$ along with $u_{1:T}$ for S time steps, starting at time $t - S$ with a resetted state. This is exactly the computation performed in the GPCLoss definition in Equation 4.1. We can use the linear operator ψ defined in Definition 8 for an alternative and succinct definition as follows.

$$\hat{x}_t(u_{1:T}, M_{1:L}) = \sum_{j=t-S}^{t-1} T_{j\to t}(B_j u_j + w_j) + \psi^M_{t-S\to t}(\mathcal{T}_T M_{1:L}). \tag{B.6}$$

Further given a cost function c_t, we can use the above definition to also define a surrogate cost

$$\hat{c}_t(u_{1:T}, M_{1:L}) = c_t\left(\hat{x}_t(u_{1:T}, M_{1:L}), u_t + \sum_{j=1}^{L} M_j w_{t-j}\right) \tag{B.7}$$

It can be observed now by the definition of Algorithm 2, the sequence $M^i_{1:T,1:L}$ played by the algorithm is chosen iteratively as follows

$$M^i_{t+1,1:L} = \text{Proj}_{\mathcal{M}_\gamma}\left(M^i_{t,1:L} - \eta_{\text{in}} \nabla_M \hat{c}_t(u^i_{1:T}, M^i_{t,1:L})\right). \tag{B.8}$$

To proceed with the proof we will need the following lemma

Lemma 10 *Consider a linear system $AB_{1:T}$ satisfying Assumption 2, a bounded disturbance sequence $w_{1:T}$ and a sequence of cost functions $c_{1:T}$ satisfying Assumption 1. Given any open loop sequence $u_{1:T} \in \mathcal{U}$ and a closed-loop matrix sequence $M_{1:T,1:L} \in \mathcal{M}_\gamma$ generated through the iteration specified in Equation B.8, we have that the following properties hold for all $t \in [T]$*

- *For all $j > t$, $\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L}) = 0$.*
- *For all $j < t$, $\|\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})\| \leq \kappa^2 G(U + \gamma LW)LW\delta^{-1}(1-\delta)^{t-j}$.*
- *For all t, $\|\nabla_{M_{1:L}} \hat{c}_t(u_{1:T}, M_{1:L})\| \leq GLW(U + \gamma LW)\left(1 + \frac{\kappa^2}{\delta^2}\right)$.*
- *Furthermore, for any $\overset{*}{M}_{1:L} \in \mathcal{M}_\gamma$ and for any t, we have that*

$$\sum_{j=t-S}^{t} \nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{j,1:L} - \overset{*}{M}_{1:L}) \leq \nabla_{M_{t,1:L}} \hat{c}_t(u_{1:T}, M_{t,1:L})(M_{t,1:L} - \overset{*}{M}_{1:L})$$
$$+ 20\eta_{\text{in}} \log^2(\eta_{\text{in}})\gamma\kappa^4\delta^{-3}\beta G^2 L^3 W^3 (U + \gamma LW)^2$$

We are now ready to analyze the inner regret term. We analyze this term for one particular rollout say i (thereby dropping i from our notation). We get the following series of calculations,

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\nabla_M c_t(u_{1:T}, M_{1:T,1:L})(M_{1:T,1:L} - \mathcal{T}_T \overset{*}{M}_{1:L})\right) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{T}\left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{j,1:L} - \overset{*}{M}_{1:L})\right) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{t}\left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{j,1:L} - \overset{*}{M}_{1:L})\right) \\
&\leq \sum_{t=1}^{T}\sum_{j=t-S}^{t}\left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{j,1:L} - \overset{*}{M}_{1:L})\right) + 2\kappa^2\gamma GLW(U+\gamma LW)\delta^{-2}(1-\delta)^S \\
&\leq \sum_{t=1}^{T}\left(\underbrace{\nabla_{M_{t,1:L}} \hat{c}_t(u_{1:T}, M_{t,1:L})}_{g_t}(M_{t,1:L} - \overset{*}{M}_{1:L})\right) + 22T\eta_{\text{in}}\log^2(\eta_{\text{in}})\gamma\kappa^4\delta^{-3}\beta G^2L^3W^3(U+\gamma LW)^2,
\end{aligned}$$

where the statements follow via repeated application of Lemma 10 and the choice of $S = \delta^{-1}\log(\eta_{\text{in}})$. To analyse further once again via a similar argument as in the case of the outer regret regarding projected gradient descent with learning rate η_{in}, we get that,

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\underbrace{\nabla_{M_{t,1:L}} \hat{c}_t(u_{1:T}, M_{1:T,1:L})}_{g_t}(M_{t,1:L} - \overset{*}{M}_{1:L})\right) \\
&\leq \sum_{t=1}^{T}\left(\frac{\eta_{\text{in}}}{2}\|g_t\|^2 + \frac{\|M_{t,1:L} - \overset{*}{M}_{1:L}\|^2 - \|M_{t+1,1:L} - \overset{*}{M}_{1:L}\|^2}{2\eta_{\text{in}}}\right) \\
&\leq \frac{\eta_{\text{in}}T}{2}\|g_t\|^2 + \frac{\|M_{1,1:L} - \overset{*}{M}_{1:L}\|^2}{2\eta_{\text{in}}}
\end{aligned}$$

Combining the above equations, Equation B.9 and the choice of η_{in}, we get that the inner regret is bounded as,

$$\sum_{t=1}^{T}\left(\nabla_M c_t(u_{1:T}, M_{1:T,1:L})(M_{1:T,1:L} - \mathcal{T}_T \overset{*}{M}_{1:L})\right) \leq \tilde{O}\left(\sqrt{T\gamma^3\kappa^4\delta^{-3}\beta G^2L^5W^3(U+\gamma LW)^2}\right)$$

Combining the outer and inner regret terms we finish the proof. ■

In the remaining subsections we prove Lemmas 9 and 10, thereby finishing the proof of Theorem 4.

B.2. Proof of Lemma 9

In this section we prove Lemma 9. Before the proof we establish some other lemmas which will be useful to us.

Lemma 11 *Given a linear system $AB_{1:T}$ satisfying Assumption 2, then the transfer matrices defined in Definition 8 are bounded as follows*

$$\forall j,k \in [T],[j+1,T] \qquad \|T_{j\to k}\| \le (1-\delta)^{k-j-1}$$

Proof [Proof of Lemma 11] If $k = j+1$ then by definition and Assumption 2,

$$\|T_{j\to k}\| = \|I\| \le 1.$$

Otherwise, again by definition and Assumption 2,

$$\|T_{j\to k}\| \le \left(\Pi_{t=j+2}^{k}\|A_t\|\right) \le (1-\delta)^{k-j-1}.$$

■

Lemma 12 *Given a linear system $AB_{1:T}$ satisfying Assumption 2, a bounded disturbance sequence $w_{1:T}$ and a cost sequence c_t satisfying Assumption 1, then for any $\gamma \ge 0, \mathcal{U}$, let $u_{1:T} \in \mathcal{U}, M_{1:T,1:L} \in \mathcal{M}_\gamma$ be two sequences, the following bounds hold for x_t, a_t for all t,*

$$\|x_t(u_{1:T}, M_{1:T,1:L})\| \le \delta^{-1}(\kappa U + \kappa\gamma LW + W),$$

$$\|a_t(u_{1:T}, M_{1:T,1:L})\| \le U + \gamma LW.$$

Furthermore we have that for all $j,t \in [T]$ we have that

$$\left\|\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial u_j}\right\| \le \begin{cases} \kappa(1-\delta)^{t-j-1} & \textit{if } j < t \\ 0 & \textit{otherwise} \end{cases}$$

$$\left\|\frac{\partial a_t(u_{1:T}, M_{1:T,1:L})}{\partial u_j}\right\| = \begin{cases} 1 & \textit{if } j = t \\ 0 & \textit{otherwise} \end{cases}$$

Furthermore we have that for $j,t \in [T]$ and $r \in [L]$, we have that

$$\left\|\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}}\right\| \le \begin{cases} \kappa W(1-\delta)^{t-j-1} & \textit{if } j < t \\ 0 & \textit{otherwise} \end{cases}$$

$$\left\|\frac{\partial a_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}}\right\| \le \begin{cases} W & \textit{if } j = t \\ 0 & \textit{otherwise} \end{cases}$$

Proof From the definition in Equation B.1 it follows that

$$\|a_t(u_{1:T}, M_{1:T,1:L})\| \le \|u_t\| + \sum_{r=1}^{L}\|M_{t,r}\|\|w_{t-r}\| \le U + \gamma LW.$$

Also from the definition it follows that

$$\left\|\frac{\partial a_t(u_{1:T}, M_{1:T,1:L})}{\partial u_j}\right\| = \|\delta_{jt} I\| = \begin{cases} 1 & \text{if } j = t \\ 0 & \text{otherwise} \end{cases}$$

From the expansion in Equation B.4, we have that

$$\|x_t(u_{1:T}, M_{1:T,1:L})\| \leq \sum_{j=1}^{t-1} (\|T_{j\to t}(B_j u_j + w_j)\|) + \|\psi^M_{1\to t}(M_{1:T,1:L})\|$$

$$\leq \sum_{j=1}^{t-1} (\|T_{j\to t}\|\|(B_j u_j + w_j)\|) + \sum_{j=1}^{t-1} \left(\|T_{j\to t}\| \left(\sum_{r=1}^{L} \|M_{j,r}\|\|w_{j-r}\|\right)\right) \quad \text{(Definition 8 \& } \Delta\text{-inequality)}$$

$$\leq (\kappa U + \kappa\gamma LW + W)) \sum_{j=1}^{t-1} (1-\delta)^{t-j-1} \quad \text{(Lemma 11 and definitions)}$$

$$\leq \frac{1}{\delta}(\kappa U + \kappa\gamma LW + W)$$

Also from the definition it follows that for $j \geq t$,

$$\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial u_j} = 0,$$

and if $j < t$, we have that

$$\left\|\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial u_j}\right\| \leq \|T_{j\to t}B_j\| \leq \kappa(1-\delta)^{t-j-1} \qquad \text{(Lemma 11)}$$

From the definition in Equation B.1 it follows that for any $r \in [L]$

$$\left\|\frac{\partial a_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}}\right\| = \|\delta_{jt} I \otimes w_{t-r}^\top\| \leq \begin{cases} W & \text{if } j = t \\ 0 & \text{otherwise} \end{cases}$$

From the expansion in Equation B.4, it follows that for any r and $j \geq t$,

$$\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}} = 0,$$

and if $j < t$, we have that

$$\left\|\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}}\right\| \leq \|T_{j\to t}B_j(I \otimes w_{j-r}^\top)\| \leq \kappa W(1-\delta)^{t-j-1} \qquad \text{(Lemma 11)}$$

■

We are now ready to prove Lemma 9.

Proof [Proof of Lemma 9] Consider the following calculations for all j, t, following from Lemma 12,

$$\|\nabla_{u_j}\left(c_t(u_{1:T}, M_{1:T,1:L})\right)\|$$

$$\leq G\max(\|x_t(u_{1:T}, M_{1:T,1:L}\|\|a_t(u_{1:T}, M_{1:T,1:L}\|)\left(\left\|\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial u_j}\right\| + \left\|\frac{\partial a_t(u_{1:T}, M_{1:T,1:L})}{\partial u_j}\right\|\right)$$

$$\leq \begin{cases} G\kappa\delta^{-1}(\kappa U + \kappa\gamma LW + W)(1-\delta)^{t-j-1} & \text{if } j < t \\ G\kappa\delta^{-1}(\kappa U + \kappa\gamma LW + W) & j = t \\ 0 & \text{otherwise} \end{cases}$$

Therefore we have that,

$$\left\|\nabla_{u_j}\left(\sum_{t=1}^{T} c_t(u_{1:T}, M_{1:T,1:L})\right)\right\| \leq 2G\kappa\delta^{-2}(\kappa U + \kappa\gamma LW + W)$$

■

B.3. Proof of Lemma 10

In this section we prove Lemma 10. To this end we will need the following lemma that is the extension of Lemma 12 to surrogate states.

Lemma 13 *Given a linear system $AB_{1:T}$ satisfying Assumption 2, a bounded disturbance sequence $w_{1:T}$ and a cost sequence c_t satisfying Assumption 1, then for any $\gamma \geq 0, \mathcal{U}$, let $u_{1:T} \in \mathcal{U}, M_{1:L} \in \mathcal{M}_\gamma$ be two sequences, then we have that for all $j, t \in [T]$,*

$$\|\hat{x}_t(u_{1:T}, M_{1:L})\| \leq \delta^{-1}(\kappa U + \kappa\gamma LW + W)$$

Furthermore we have that for $t \in [T]$ and $r \in [L]$, we have that

$$\left\|\frac{\partial \hat{x}_t(u_{1:T}, M_{1:L})}{\partial M_r}\right\| \leq \delta^{-1}\kappa W$$

Proof From the expansion in Equation B.6, we have that

$$\begin{aligned}
&\|x_t(u_{1:T}, M_{1:L})\| \leq \sum_{j=t-S}^{t-1} (\|T_{j\to t}(B_j u_j + w_j)\|) + \|\psi^M_{t-S\to t}(\mathcal{T}_T M_{1:L})\| \\
&\leq \sum_{j=t-S}^{t-1} (\|T_{j\to t}\|\|B_j u_j + w_j\|) + \sum_{j=t-S}^{t-1}\left(\|T_{j\to t}\|\left(\sum_{r=1}^{L}\|M_r\|\|w_{j-r}\|\right)\right) && \text{(Definition 8 \& Δ-inequality)} \\
&\leq (\kappa U + \kappa\gamma LW + W)\sum_{j=t-S}^{t-1}(1-\delta)^{t-j-1} && \text{(Lemma 11 and Definitions)} \\
&\leq \frac{1}{\delta}(\kappa U + \kappa\gamma LW + W)
\end{aligned}$$

From the expansion in Equation B.6, it follows that

$$\left\|\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_r}\right\| \leq \left\|\sum_{j=t-S}^{t-1} T_{j\to t} B_j I \otimes w_{j-r}^\top\right\| \leq \delta^{-1}\kappa W \qquad \text{(Lemma 11)}$$

■

We are now ready to prove Lemma 10.

Proof [Proof of Lemma 10] Since for any $j > t$, by Lemma 12, we have that

$$\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,1:L}} = 0, \frac{\partial a_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,1:L}} = 0,$$

it immediately follows that for all $j > t$,

$$\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L}) = 0.$$

Furthermore again from Lemma 12, we have that for all $j < t$ and for all $r \in [L]$,

$$\left\| \frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}} \right\| \leq \kappa W (1-\delta)^{t-j-1}$$

and further if $j < t$ and for all $r \in [L]$,

$$\frac{\partial a_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}} = 0$$

Therefore, since the cost function c_t satisfies the Assumption 1, using Lemma 12, we have that for all $j < t$ and for any $r \in [L]$

$$\begin{aligned} \left\| \nabla_{M_{j,r}} c_t(u_{1:T}, M_{1:T,1:L}) \right\| &\leq G \|x_t(u_{1:T}, M_{1:T,1:L})\| \left\| \frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,r}} \right\| \\ &\leq G\kappa\delta^{-1} W(\kappa U + \kappa\gamma LW + W)(1-\delta)^{t-j} \end{aligned} \tag{B.9}$$

Using Lemma 13 for the surrogate states and using Assumption 1, we have that for all t and for all $r \in [L]$,

$$\|\nabla_{M_r} \hat{c}_t(u_{1:T}, M_{1:L})\| \leq 2G\kappa\delta^{-2} W(\kappa U + \kappa\gamma LW + W)$$

Since the gradient is bounded according to the above calculation and the $M_{t,1:L}$ are generated via gradient descent with a learning rate η_{in}, it is immediate that for any $j, k \in [T]$ and for any $r \in [L]$,

$$\|M_{j,r} - M_{k,r}\| \leq \eta_{\text{in}} |j-k| \cdot 2G\kappa\delta^{-2} W(\kappa U + \kappa\gamma LW + W) \tag{B.10}$$

Given the above we show that for any execution the surrogate states and the real states are close to each other. To this end consider the following calculations.

$$\begin{aligned} &\|x_t(u_{1:T}, M_{1:T,1:L}) - \hat{x}_t(u_{1:T}, M_{t,1:L})\| \\ &\leq \left\| \sum_{j=1}^{t-1} T_{j\to t}\left(B_j u_j + w_j\right) + \psi^M_{1\to t}(M_{1:T,1:L}) - \sum_{j=t-S}^{t-1} T_{j\to t}\left(B_j u_j + w_j\right) - \psi^M_{t-S\to t}(\mathcal{T}_T M_{t,1:L}) \right\| \\ &= \left\| \sum_{j=1}^{t-S-1} \left(T_{j\to t}\left(B_j u_j + w_j + \sum_{r=1}^{L} M_{j,r} w_{j-r} \right)\right) + \sum_{j=t-S}^{t-1} \left(T_{j\to t}\left(\sum_{r=1}^{L} (M_{j,r} - M_{t,r}) w_{j-r} \right)\right) \right\| \\ &\leq (\kappa U + \kappa\gamma LW + W)\left(\delta^{-1}(1-\delta)^S + 2\eta_{\text{in}}\kappa\delta^{-2} S^2 GLW^2\right) \end{aligned} \tag{B.11}$$

Furthermore, note by definitions that

$$\sum_{j=t-S}^{t-1} \frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,1:L}} = \frac{\partial \hat{x}_t(u_{1:T}, M_{t,1:L})}{\partial M_{t,1:L}} \tag{B.12}$$

Before moving further, consider the following calculations

$$
\begin{aligned}
&\sum_{j=t-S}^{t-1} \left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})\right) \\
&= \sum_{j=t-S}^{t-1} \left(\frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,1:L}} \nabla_x c_t(x_t(u_{1:T}, M_{1:T,1:L}), a_t(u_{1:T}, M_{1:T,1:L}))\right) \\
&= \sum_{j=t-S}^{t-1} \frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,1:L}} \left((\nabla_x c_t(\hat{x}_t(u_{1:T}, M_{t,1:L}), a_t(u_{1:T}, M_{1:T,1:L})) + v\right)\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\|v\| \triangleq \|\nabla_x c_t(x_t(u_{1:T}, M_{1:T,1:L}), a_t(u_{1:T}, M_{1:T,1:L})) - \nabla_x c_t(\hat{x}_t(u_{1:T}, M_{t,1:L}), a_t(u_{1:T}, M_{1:T,1:L}))\| \\
\leq \beta(\kappa U + \kappa\gamma LW + W)\left(\delta^{-1}(1-\delta)^S + 2\eta_{\text{in}}\kappa\delta^{-2}S^2GLW^2\right) \quad \text{(B.13)}
\end{aligned}
$$

using Equation B.11 and the β-smoothness of c_t via Assumption 1. Using Equation B.12 and Lemma 12 we now get that

$$
\begin{aligned}
&\sum_{j=t-S}^{t-1} \left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})\right) \\
&= \left(\sum_{j=t-S}^{t-1} \frac{\partial x_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{j,1:L}}\right) \left(\nabla_x c_t(\hat{x}_t(u_{1:T}, M_{t,1:L}), a_t(u_{1:T}, M_{1:T,1:L})) + v\right) \\
&= \frac{\partial \hat{x}_t(u_{1:T}, M_{1:T,1:L})}{\partial M_{t,1:L}} \left(\nabla_x c_t(\hat{x}_t(u_{1:T}, M_{t,1:L}), a_t(u_{1:T}, M_{1:T,1:L}))\right) + v' \qquad \text{(B.14)}
\end{aligned}
$$

where v' is a vector whose norm using Equation B.13 and Lemma 13 can be bounded as follows

$$
\beta\delta^{-1}LW(\kappa U + \kappa\gamma LW + W)\left(\delta^{-1}(1-\delta)^S + 2\eta_{\text{in}}\kappa\delta^{-2}S^2GLW^2\right). \qquad \text{(B.15)}
$$

Now, consider the following computation which follows from Equation B.14 and using the defintiions for the $j = t$ case,

$$
\sum_{j=t-S}^{t} \left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})\right) = \nabla_{M_{t,1:L}} \hat{c}_t(u_{1:T}, M_{t,1:L}) + v'. \qquad \text{(B.16)}
$$

We can now perform the calculation to relate the gradient inner products for surrogate cost to those of real cost.

$$
\begin{aligned}
&\sum_{j=t-S}^{t} \left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{j,1:L} - \overset{*}{M}_{1:L})\right) \\
&= \sum_{j=t-S}^{t} \left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{t,1:L} - \overset{*}{M}_{1:L}) + \nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{j,1:L} - M_{t,1:L})\right) \\
&\le \sum_{j=t-S}^{t} \left(\nabla_{M_{j,1:L}} c_t(u_{1:T}, M_{1:T,1:L})(M_{t,1:L} - \overset{*}{M}_{1:L})\right) + \eta_{\text{in}} 2G^2 L S^2 \kappa^2 \delta^{-3} W^2 (\kappa U + \kappa\gamma L W + W)^2 \\
&\le \nabla_{M_{t,1:L}} \hat{c}_t(u_{1:T}, M_{t,1:L})(M_{t,1:L} - \overset{*}{M}_{1:L}) + \\
&\qquad\qquad \beta\delta^{-1}\gamma L^2 W(\kappa U + \kappa\gamma L W + W)^2 \left(\delta^{-1}(1-\delta)^S + 4\eta_{\text{in}}\kappa^2\delta^{-2} S^2 G^2 L W^2\right) \\
&\le \nabla_{M_{t,1:L}} \hat{c}_t(u_{1:T}, M_{t,1:L})(M_{t,1:L} - \overset{*}{M}_{1:L}) + 5\eta_{\text{in}} \log^2(\eta_{\text{in}})\gamma\kappa^2\delta^{-3}\beta G^2 L^3 W^3 (\kappa U + \kappa\gamma L W + W)^2
\end{aligned}
$$

where the first inequality follows from applying Equations B.9, B.10 and Lemma 12, the second last inequality follows from Equations B.15 and B.16 and the last inequality follows from the choice of the parameter $S = \delta^{-1}\log(\eta_{\text{in}})$. This finishes the proof. ■

Appendix C. Adaptation of Algorithm to General Policies

In this section we provide a more general version of our algorithms 1 and 2, defined for any base outer policy class Π. Note that our formal results dont cover this generalization and it is provided with practical use in mind.

Algorithm 4 iGPC Algorithm

Require: [Online] $f_{1:T}^{1:N}$: Dynamical Systems, $w_{1:T}^{1:N}$: Disturbances
Parameters: Policy class: Π, η_{out} : Learning Rate

Initialize $\pi_{1:T}^1 \in \Pi$.
for $i = 1 \ldots N$ **do**
    Receive the dynamical system $f_{1:T}^i$ for the next rollout.
    Rollout: Collect trajectory data by rolling out policy $\pi_{1:T}^i$ with GPC ▷ (Algorithm 2)

$$\text{TrajData}^i = \{x_{1:T}^i, a_{1:T}^i, w_{1:T}^i, o_{1:T}^i\} \leftarrow \text{GPCRollout}(f_{1:T}^i, \pi_{1:T}^i)$$

    Update: Compute update to the policy

$$\pi_{1:T}^{i+1} = \text{Proj}_{\Pi}\left(\pi_{1:T}^i - \eta_{\text{out}} \nabla_{\pi_{1:T}} J(\pi_{1:T}^i + \pi(o_{1:T}^i) | f_{1:T}^i, w_{1:T}^i)\right)$$

end for

Algorithm 5 GPCRollout

Require: $f_{1:T}$: dynamical system, $\pi_{1:T}$: input policy, [Online] $w_{1:T}$: disturbances.
Parameters: L:Window, η_{in}: Learning rate, γ: Feedback bound, S: Lookback

1: Initialize $M_{1,1:L} = \{M_{1,j}\}_{j=1}^{L} \in \mathcal{M}_\gamma$.
2: Set $w_i = 0$ for any $i \leq 0$.
3: **for** $t = 1 \ldots T$ **do**
4: Compute GPC Offset
$$o_t = M_{t,1:L} \cdot w_{t-1:t-L}.$$
5: Play action
$$a_t = \pi_t(\cdot) + o_t$$
6: Observe state x_{t+1}.
7: Compute perturbation
$$w_t = x_{t+1} - f_t(x_t, a_t).$$
8: Update $M_{t+1,1:T}$ for the next round as:
$$M_{t+1,1:L} = \text{Proj}_{\mathcal{M}_\kappa}\left(M_{t,1:L} - \eta_{\text{inner}} \nabla_{M_{1:L}} \text{GPCLoss}(M_{t,1:L}, \pi_{t-S+1:t}, w_{t-S-L+1:t-1})\right)$$
▷ GPCLoss defined in Equation 4.1
9: **end for**
10: **return** $x_{1:T}, a_{1:T}, w_{1:T}, o_{1:T}$.

Appendix D. Details of ILQR/ILC/IGPC Algorithms

To succinctly state the algorithms define the following policy which takes as arguments a nominal trajectory $\mathring{x}_{1:T} \in \mathbb{R}^{d_x}, \mathring{u}_{1:T} \in \mathbb{R}^{d_u}$, open-loop gain seqeunce $k_{1:T}$ and closed-loop gain sequence $K_{1:T}$ and a parameter α. The policy defined as $\pi(\alpha, x_{1:T}, k_{1:T}, K_{1:T})$, in the sequel executes the following *standard* rollout on a dynamical system $f_{1:T}$.

$$a_t = \mathring{u}_t + \alpha k_t + K(x_{t-1} - \mathring{x}_{t-1})$$
$$x_{t+1} = f_t(x_t, a_t)$$

Before stating the algorithm we also need the following quadratic approximation of the cost function c around pivots x_0, u_0

$$\begin{aligned} Q(c, x_0, u_0)(x, u) \triangleq \nabla c_x(x_0, u_0)(x - x_0) + \nabla c_u(x_0, u_0)(u - u_0) \\ + \frac{1}{2}([x, u] - [x_0, u_0])^\top \nabla^2 c(x, u)([x, u] - [x_0, u_0]) \end{aligned} \tag{D.1}$$

Algorithm 6 now presents a combined layout for ILQG,ILC and IGPC.

Algorithm 6 Iterative Planning Algorithm

Require: $g_{1:T}$ Real Dynamical Systems, $g_{1:T}$ Simulator.

1: Initialize starting sequence of actions $u^0_{1:T}$
2: Initialize sequence of open loop $k^0_{1:T} = 0$ and closed loop gains $K^0_{1:T} = 0$.
3: **for** $i = 1 \dots N$ **do**
4: **Rollout the Policy**:

- **ILQG:** Standard Rollout on $f_{1:T}$.

$$x^i_{1:T}, u^i_{1:T} = \text{Rollout}(f_{1:T}, \pi(\alpha, x^{i-1}_{1:T}, u^{i-1}_{1:T}, k^{i-1}_{1:T}, K^{i-1}_{1:T}))$$

- **ILC:** Standard Rollout on $g_{1:T}$.

$$x^i_{1:T}, u^i_{1:T} = \text{Rollout}(g_{1:T}, \pi(\alpha, x^{i-1}_{1:T}, u^{i-1}_{1:T}, k^{i-1}_{1:T}, K^{i-1}_{1:T}))$$

.

- **IGPC:** GPCRollout on $g_{1:T}$,

$$x^i_{1:T}, u^i_{1:T} = \text{GPCRollout}(g_{1:T}, \pi(\alpha, x^{i-1}_{1:T}, u^{i-1}_{1:T}, k^{i-1}_{1:T}, K^{i-1}_{1:T}))$$

5: **Update**: Obtain $k^i_{1:T} \in \mathbb{R}^{d_u}, K^i_{1:T} \in \mathbb{R}^{d_u \times d_x}$ as the optimal non-stationary affine policy to the following LQG problem.

$$\min \mathbb{E}_z \left[\sum_{t=1}^{T} Q(c_t, x^i_t, u^i_t)(x_t, u_t) \right]$$

$$\text{subject to} \qquad x_{t+1} - x^i_{t+1} = \frac{\partial f_t(x^i_t, u^i_t)}{\partial x^i_t}(x_t - x^i_t) + \frac{\partial f_t(x^i_t, u^i_t)}{\partial u^i_t}(u_t - u^i_t) + z_t$$

where z_t are independent Gaussians of any non-zero variance.
6: **end for**
