Online Adaptive Controller Selection in Time-Varying Systems: No-Regret via Contractive Perturbations

YIHENG LIN, California Institute of Technology, United States
JAMES PREISS, California Institute of Technology, United States
EMILE ANAND, California Institute of Technology, United States
YINGYING LI, California Institute of Technology, United States
YISONG YUE, California Institute of Technology, United States
ADAM WIERMAN, California Institute of Technology, United States

We study the problem of online controller selection in systems with time-varying costs and dynamics. We focus on settings where the closed-loop dynamics induced by the policy class satisfy a contractive perturbation property that generalizes an established property of disturbance-feedback controllers. When the policy class is continuously parameterized, under the additional assumption that past dynamics and cost Jacobians are known, we propose Gradient-based Adaptive Policy Selection (GAPS), which achieves a time-averaged adaptive policy regret of $O(1/\sqrt{T})$. Compared with previous work on disturbance feedback controllers in linear systems, our result applies to a more general setting and improves the regret bound by a factor of $\log T$. When the policy class is finite, we propose Bandit-based Adaptive Policy Selection (BAPS), which achieves a time-averaged policy regret of $O(T^{-1/3})$. We apply the proposed algorithms to the setting of Model Predictive Control (MPC) with unreliable predictions to optimize continuous “confidence” parameters (GAPS) and select the MPC horizon (BAPS), and demonstrate good results for GAPS on a nonlinear system that does not fully satisfy the conditions of our regret bound.

1 INTRODUCTION

We study the problem of online adaptive controller selection for nonlinear time-varying discrete-time dynamical systems, where an online controller selects its parameter at every time step based on its historical experience. In a dynamical system $x_{t+1} = g_t(x_t, u_t)$ with state $x_t$ and control input $u_t$ at time $t$, the controller class is a mapping $\text{ALG}$ from a tuple $(x_t, \varphi_t, \theta_t)$ to a control input $u_t$ (i.e., $u_t = \text{ALG}(x_t, \varphi_t, \theta_t)$), where $\theta_t$ is the controller parameter and $\varphi_t$ is a context provided by the environment. At every time step $t$, the online controller incurs a stage cost $f_t(x_t, u_t)$ that depends on the current state and control input. The goal of online adaptive controller selection is to pick the parameter $\theta_t$ online to minimize the total stage costs over a finite horizon $T$.

Online adaptive controller selection has received significant attention recently [2, 3, 12, 17, 25, 27, 32, 54] because a wide range of control tasks require running the controller on a single trajectory, as opposed to restarting the episode to evaluate a different controller from the same initial state. The online controller’s ability to adapt is also important for systems where the dynamics and cost functions are time-varying. For example, in robotics applications the dynamics change when a manipulator picks up different objects, or when a ground robot moves between surfaces with different friction characteristics. Time-varying costs arise when an operator or planning algorithm specifies a sequence of low-level control objectives to achieve a high-level goal.

A promising approach is to apply techniques from the field of online learning, which provides powerful tools for finite-sample performance guarantees in time-varying settings such as Multi-Armed Bandits (MAB) and Online Convex Optimization (OCO); see Hazan et al. [26] for a
survey. However, a critical difference from control that must be bridged is the bounded memory assumption in online learning. Specifically, in OCO with Memory, the impact of a decision at time \( t \) only exists in the system until time \( t + H - 1 \), where \( H \) is called the memory length. The regret bound also depends on \( H \), and the bound can deteriorate to the order of \( \Omega(T) \) as \( H \) grows to \( T \) [5, 6]. However, in online control the impact of the actions taken at time \( t \) may in general affect the system indefinitely, because the stage cost at time step \( t' > t \) depends on the state \( x_{t'} \), which is impacted by the historical control input \( u_t \).

A natural strategy is to consider algorithms with bounded memory, i.e. online controller selection that only uses a truncated history of length \( H \) [2, 3, 27, 41]. Intuitively, such an approximation may still yield strong guarantees if the impact of historical decisions decays quickly over time. This intuition has been formalized in a few instances, such as for disturbance-based controllers in linear systems [3], where the authors reduce the online controller selection to OCO with Memory of length \( H \). That reduction incurs negligible approximation error once \( H \) is \( \Omega(\log T) \). Leveraging results for OCO with Memory gives \( O(\log T / \sqrt{T}) \) time-averaged policy regret bounds for online adaptive controller selection. Although this reduction approach is intuitive and has been used by many recent works [1, 22, 25, 27], important theoretical questions remain. First, it is unclear whether the extra \( \log T \) factor is necessary due to the difficulty of online control, or an inefficiency in existing reductions (OCO with Memory achieves \( O(1/\sqrt{T}) \) time-averaged regret by itself). Second, in the discrete-parameter setting, proving a regret bound for online controller selection is still open.

In this work, we propose a general exponentially decaying, or “contractive”, perturbation property on the closed-loop dynamics induced by the controller class. Intuitively, the contractive property holds if the impact of a current state on future states decays exponentially with time, which makes the online adaptive controller selection problem more like online learning with bounded memory. Our contractive perturbation property also generalizes many such previously proposed properties [2, 3, 27, 41]. Under the assumption that this property holds, we ask the following question:

**Can we achieve the same regret bounds in online adaptive controller selection as in online learning?**

**Contributions.** In this paper, we provide affirmative answers to this question for controller classes with either compact continuous or discrete policy parameters. More specifically, we summarize our major contributions below.

First, we propose a general exponentially decaying, or “contractive”, perturbation property (Definition 2.2) on the controller-induced closed-loop dynamics. This property has two versions: the strong version enables the controller selection over a continuous parameter space and is satisfied by controllers considered in many previous works (e.g., [2, 35, 54]); the weak version holds for broader controller classes and still enables online controller selection from a finite set of controllers.

Second, under strong contractive perturbation, we propose an efficient controller selection algorithm, Gradient-based Adaptive Policy Selection (GAPS, Algorithm 1), that applies to a general continuous-parameterized controller class. Even for nonlinear dynamics and nonconvex costs, GAPS can mimic the update of an ideal online gradient descent (OGD) algorithm [26, 60] with the oracle knowledge on how the current policy parameter \( \theta_t \) would have performed if used exclusively over the whole trajectory. Under additional convexity assumptions, we show that GAPS can achieve an \( O(1/\sqrt{T}) \) time-averaged adaptive regret (Theorem 3.1), which is a stronger benchmark for time-varying systems and implies the same policy regret bound. This improves the best existing regret bound by a factor of \( \log T \) and shows that it is possible to achieve a regret bound for adaptive controller selection that matches the bound in online learning. To enable this improvement, we adopt a novel proof technique based on the contractive perturbation property and OGD with biased gradients. GAPS is also more computationally efficient than the method of Agarwal et al. [2] when the controllers are computationally expensive, e.g. MPC with a long prediction horizon.
Third, under weak contractive perturbation, we propose another algorithm, Bandit-based Adaptive Policy Selection (BAPS, Algorithm 2), that is inspired by the EXP3 algorithm [11] for non-stochastic multi-armed bandits and applies to finite controller classes. We show that BAPS achieves an $O(T^{-1/3})$ expected time-averaged policy regret (Theorem 4.1). Although this guarantee is weaker than that of GAPS, it also requires less information, namely the dynamics Jacobians and cost gradients. Our regret upper bound for BAPS also matches the regret lower bound of the closely related online learning setting, MAB with switching costs [20], although this lower bound does not apply directly to our setting (see Section 4 for discussion).

Lastly, we deploy GAPS and BAPS in three numerical simulations (Section 5), demonstrating that 1) GAPS can adapt more quickly than an existing benchmark [35] for robustness/consistency confidence tuning, 2) GAPS performs well empirically in a nonlinear setting that does not satisfy the contractiveness and convexity assumptions required by our theoretical bounds, and 3) GAPS and BAPS can both be used to handle horizon-varying prediction accuracy in MPC, with GAPS converging faster by leveraging more information and structure. Inspired by these results, we find deriving theoretical bounds for GAPS under weaker assumptions an interesting future direction.

Related Work. The work most related to ours is online control with adversarial disturbances and regret guarantees [2, 17, 21, 25, 44, 54]. For example, there is rich literature on policy regret bounds for linear time-invariant (LTI) dynamics [2, 17, 21, 54]. Further, there is a growing interest in adaptive regret, which is a stronger performance metric than policy regret, for linear time-varying (LTV) dynamics [25, 44]. In this work, we consider online controller selection with time-varying dynamics and controller classes that satisfy the contractive perturbation property (see Definition 2.2). Our setting includes LTI and LTV dynamics with many common controller classes as special cases. Moreover, our continuous policy selection algorithm GAPS improves the state-of-the-art bounds for both the policy regret for LTI dynamics [2] and the adaptive regret for LTV dynamics [25]. Our result also closes the gap between the upper and lower bounds of adaptive regret for the online control problem studied in [25].

More broadly, this work also relates to other topics in online control, such as online control with stochastic disturbances [3, 14, 33, 37, 55], online control with unknown dynamics [17, 54], learning-based control with regret analysis [18, 19, 33, 38], and the analysis of other performance metrics, e.g., dynamic regret [36, 41, 56] and competitive ratio [23, 50, 53, 57]. In particular, it is worth mentioning the recent efforts on online nonlinear control, which considers different settings from ours, e.g., stochastic disturbances [14, 33, 55], performance metrics other than regrets [32, 52]. Furthermore, our contractive perturbation assumption is related to the incremental stability assumption considered in online nonlinear control literature [14, 52].

The related work discussed so far applies to continuously-parameterized policy classes. For discrete policy classes, there has been research on switching-based supervisory control [4, 9, 31, 47], which focuses on stabilizing the system by switching policies from a finite policy class, but they do not discuss regret bounds. Further, there are recent papers on reinforcement learning that also consider switching algorithms/policies from a finite pool. For example, Laroche and Feraud [34] provide a heuristic method for switching between reinforcement learning algorithms in generic systems, and Mazumdar et al. [43] address expert switching for finite-state systems instead of continuous-state systems as considered in our paper.

In addition to online control literature, this work is also related to online learning [11, 26, 60], especially the subfield of online learning with memory and/or switching costs. This is because, in both this subfield and our work, the cost at each time step depends on the history decisions, which are also called the memory. More specifically, our continuous policy selection problem is related to online convex optimization (OCO) with memory/switching costs [6, 10, 16, 24, 39, 40, 48, 53], and
our discrete policy selection problem is related to the multi-armed bandit with memory/switching costs \[10, 15, 20\]. However, online learning with memory/switching costs only considers a finite memory/history, while this work studies infinite memory/history.

Finally, our work also contributes to model predictive control (MPC), which is a classic method in the control literature \[46\] and, recently, has been studied from the online learning perspective \[41, 59\]. Specifically, our algorithms provide learning-based and theoretically justified approaches for tuning the parameters of MPC, which are often tuned manually in the literature.

**Notation.** We use \([t_1 : t_2]\) to denote the sequence \([t_1, \ldots, t_2]\) and \(a_{t_1:t_2}\) \((t_1 \leq t_2)\) to denote the sequence \([a_{t_1}, a_{t_1+1}, \ldots, a_{t_2}]\). Further, \(a_{\times r}\) \((r \geq 0)\) denotes the sequence \([a, a, \ldots, a]\) with the same entry \(a\) repeated \(r\) times. We use \(q(x, Q)\) to denote the quadratic form \(x^T Q x\) and \(0\) to denote the vector or matrix of all ones and all zeros respectively, with dimensionality implied by context.

The \(n \times n\) identity matrix is denoted by \(I_n\). For any set \(\Theta\), set \(\Theta'\) denotes the Cartesian product of \(t\) copies of set \(\Theta\). The norm \(||\cdot||\) refers to the Euclidean norm. We use \(B_n(0, R)\) to denote a ball centered at \(0\) with radius \(R\) in space \(\mathbb{R}^n\).

## 2 PRELIMINARIES

We consider a general problem of online controller selection on a single trajectory. (We use the terms controller selection and policy selection interchangeably.) The setting is a discrete-time dynamical system with state \(x_t \in \mathcal{X} \subseteq \mathbb{R}^n\), \(t \in [0 : T − 1]\), where \(n\) is the dimension of the state space. At every time step \(t\) in a finite horizon \(T\), after the controller picks a control action \(u_t \in \mathcal{U} \subseteq \mathbb{R}^m\), the next state and the incurred cost are given by:

**Dynamics:** \(x_{t+1} = g_t(x_t, u_t)\),  
**Cost:** \(f_t(x_t, u_t)\),

respectively, where \(g_t : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{X}\) is a time-varying dynamics function and \(f_t : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}\) is the incurred stage cost every time step \(t \in [0 : T − 1]\). The online controller’s goal is to minimize the total cost \(\sum_{t=0}^{T-1} f_t(x_t, u_t)\) incurred over the whole horizon \([0 : T − 1]\). We assume that the dynamics and cost function sequences \(f_0:T−1\) and \(g_0:T−1\) are oblivious, meaning they do not adapt to the state and control trajectories. This assumption can be made without the loss of generality for deterministic controllers, and is also common for randomized online learning algorithms \[2, 27, 59\].

We also assume that the environment provides context \(\varphi_t \in \Phi\) at every time step that the policy can use. For example, the context may contain the predictions of future cost functions and/or dynamics as well as past disturbances. We give three specific examples in Section 2.3.

### 2.1 Policy Class and Performance Metrics

We consider parameterized controllers in the form of \(u_t = \text{ALG}(x_t, \varphi_t, \theta_t)\), where \(x_t \in \mathcal{X}\) is the current state at the \(t\)th time step, \(\varphi_t \in \Phi\) is current context provided by the environment and is included for generality, and \(\theta_t \in \Theta\) is the parameter of the current controller. We consider oblivious context \(\varphi_t\) and sometimes omit the notation \(\varphi_t\) when it is clear from the context.

We consider two important policy parameter classes in this paper:

- **Continuous parameter class:** class \(\Theta \in \mathbb{R}^d\) is a convex compact set;
- **Finite parameter class:** class \(\Theta = \{1, 2, \ldots, k\}\) is a finite set.

Both policy classes are generic and could express, e.g., linear feedback controllers and neural network controllers. Later in this paper, we will impose additional structures on the parameter classes for meaningful theoretical guarantees. The most critical assumption is the contractive perturbation property, which will be detailed in Section 2.2. Other assumptions on the parameter classes as well as on the system dynamics and cost functions are deferred to Section 3 and Section 4.

To measure the performance of an online controller with parameter learning, we adopt the objective of time-averaged policy regret, which has been adopted by \[17, 26\]. We use \((x_{0:T−1}, u_{0:T−1})\) to
denote the trajectory of the online controller with parameter learning, and use \((\hat{x}_{0:T-1}(\theta), \hat{u}_{0:T-1}(\theta))\) to denote the trajectory of a controller that keeps using ALG with fixed parameter \(\theta\) for all time steps. The time-averaged static policy regret \(R(T)\) is defined as the difference between the time-averaged cost of the online controller with parameter learning and the time-averaged cost of a controller with the clairvoyant optimal fixed policy parameter, i.e.,

\[
R(T) = \frac{1}{T} \left( \sum_{t=0}^{T-1} f_t(x_t, u_t) - \min_{\theta \in \Theta} \sum_{t=0}^{T-1} f_t(\hat{x}_t(\theta), \hat{u}_t(\theta)) \right).
\]

(1)

An online controller is called no-regret if its time-averaged policy regret satisfies \(\lim_{T \to \infty} R(T) = 0\). In what follows, we sometimes use \(J(\theta) := \frac{1}{T} \sum_{t=0}^{T-1} f_t(\hat{x}_t(\theta), \hat{u}_t(\theta))\) for brevity.

A stronger benchmark more suited to time-varying environments is the time-averaged adaptive regret [25, 28], which is defined below.

\[
R^A(T) = \frac{1}{T} \sup_{I = [r,s] \subset [0:T-1]} \left( \sum_{t=r}^{s} f_t(x_t, u_t) - \min_{\theta_t \in \Theta} \sum_{t=r}^{s} f_t(\hat{x}_t(\theta_t), \hat{u}_t(\theta_t)) \right).
\]

(2)

Note that \(R^A(T)\) is scaled to be consistent with \(R(T)\). Adaptive regret is strictly stronger than the static policy regret (1) because we can recover the static policy regret \(R(T)\) by setting \(r = 0, s = T - 1\). Another regret metric that can work well for time-varying systems is the dynamic regret (e.g., [36, 41, 59]), which allows the benchmark policy to change arbitrarily rather than fixed. Although we do not consider dynamic regret in this work, including it is a direction of our future research.

### 2.2 Contractive Perturbation

The core property that enables us to achieve no-regret guarantees in online controller selection is an exponentially decaying, or "contractive", perturbation property of the closed-loop dynamics induced jointly by the dynamical system and the controller class. To formalize this property, we will first introduce some useful definitions.

Especially, we introduce the definitions of multi-step dynamics below, which can represent how the current state \(x_t\) affects future states in a compact way.

**Definition 2.1 (Multi-Step Dynamics).** The multi-step dynamics \(g_{t|\tau}\) between two time steps \(\tau\) and \(t (\tau \leq t)\) specifies how the state \(x_t\) behaves as a function of the previous state \(x_{\tau}\) and previous policies \(\theta_{t-\tau}\). Mathematically, it is recursively defined as \(g_{t|\tau}(x_{\tau}) := x_t\), and

\[
g_{t+1|\tau}(x_{\tau}, \varphi_{t-\tau}, \theta_{t-1}) := g_t(g_{t|\tau}(x_{\tau}, \varphi_{t-\tau-1}, \theta_{t-1}), \text{ALG}(g_{t|\tau}(x_{\tau}, \varphi_{t-\tau-1}, \theta_{t-1}), \varphi_t, \theta_t)), \forall t \geq \tau.
\]

With the multi-step dynamics defined in Definition 2.1, we can formally define the contractive perturbation properties as the perturbation bounds of the multi-step dynamics function \(g_{t|\tau}\) with respect to the current state \(x_t\), which is detailed below.

**Definition 2.2 (Strong/Weak Contractive Perturbation).** We say the strong contractive perturbation property for the controller class \(\text{ALG}\) holds for some constants \(C_0 > 0\) and decay factor \(\rho \in (0, 1)\) if for any time steps \(\tau \leq t\), the inequality

\[
\|g_{t|\tau}(x_{\tau}, \varphi_{t-\tau-1}, \theta_{t-1}) - g_{t|\tau}(x'_{\tau}, \varphi_{t-\tau-1}, \theta_{t-1})\| \leq C_0 \rho^{t-\tau} \|x_t - x'_t\|
\]

holds for arbitrary \(x_{\tau}, x'_{\tau} \in X, \varphi_{t-\tau-1} \in \Phi^{t-\tau}, \theta_{t-1} \in \Theta^{t-\tau}\). We say the weak contractive perturbation property for the controller class \(\text{ALG}\) holds for some constants \(C_0 > 0\) and \(\rho \in (0, 1)\) if for any time steps \(\tau \leq t\), the inequality

\[
\|g_{t|\tau}(x_{\tau}, \varphi_{t-\tau-1}, \theta_{k(t-\tau)}) - g_{t|\tau}(x'_{\tau}, \varphi_{t-\tau-1}, \theta_{k(t-\tau)})\| \leq C_0 \rho^{t-\tau} \|x_t - x'_t\|
\]

holds for arbitrary \(x_{\tau}, x'_{\tau} \in X, \varphi_{t-\tau-1} \in \Phi^{t-\tau}, \theta \in \Theta\).
Intuitively, when strong contractive perturbation holds, the impact of the current state \(x_t\) at time step \(\tau\) on a future state \(x_{\tau + 1}\) at time \(t\) decays quickly with respect to \(t - \tau\) as long as all the controllers from \(s \in [\tau : t]\) have parameters \(\theta_s \in \Theta\). Though this condition may seem strong, this property holds for several important controller classes in the literature, e.g., [2, 59]; see Section 2.3 for concrete examples. In our results for finite policy parameter sets (Section 4), we can relax strong contractive perturbation to its weak counterpart, which only considers time-invariant controllers. Weak contractive perturbation holds for a broader range of dynamics and controllers, e.g. Example 2.3.

**Remark 2.1 (Relation to other notions of contraction/stability).** Strong contractive perturbation is closely related to incremental exponential stability in nonlinear control theory [8, 42, 49], which considers a time-varying dynamical system \(x_{t+1} = h_t(x_t)\) and requires the following convergent property: \(\|x_t - x'_t\| \leq C_0 \rho^{t-\tau}\|x_\tau - x'_\tau\|\) for any \(x_\tau, x'_\tau \in \mathcal{X}\), \(x'_{\tau+1} = h_t(x'_\tau)\), and \(t \geq \tau\). Basically, strong contractive perturbation requires that any closed-loop dynamics \(x_{t+1} = g_t(x_t, \text{ALG}(x_t, \varphi_t, \theta_t))\) generated by controllers \(\theta_t \in \Theta\) will guarantee incremental exponential stability. A sufficient condition to verify the incremental exponential stability is the contraction analysis. See [42] for more details.

Similar with the multi-step dynamics, a future stage cost \(f_t(x_t, u_t)\) can also be represented in a compact way using the current state \(x_t\) and the policy sequence \(\theta_{t:t}\). We define this compact representation as multi-step costs \(f_{t|\tau}\) in Definition 2.3. Another concept that we will use heavily in this work is the surrogate costs \(F_t\), which characterizes how good \(\theta\) should be at time \(t\) if we had applied policy \(\bar{\theta}\) for every step since the start of the game. It is defined formally in Definition 2.3.

**Definition 2.3 (Multi-Step Costs and Surrogate Costs).** The multi-step cost function \(f_{t|\tau}\) between two time steps \(\tau\) and \(t\) (\(\tau \leq t\)) specifies how the stage cost \(f_t\) behaves as a function of the previous state \(x_t\) and previous policies \(\theta_{t:t}\):

\[
f_{t|\tau}(x_t, \varphi_{t:t}, \theta_{t:t}) := f_t(g_{t|\tau}(x_t, \varphi_{t:t-1}, \theta_{t:t-1}), \text{ALG}(g_{t|\tau}(x_t, \varphi_{t:t-1}, \theta_{t:t-1}), \varphi_t, \theta_t))\)

The surrogate stage cost function \(F_t : \Theta \rightarrow \mathbb{R}\) is defined as \(F_t(\theta) := f_{t|0}(x_0, \varphi_{0:t}, \theta_{x(t+1)})\).

### 2.3 Motivating Examples

We now provide three motivating examples. Our first example, the disturbance feedback controller, has been widely used in recent online control literature [2, 17, 27]. It satisfies strong contractive perturbation (Definition 2.2). When applied to linear dynamics, the controller class renders the multi-step cost functions \(f_{t|\tau}\) convex in the controller parameters, which enables the reduction to OCO with Memory [2, 27]. It is also shown to be sufficient to approximate any strongly stable state feedback controller when the memory length \(h\) is large enough.

**Example 2.1 (Disturbance Feedback Controllers in LTV Systems).** Consider a linear time-varying (LTV) system \(g_t(x_t, u_t) = A_t x_t + B_t u_t + w_t\), where \(w_t\) is a potentially adversarial disturbance at time \(t\). The context at time \(t\) is given by \(\varphi_t = w_{t-h:t-1}\). The disturbance feedback controller is:

\[
\text{ALG}(x_t, \varphi_t, \theta_t) = -K_t^{stb} x_t + \sum_{i=1}^{h} M_t^{[i]} w_{t-i}, \text{ where } \theta_t = \left(M_t^{[0]}, M_t^{[1]}, \ldots, M_t^{[h-1]}\right).
\]

Here, \(K_t^{stb}\) is a sequence of pre-determined feedback matrices such that, for any \(t_2 > t_1\), there exists constants \(C'_0 > 0, \rho' \in (0, 1)\) such that \(\|\prod_{t=t_1}^{t_2} (A_t - B_t K_t^{stb})\| \leq C'_0 \rho'^{t_2-t_1}\). The parameter set \(\Theta\) is given by \(\{M_t^{[0]}, M_t^{[1]}, \ldots, M_t^{[h-1]}\} \mid \sum_{i=0}^{h-1} \|M_t^{[i]}\| \leq R_M\} \) for some constant \(R_M > 0\). Under these assumptions, the disturbance feedback controller class satisfies the strong contractive perturbation property with \((C_0, \rho) = (C'_0, \rho')\).
Online Adaptive Controller Selection

The second example, Model Predictive Control (MPC) with Confidence Coefficients, generalizes the \( \lambda \)-confident controller proposed by Li et al. [35]. In this setting, some policy parameter \( \theta^{(\ell)} \in \Theta \) can achieve near-optimal performance when the predictions of the future are accurate (consistency), and another policy parameter \( \theta^{(r)} \in \Theta \) has a worst-case guarantee even when the predictions are unreliable (robustness). Then, we can achieve the robustness and the consistency simultaneously if we can compete against the clairvoyant optimal policy parameter \( \theta^* \) (e.g., \( R(T) \) is small) since \( \theta^* \) is not worse than either \( \theta^{(c)} \) or \( \theta^{(r)} \).

Example 2.2 (Model Predictive Control (MPC) with Confidence Coefficients). Consider the problem of predictive control in a linear time-invariant (LTI) system \( g_t(x_t, u_t) = Ax_t + Bu_t + w_t \) with time-varying cost functions \( f_t(x_t, u_t) = q(x_t, Q_t) + q(u_t, R_t) \). The context \( q_t \) at time step \( t \) is given by \( q_t = \{ A, B, Q_{t:t+k-1}, R_{t:t+k-1}, w_{t:t+k-1}[t] \} \), where \( w_{t:t}[t] \) denotes the prediction of the future disturbance \( w_t \) at time step \( t \). Alg(\( x_t, q_t, \theta_t \)) commits the \( u_t \) entry of the optimal solution

\[
\psi_{t:t+k}(x_t, q_t, \theta_t) := \arg \min_{u_{t:t+k-1}} \sum_{t=0}^{t+k-1} (q(x_t[t], Q_t) + q(u_t[t], R_t)) + q(x_{t+k}[t], \hat{Q})
\]

s.t. \( x_{t+1}[t] = Ax_t[t] + Bu_t[t] + \bar{w}_{t}^{[t-t]} \) \( w_t \), for \( t = t + k - 1 \),

\[ x_{t+1} = x_t, \tag{3} \]

where \( \theta_t = (\lambda_t^{(0)}, \lambda_t^{(1)}, \ldots, \lambda_t^{(k-1)}) \) and \( \Theta \subseteq [0,1]^k \). Under the assumption that the pair \( (A,B) \) is stabilizable and \( 0 < Q_{\min} \leq Q_t \leq Q_{\max}, 0 < R_{\min} \leq R_t \leq R_{\max} \) holds for all time \( t \), Proposition 3 in [59] shows that when \( \hat{Q} = P_{\max} \), this policy class satisfies strong contractive perturbation with:

\[
C_0 = \sqrt{\lambda_{\max}(P_{\max})/\lambda_{\min}(Q_{\min})}, \quad \rho = \sqrt{1 - \lambda_{\min}(Q_{\min})/\lambda_{\max}(P_{\max})},
\]

where \( P_{\max} \) denotes the solution to the discrete time algebraic Riccati equation (DARE)

\[
P = Q_{\max} + A^T PA - A^T PB(R_{\max} + B^T PB)^{-1} B^T PA,
\]

and \( \lambda_{\min} (\lambda_{\max}) \) denotes the minimum (maximum) eigenvalue of a symmetric matrix.

Some common controller classes satisfy the weak contractive perturbation property but not the strong one. Our next example highlights this distinction.

Example 2.3 (State feedback Controller). Consider a LTI system \( g_t(x_t, u_t) = Ax_t + Bu_t + w_t \) and the context \( q_t = \{ A, B \} \). The policy class Alg(\( x_t, q_t, \theta_t \)) = \( -K_t x_t \) where \( \theta_t = K_t \) and for some constants \( C_0 > 0, \rho' \in (0,1), \Theta := \{ K | \| (A - BK') \|^2 \leq C_0'(\rho') \}, \forall t \in \mathbb{Z}_{\geq 0} \) satisfies weak contractive perturbation with \( (C_0, \rho) = (C_0', \rho') \). However, [45] shows that this policy class does not satisfy strong contractive perturbation. A simple controller selection question is to choose from a set \( \Theta \) of state feedback controllers adaptively in a time-varying system to minimize the quadratic costs.

3 CONTROLLER LEARNING WITH CONTINUOUS PARAMETERS

We begin with the continuous policy class setting. We first present an efficient online controller selection algorithm, Gradient-based Adaptive Policy Selection (GAPS, Algorithm 1), and discuss its intuition in Section 3.1. Then, in Section 3.2, we present our main theoretical result (Theorem 3.1), which shows that GAPS mimics the updates and incurs similar costs with an ideal online gradient descent (OGD) algorithm that operates on the surrogate stage costs \( F_0,T-1 \). Further, if we assume that the surrogate stage costs are convex in policy parameters, GAPS achieves a time-averaged adaptive policy regret of \( O(1/\sqrt{T}) \) that improves the best existing regret bound by a factor of \( \log T \). Finally, in Section 3.3, we present an outline of our proof of Theorem 3.1.
3.1 Gradient-based Adaptive Policy Selection (GAPS)

Our algorithm, Gradient-Based Adaptive Policy Selection (GAPS), is a variant of the classic online gradient descent (OGD) approach [13, 26], with a key additional step of approximating the gradient of the surrogate stage cost $F_t$. Intuitively, OGD works as follows: At each time step $t$, the current stage cost characterizes how good the learner’s current decision $\theta_t$ is. The learner updates its decision by taking a step in the direction of the gradient of the current stage cost. Mapping this intuition to the online controller selection problem, the ideal OGD update rule should be $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta \nabla F_t(\theta_t))$, because the surrogate stage cost $F_t$ (Definition 2.3) characterizes how good $\theta_t$ is at time $t$ if we had applied from the start, i.e., without the impact of other historical policy parameters $\theta_{0:t-1}$. However, it is very difficult, if not intractable, for the online controller to compute $\nabla F_t$ exactly. Although impractical, the ideal OGD is an important benchmark that we rely on to design our algorithm and develop the analysis, so we define it formally in Definition 3.1.

**Definition 3.1 (Ideal OGD Update).** At time step $t$, update $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta \nabla F_t(\theta_t))$.

As outlined in Algorithm 1, GAPS uses $G_t$ (line 5) to approximate $\nabla F_t(\theta_t)$ efficiently. To see this, we compare the decompositions, with the key differences highlighted in colored text:

$$
\nabla F_t(\theta_t) = \sum_{b=0}^{t} \frac{\partial f_{t|0}}{\partial \theta_{t-b}} |_{x_0, \phi_{0:t} (\theta_t)_{x(t+1)}} \quad \text{and} \quad G_t = \sum_{b=0}^{\min\{B-1,t\}} \frac{\partial f_{t|0}}{\partial \theta_{t-b}} |_{x_0, \phi_{0:t}, \theta_{0:t}}. \tag{6}
$$

GAPS uses the following two techniques to give an efficient approximation of $\nabla F_t(\theta_t)$:

1. **Replace the policy parameter sequence $(\theta_t)_{x(t+1)}$ by the actual sequence $\theta_{0:t}$.** This enables us to compute the gradients along the actual trajectory experienced by the online controller without re-simulating the trajectory using $\theta_t$; see Algorithm 3 for implementation details.

2. **Truncate the whole historical dependence from time step 0 to at most $B$ steps from the current time $t$.** This keeps the memory used by GAPS bounded. Strong contractive perturbation implies that the approximation error is small when $B$ is sufficiently large.

**Algorithm 1 Gradient-based Adaptive Policy Selection (GAPS)**

**Require:** A sequence of learning rates $\{\eta_t\}$, buffer length $B$, initial parameter $\theta_0$.

1. **for** $t = 0, \ldots, T-1$ **do**
2. Observe the current state $x_t$ and context $\phi_t$.
3. Pick the control action $u_t = \text{ALG}(x_t, \phi_t, \theta_t)$.
4. Incur the stage cost $f_t(x_t, u_t)$.
5. Compute the approximated gradient: $G_t = \sum_{b=0}^{\min\{B-1,t\}} \frac{\partial f_{t|0}}{\partial \theta_{t-b}} |_{x_0, \phi_{0:t}, \theta_{0:t}}$.
6. Perform the projected gradient update $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta_t G_t)$.
7. **end for**

While Algorithm 1 presents GAPS in its simplest form, it does not clarify how to efficiently compute the partial derivatives that form $G_t$. The computation can be done efficiently by maintaining a buffer of length $B$ that stores the partial derivatives computed in previous time steps. We provide a practical description of GAPS in Algorithm 3 in Appendix A.

Many previous works on selecting the clairvoyant optimal disturbance feedback controllers [2, 17, 27] take a different perspective of reducing the online control problem to OCO with Memory [6] by completely removing the dependence on policy parameters before time step $t - B$ for a fixed memory length $B$. In this reduction-based approach, one must “imaginarily” reset the state at time
$t - \tau$ to be zero and then use the $B$-step truncated multi-step cost function $f_t[t - B](0, \varphi_t[0 : B], \theta_t[0 : B])$ in the OGD with Memory algorithm [2]. When applied to our setting, this is equivalent to replacing $G_t$ in line 5 of Algorithm 1 by $G_t' = \sum_{b=0}^{B-1} \frac{\partial f_t[t - b, \theta_t[t-b]]}{\partial \theta_t[t-b]}|_{0,\varphi_t[t-b],(\theta_t[t-b])_{0:B}}$. However, the estimator $G_t'$ is less computationally efficient than $G_t$ in GAPS, especially when the controller is expensive to execute. Taking MPC in Example 2.2 as an example, to compute $G_t'$ at every time step, one must solve $B$ MPC optimization problems when re-simulating the system from time step $t - B$, where $B = \Omega(\log T)$. In contrast, computing $G_t$ in GAPS only requires solving one MPC optimization problem and doing $O(B)$ matrix multiplications to update the stored partial derivatives.

3.2 Assumptions and Main Result for GAPS

We now present our main results for GAPS and the assumptions required to derive them. Our main results contains two parts. In the first part, we show that the actual stage cost $f_t(x_t, u_t)$ incurred by GAPS is close to the ideal surrogate cost $F_t(\theta_t)$, and the approximated gradient $G_t$ is close to the ideal gradient $\nabla F_t(\theta_t)$, so that the policy update of GAPS mimics the ideal OGD update (Definition 3.1). Besides the strong contractive property in Definition 2.2, we also need additional assumptions on the stability, dynamics, controllers, and costs. Before introducing the assumption, we first define the definitions of Lipschitz and smooth functions in Definition 3.2.

**Definition 3.2.** We say a function $h : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}^{n_3}$ is $(L_1, L_2)$-Lipschitz on a set $\mathcal{D} \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ if for any $(y, z), (y', z') \in \mathcal{D}, \|h(y, z) - h(y', z')\| \leq L_1 \|y - y'\| + L_2 \|z - z'\|$. We say $h$ is $(\ell_1, \ell_2)$-smooth on $\mathcal{D}$ if $h$ is differentiable and for any $(y, z), (y', z') \in \mathcal{D}, \|\nabla h(y, z) - \nabla h(y', z')\| \leq \ell_1 \|y - y'\| + \ell_2 \|z - z'\|.$

**Assumption 3.1.** For any possible initial state $x_0$ and context sequence $\varphi_{0:T-1}$, we define the reachable state/action set of the policy class $\Theta$ as:

$$D_{\Theta,t} := \{(x_t, u_t) \mid \varphi_{0:t} \in \Phi^{t+1}, \theta_{0:t} \in \Theta^{t+1}, x_t = g_t[0](x_0, \varphi_{0:t-1}, \theta_{0:t-1}), u_t = ALG(x_t, \varphi_t, \theta_t)\}.$$  

We assume that the closed-loop dynamics induced by any possible sequence of policy parameters $\theta_{0:T-1}$ is **stable**, i.e., for every time step $t$, $D_{\Theta,t} \subseteq B_n(0, R_x) \times B_m(0, R_u)$ where $R_x$ and $R_u$ are bounded constants. We also assume the following on the dynamics $g_t$, controllers ALG, and cost functions $f_t$ respectively:

1. $g_t(x, u)$ is $(L_{g,x}, L_{g,u})$-Lipschitz and $(\ell_{g,x}, \ell_{g,u})$-smooth in $(x, u)$ on $B_n(0, R_x) \times B_m(0, R_u)$.
2. $ALG(x, y, \theta)$ is $(\ell_{ALG,x}, \ell_{ALG,\theta})$-Lipschitz and $(\ell_{ALG,x}, \ell_{ALG,\theta})$-smooth in $(x, u)$ on $B_n(0, R_x) \times \Theta$.
3. $f_t(x, u)$ is $(L_f, L_f)$-Lipschitz in $(x, u)$ and $(\ell_{f,x}, \ell_{f,u})$-smooth in $(x, u)$ on $B_n(0, R_x) \times B_m(0, R_u)$.

The most important part of Assumption 3.1 is the stability of the closed-loop dynamics induced by any possible sequence of parameters $\theta_{0:t}$. The goal of this assumption is to ensure the system will stay stable no matter how we explore over the policy parameter set $\Theta$. We show that Examples 2.1 and 2.2 satisfy this assumption in Appendix C. Actually, we only need the stability for slowly time-varying parameter sequence $\theta_{0:t}$ when the learning rate $\eta$ is small, and in simulations we find that GAPS can still perform well in a setting where this stability assumption is violated (Section 5.2). Further relaxing this assumption is a direction of our future work.

Given that the stability assumption ($D_{\Theta,t} \subseteq B_n(0, R_x) \times B_m(0, R_u)$) holds, the other three assumptions on dynamics, controllers, and costs are all standard and can be easily satisfied even for nonlinear systems, such as the inverted pendulum in Section 5.2, and/or nonconvex (but smooth) cost functions.

The second part of our main result is on the time-averaged adaptive policy regret of GAPS. To derive this result, we need additional assumptions on the policy class and the multi-step cost functions, which we summarize in Assumption 3.2.
Assumption 3.2. The policy parameter set $\Theta \subset \mathbb{R}^d$ is a convex compact set with diameter $\leq D$, and the surrogate stage cost $F_t: \Theta \rightarrow \mathbb{R}$ in Definition 2.3 is convex for all time step $t$.

Both Assumptions 3.1 and 3.2 are satisfied by many common controller classes and systems including the disturbance feedback controllers (Example 2.1) and MPC with Confidence Coefficients (Example 2.2) with some moderate additional assumptions. We provide a detailed proof in Appendix C.

This convexity assumption is also required by previous works [2, 27]. Compared with the policy regret of $O(T)$, guarantees that the ideal OGD update with constant step size $\eta$ achieves the adaptive regret bound $O(\sqrt{T})$ [26]. By taking the biases on the stage costs and the gradients into consideration (see Section 3.3 for details), we derive the regret bound in Theorem 3.1.

Theorem 3.1. Suppose strong contractive perturbation (Definition 2.2) holds for $(C_0, \rho)$. Consider GAPS (Algorithm 1) with learning rate $\eta_t = \eta$ and buffer size $B$. Let $\{(x_t, u_t, \theta_t)\}_{t \in [0: T-1]}$ denote the trajectory of Algorithm 1. Under Assumption 3.1, the following inequalities hold:

$$|f_t(x_t, u_t) - F_t(\theta_t)| = O((1 - \rho)^{-3}\eta), \text{ and } \|G_t - \nabla F_t(\theta_t)\| = O((1 - \rho)^{-5}\eta + (1 - \rho)^{-1}\rho B).$$

(7)

If Assumption 3.2 also holds, then Algorithm 1 achieves the adaptive regret bound

$$R^A(T) = O\left((1 - \rho)^{-5}\eta + (\eta T)^{-1} + (1 - \rho)^{-1}\rho B + (1 - \rho)^{-10}\eta^3 + (1 - \rho)^{-2}\rho^2 B^2\eta\right).$$

(8)

Further, if we assume $T \gg \frac{1}{1 - \rho}$ and set $\eta = (1 - \rho)^{\frac{5}{2}}T^{-\frac{1}{2}}$ and $B \geq \frac{1}{2}\log(T)/\log(1/\rho)$, the time-averaged regret is $R^A(T) = O((1 - \rho)^{-\frac{5}{2}}T^{-\frac{1}{2}})$.

Note that the big-$O$ notation in Theorem 3.1 hides coefficients that depend on the Lipschitz and smoothness constants defined in Assumption 3.1, the constant $D$ in Assumption 3.2, and $C_0$. As a remark, the first part of our result (7), which says that GAPS mimics the update of ideal OGD, only relies on Assumption 3.1. Therefore, it also applies to nonlinear dynamics and/or nonconvex cost functions. However, the second part of our result (8) requires the multi-step costs to be convex. This convexity assumption is also required by previous works [2, 27]. Compared with the policy regret bounds derived in [2, 27], our regret bound is tighter by a factor of $\log T$. This improvement is enabled by doing a more careful analysis of how a past policy $\theta_{t-B}$ ($b \in [0 : B - 1]$) impacts the current stage cost rather than treating $\theta_{t-B+1:t}$ as a whole. The contractive perturbation bounds on the partial derivatives $\frac{\partial f_{t,b}}{\partial \theta_{t,b}}$ avoids the extra factor of $B = O(\log T)$ as in [2, 27].

A limitation of our regret bound is that Assumption 3.2 requires the ideal cost function $F_t$ to be convex, which rules out many nonlinear dynamics and nonconvex stage costs. An interesting future direction is to relax the convexity assumption, possibly using recent works on the regret bounds for online nonconvex optimization [29, 30].

3.3 Proof Outline for GAPS

In this section, we provide a proof outline of the adaptive regret bound for GAPS. As we discussed in Section 3.1, the intuition behind GAPS is to mimic the ideal OGD update $\theta_{t+1} = \prod_\Theta(\theta_t - \eta_t \nabla F_t(\theta_t))$ with limited memory buffer and computational complexity. While the existing literature of OCO guarantees that the ideal OGD update with constant step size $\eta = \Theta(1/\sqrt{T})$ achieves a time-averaged policy regret of $O\left(1/\sqrt{T}\right)$, GAPS incurs an approximation error at every time step since it uses...
G_t (Algorithm 1 Line 5) instead of \( \nabla F_t(\theta_t) \) to implement gradient descent. We characterize how a per-step bias in the gradient estimation may affect the regret guarantee of the OGD in Theorem 3.2.

**Theorem 3.2.** Consider the update rule \( \theta_{t+1} = \prod_0(\theta_t - \eta G_t) \). If Assumption 3.2 holds, and \( \|\nabla F_t(\theta)\| \leq W \) for all \( \theta \in \Theta \), and \( \|\nabla F_t(\theta_t) - G_t\| \leq \epsilon \) holds for all time steps \( t \), we have that

\[
\frac{1}{T} \sup_{t \in [T]} \left( \sum_{\tau=t}^{T-1} F_t(\theta_t) - \min_{\theta \in \Theta} \sum_{\tau=t}^{T-1} F_t(\theta_t) \right) \leq \epsilon D + (W^2 + \epsilon^2)\eta + \frac{D^2}{2\eta T}.
\]

If \( \eta = D/(\sqrt{2(W^2 + \epsilon^2)}T) \), the right hand side can be upper bounded by \( (\sqrt{2(W^2 + \epsilon^2)}T + \epsilon)D \).

With Theorem 3.2, obtaining \( O(1/\sqrt{T}) \) time-averaged policy regret for GAPS reduces to showing:

\[
|F_t(x_t, u_t) - F_t(\theta_t)| = O(1/\sqrt{T}), \quad \text{and} \quad \|\nabla F_t(\theta_t) - G_t\| = O(1/\sqrt{T}),
\]

where we only consider the order of magnitude with respect to the horizon \( T \) for clarity here. As we show in Theorem 3.4 and Theorem 3.5, both of these quantities are in the order of \( O(\eta) \) when GAPS adopts a constant learning rate of \( \eta \). To obtain these results, we first show a lemma about the contractive property of the partial derivatives of the multi-step dynamics.

**Lemma 3.3 (Lipschitzness/Smoothness of the Multi-Step Dynamics).** Suppose strong contractive perturbation (Definition 2.2) holds for \( (C_0, \rho) \). Given two time steps \( t > \tau \), under Assumption 3.1, for any \( x_t, x_{\tau} \in D_{\theta_{\tau}} \) and \( \theta_t, \theta_{\tau} \in \Theta, \theta_{t+1:1-1} \in \Theta^{t-1} \), the multi-step dynamical function \( g_{t|\tau}(x_t, \varphi_{t-1}, \theta_{t-1}) \) is \( O(\rho^{t-\tau}) \)-smooth in \( x_t \) (and \( \theta_t \)) when all other variables are fixed, i.e.,

\[
\left\| \frac{\partial g_{t|\tau}}{\partial x_t} |_{x_t, \theta_{t-1}} \right\| \leq C_{L,g,x} \rho^{t-\tau}, \quad \left\| \frac{\partial g_{t|\tau}}{\partial \theta_t} |_{x_t, \theta_{t-1}} \right\| \leq C_{L,g,\theta} \rho^{t-\tau},
\]

\[
\left\| \frac{\partial g_{t|\tau}}{\partial x_t} |_{x_t, \theta_{t+1}, \theta_{t+1-1}} \right\| \leq C_{L,g,(x,\theta)} \rho^{t-\tau} \left\| x_t - x_\tau \right\| + C_{L,g,(x,\theta)} \rho^{t-\tau} \left\| \theta_t - \theta_\tau \right\|,
\]

\[
\left\| \frac{\partial g_{t|\tau}}{\partial \theta_t} |_{x_t, \theta_{t+1}, \theta_{t+1-1}} \right\| \leq C_{L,g,(\theta,\theta)} \rho^{t-\tau} \left\| x_t - x_\tau \right\| + C_{L,g,(\theta,\theta)} \rho^{t-\tau} \left\| \theta_t - \theta_\tau \right\|,
\]

where \( C_{L,g,x} = C_0, C_{L,g,\theta} = \frac{C_{L_u,g,L_{ALG,\theta}}}{\rho} \), and

\[
C_{L,g,(x,\theta)} = ((1 + L_{ALG,x})(\ell_{x} + \ell_{u} L_{ALG,x}) + L_{x}(\ell_{ALG,x}) C_0 \rho^{-1}(1 - \rho)^{-1},
\]

\[
C_{L,g,(\theta,\theta)} = ((1 + L_{ALG,\theta})(\ell_{x} + \ell_{u} L_{ALG,\theta}) + L_{x}(\ell_{ALG,\theta}) C_0 \rho^{-1}(1 - \rho)^{-1},
\]

\[
(1 + L_{ALG,\theta}) \ell_{u} L_{ALG,\theta} + L_{u}(\ell_{ALG,\theta}) C_0 \rho^{-1}(1 - \rho)^{-1},
\]

\[
C_{L,g,(\theta,\theta)} = ((1 + L_{ALG,\theta})(\ell_{x} + \ell_{u} L_{ALG,x}) + L_{x}(\ell_{ALG,\theta}) + L_{u}(\ell_{ALG,\theta}) C_0 \rho^{-2}(1 - \rho)^{-1},
\]

\[
(1 + L_{ALG,\theta}) \ell_{u} L_{ALG,\theta} + L_{u}(\ell_{ALG,\theta}) C_0 \rho^{-2}(1 - \rho)^{-1},
\]

\[
C_{L,g,(x,\theta)} = ((1 + L_{ALG,x})(\ell_{x} + \ell_{u} L_{ALG,x}) + L_{x}(\ell_{ALG,\theta}) \ell_{u} L_{ALG,\theta} + L_{u}(\ell_{ALG,\theta}) \ell_{u} L_{ALG,\theta} C_0 \rho^{-2}(1 - \rho)^{-1},
\]

\[
(1 + L_{ALG,\theta}) \ell_{u} L_{ALG,\theta} + L_{u}(\ell_{ALG,\theta}) \ell_{u} L_{ALG,\theta} C_0 \rho^{-2}(1 - \rho)^{-1},
\]

Intuitively, Lemma 3.3 shows that the dependence of the state \( x_t \) on the previous state \( x_\tau \) and \( \theta_\tau \) decays exponentially with respect to their time distance \( t - \tau \). Specifically, recall that the multi-step dynamics \( g_{t|\tau} \) writes \( x_t \) as a function of \( x_\tau \) and \( \theta_{t-1} \). When other variables are fixed, the Lipschitzness and smoothness constants with respect to \( x_t \) and \( \theta_t \) are both \( O(\rho^{t-\tau}) \), where \( \rho \) is the decay factor specified in Assumption 3.1 and Definition 2.2. While the contractive Lipschitzness on \( x_t \) is automatically guaranteed by strong contractive perturbation (Definition 2.2), we use this property and the chain rule decomposition to show the Lipschitzness on \( \theta_t \) and the smoothness.
where we use the notation \( x'_{t'} = g_{t'}(x_t, \theta_{t+1}; x_{t-1}) \) and \( x''_{t'} = g_{t'}(x'_{t'}, \theta_{t+1}; x_{t-1}) \) for \( t' \in [t+1, t-1] \).

Note that for any \( i \in [1 : t - \tau] \) and any \( \theta_{t-i} \in \Theta \), we have the decomposition

\[
\frac{\partial g_{t-i+1}[r-i]}{\partial x_{t-i}} |_{x_{t-i}, \theta_{t-i}} - \frac{\partial g_{t-i+1}[r-i]}{\partial x_{t-i}} |_{x'_{t-i}, \theta'_{t-i}}
\]

\[
= \frac{\partial g_{t-i}}{\partial x_{t-i}} |_{x_{t-i}, u_{t-i}} - \frac{\partial g_{t-i}}{\partial x_{t-i}} |_{x'_{t-i}, u'_{t-i}} + \frac{\partial g_{t-i}}{\partial u_{t-i}} |_{x_{t-i}, u_{t-i}} \frac{\partial \text{ALG}}{\partial x_{t-i}} |_{x_{t-i}, \theta_{t-i}} - \frac{\partial g_{t-i}}{\partial u_{t-i}} |_{x'_{t-i}, u'_{t-i}} \frac{\partial \text{ALG}}{\partial x_{t-i}} |_{x'_{t-i}, \theta'_{t-i}}
\]

\[
+ \frac{\partial g_{t-i}}{\partial u_{t-i}} |_{x'_{t-i}, u'_{t-i}} \left( \frac{\partial \text{ALG}}{\partial x_{t-i}} |_{x_{t-i}, \theta_{t-i}} - \frac{\partial \text{ALG}}{\partial x_{t-i}} |_{x'_{t-i}, \theta'_{t-i}} \right),
\]

where we use the notation \( u_{t-i} = \text{ALG}(x_{t-i}, \theta_{t-i}), u'_{t-i} = \text{ALG}(x'_{t-i}, \theta'_{t-i}) \). Taking norms on both sides of the equation and applying the triangle inequality give that

\[
\left\| \frac{\partial g_{t-i+1}[r-i]}{\partial x_{t-i}} |_{x_{t-i}, \theta_{t-i}} - \frac{\partial g_{t-i+1}[r-i]}{\partial x_{t-i}} |_{x'_{t-i}, \theta'_{t-i}} \right\|
\]

\[
\leq \ell_{g,x} \left\| x_{t-i} - x'_{t-i} \right\| + \ell_{g,u} \left\| \text{ALG}(x_{t-i}, \theta_{t-i}) - \text{ALG}(x'_{t-i}, \theta'_{t-i}) \right\|
\]

\[
+ L_{\text{ALG},x} \left( \ell_{g,x} \left\| x_{t-i} - x'_{t-i} \right\| + \ell_{g,u} \left\| \text{ALG}(x_{t-i}, \theta_{t-i}) - \text{ALG}(x'_{t-i}, \theta'_{t-i}) \right\| \right)
\]

\[
+ L_{g,x} \cdot \left( \ell_{g,x} \left\| x_{t-i} - x'_{t-i} \right\| + \ell_{g,u} \left\| \text{ALG}(x_{t-i}, \theta_{t-i}) - \text{ALG}(x'_{t-i}, \theta'_{t-i}) \right\| \right)
\]

\[
\leq (1 + L_{\text{ALG},x}) \left( \ell_{g,x} + \ell_{g,u} \cdot L_{\text{ALG},x} \right) + L_{g,x} \cdot \ell_{g,x} \left\| x_{t-i} - x'_{t-i} \right\|
\]

\[
+ ((1 + L_{\text{ALG},x}) \ell_{g,u} \cdot L_{\text{ALG},\theta} + L_{g,u} \cdot \ell_{g,x} \cdot L_{\text{ALG},x}) \left\| \theta_{t-i} - \theta'_{t-i} \right\|
\]

\[
\leq \left( (1 + L_{\text{ALG},x}) \ell_{g,x} + \ell_{g,u} \cdot L_{\text{ALG},x} \right) \left\| x_{t-i} - x'_{t-i} \right\|
\]

\[
+ \left( (1 + L_{\text{ALG},x}) \ell_{g,u} \cdot L_{\text{ALG},\theta} + L_{g,u} \cdot \ell_{g,x} \cdot L_{\text{ALG},x} \right) \left\| \theta_{t-i} - \theta'_{t-i} \right\|
\]

where we use Assumption 3.1 and the definition of \( u_{t-i}, u'_{t-i} \) in (11a); and Assumption 3.1 in (11b). Therefore, by (10) and (11), we see that
where we use the strong contractive perturbation property and (11) in (12a); we use the first two inequalities to bound \( \|x_{t-i} - x'_{t-i}\| \leq C_0 \rho^{t-i} \|x_t - x'_t\| + \frac{C_{d, u}}{\rho} \cdot \rho^{t-i} \|\theta_t - \theta'_t\| \) in (12b).

Note that for any \( t < t \), the previous state \( x_t \) and previous policy parameter \( \theta_t \) can only affect the current stage cost \( f_t \) by affecting the current state \( x_t \). Thus, the contractive Lipschitzness/smoothness of the multi-step cost function \( f_{t|t} \) is an implication of Lemma 3.3. We formalize this result in Corollary 3.3.1 and defer the detailed proof to Appendix E.

**Corollary 3.3.1 (Lipschitzness/Smoothness of the Multi-Step Costs).** Suppose strong contractive perturbation (Definition 2.2) holds for \( (C_0, \rho) \). Given two time steps \( t \geq t \), under Assumption 3.1, for any \( x_t, x'_t \in D_{\Theta_t} \) and \( \theta_t, \theta'_t \in \Theta, \theta_{t+1:t-1} \in \Theta^{t-1} \), the multi-step cost function \( f_{t|t}(x_t, \theta_{t:t-1}, \theta_t-1) \) is \( (\rho^{t-i}) \)-smooth in \( x_t \) and \( \theta_t \) when all other variables are fixed, i.e.,

\[
\left\| \frac{\partial f_{t|t}}{\partial x_t} \right\|_{x_t, \theta_t} \leq C_{L,f,x} \rho^{t-i}, \quad \left\| \frac{\partial f_{t|t}}{\partial \theta_t} \right\|_{x_t, \theta_t} \leq C_{L,f,\theta} \rho^{t-i},
\]

\[
\left\| \frac{\partial f_{t|t}}{\partial x_t} \right\|_{x_t, \theta_t, \theta_{t+1:t}} \leq C_{L,f,(x, \theta)} \rho^{t-i} \left\| x_t - x'_t \right\| + C_{f,(x, \theta, \theta_{t+1:t})} \rho^{t-i} \left\| \theta_t - \theta'_t \right\|,
\]

where \( C_{L,f,x} = L_f C_0 (1 + L_{\text{ALG}, x}), C_{L,f,\theta} = L_f \max \{C_{L,g,\theta}(1 + L_{\text{ALG}, \theta}), L_{\text{ALG}, \theta}\} \), and

\[
C_{f,(x, \theta)} = L_f (1 + L_{\text{ALG}, x}) C_{L,g,(x, \theta)} + ((\ell_f, x + L_{\text{ALG}, x})(1 + L_{\text{ALG}, x}) + L_{\text{ALG}, x} C_{L,g,x}) C^2_{L,g,x},
\]

\[
C_{f,(x, \theta, \theta_{t+1:t})} = L_f (1 + L_{\text{ALG}, x}) C_{L,g,(x, \theta, \theta_{t+1:t})} + ((\ell_f, x + \ell_f, u_{\text{ALG}, x})(1 + L_{\text{ALG}, x}) + L_{\text{ALG}, x} C_{L,g,x} C_{L,g,\theta}) C^2_{L,g,x} C_{L,g,\theta},
\]

With the help of Lemma 3.3 and Corollary 3.3.1, we show the first target of (9) in Theorem 3.4. This inequality bounds the difference between the actual stage cost \( f_t(x_t, u_t) \) incurred by GAPS and the ideal cost \( F_t(\theta_t) \). Besides this inequality, in Theorem 3.4, we also bound the distance between GAPS’ trajectory and the imaginary trajectory if the same policy parameter \( \theta_t \) had been used from time 0 to time \( t \), which will be useful for showing Theorem 3.5 later in this section.

**Theorem 3.4.** Suppose strong contractive perturbation (Definition 2.2) holds for \( (C_0, \rho) \). Under Assumption 3.1, let \( x_{0:T-1}, u_{0:T-1} \) and \( \theta_{0:T-1} \) be the states, control inputs, and policy parameters on the trajectory of Algorithm 1 with parameter \( \eta_t \equiv \eta \). Then, both \( \|G_t\| \) and \( \|F_t(\theta_t)\| \) are upper bounded by \( \frac{C_{L,f,\theta}}{1-\rho} \), and the following inequalities holds for any two time steps \( t, t' \) (\( t \leq t' \)):

\[
\|\theta_t - \theta_t\| \leq \frac{C_{L,f,\theta}}{1-\rho} \cdot (t - t') \eta_t, \quad \text{and} \quad \|x_t - \hat{x}_t(\theta_t)\| \leq \frac{C_{L,f,\theta}}{1-\rho} \cdot \left(\frac{t - t'}{1-\rho}\right) \cdot \eta_t,
\]

where recall that \( \hat{x}_t(\theta) := g_{0,t}(x_0, \theta_0, t, \theta_{x(t+1)}), \forall \theta \in \Theta \). Further, we have that

\[
\|f_t(x_t, u_t) - F_t(\theta_t)\| \leq C_{f,\theta}^2 (1 - \rho)^{-2} \cdot \eta_t.
\]
To show Theorem 3.4, we first derive a uniform upper bound on the norm the estimated gradient $G_t$, which implies that the policy parameter sequence does not vary too quickly, i.e. it is in the same order as the constant learning rate $\eta$. We then leverage strong contractive perturbation to bound $\|x_t - \hat{x}_t(\theta_t)\|$ and use it to bound $|f_t(x_t, u_t) - F_t(\theta_t)|$ by the Lipschitzness of $f_t$. To highlight the key technical step, we discuss here bounding $\|x_t - \hat{x}_t(\theta_t)\|$ using $\|\theta_t - \theta_t\|$, and defer the complete proof to Appendix F.

Recall that we have $x_t = g_{t|0}(x_0, \theta_{0:T-1}), \hat{x}_t(\theta_t) =: g_{t|0}(x_0, (\theta_t)_{x_T})$. Therefore:

\[
\|x_t - \hat{x}_t(\theta_t)\| = \|g_{t|0}(x_0, \theta_{0:T-1}) - g_{t|0}(x_0, (\theta_t)_{x_T})\|
\]

\[
\leq \sum_{t'=0}^{t-1} \|g_{t'|0}(x_0, \theta_{0:t'}, (\theta_t)_{x_{t'-1}}) - g_{t'|0}(x_0, \theta_{0:t'-1}, (\theta_t)_{x_{t'-1}})\| \tag{13a}
\]

\[
\leq \sum_{t'=0}^{t-1} \|g_{t'|t'}(x_t, \theta_t, (\theta_t)_{x_{t'-1}}) - g_{t'|t'}(x_{t'}, \theta_{t'}, (\theta_t)_{x_{t'-1}})\| \tag{13b}
\]

\[
\leq \sum_{t'=0}^{t-1} C_{L,g,\theta} \rho^{t-t'} \|\theta_t - \theta_{t'}\| \tag{13c}
\]

\[
\leq C_{L,f,\theta} \rho C_{L,g,\theta} \eta \frac{t-1}{1-\rho} \sum_{t'=0}^{t-1} (t-t') \rho^{t-t'} \leq C_{L,f,\theta} \rho C_{L,g,\theta} \eta \left( (1-\rho) \sum_{t'=0}^{t-1} \rho^{t-t'} \right) \|F_t(\theta_t)\| \tag{13d}
\]

where we use the triangle inequality in (13a), the definition of multi-step dynamics in (13b), Lemma 3.3 in (13c), and the bound on parameter difference in (13d).

In Theorem 3.5 below, we bound the difference between the estimated gradient $G_t$ used by GAPS and the ideal gradient $\nabla F_t(\theta_t)$ used by the ideal OGD.

**Theorem 3.5 (Gradient Bias).** Suppose strong contractive perturbation (Definition 2.2) holds for $(C_0, \rho)$. Under Assumption 3.1, let $x_{0:T}$ and $\theta_{0:T}$ be the states and policy parameters on the trajectory of Algorithm 1 with $\eta_t \equiv \eta$. Then, the following holds for all $t \leq T$:

\[
\left\| \frac{\partial f_{t|0}}{\partial \theta_t} |_{x_0, \theta_{0:T}} - \frac{\partial f_{t|0}}{\partial \theta_t} |_{\theta_t(x_t), (\theta_t)_{x_{t+1}}} \right\| \leq \left( \hat{C}_0 + \hat{C}_1 (t-\tau) + \hat{C}_2 (t-\tau)^2 \right) \rho^{t-\tau} \cdot \eta,
\]

for $\hat{C}_0 = \frac{\rho_{L,f,\theta} C_{L,g,\theta} C_{L,f,\theta}}{(1-\rho)^2}$, $\hat{C}_1 = \left( 1-\rho \right) C_{L,f,\theta} C_{L,g,\theta} C_{L,f,\theta} + \rho_{L,f,\theta} C_{L,g,\theta} C_{L,f,\theta}$, $\hat{C}_2 = \frac{C_{L,f,\theta} C_{L,g,\theta} C_{L,f,\theta}}{1-\rho}$. Next,

\[
\|G_t - \nabla F_t(\theta_t)\| \leq \left( \hat{C}_0 (1-\rho)^{-1} + (\hat{C}_1 + \hat{C}_2) (1-\rho)^{-2} + \hat{C}_2 (1-\rho)^{-3} \right) \eta + C_{L,f,\theta} (1-\rho)^{-1} \cdot \rho^B.
\]

where recall that $F_t(\theta) := f_{t|0}(x_0, \varphi_{0:t}, \theta_{x_{t+1}})$, $\forall \theta \in \Theta$.

The key technique we used to show Theorem 3.5 is a sequential decomposition of the error based on the triangle inequality. Specifically, note that Corollary 3.3.1 only allows us to compare the partial derivatives when $\theta_{t+1}$ are fixed and the only perturbations are on $x_t$ and $\theta_t$. To compare the partial derivatives realized on two trajectory instances $(\hat{x}_t(\theta_t), (\theta_t)_{x_{t+1}})$ and $(x_t, \theta_{t+1})$, we change the parameters sequentially one by one, following the path

$(\hat{x}_t(\theta_t), (\theta_t)_{x_{t+1}}) \rightarrow (x_t, \theta_t, (\theta_t)_{x_{t+1}}) \rightarrow (x_t, \theta_{t+1}, (\theta_t)_{x_{t+1}}) \rightarrow \cdots \rightarrow (x_t, \theta_{t+1})$.

Specifically, we use the triangle inequality to do the decomposition

\[
\left\| \frac{\partial f_{t|0}}{\partial \theta_t} |_{x_0, \theta_{0:T}} - \frac{\partial f_{t|0}}{\partial \theta_t} |_{\theta_t(x_t), (\theta_t)_{x_{t+1}}} \right\| \leq \left\| \frac{\partial f_{t|t+1}}{\partial \theta_t} |_{x_t, \theta_t, (\theta_t)_{x_{t+1}}} - \frac{\partial f_{t|t+1}}{\partial \theta_t} |_{\hat{x}_t, (\theta_t)_{x_{t+1}}} \right\|.
\]
Online Adaptive Controller Selection

We now consider the problem of selecting the clairvoyant optimal controller from a finite (discrete) set \( \Theta = \{1, 2, \ldots, k\} \), which requires significantly different tools than the continuous case. To see this, note that the design of GAPS depends on the property that two similar parameters \( \theta, \theta' \approx \theta'' \) would give similar trajectories and incur similar costs. Therefore, we can approximately evaluate the performance of (an updated) \( \theta' \) on the realized trajectory of (the current) \( \theta \), as long as the learning rate to update \( \theta' \) from \( \theta \) is sufficiently small. However, such a structural property based on continuity does not exist when \( \Theta = \{1, 2, \ldots, k\} \) and each parameter represents an independent controller candidate. Evaluating one parameter does not provide additional information for any other parameter. To address the challenge in this setting, we propose the Bandit-based Adaptive Policy Selection (BAPS, Algorithm 2) that can be applied when \( \Theta = \{1, \ldots, k\} \) with a finite \( k \).

### Algorithm 2 Bandit-based Adaptive Policy Selection (BAPS)

**Require:** Policy parameter set \( \{1, \ldots, k\} \), batch size \( b \), learning step size \( \eta \).

1. Set the initial distribution to be the uniform distribution \( s_0 = (1/k)_{\chi_k} \).
2. for \( m = 0, \ldots, T/b - 1 \) do
3. Sample \( j_m \sim s_m \) and for all \( t \in [mb : (m + 1)b - 1] \), pick \( u_t = \text{ALG}(x_t, q_t, j_m) \).
4. Compute the batched loss vector \( \hat{p}_m \in \mathbb{R}^k \):
   \[
   \hat{p}_m(j) = \begin{cases} 
   \frac{1}{s_m(j_m)} \cdot \sum_{t=mb}^{(m+1)b-1} f_t(x_t, u_t), & j = j_m \\
   0, & \text{otherwise}.
   \end{cases}
   \]
5. Set \( s_{m+1}(j) = s_m(j) e^{-\eta \hat{p}_m(j)}, \forall j \in \{1, \ldots, k\} \) and do the normalization \( s_{m+1} \leftarrow \frac{s_{m+1}}{\|s_{m+1}\|_1} \).
6. end for

The intuition is that whenever we switch the controller parameter, we wait until it “mixes” well before evaluating its performance. Specifically, we use the same policy parameter for a batch of \( b \).
time steps before possible switches to other policy parameters. By enforcing these batches and choosing a sufficiently large batch size $b$, we can guarantee that the difference between the total actual cost over a batch and the total surrogate cost ($F_b$) is negligible compared with the batch size $b$. To achieve this, we not only need weak contractive perturbation (Definition 2.2), but also an assumption on the stability of the controller candidates, which we state formally in Assumption 4.1.

**Assumption 4.1.** Suppose $\Theta = \{1, \ldots, k\}$. There exists some positive integer $\tau_0$ and positive constants $R'_x \geq R_x > 0$ such that for any time step $t$ and policy parameter $\theta$, if $x_t \in B_n(0, R_x)$, we have

$$g_{\tau}(x_t, \varphi_{t, \tau - 1}, \theta_{x_t(t - \tau)}) \in \begin{cases} B_n(0, R'_x) & \text{if } t \leq \tau < t + \tau_0, \\ B_n(0, R_x) & \text{if } t + \tau_0 \leq \tau, \end{cases}$$

for all $\varphi_{t, \tau - 1} \in \Phi^{t - \tau}$.

Further, for any $x, x' \in B(0, R'_x)$, $\varphi, \Omega, \theta \in \Theta$, the following inequalities hold for all time step $t$:

$$|f_{\tau}(x, \varphi, \theta)| \leq D_0, \text{ and } |f_{\tau}(x, \varphi, \theta) - f_{\tau}(x', \varphi, \theta)| \leq L_0||x - x'||.$$

Intuitively, Assumption 4.1 assumes there is a bounded ball $B_n(0, R_x)$ such that, if it starts from any state inside this ball at time $t$, the controller with a fixed parameter $\theta$ can come back into this bounded ball and never leave again in no more than $\tau_0$ steps. We also require that in the process of coming back into the ball $B_n(0, R_x)$, the controller never leaves a larger ball $B(0, R'_x)$ with $R'_x \geq R_x$. A simple example that satisfies Assumption 4.1 is Example 2.3 we discussed in Section 2.3. In this case, we can set $R_x = 1$ and $R'_x = C_0$, with $\tau_0 = -\log C_0'/\log \rho'_0$.

With Assumption 4.1 in hand, we present our main result for BAPS in Theorem 4.1.

**Theorem 4.1.** Suppose Assumption 4.1 and weak contractive perturbation holds for $(C_0, \rho)$. Then, the expected time-averaged policy regret of Algorithm 2 with batch size $b \geq \tau_0$ and learning step size $\eta > 0$ satisfies:

$$\mathbb{E}[R(T)] \leq C_0D_0 \cdot (1 - \rho)^{-1}b^{-1} + kD_0\eta b + \log k \cdot \eta^{-1}T^{-1}.$$

Furthermore, if we set $b = \left(\frac{C_0^2D_0T}{(1 - \rho)^2k \log k}\right)^{\frac{1}{2}}$ and $\eta = \left(\frac{(1 - \rho)(\log k)^2}{C_0D_0kT^2}\right)^{\frac{1}{2}}$, we have $\mathbb{E}[R(T)] = 3\left(\frac{C_0D_0^2k \log k}{(1 - \rho)T}\right)^{\frac{1}{2}}$.

The proof of Theorem 4.1 is inspired by the proof of Lemma 6.3 in [26]. We defer the details to Appendix I. Note that the $O(T^{-1/3})$ time-averaged static policy regret bound is worse than the $O(T^{-1/2})$ bound for GAPS in Theorem 3.1 in the dependence on the horizon $T$, but it already matches the regret lower bound for MAB with switching costs [20], which we conjectured to be the parallel online learning setting for online controller selection. The intuition is that the possible extra cost incurred when we wait for the trajectory to converge to $\hat{x}_t(\theta)$ after switching to a new policy parameter $\theta$ can be modeled as an unknown switching cost with a constant upper bound. An interesting open question is verify this intuition and see whether the dependence of Theorem 4.1 on $T$ is tight without any convexity or continuity assumptions and knowledge of the dynamics.

5 NUMERICAL EXPERIMENTS

We present three simulation experiments to validate our methods. First, Section 5.1 compares GAPS to a recently proposed follow-the-leader (FTL)-type method [35] for tuning a scalar confidence parameter in online control with predictions. Our results show that GAPS can adapt more quickly to a sudden change in prediction accuracy. Second, Section 5.2 compares GAPS to a realistic baseline for tuning feedback controller gains in a time-varying nonlinear system. Although this setting violates multiple assumptions of our regret bound, GAPS still performs well, outperforming the baseline when the disturbance is not i.i.d.. Finally, Section 5.3 considers model-predictive control when near-future disturbance predictions are more accurate than far-future ones. We compare BAPS and GAPS, highlighting that BAPS can effectively select between full-confidence MPC controllers.
of different horizons, and illustrating the efficiency that GAPS gains by exploiting the problem structure.

In this section, we will use the concept of LQR, which refers to the infinite-horizon discrete-time linear quadratic regulator optimal control problem, defined by LTI dynamics \( x_{t+1} = Ax_t + Bu_t \) and objective \( \sum_{t=0}^{\infty} q(x_t, Q) + q(u_t, R) \), where \( Q \succeq 0, R > 0 \).

5.1 Adapting a scalar confidence parameter

We compare our GAPS method against the “robust and consistent” method of Li et al. [35], which tunes a scalar confidence parameter for model-predictive control in a LTI system with imperfect disturbance predictions. The setting of [35] is similar to Example 2.2 but has several restrictions:
1) the cost matrices \( Q, R \) are time-invariant, 2) the disturbance predictions \( \{ \bar{w}_t \}_{t=0}^{T-1} \) are static and known before the game begins, and 3) the controller is parameterized by a single confidence parameter \( \lambda \in [0, 1] \). Li et al. [35] propose a “follow-the-leader”-style algorithm where the entire trajectory history is used to select \( \lambda \) at each time step (described in Appendix J.1 for completeness).

We compare the ability of each algorithm to quickly adapt to a change in prediction accuracy. We consider the unstable scalar system \( x_{t+1} = 2x_t + u_t + w_t \) with the LQR costs \( Q = R = 1 \). We construct sinusoidal disturbances \( w_t \) and noise-corrupted disturbance predictions \( \tilde{w}_t \) following

\[
    w_t = \sin(t/10), \quad \tilde{w}_t = w_t + n_t, \quad n_t \sim \begin{cases} \text{Uniform}(−2, 2) & : t \leq T/4 \\ \text{Uniform}(−0.02, 0.02) & : \text{otherwise}. \end{cases}
\]

In other words, the predictions are initially noisy but become accurate after \( t = T/4 \).

Figure 1 shows the values of the confidence parameter \( \lambda \) (left) generated by each algorithm, along with the per-timestep LQR costs \( x_t^2 + u_t^2 \) (right). Both methods are initialized with \( \lambda = 1 \), which is not optimal for the large prediction noise. The method of Li et al. [35] rapidly adjusts to an appropriate confidence level at first. However, when the predictions become accurate, our method adapts more quickly and obtains smaller LQR costs towards the end of the simulation.

5.2 Nonlinear system

We evaluate the performance of GAPS for tuning feedback control gains in a nonlinear dynamical system. Our regret bounds do not apply to this setting. The goal here is to test if GAPS can still behave well in such settings. We compare with a strong LQR baseline built from standard optimal control methods. We find that GAPS nearly matches the baseline performance when the system is disturbed by i.i.d. noises, and significantly outperforms the baseline when the disturbance is a random walk.
5.2.1 Setting. We consider a standard inverted pendulum system. The state is \( x = (\phi, \dot{\phi}) \), where \( \phi \) is the pivot angle and \( \dot{\phi} \) is its velocity. The input \( u \) is a torque about the pivot point. The dynamics are time-varying due to a changing mass parameter \( m \). Details on the inverted pendulum are given in Appendix J.2. We apply GAPS to tune the (positive) parameters \( \theta = (k_p, k_d) \) of a linear feedback controller \( u = -k_p \phi - k_d \dot{\phi} \), also known as a proportional-derivative (PD) controller. Appendix J.2.1 verifies that this setting violates strong contractive perturbation. We simulate time-varying pendulum mass \( m_t \) with regularly-spaced step changes. We measure controller performance by the LQR cost \( f_t(x_t, u_t) = \Delta_t \cdot (\dot{\phi}_t^2 + \phi_t^2 + 0.1u_t^2) \), where \( \Delta_t \) is the discretization interval.

We compare GAPS to a natural baseline based on an infinite-horizon LQR that deploys the policy \( u_t = K^*(m_t)x_t \), where \( K^* \) is the optimal PD controller for the infinite-horizon summation of the LQR cost above for the linearized and discretized pendulum dynamics with mass \( m_t \). Although this is not the exact offline optimal for the finite-horizon time-varying problem, it is a good approximation when mass \( m_t \) changes infrequently, and is practical and realistic.

We model disturbances via an Ornstein-Uhlenbeck (O-L) random walk with state \( s_t \in \mathbb{R} \) and dynamics \( s_{t+1} = y s_t + \delta_t \), where \( \delta_t \sim \mathcal{N}(0, \sigma^2) \) and \( y \in [0, 1] \). When \( y = 0 \), the O-L process becomes i.i.d. Gaussian noises. The disturbance enters as a perturbation of the velocity, i.e. the system is driven by the true input \( u_t + m^\epsilon s_t \). We test two scenarios: the Gaussian case \((y = 0, \sigma = 8)\) and the random walk \((y = 0.95, \sigma = 0.5)\). These values lead to similar closed-loop state magnitudes.

5.2.2 Results. We deploy GAPS with buffer size \( B = 100 \) and learning rate \( \eta = 0.01 \). In Figure 2a, we compare the evolution of the parameters \( k_p, k_d \) induced by GAPS for each disturbance model to those of \( K^*(m_t) \). Figure 2b shows the cumulative cost difference between the methods. We see significant differences between the two disturbances. Under i.i.d. zero-mean disturbances, the individual controllers used by the LQR baseline are optimal for their respective infinite-horizon problems [7], so the LQR baseline is strong. GAPS closely tracks the LQR-optimal \( k_d \), but accumulates somewhat more cost. Under random-walk disturbances, however, the optimality no longer applies. The \( k_d \) gain under GAPS diverges from the LQR-optimal value, and GAPS substantially outperforms LQR. (The \( k_p \) parameter is insensitive to the disturbance and does not track the LQR-optimal value, suggesting that it may be unimportant in this setting.)

5.3 Comparing discrete and continuous methods for MPC horizon selection

We compare GAPS and BAPS in the setting of MPC with confidence coefficients (Example 2.2) when predictions are accurate in the near future and inaccurate in the far future. GAPS can be used to tune the confidence parameters \( \theta_t \). BAPS can select from instances of MPC with full confidence but
varying horizons. We denote by $\text{MPC}_j(\theta)$ the MPC algorithm (3) with horizon $j$ and confidences $\theta \in [0, 1]^j$. When the MPC terminal cost matrix is the solution of the DARE for the associated infinite-horizon LQR problem, selecting from $\{\text{MPC}_0(1), \ldots, \text{MPC}_k(1)\}$ is equivalent to selecting from the confidences $\Theta_{\text{cts}} = \{[0 \cdots 0], [1 \cdots 0], [1 1 \cdots 0], \ldots, [1 \cdots 1]\}$ for $\text{MPC}_k$. For the continuous parameter set $\Theta_{\text{cts}} = [0, 1]^k$ we have $\Theta_{\text{dis}} \subseteq \Theta_{\text{cts}}$, so in the long term GAPS should outperform BAPS, but it is not clear which will adapt faster.

5.3.1 Setting. We consider a two-dimensional discretized double integrator system subject to disturbances in both velocity and position, with LQR cost objective. Details are given in Appendix J.3. The true disturbances $\{w_t\}_{t=0}^{H-1} \in \mathbb{R}^4$ are sampled i.i.d. from a uniform distribution. To model the idea that the predictions of the near future are more accurate than the predictions of the far future, we construct noisy predictions using the process $w_{t+\tau} := w_{t+\tau} + \sum_{\tau=1}^{H-1} \epsilon_{\tau}$, where $\{\epsilon_{\tau}\}_{\tau=0}^{H-1}$ are sampled i.i.d. from another uniform distribution. We tune the cost and noise scale factors such that the total cost of MPC is strongly affected by the MPC horizon / confidence weights. The distributions of per-step costs are shown in Figure 3a. We use the GAPS/BAPS learning rates $\eta$ and BAPS batch size $B$ suggested by our regret bound, with the choice of decay constants $(C, \rho)$ detailed in Appendix J.3.

5.3.2 Results. Figure 3 visualizes the behavior of BAPS on the controllers $\{\text{MPC}_0(1), \ldots, \text{MPC}_k(1)\}$. A dot at the coordinate $(\beta, j)$ indicates that BAPS selected $\text{MPC}_j(\beta)$ during the $\beta$th batch. The selections eventually concentrate on the optimal $\text{MPC}_3(1)$ along with the nearly-optimal $\text{MPC}_2(1)$ and $\text{MPC}_4(1)$. However, even towards the end of the simulation, BAPS still selects suboptimal MPC horizons fairly often. This is inevitable with an adversarial bandit algorithm.

For GAPS, the control policy is $\text{MPC}_k$ with initial trust parameters $\theta_0 = 0$. Figure 4 shows the evolution of $\theta_t$. We see it quickly converging to values that are roughly inversely proportional to the magnitude of the prediction noises.

Due to the aforementioned gap between $\Theta_{\text{cts}}$ and $\Theta_{\text{dis}}$, we cannot evaluate the regret of each algorithm with respect to the same optimum. In this example, we have $J(\theta_{T-1}) = 0.89 \min_{\theta \in \Theta_{\text{cts}}} J(\theta)$, where $\theta_{T-1}$ is the final parameter value selected by GAPS. Therefore, we plot separate regret curves in Figure 5. Since we cannot directly compute $\min_{\theta \in \Theta_{\text{cts}}} J(\theta)$, we instead evaluate GAPS against $\hat{x}_t(\theta_{T-1}), \hat{u}_t(\theta_{T-1})$, under the assumption that $\theta_{T-1}$ is approximately optimal due to time-invariance. Over the simulation timeframe, BAPS exhibits characteristic $T^{2/3}$ regret while GAPS reaches approximately constant regret. Since GAPS also competes against a stronger baseline, we see the significant benefit of exploiting the continuous problem structure.
6 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we study the problem of online adaptive controller selection under a general contractive perturbation property (Definition 2.2). For continuous policy parameters, we propose the GAPS algorithm (Algorithm 1), which can be viewed as an efficient approximation of the intractable ideal OGD update. Under additional stability and convexity assumptions, we show that GAPS achieves a time-averaged adaptive policy regret of $O(1/\sqrt{T})$ (Theorem 3.1), which closes the log $T$ gap between online control and OCO left open by previous results [2, 27, 58]. For discrete and finite policy parameters, we propose the BAPS algorithm (Algorithm 2) and show it achieves a static time-averaged regret bound of $O(T^{-\frac{3}{2}})$. To the best of our knowledge, this is the first regret bound for online adaptive controller selection with a finite policy class. We perform numerical simulations to demonstrate the effectiveness of our proposed algorithms.

Many interesting research directions are motivated by this work. For example, relaxing the current strong contractive perturbation assumption for GAPS to weak contractive perturbation. Besides, extending the regret bound to nonlinear dynamics and nonconvex functions is another interesting question, which may require a different notion of regret or a carefully chosen class of dynamics/costs. Another direction is to address a limitation of GAPS: it requires the exact Jacobians of the dynamics $g_t$ and costs $f_t$ during implementation. Replacing them with the approximated Jacobians will make GAPS more practical in unknown systems. Further, both GAPS and BAPS assume that all candidate controllers stabilize the system and satisfy contractive perturbation. An interesting open question is what regret guarantees can be achieved for online controller selection when not all of the candidate controllers can satisfy these assumptions.

REFERENCES

[1] Naman Agarwal, Nataly Brukhim, Elad Hazan, and Zhou Lu. 2019. Boosting for dynamical systems. arXiv preprint arXiv:1906.08720 (2019).
[2] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. 2019. Online control with adversarial disturbances. In International Conference on Machine Learning. PMLR, 111–119.
[3] Naman Agarwal, Elad Hazan, and Karan Singh. 2019. Logarithmic regret for online control. Advances in Neural Information Processing Systems 32 (2019).
[4] Ibrahim Al-Shyoukh and Jeff S Shamma. 2009. Switching supervisory control using calibrated forecasts. IEEE transactions on automatic control 54, 4 (2009), 705–716.
[5] Oren Anava, Elad Hazan, and Shie Mannor. 2013. Online convex optimization against adversaries with memory and application to statistical arbitrage. arXiv preprint arXiv:1302.6937 (2013).
[6] Oren Anava, Elad Hazan, and Shie Mannor. 2015. Online learning for adversaries with memory: price of past mistakes. Advances in Neural Information Processing Systems 28 (2015).
[7] Brian DO Anderson and John B Moore. 2012. *Optimal filtering*. Courier Corporation.
[8] David Angeli. 2002. A Lyapunov approach to incremental stability properties. *IEEE Trans. Automat. Control* 47, 3 (2002), 410–421.
[9] David Angeli and Edouard Mosca. 2004. Adaptive switching supervisory control of nonlinear systems with no prior knowledge of noise bounds. *Automatica* 40, 3 (2004), 449–457.
[10] Raman Arora, Ofer Dekel, and Ambuj Tewari. 2012. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 1747–1754.
[11] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32, 1 (2002), 48–77.
[12] Dheeraj Baby and Yu-Xiang Wang. 2022. Optimal Dynamic Regret in LQR Control. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[13] Nikhil Bansal and Anupam Gupta. 2019. Potential-function proofs for gradient methods. *Theory of Computing* 15, 1 (2019), 1–32.
[14] Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. 2021. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*. PMLR, 471–483.
[15] Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. 2013. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems* 26 (2013).
[16] Nianqiu Chen, Gautam Goel, and Adam Wierman. 2018. Smoothed online convex optimization in high dimensions via online balanced descent. In *Conference On Learning Theory*. PMLR, 1574–1594.
[17] Xinyi Chen and Elad Hazan. 2021. Black-box control for linear dynamical systems. In *Conference on Learning Theory*. PMLR, 1114–1143.
[18] Alon Cohen, Tomer Koren, and Yishay Mansour. 2019. Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret. In *International Conference on Machine Learning*. PMLR, 1300–1309.
[19] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. 2018. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems* 31 (2018).
[20] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. 2014. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 459–467.
[21] Dylan Foster and Max Simchowitz. 2020. Logarithmic Regret for Adversarial Online Control. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 3211–3221. https://proceedings.mlr.press/v119/foster20b.html
[22] Udaya Ghati, Udari Madhushani, Naomi Leonard, and Elad Hazan. 2022. A Regret Minimization Approach to Multi-Agent Control. *arXiv preprint arXiv:2201.13288* (2022).
[23] Gautam Goel and Babak Hassibi. 2021. Competitive control. *arXiv preprint arXiv:2107.13657* (2021).
[24] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. 2019. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Advances in Neural Information Processing Systems* 32 (2019).
[25] Paula Gradu, Elad Hazan, and Edgar Minasyan. 2020. Adaptive regret for control of time-varying dynamics. *arXiv preprint arXiv:2007.04393* (2020).
[26] Elad Hazan et al. 2016. Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2, 3-4 (2016), 157–325.
[27] Elad Hazan, Sham Kakade, and Karan Singh. 2020. The nonstochastic control problem. In *Algorithmic Learning Theory*. PMLR, 408–421.
[28] Elad Hazan and Comandur Seshadri. 2007. Adaptive algorithms for online decision problems. In *Electronic colloquium on computational complexity (ECCC)*, Vol. 14.
[29] Elad Hazan, Karan Singh, and Cyril Zhang. 2017. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*. PMLR, 1433–1441.
[30] Amélie Héliou, Matthieu Martin, Panayotis Mertikopoulos, and Thibaud Rahier. 2020. Online non-convex optimization with imperfect feedback. *Advances in Neural Information Processing Systems* 33 (2020), 17224–17235.
[31] Joao P Hespanha, Daniel Liberzon, and A Stephen Morse. 2003. Overcoming the limitations of adaptive control by means of logic-based switching. *Systems & control letters* 49, 1 (2003), 49–65.
[32] Dimitar Ho, Hoang Le, John Doyle, and Yisong Yue. 2021. Online robust control of nonlinear systems with large uncertainty. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3475–3483.
[33] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. 2020. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems* 33 (2020), 15312–15325.
[34] Romain Laroche and Raphael Feraud. 2018. Reinforcement Learning Algorithm Selection. In *International Conference on Learning Representations*. 
[35] Tongxin Li, Ruixiao Yang, Guannan Qu, Guanya Shi, Chenkai Yu, Adam Wierman, and Steven Low. 2021. Robustness and consistency in linear quadratic control with predictions. arXiv preprint arXiv:2106.09659 (2021).

[36] Yingying Li, Xin Chen, and Na Li. 2019. Online optimal control with linear dynamics and predictions: Algorithms and regret analysis. Advances in Neural Information Processing Systems 32 (2019).

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

[38] Yingying Li, Subhro Das, Jeff Shamma, and Na Li. 2021. Safe Adaptive Learning-based Control for Constrained Linear Quadratic Regulators with Regret Guarantees. arXiv preprint arXiv:2111.00411 (2021).

[39] Yingying Li, Guannan Qu, and Na Li. 2020. Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit. IEEE Trans. Automat. Control 66, 10 (2020), 4761–4768.

[40] Yingying Li, Gautam Goel, and Adam Wierman. 2020. Online optimization with predictions and non-convex losses. Proceedings of the ACM on Measurement and Analysis of Computing Systems 4, 1 (2020), 1–32.

[41] Yingying Li, Yang Hu, Haoyuan Sun, Guanya Shi, Guannan Qu, and Adam Wierman. 2021. Perturbation-based Regret Analysis of Predictive Control in Linear Time Varying Systems. arXiv preprint arXiv:2106.10497 (2021).

[42] Winfried Lohmiller and Jean-Jacques E Slotine. 1998. On contraction analysis for non-linear systems. Automatica 34, 6 (1998), 683–696.

[43] Eric Mazumdar, Roy Dong, Vicenţu Râbies Royo, Claire J. Tomlin, and S. Shankar Sastry. 2017. A Multi-Armed Bandit Approach for Online Expert Selection in Markov Decision Processes. CoRR abs/1707.05714 (2017). arXiv:1707.05714 http://arxiv.org/abs/1707.05714

[44] Edgar Minasyan, Paula Gradu, Max Simchowitz, and Elad Hazan. 2021. Online control of unknown time-varying dynamical systems. Advances in Neural Information Processing Systems 34 (2021), 15934–15945.

[45] Guannan Qu, Yuanyuan Shi, Sahin Lale, Anima Anandkumar, and Adam Wierman. 2021. Stable Online Control of Linear Time-Varying Systems. In Proceedings of the 3rd Conference on Learning for Dynamics and Control (Proceedings of Machine Learning Research, Vol. 144), Ali Jadbabaie, John Lygeros, George J. Pappas, Pablo A. &nbsp; Parrilo, Benjamin Recht, Claire J. Tomlin, and Melanie N. Zeilinger (Eds.), PMLR, 742–753. https://proceedings.mlr.press/v144/qu21a.html

[46] James B. Rawlings. 2000. Tutorial overview of model predictive control. IEEE control systems magazine 20, 3 (2000), 38–52.

[47] Paulo Rosa, Jeff S Shamma, Carlos Silvestre, and Michael Athans. 2009. Stability overlay for linear and nonlinear time-varying plants. In Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference. IEEE, 2435–2440.

[48] Cloë Rouyer, Yevgeny Seldin, and Nicolò Cesa-Bianchi. 2021. An algorithm for stochastic and adversarial bandits with switching costs. In International Conference on Machine Learning. PMLR, 9127–9135.

[49] Björn S Rüffer, Nathan Van De Wouw, and Markus Mueller. 2013. Convergent systems vs. incremental stability. Systems & Control Letters 62, 3 (2013), 277–285.

[50] Oron Sabag, Sahin Lale, and Babak Hassibi. 2022. Optimal Competitive-Ratio Control. arXiv preprint arXiv:2206.01782 (2022).

[51] Rolf Schneider. 2014. Convex bodies: the Brunn–Minkowski theory. Number 151. Cambridge university press.

[52] Guanya Shi, Kamyar Azizzadenesheli, Michael O’Connell, Soon-Jo Chung, and Yisong Yue. 2021. Meta-adaptive nonlinear control: Theory and algorithms. Advances in Neural Information Processing Systems 34 (2021), 10013–10025.

[53] Guanya Shi, Yiheng Lin, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2020. Online optimization with memory and competitive control. Advances in Neural Information Processing Systems 33 (2020), 20636–20647.

[54] Max Simchowitz, Karan Singh, and Elad Hazan. 2020. Improper learning for non-stochastic control. In Conference on Learning Theory. PMLR, 3320–3436.

[55] Kim P Wabersich and Melanie N Zeilinger. 2020. Performance and safety of Bayesian model predictive control: Scalable model-based RL with guarantees. arXiv preprint arXiv:2006.03483 (2020).

[56] Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2020. The power of predictions in online control. Advances in Neural Information Processing Systems 33 (2020), 1994–2004.

[57] Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2022. Competitive control with delayed imperfect information. In 2022 American Control Conference (ACC). IEEE, 2604–2610.

[58] Lijun Zhang, Tie-Yan Liu, and Zhi-Hua Zhou. 2019. Adaptive regret of convex and smooth functions. In International Conference on Machine Learning. PMLR, 7414–7423.

[59] Runyu Zhang, Yingying Li, and Na Li. 2021. On the regret analysis of online LQR control with predictions. In 2021 American Control Conference (ACC). IEEE, 697–703.

[60] Martin Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. In Proceedings of the 20th international conference on machine learning (icml-03). 928–936.
A  PRACTICAL ALGORITHM IMPLEMENTATION

In this section, we present an equivalent version of the GAPS (Algorithm 1) for practical implementation that specifies how to compute the partial derivative $\frac{\partial f_{10}}{\partial x_{i-b}} |_{x_0, \theta_0, \theta_1}$ efficiently using the buffered information.

Algorithm 3 Gradient-based Adaptive Policy Selection (For practical implementation)

Require: A sequence of learning rates $\{\eta_t\}$, buffer length $B$, initial parameter $\theta_0$.

1: for $t = 0, \ldots, T-1$ do
2: Observe the current state $x_t$ and context $\varphi_t$.
3: Pick control action $u_t = \text{ALG}(x_t, \varphi_t, \theta_t)$.
4: Incur the stage cost $f_t(x_t, u_t)$.
5: If $t \geq 1$, set $\frac{\partial x_t}{\partial x_{t-1}} := \frac{\partial x_{t-1}}{\partial x_{t-1}} |_{x_{t-1}, u_{t-1}} \cdot \frac{\partial \text{ALG}}{\partial \theta_{t-1}} |_{x_{t-1}, \varphi_{t-1}, \theta_{t-1}}$.
6: If $t \geq 1$, set $\frac{\partial x_t}{\partial x_{t-1}} := \frac{\partial x_{t-1}}{\partial x_{t-1}} |_{x_{t-1}, u_{t-1}} + \frac{\partial \text{ALG}}{\partial \theta_{t-1}} |_{x_{t-1}, \varphi_{t-1}, \theta_{t-1}}$.
7: for $b = 2, \ldots, B-1$ do
8: If $t \geq b$, set $\frac{\partial x_t}{\partial x_{t-b}} := \frac{\partial x_{t-b}}{\partial \theta_{t-b}}$, where $\frac{\partial x_{t-b}}{\partial \theta_{t-b}}$ is in the buffer.
9: end for
10: Compute the approximated gradient (ignore $\frac{\partial x_t}{\partial x_{t-b}}$ if $t < b$):

$$G_t = \left( \frac{\partial f_t}{\partial x_t} |_{x_t, u_t} + \frac{\partial f_t}{\partial u_t} |_{x_t, u_t} \cdot \frac{\partial \text{ALG}}{\partial x_t} |_{x_t, \varphi_t, \theta_t} \right) \cdot \sum_{b=1}^{B-1} \frac{\partial x_t}{\partial x_{t-b}} + \frac{\partial f_t}{\partial u_t} |_{x_t, u_t} \cdot \frac{\partial \text{ALG}}{\partial \theta_t} |_{x_t, \varphi_t, \theta_t}$$
11: Perform the projected gradient update $\theta_{t+1} = \Pi_{\Theta}(\theta_t - \eta_t G_t)$.
12: Empty the buffer, and save $\left[ \frac{\partial x_t}{\partial \theta_1}, \frac{\partial x_t}{\partial \theta_2}, \ldots, \frac{\partial x_t}{\partial \theta_{B-1}} \right]$ (if the term exists) into the buffer.
13: end for

B  PROOF OF THEOREM 3.2

Our proof is inspired by the proof of Theorem 2.1 in [13]. For a fixed time interval $I = [r : s] \subseteq [0 : T-1]$ and $\theta_t \in \Theta$, we consider the potential function $\Phi_t = \frac{1}{2\eta} ||\theta_t - \theta||^2$. Note that $\theta_t$ satisfies $||\theta_t - \theta|| \leq D$ because we assume $\text{diam}(\Theta) \leq D$. To simplify the notation, we define $\theta'_{t+1} = \theta_t - \eta G_t$.

By proposition 2.2 in [13], we see that

$$\frac{1}{2} (||\theta'_{t+1} - \theta_t||^2 - ||\theta_t - \theta||^2) \leq \frac{1}{2} (||\theta'_{t+1} - \theta_t||^2 - ||\theta_t - \theta||^2)
= \langle \theta'_{t+1} - \theta_t, \theta_t - \theta_t \rangle + \frac{1}{2} ||\theta'_{t+1} - \theta_t||^2
= \eta \langle G_t, \theta_t - \theta_t \rangle + \frac{\eta^2}{2} ||G_t||^2.$$  

Using this inequality, we see that

$$F_t(\theta_t) - F_t(\theta_t) + \Phi_{t+1} - \Phi_t$$
$$= F_t(\theta_t) - F_t(\theta_t) + \langle G_t, \theta_t - \theta_t \rangle + \frac{\eta}{2} ||G_t||^2$$
$$= F_t(\theta_t) - F_t(\theta_t) + \langle \nabla F_t(\theta_t), (G_t - \nabla F_t(\theta_t)), \theta_t - \theta_t \rangle + \frac{\eta}{2} ||G_t - \nabla F_t(\theta_t) + (G_t - F_t(\theta_t))||^2$$
$$\leq F_t(\theta_t) - F_t(\theta_t) + \langle \nabla F_t(\theta_t), \theta_t - \theta_t \rangle + \langle G_t - \nabla F_t(\theta_t), \theta_t - \theta_t \rangle$$
$$+ \eta ||\nabla F_t(\theta_t)||^2 + \eta ||G_t - \nabla F_t(\theta_t)||^2$$  

(14a)
\[\begin{align*}
\leq 0 + \|G_t - \nabla F_t(\theta_t)\| \cdot \|\theta_t - \theta_t\| + \eta \|\nabla F_t(\theta_t)\|^2 + \eta \epsilon^2 \quad (14b) \\
\leq \epsilon D + W^2 \eta + \eta \epsilon^2, \quad (14c)
\end{align*}\]

where we used the triangle inequality and the AM-GM inequality in (14a); we used the assumption that \(F_t\) is convex, \(\|G_t - \nabla F_t(\theta_t)\| \leq \epsilon\), and the Cauchy-Schwarz inequality in (14b); we used the assumptions \(\|G_t - \nabla F_t(\theta_t)\| \leq \epsilon\), \(diam(\Theta) \leq D\), \(\|\nabla F_t(\theta_t)\| \leq W\).

Summing (14) over the time interval \([r : s]\) gives that
\[\frac{1}{T} \sum_{t=r}^{s} (F_t(\theta_t) - F_t(\theta_s)) \leq \frac{s-r}{T} \cdot (\epsilon D + W^2 \eta + \eta \epsilon^2) + \frac{\Phi_r - \Phi_{s+1}}{T} \leq (\epsilon D + W^2 \eta + \eta \epsilon^2) + \frac{D^2}{2\eta T} ,\]

where we used \(diam(\Theta) \leq D\) and \(\Phi_{s+1} \geq 0\) in the last inequality. Since this inequality holds for any time interval \(I = [r : s]\) and \(\theta_t \in \Theta\), this finishes the proof of Theorem 3.2.

C DISCUSSION OF EXAMPLE 2.1 AND EXAMPLE 2.2

In this section, we discuss why Example 2.1 and 2.2 satisfy the strong contractive perturbation (Definition 2.2) as well as Assumption 3.1 and 3.2.

We first look at Example 2.1, where the controller class is given by
\[ALG(x_t, q_t, \theta_t) = -K^{stb}_t x_t + \sum_{i=1}^{h} M_t^{[i-1]} w_{t-i}, \text{ where } \theta_t = (M_t^{[0]}, M_t^{[1]}, ..., M_t^{[h-1]}),\]
and \(\Theta = \{(M_t^{[0]}, M_t^{[1]}, ..., M_t^{[h-1]})) \mid \sum_{i=0}^{h-1} \|M_t^{[i]}\| \leq R_M\}.\) For this example to satisfy Assumption 3.1, we additionally assume that there exists positive constants \(a, b, c, d\) such that \(\|A_t\| \leq a, \|B_t\| \leq b, \|K^{stb}_t\| \leq c, \|w_t\| \leq d\) holds for all \(t\). The cost functions \(f_t\) is \(\ell_f\)-smooth and \(arg\min f_t = (0, 0)\).

To see why the stability in Assumption 3.1 holds, define the state transition matrix as
\[\Psi_{t|\tau} := \begin{cases} 
\prod_{r=1}^{\tau} (A_{t-r} - B_{t-r} K^{stb}_{t-r}) & \text{if } t > \tau, \\
I & \text{otherwise.}
\end{cases}\]

By assumption, we know that \(\|\psi_{t|\tau}\| \leq C_0' (\rho_0')^{t-\tau}\). We see that
\[x_t = \psi_{t|0} x_0 + \sum_{t=0}^{t} \psi_{t|t} \left( \sum_{i=1}^{h} M_t^{[i-1]} w_{t-i} \right). \quad (15)\]

Taking norms on both side of the equation and applying the triangle inequality gives that
\[\|x_t\| \leq \|\psi_{t|0}\| \cdot \|x_0\| + \sum_{t=0}^{t} \|\psi_{t|t}\| \left( \sum_{i=1}^{h} M_t^{[i-1]} \right) \cdot \|w_{t-i}\| \leq C_0' (\rho_0')^t \|x_0\| + \sum_{t=0}^{t} C_0' (\rho_0')^{t-\tau} M_{Rd} \leq C_0' \|x_0\| + \frac{C_0' M_{Rd}}{1 - \rho_0'},\]

where we use the assumption that \(\|\psi_{t|\tau}\| \leq C_0' (\rho_0')^{t-\tau}\), the definition of the policy class, and the bound on disturbances in the second inequality. This inequality implies that \(D_{\Theta, t} \in B(0, \hat{R}) \times \hat{B}(0, c\hat{R})\), where \(\hat{R} = (1 + c) \left( C_0' \|x_0\| + \frac{C_0' M_{Rd}}{1 - \rho_0'} \right)\). Then, it is straightforward to see Assumption 3.1 (1, 2, 3)
holds for (1) $L_{g,x} = a, L_{g,u} = b, t_{g,x} = t_{g,u} = 0$; (2) $L_{\text{ALG},x} = c, L_{\text{ALG},\theta} = R_M, t_{\text{ALG},x} = t_{\text{ALG},\theta} = 0$; (3) $L_f = R_f, t_{f,x} = t_{f,u} = t_f$. Note that the composition of a convex function and an affine function is still convex, thus Assumption 3.2 is also satisfied because $\Theta$ is bounded and the multi-step dynamics is an affine function of policy parameters by (15).

Now, we look at Example 2.2. We additionally assume that $\|A\|, \|B\|, \|w_f\|, \|w_r\|$ are uniformly bounded by constants, and the pair $(A, B)$ is controllable. To simplify the discussion, we will use the big $O$ notation to hide polynomial dependence on $\|A\|, \|B\|, \lambda_{\max}(Q), \lambda_{\min}(Q), \lambda_{\max}(R), \lambda_{\min}(R)$. By the results in [59], we know $\text{ALG}(x_t, \varphi_t, \theta_t)$ can be expressed equivalently as
\[
\text{ALG}(x_t, \varphi_t, \theta_t) = -\bar{K}_t x_t - \sum_{t=1}^{T-1} \lambda_t^{[t-\tau]} \bar{K}_t^{d,\tau} w_{t|t},
\]
where the matrices $\bar{K}_t, \bar{K}_t^{d,\tau}$ only depends on $A, B, Q_{0:T-1}, R_{0:T-1}, Q_{\min}, Q_{\max}, R_{\min}, R_{\max}$. Therefore, the strong contractive perturbation property holds because Proposition 3 in [59] gives that
\[
\| (A - B \bar{K}_t_{t-1}) \cdots (A - B \bar{K}_t_t) \| = O(\rho^{t_2-t_1}), \text{ for any time step } t_2 > t_1.
\]
By the results in [59], we also know that $\| \bar{K}_t^{d,\tau} \| = O(\rho^{t-t})$. Therefore, using a similar approach with Example 2.1, we can show the stability assumption in Assumption 3.1 is satisfied with $\bar{K}_t$ and $R_u$ in $O((1 - \rho)^{-2})$. For Assumption 3.1 (1, 2), we have $L_{g,x}, L_{g,u}, L_{\text{ALG},\theta}$ are in $O(1)$, and $L_{\text{ALG},x}$ is also in $O(1)$ because $\| \bar{K}_t \| = O(1)$ by the perturbation bounds derived in [41]. We also have that $t_{g,x} = t_{g,u} = t_{\text{ALG},x} = t_{\text{ALG},\theta} = 0$ because the dynamics and the controller are affine. For Assumption 3.1 (3), we have $L_f = O((1 - \rho)^{-2})$ by the stability bound and $t_{f,x}, t_{f,u}$ are in $O(1)$.

Assumption 3.2 is also satisfied because $\Theta$ is bounded and the multi-step dynamics is an affine function of policy parameters by (16).

D PROOF OF INEQUALITIES 1, 2, AND 4 IN LEMMA 3.3

The first inequality directly follows from Assumption 3.1. For the second inequality, when $t = \tau + 1$, note that $\frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_t,\theta_t} = \frac{\partial g_{r}|_{x,\theta} \cdot \frac{\partial \text{ALG}}{\partial \theta}}{\partial \theta} |_{x_t,\theta_t}$, where $u_t = \text{ALG}(x_t, \theta_t)$. Taking norms of both sides of the equation gives
\[
\left\| \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_t,\theta_t} \right\| = \frac{\partial g_{r}|_{x,\theta} \cdot \frac{\partial \text{ALG}}{\partial \theta}}{\partial \theta} |_{x_t,\theta_t} \leq \frac{\partial g_{r}|_{x,\theta} \cdot \frac{\partial \text{ALG}}{\partial \theta}}{\partial \theta} |_{x_t,\theta_t} \leq L_{g,u} L_{\text{ALG},\theta}.
\]
When $t > \tau + 1$, we see that
\[
\left\| \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_t,\theta_t} \right\| = \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_{t+1},\theta_{t+1}} \cdot \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_t,\theta_t} \leq \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_{t+1},\theta_{t+1}} \cdot \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_t,\theta_t} \leq \frac{C_0 L_{g,u} L_{\text{ALG},\theta}}{\rho^{t-\tau}},
\]
where $x_{t+1} = g_{r+1}|_{x_t, \theta_t}$.

For the last inequality of Lemma 3.3, when $t = \tau + 1$, we see that
\[
\left\| \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_t,\theta_t} - \frac{\partial g_{r+1}|_{x,\theta}}{\partial \theta} |_{x_t',\theta_t'} \right\| = \left\| \frac{\partial g_{r}|_{x,\theta} \cdot \frac{\partial \text{ALG}}{\partial \theta}}{\partial \theta} |_{x_t,\theta_t} - \frac{\partial g_{r}|_{x,\theta} \cdot \frac{\partial \text{ALG}}{\partial \theta}}{\partial \theta} |_{x_t,\theta_t} \right\| \leq \left( \left\| \frac{\partial g_{r}}{\partial \theta} |_{x_t,\theta_t} - \frac{\partial g_{r}}{\partial \theta} |_{x_t',\theta_t'} \right\| + \left\| \frac{\partial g_{r}}{\partial \theta} |_{x_t,\theta_t} - \frac{\partial g_{r}}{\partial \theta} |_{x_t',\theta_t'} \right\| + \left\| \frac{\partial g_{r}}{\partial \theta} |_{x_t,\theta_t} - \frac{\partial g_{r}}{\partial \theta} |_{x_t',\theta_t'} \right\| \right) \leq L_{\text{ALG},\theta} (\ell_{g,x} \| x_t - x_t' \| + \ell_{g,u} \| u_t - u_t' \|) + L_{g,u} (\ell_{\text{ALG},x} \| x_t - x_t' \| + \ell_{\text{ALG},\theta} \| \theta_t - \theta_t' \|)\]
(18a)

(18b)
\[ \leq (L_{\text{ALG}, \theta}(\ell_{g,x} + \ell_{g,u} L_{\text{ALG}, x}) + L_{g,u} L_{\text{ALG}, x}) \| x_t - x_t' \| + (t^2 \| L_{\text{ALG}, \theta} \ell_{g,u} + L_{g,u} L_{\text{ALG}, \theta}) \| \theta_t - \theta_t' \|. \]  

where we use the notations \( u_t = \text{ALG}(x_t, \theta_t), u_t' = \text{ALG}(x_t, \theta_t') \). We use the triangle inequality in (18a); We use Assumption 3.1 in both (18b) and (18c).

When \( t > \tau \), we see that

\[ \left( \frac{\partial g_{t|\tau}}{\partial x_{t+1}} \right) |_{x_t, \theta_t, \theta_{t+1}} \leq \left( \frac{\partial g_{t|\tau}}{\partial x_{t+1}} \right) |_{x_t, \theta_t} - \left( \frac{\partial g_{t|\tau}}{\partial x_{t+1}} \right) |_{x'_t, \theta_t} \]  

\[ + \left( \frac{\partial g_{t+1|\tau}}{\partial x_{t+1}} \right) |_{x'_t, \theta_t} \) \]  

\[ \leq \left( (1 + L_{\text{ALG}, x})(\ell_{g,x} + \ell_{g,u} \cdot L_{\text{ALG}, x}) + L_{g,u} L_{\text{ALG}, x}) \right) C_0 \cdot \rho^{t-\tau} \cdot \| x_{t+1} - x_t' \| \cdot L_{g,u} L_{\text{ALG}, \theta} \]  

\[ + C_0 \cdot \rho^{t-\tau} \cdot (L_{\text{ALG}, \theta}(\ell_{g,x} + \ell_{g,u} L_{\text{ALG}, x}) + L_{g,u} L_{\text{ALG}, x}) \| x_t - x_t' \| \]  

\[ + C_0 \cdot \rho^{t-\tau} \cdot (L_{\text{ALG}, \theta}(\ell_{g,x} + \ell_{g,u} L_{\text{ALG}, x}) \| \theta_t - \theta_t' \| \]  

\[ \leq C_{\ell,g, \theta}(\theta, x) \rho^{t-\tau} \| x_t - x_t' \| + C_{\ell,g, \theta}(\theta, \theta, \rho) \rho^{t-\tau} \| \theta_t - \theta_t' \|, \]  

where we use the notations \( x_{t+1} = g_{t+1|\tau}(x_t, \theta_t), x_t' = g_{t+1|\tau}(x_t, \theta_t') \). We use the chain rule decomposition in (19a); We use the triangle inequality in (19b); We use the first and the third inequality of Lemma 3.3 as well as (18) in (19c); We use the first two inequalities of Lemma 3.3 in (19d).

E  PROOF OF COROLLARY 3.3.1

To show the first inequality, note that

\[ \frac{\partial f_t|\tau}{\partial x_{t+1}} |_{x_t, \theta_t} = \frac{\partial f_t}{\partial x_{t+1}} |_{x_t, u_t} + \frac{\partial f_t}{\partial u_t} |_{x_t, u_t} \cdot \frac{\partial \text{ALG}}{\partial x_{t+1}} |_{x_t, \theta_t} \cdot \frac{\partial g_{t|\tau}}{\partial x_{t+1}} |_{x_t, \theta_t}, \]  

where \( x_t = g_{t|\tau}(x_t, \theta_{t-1}), u_t = \text{ALG}(x_t, \theta_t) \). Thus, by Lemma 3.3, we see that

\[ \left\| \frac{\partial f_t|\tau}{\partial x_{t+1}} |_{x_t, \theta_t} \right\| \leq \left( \left\| \frac{\partial f_t}{\partial x_{t+1}} |_{x_t, u_t} \right\| + \left\| \frac{\partial f_t}{\partial u_t} |_{x_t, u_t} \cdot \frac{\partial \text{ALG}}{\partial x_{t+1}} |_{x_t, \theta_t} \right\| \right) \cdot \left\| \frac{\partial g_{t|\tau}}{\partial x_{t+1}} |_{x_t, \theta_t} \right\| \]  

\[ \leq L_f (1 + L_{\text{ALG}, x}) \cdot C_0 \lambda^{t-\tau}. \]  

For the second inequality, when \( t = \tau \), we see that

\[ \left\| \frac{\partial f_t|\tau}{\partial \theta_{\tau+1}} |_{x_t, \theta_{\tau+1}} \right\| \leq \left\| \frac{\partial f_t}{\partial \theta_{\tau+1}} |_{x_t, u_t, \theta_t} \right\| \cdot \left\| \frac{\partial \text{ALG}}{\partial \theta_{\tau+1}} |_{x_t, \theta_t} \right\| \leq L_f L_{\text{ALG}, \theta}. \]  

When \( t > \tau \), the second inequality can be shown similarly with the first inequality in Corollary 3.3.1 because we have the chain-rule decomposition

\[ \frac{\partial f_t|\tau}{\partial \theta_{\tau+1}} |_{x_t, \theta_{\tau+1}} = \left( \frac{\partial f_t}{\partial x_{t+1}} |_{x_t, u_t} + \frac{\partial f_t}{\partial u_t} |_{x_t, u_t} \cdot \frac{\partial \text{ALG}}{\partial x_{t+1}} |_{x_t, \theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_{t+1}} |_{x_t, \theta_t}. \]  

(20)
Applying Lemma 3.3 gives that \( C_{L,f,0} = L_f C_{L,g,0} (1 + L_{ALG,x}) \).

For the third inequality, using (20), we see that

\[
\begin{align*}
&\left\| \frac{\partial f_{l,t}}{\partial \ell_t} |_{x_t, \theta_t, \theta_{t+1,t}} - \frac{\partial f_{l,t}}{\partial \ell_t} |_{x'_t, \theta'_t, \theta_{t+1,t}} \right\| \\
\leq & \left\| \frac{\partial f_{l,t}}{\partial \ell_t} |_{x_t, u_t} \cdot \left( \frac{\partial g_{l,t}}{\partial \ell_t} |_{x_t, \theta_t, \theta_{t+1,t}} - \frac{\partial g_{l,t}}{\partial \ell_t} |_{x'_t, \theta'_t, \theta_{t+1,t}} \right) \right\| \\
& + \left\| \frac{\partial f_{l,t}}{\partial \ell_t} |_{x_t, u_t} \cdot \frac{\partial \text{ALG}}{\partial \ell_t} |_{x_t, \theta_t} \cdot \left( \frac{\partial g_{l,t}}{\partial \ell_t} |_{x_t, \theta_t, \theta_{t+1,t}} - \frac{\partial g_{l,t}}{\partial \ell_t} |_{x'_t, \theta'_t, \theta_{t+1,t}} \right) \right\| \\
& + \left\| \frac{\partial f_{l,t}}{\partial \ell_t} |_{x_t, u_t} \cdot \left( \text{ALG} \left( \frac{\partial g_{l,t}}{\partial x_t} |_{x_t, \theta_t} \right) \right) \right\| \right\| \\
& + \left\| \frac{\partial f_{l,t}}{\partial \ell_t} |_{x_t, u_t} \cdot \left( \text{ALG} \left( \frac{\partial g_{l,t}}{\partial x_t} |_{x_t, \theta_t} \right) \right) \right\| \\
& + \left\| \frac{\partial f_{l,t}}{\partial \ell_t} |_{x_t, u_t} \cdot \left( \text{ALG} \left( \frac{\partial g_{l,t}}{\partial x_t} |_{x_t, \theta_t} \right) \right) \right\| \\
& + \left\| \frac{\partial f_{l,t}}{\partial \ell_t} |_{x_t, u_t} \cdot \left( \text{ALG} \left( \frac{\partial g_{l,t}}{\partial x_t} |_{x_t, \theta_t} \right) \right) \right\| \\
\leq & L_f \lambda^{-\tau} (C_{l,g,(x_t,x'_t)} \|x_t - x'_t\| + C_{l,g,(x_t,\theta)} \|\theta_t - \theta'_t\|) + (\ell_{f,x} \|x_t - x'_t\| + \ell_{f,u} \|u_t - u'_t\|) \cdot C_{L,g,x} \lambda^{-\tau} \\
& + L_f L_{ALG,x} \lambda^{-\tau} (C_{l,g,(x_t,x'_t)} \|x_t - x'_t\| + C_{l,g,(x_t,\theta)} \|\theta_t - \theta'_t\|) \\
& + L_f \ell_{ALG,x} \|x_t - x'_t\| \cdot C_{L,g,x} \lambda^{-\tau} + (\ell_{f,x} \|x_t - x'_t\| + \ell_{f,u} \|u_t - u'_t\|) \cdot L_{ALG,x} L_{ALG,x} \lambda^{-\tau},
\end{align*}
\]"
We use (21) and the triangle inequality in (24a); We use Assumption 3.1 and Lemma 3.3 in (24b). Taking norm on both sides of the equation, we see that

\[ \|\nabla \| \text{technique to show} \]

Substituting these into (23) finishes the proof of the fourth inequality.

\[ \leq L_f \lambda^{t-r} (C_{f,g}(\theta,\theta)) \|x_t - x_t'\| + C_{f,g}(\theta,\theta) \|\theta_t - \theta_t'\| \]

\[ + (\ell_{f,x}\|x_t - x_t'\| + \ell_{f,u}\|u_t - u_t'\|) \cdot C_{L,g,\theta} \lambda^{t-r} \]

\[ + L_f L_{\text{ALG},\lambda^{t-r}} (C_{f,g}(\theta,x)) \|x_t - x_t'\| + C_{f,g}(\theta,\theta) \|\theta_t - \theta_t'\| \]

\[ + L_f \ell_{\text{ALG},x} \|x_t - x_t'\| \cdot C_{L,g,\theta} \lambda^{t-r} + (\ell_{f,x}\|x_t - x_t'\| + \ell_{f,u}\|u_t - u_t'\|) \cdot L_{\text{ALG},x} C_{L,g,\theta} \lambda^{t-r}, \]

where we use the notations

\[ x_t = g_{t|t}(x_t, \theta_t, \theta_{t+1:t-1}), x_t' = g_{t|t}(x_t', \theta_t', \theta_{t+1:t-1}), u_t = \text{ALG}(x_t, \theta_t), u_t' = \text{ALG}(x_t', \theta_t). \]

We use (21) and the triangle inequality in (24a); We use Assumption 3.1 and Lemma 3.3 in (24b). Note that by the first two inequalities in Lemma 3.3, we have

\[ \|x_t - x_t'\| \leq \lambda^{t-r} (C_{L,g,x}\|x_t - x_t'\| + C_{L,g,\theta} \|\theta_t - \theta_t'\|), \]

\[ \|u_t - u_t'\| \leq L_{\text{ALG},x} \lambda^{t-r} (C_{L,g,x}\|x_t - x_t'\| + C_{L,g,\theta} \|\theta_t - \theta_t'\|). \]

Substituting these into (23) finishes the proof of the fourth inequality.

F PROOF OF THEOREM 3.4

To simplify the notation, we omit the sequence \{\varphi_t\} throughout the proof since it is clear from the context. We first derive an upper bound of \( G_t \) in order to bound the difference between \( \theta_t \) and \( \theta_{t+1} \).

Recall that

\[ G_t := \min_{t,B-1} \sum_{t=0}^{T_B-1} \left. \frac{\partial f_t}{\partial \theta_{t-r}^0} \right|_{x_{0,\theta_0}} = \min_{t,B-1} \sum_{t=0}^{T_B-1} \left. \frac{\partial f_t}{\partial \theta_{t-r}} \right|_{x_{t-r,\theta_{t-r}}, \Theta} \cdot \]

Taking norm on both sides of the equation, we see that

\[ \|G_t\| = \left\| \min_{t,B-1} \sum_{t=0}^{T_B-1} \left. \frac{\partial f_t}{\partial \theta_{t-r}} \right|_{x_{t-r,\theta_{t-r}}} \right\| \]

\[ \leq \sum_{t=0}^{T_B-1} \left. \left| \frac{\partial f_t}{\partial \theta_{t-r}} \right|_{x_{t-r,\theta_{t-r}}} \right\| \]

\[ \leq \sum_{t=0}^{T_B-1} C_{L,f} \rho^r \]

\[ \leq \frac{C_{L,f} \rho}{1 - \rho}, \]

where we use the triangle inequality in (25a) and Corollary 3.3.1 in (25b). One can use the same technique to show \( \|\nabla F_t(\theta)\| \leq \frac{C_{L,\theta}}{1 - \rho} \). Since the projection onto the set \( \Theta \) is a contraction (see
Theorem 1.2.1 in [51]), we obtain that for any $t > \tau$,

$$\|\theta_t - \hat{\theta}_t\| \leq \frac{C_{L,f,\theta} \eta (t - \tau)}{1 - \rho}. \quad (26)$$

Now we bound the distance between $x_t$ and $\hat{x}_t(\theta_t)$ for $\tau \leq t$. Recall that by definition, we have

$$x_t = g_{t|0}(x_0, \theta_{0:t-1}), \hat{x}_t(\theta_t) = g_{t|0}(x_0, (\theta_t)_{x:t}).$$

Therefore, by taking norm on both sides of the equation, we see that

$$\|x_t - \hat{x}_t(\theta_t)\| = \|g_{t|0}(x_0, \theta_{0:t-1}) - g_{t|0}(x_0, (\theta_t)_{x:t})\|$$

\begin{align*}
&\leq \sum_{t'=0}^{t-1} \|g_{t'|0}(x_0, \theta_{0:t'-1}, (\theta_t)_{x(t-t'-1)}) - g_{t'|0}(x_0, \theta_{0:t'-1}, (\theta_t)_{x(t-t')})\| \quad (27a) \\
&\leq \sum_{t'=0}^{t-1} \|g_{t'|0}(x_{t'}, \theta_{t'}, (\theta_t)_{x(t-t'-1)}) - g_{t'|0}(x_{t'}, (\theta_t)_{x(t-t')})\| \quad (27b) \\
&\leq \sum_{t'=0}^{t-1} C_{L,g,\theta} \rho^{t-t'} \|\theta_t - \theta_t\| \quad (27c) \\
&\leq \frac{C_{L,f,\theta} C_{L,g,\theta} \eta}{1 - \rho} \sum_{t'=0}^{t-1} (t - t') \rho^{t-t'} \quad (27d)
\end{align*}

where we use the triangle inequality in (27a); we use the definition of multi-step costs in (27b); we use Lemma 3.3 in (27c); we use (26) in (27d).

Similarly, we also have that

$$\|f_{t|0}(x_0, \theta_{0:t}) - f_{t|0}(x_0, (\theta_t)_{x(t+1)})\|$$

\begin{align*}
&\leq \sum_{t'=0}^{t-1} \|f_{t'|0}(x_0, \theta_{0:t'}, (\theta_t)_{x(t-t'-1)}) - f_{t'|0}(x_0, \theta_{0:t'-1}, (\theta_t)_{x(t-t'-1)})\| \quad (28a) \\
&\leq \sum_{t'=0}^{t-1} \|f_{t'|0}(x_{t'}, \theta_{t'}, (\theta_t)_{x(t-t'-1)}) - f_{t'|0}(x_{t'}, (\theta_t)_{x(t-t'-1)})\| \quad (28b) \\
&\leq \sum_{t'=0}^{t-1} C_{L,f,\theta} \rho^{t-t'} \|\theta_t - \theta_t\| \quad (28c) \\
&\leq \frac{C_{L,f,\theta}^2 \eta}{1 - \rho} \sum_{t'=0}^{t-1} (t - t') \rho^{t-t'} \quad (28d)
\end{align*}

where we use the triangle inequality in (28a); we use the definition of multi-step costs in (28b); we use Lemma 3.3 in (28c); we use (26) in (28d).
G PROOF OF THEOREM 3.5

To simplify the notation, we define $\hat{\tau}(\theta) := g_{t,0}(x, \varphi_{0,\tau}, \theta_{X(\tau+1)})$. As we discussed below Theorem 3.5 in the main body, we use the triangle inequality to do the decomposition

$$\left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}} \leq \left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}} + \left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}} + \sum_{t'=t}^{\infty} \left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}}$$

For the first term in (29), we see that

$$\left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}} \leq (1 - \rho)^{t-1} \rho^{t'-t} \cdot \eta$$

For any $t' \in [t + 1, t + 1]$, we see that

$$\left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}} \leq \left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}} + \left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}} + \sum_{t'=t}^{\infty} \left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}}$$

where we use Corollary 3.3.1 in (30a) and Theorem 3.4 in (30b).

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

$$\left\| G_t - \nabla \ell_t(\theta_t) \right\| = \sum_{t=0}^{\min\{t, B-1\}} \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} |_{x_0, \theta_{t-1}}, \nabla \ell_t(\theta_t) = \sum_{t=0}^{\min\{t, B-1\}} \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} |_{x_0, \theta_{t-1}}$$

Therefore, we see that

$$\left\| G_t - \nabla \ell_t(\theta_t) \right\| = \sum_{t=0}^{\min\{t, B-1\}} \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} |_{x_0, \theta_{t-1}} + \sum_{t=0}^{\min\{t, B-1\}} \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} |_{x_0, \theta_{t-1}} + \sum_{t'=t}^{\infty} \left\| \frac{\partial f_t |_{x, \theta_{t-1}}}{\partial \theta_{t-1}} \right\|_{x_0, \theta_{t-1}}$$

where we use Lemma 3.3 and Corollary 3.3.1 in (31a); we use Theorem 3.4 in (31b). Substituting (30) and (31) into (29) finishes the proof of the first inequality.
\[ \leq \sum_{\tau=0}^{B-1} (\hat{C}_0 + \hat{C}_1 \tau + \hat{C}_2 \tau^2) \rho^\tau \eta + \sum_{\tau=B}^{T} C_{L,f,\theta} \rho^\tau \]

\[ \leq \left( \frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3} \right) \eta + \frac{C_{L,f,\theta}}{1-\rho} \cdot \rho^B, \]

where we use the triangle inequality in (32a); we use the first inequality in Theorem 3.5 that we have shown and Corollary 3.3.1 in (32b).

H THEOREM 3.1 WITH SPECIFIC BOUNDS AND ITS PROOF

In this section, we first present Theorem 3.1 with specific expressions of the bounds in Theorem H.1. Then, we show Theorem H.1 using the results we discussed in Section 3.3.

**Theorem H.1.** Suppose strong contractive perturbation (Definition 2.2) holds. Consider GAPS (Algorithm 1) with learning rate \( \eta_t = \eta \) and buffer size \( B \). Let \( \{x_t, u_t, \theta_t\}_{t \in [0:T-1]} \) denote the trajectory of Algorithm 1. Under Assumption 3.1, the following inequalities hold

\[ |f_t(x_t, u_t) - F_t(\theta_t)| \leq \frac{C_{L,f,\theta}^2 \lambda}{(1-\lambda)^3} \cdot \eta, \]

\[ \|G_t - \nabla F_t(\theta_t)\| \leq \left( \frac{\hat{C}_0}{1-\lambda} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\lambda)^2} + \frac{\hat{C}_2}{(1-\lambda)^3} \right) \eta + \frac{C_{L,f,\theta}}{1-\lambda} \cdot \lambda^B, \]

where the constants \( C_{L,f,\theta}, \hat{C}_0, \hat{C}_1, \hat{C}_2 \) are given in Corollary 3.3.1 and Theorem 3.5. If Assumption 3.2 also holds, the following adaptive regret holds

\[ R^A(T) \leq \left( \frac{C_{L,f,\theta}^2}{(1-\lambda)^3} + \left( \frac{\hat{C}_0}{1-\lambda} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\lambda)^2} + \frac{\hat{C}_2}{(1-\lambda)^3} \right) D \right) \eta + \frac{D^2}{2\eta T} + \frac{DC_{L,f,\theta}}{1-\lambda} \cdot \lambda^B 
+ 2 \left( \frac{\hat{C}_0}{1-\lambda} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\lambda)^2} + \frac{\hat{C}_2}{(1-\lambda)^3} \right)^2 D^2 \eta^3 + \frac{2C_{L,f,\theta}^2 \lambda^2 B \eta.} \]

Further, if we assume \( T \gg \frac{1}{\lambda^2} \) and set \( \eta = (1-\lambda)^{\frac{3}{2}} T^{-\frac{1}{2}} \), \( B \geq \frac{1}{2} \log(T)/\log(1/\lambda) \), the time-averaged regret is \( R^A(T) = O\left( (1-\lambda)^{-\frac{3}{2}} T^{-\frac{1}{2}} \right) \).

**Proof of Theorem H.1.** The first two inequalities are shown in Theorem 3.4 and Theorem 3.5. Thus, we focus on the adaptive regret part here in the proof.

Fix a time interval \( I = [r:s] \subseteq [0:T-1] \) and let \( \theta_I \) be an arbitrary policy parameter in \( \Theta \). By Theorem 3.2 and Theorem 3.5, we see that the sequence of policy parameters of the online controller satisfies that

\[ \frac{1}{T} \sum_{t=r}^{s} F_t(\theta_t) - \sum_{t=r}^{s} F_t(\theta_I) \leq (W^2 + e^2) \eta + \frac{D^2}{2\eta T} + \epsilon D, \]

where \( W = \frac{C_{L,f,\theta}}{1-\lambda} \) by Theorem 3.4 and \( e = \left( \frac{\hat{C}_0}{1-\lambda} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\lambda)^2} + \frac{\hat{C}_2}{(1-\lambda)^3} \right) \eta + \frac{C_{L,f,\theta}}{1-\lambda} \cdot \lambda^B. \)

Note that by definition, we have \( F_t(\theta_I) = f_t(\hat{x}_t(\theta_I), \hat{u}_t(\theta_I)) \) and by Theorem 3.4, we have

\[ |f_t(x_t, u_t) - F_t(\theta_I)| \leq \frac{C_{L,f,\theta}^2 \lambda}{(1-\lambda)^3} \cdot \eta. \]
Substituting these into (33) gives that

\[ \frac{1}{T} \left( \sum_{i=r}^{s} f_i(x_t, u_t) - \sum_{i=r}^{s} f_i(\hat{x}_t(\theta_t), \hat{u}_t(\theta_t)) \right) \]

\[ \leq \frac{1}{T} \left( \sum_{i=r}^{s} F_i(\theta_t) - \sum_{i=r}^{s} F_i(\theta_t) \right) + \frac{1}{T} \sum_{i=r}^{s} |f_i(x_t, u_t) - F_i(\theta_t)| \]

\[ \leq \frac{D^2}{2\eta T} + \epsilon D + \left( W^2 + \epsilon^2 + \frac{C_{L,f,\theta}^2}{(1 - \lambda)^3} \right) \cdot \eta \]

\[ \leq \left( \frac{C_{L,f,\theta}^2}{(1 - \lambda)^3} + \left( \frac{\hat{C}_0}{1 - \lambda} + \frac{\hat{C}_1 + \hat{C}_2}{(1 - \lambda)^2} + \frac{\hat{C}_2}{(1 - \lambda)^3} \right) \cdot D \right) \eta + \frac{D^2}{2\eta T} + \frac{DC_{L,f,\theta}^2}{1 - \lambda} \cdot \lambda^B \]

\[ + 2 \left( \frac{\hat{C}_0}{1 - \lambda} + \frac{\hat{C}_1 + \hat{C}_2}{(1 - \lambda)^2} + \frac{\hat{C}_2}{(1 - \lambda)^3} \right)^2 D^2 \eta^3 + \frac{2C_{L,f,\theta}^2}{(1 - \lambda)^2} \lambda^B \eta, \]

where we used (33) and (34) in the second inequality. □

I PROOF OF THEOREM 4.1

By Theorem 1.5 in [26], the following inequality holds for any \( j \in \Theta \):

\[ \sum_{m=0}^{T/b-1} s_m^T \tilde{p}_m^b \leq \sum_{m=0}^{T/b-1} \tilde{p}_m^b(j) + \eta \sum_{m=0}^{T/b-1} s_m^T (\tilde{p}_m^b)^2 + \frac{\log k}{\eta}. \]  \hspace{1cm} (35)

Taking the expectation on both sides gives that

\[ \sum_{m=0}^{T/b-1} E \left[ s_m^T \tilde{p}_m^b \right] \]

\[ \leq \sum_{m=0}^{T/b-1} E \left[ \tilde{p}_m^b(j) \right] + \eta \sum_{m=0}^{T/b-1} E \left[ s_m^T (\tilde{p}_m^b)^2 \right] + \frac{\log k}{\eta} \]

\[ = \sum_{m=0}^{T/b-1} E \left[ \tilde{p}_m^b(j) \mid j_{0:m-1} \right] + \eta \sum_{m=0}^{T/b-1} E \left[ s_m^T (\tilde{p}_m^b)^2 \mid j_{0:m-1} \right] + \frac{\log k}{\eta} \]

\[ = \sum_{m=0}^{T/b-1} \left[ \sum_{t=mb}^{(m+1)b-1} f_{t\mid mb}(x_{mb}, f_{x(t-mb)}) \right] + \eta \sum_{m=0}^{T/b-1} E \left[ \left( \sum_{t=mb}^{(m+1)b-1} f_{t\mid mb}(x_{mb}, q_{x(t-mb)}) \right)^2 \right] + \frac{\log k}{\eta} \]

\[ \leq \sum_{j=0}^{T/b-1} F_j(j) + 2C_0 L_0 R_x \cdot \frac{T}{1 - \rho} + \eta b D_0^2 k T + \frac{\log k}{\eta}. \]  \hspace{1cm} (36)

Note that the left hand side of (36) is equal to the expected cost incurred by Algorithm 2, i.e.,

\[ E \left[ \sum_{t=0}^{T-1} f_t(x_t, u_t) \right]. \]

Therefore, we see that for any \( j \in \Theta \), we have

\[ E \left[ \sum_{t=0}^{T-1} f_t(x_t, u_t) \right] \leq \sum_{t=0}^{T-1} F_t(j) + 2C_0 L_0 R_x \cdot \frac{T}{1 - \rho} + \eta b D_0^2 k T + \frac{\log k}{\eta}. \]

Since the cost functions and dynamics are oblivious, this finishes the proof of Theorem 4.1.


J NUMERICAL EXPERIMENT DETAILS

In this section, we give the detailed problem settings of the three experiments discussed in Section 5.

J.1 Details of $\lambda$-confident controller

In this appendix, we provide details on the controller of Li et al. [35] used as a baseline for comparison in Section 5.1 for completeness. Their method commits the $u_{t|t}$ entry of the optimal solution

$$\arg\min_{u_{t:t+k-1}} \sum_{t=0}^{t+k-1} \left[ q(x_{t|t}, Q) + q(u_{t|t}, R) \right] + q(x_{t+k|t}, P)$$

s.t. $x_{t+1}|t = Ax_{t|t} + Bu_{t|t} + \lambda_t \tilde{w}_t$, for $\tau = [t : t + k - 1]$,

$$x_{t|t} = x_t,$$

where $P$ solves the discrete-time algebraic Riccati equation (DARE) for $A, B, Q, R$. The confidence parameter $\lambda_t$ is determined according to

$$\lambda_t = \sum_{s=0}^{t} (\eta(w,s,t-1))^T H(\eta(ws,s,t-1)),$$

where $\eta(w,s,t) = \sum_{r=s}^{t} (F^r)^{t-s} P w_r$, where $H = B(R + B^T PB)^{-1} B^T$ and $F = A - BK^* = A - B(R + B^T PB)^{-1} B^T PA$ is the closed-loop linear dynamics matrix of the LQR-optimal infinite-horizon linear controller.

J.2 Details of inverted pendulum system

Fig. 6. The inverted pendulum system.

In this appendix, we provide details on the nonlinear inverted pendulum system discussed in Section 5.2 and the baseline controller based on infinite-horizon LQR controllers. We consider the inverted pendulum illustrated in Figure 6, comprising a point mass on a massless rod with a stationary and frictionless pivot and a torque input. The continuous-time nonlinear dynamics are given by

$$\ddot{\phi} = \frac{G}{\ell} \sin \phi + \frac{u}{m\ell^2},$$

where $\phi$ is the pivot angle, $u$ is the applied torque, $m$ is the mass, $\ell$ is the rod length, and $G$ is the gravitational constant. Near the unstable equilibrium of $(\phi, \dot{\phi}) = 0$, the system is well-approximated by its state-space linearization:

$$\frac{d}{dt} x \approx \begin{pmatrix} 0 & 1 \\ \frac{G}{\ell} & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1/m\ell^2 \end{pmatrix} u, \quad x = \begin{pmatrix} \phi \\ \dot{\phi} \end{pmatrix}. \quad (37)$$

A linear feedback controller takes the form $u = -k_p \phi - k_d \dot{\phi}$ for the positive constants $k_p$, $k_d$, also known as a proportional-derivative (PD) controller.

To implement the baseline LQR policy, we must compute the LQR-optimal feedback gains $K^*(m)$ for an arbitrary mass $m$. We do so by discretizing the linearized dynamics (37) with a zero-order hold and solving the DARE for the associated infinite-horizon LQR problem.
\[ x_{t+1} = Ax_t + Bu_t := \begin{pmatrix} 1 & \Delta_t \\ \Delta_t & 1 \end{pmatrix} x_t + \begin{pmatrix} 0 \\ \Delta_t \end{pmatrix} u_t. \]

We seek \( \theta = (k_p, k_d) \) and \( \hat{\theta} = (\hat{k}_p, \hat{k}_d) \) such that, with the parameter set \( \Theta = \{ \theta, \hat{\theta} \} \), the weak exponentially decaying perturbation property holds but the strong exponentially decaying perturbation property is violated. The weak property can be verified from the eigenvalues of the matrices \( F := A - B[k_p \ k_d] \) and \( \hat{F} := A - B[\hat{k}_p \ \hat{k}_d] \). On the other hand, we consider the parameter sequence where \( \theta_t \) alternates between \( \theta \) and \( \theta' \). The growth/decay behavior of this sequence is captured by the eigenvalues of the matrix \( \hat{F} \). Aided by numerical search, we find that the values \( \Delta_t = 1/1000 \), \( k_p = 2000, k_d = 2.5, \hat{k}_p = 1000, \hat{k}_d = 1.25 \), lead to \( r(F) < 1 \), \( r(\hat{F}) < 1 \) but \( r(F \hat{F}) > 1 \), where \( r \) denotes the spectral radius. This example confirms that the pendulum system violates the strong exponentially decaying perturbation property.

### J.3 Details of double-integrator system

In this appendix we give details on the double-integrator system used to compare GAPS and BAPS in Section 5.3. We consider a two-dimensional double integrator system discretized with forward Euler integration over the interval \( \Delta_t \), with dynamics and cost matrices

\[
A = \begin{pmatrix} I_2 & \Delta t I_2 \\ 0 & I_2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ \Delta t I_2 \end{pmatrix}, \quad Q = s_n \begin{pmatrix} I_2 & 0 \\ 0 & s_Q I_2 \end{pmatrix}, \quad R = s_n s_R I_2,
\]

with cost scale factors \( 0 < s_Q, s_R, s_n \) to balance the objectives and ensure that the stage costs satisfy the assumptions of Algorithm 2. Note that the optimal policy is invariant to changes in \( s_n \).

The true disturbances \( \{ w_t \}_{t=0}^{T-1} \) are sampled i.i.d. from \( (U[-\Delta_t s_w, \Delta_t s_w])^4 \), where \( s_w \) is a scale factor. To model the idea that predictions of the near future are more accurate than predictions of the distant future, we construct noisy predictions using the process

\[
w_{t+\tau|t} := w_{t+\tau} + s_\epsilon \Delta t \sum_{i=t}^{t+\tau} \epsilon_i,
\]

where \( s_\epsilon \) is a scale factor and \( \{ \epsilon_i \}_{i=0}^{T-1} \) are i.i.d. from \( (U[-1, 1])^4 \).

In this experiment, we select the values \( \Delta t = 0.1, s_Q = 0.1, s_R = 0.01, s_w = 1, \) and \( s_\epsilon = 0.25 \). These choices ensure that the optimal cost is strongly affected by the MPC horizon / confidence weights. To compute the cost normalization \( s_n \), we evaluate each fixed-horizon MPC controller and set

\[
s_n = \left( \max_{j \in \{0, \ldots, k\}} \frac{1}{T} \sum_{t=0}^{T-1} f_t(\tilde{x}_t(j), \tilde{u}_t(j)) \right)^{-1}.
\]

The distribution of per-step costs after normalization are shown in Figure 3a.

Our regret bounds are realized by particular choices of learning rate and EXP3 batch size that depend on the decay constants \( (C, \rho) \) induced by the MPC controllers. The conservative bounds on \( (C, \rho) \) yielded by (4) will lead to very slow learning rates and long batches. Therefore, we appeal to the fact that MPC is solving the same optimization as the infinite-horizon LQR, and use the constants of the optimal linear controller \( K^* \). We compute \( \rho \) empirically from the spectrum of \( A - BK^* \) and let

\[
C = \max \{ \rho^{-n} \| (A - BK)^n \| \}_{n=0}^{N}
\]

for sufficiently large \( N \).