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

Yiheng Lin
California Institute of Technology
Pasadena, CA, USA
yihengl@caltech.edu

James A. Preiss
California Institute of Technology
Pasadena, CA, USA
japreiss@caltech.edu

Emile Anand
California Institute of Technology
Pasadena, CA, USA
eanand@caltech.edu

Yingying Li
University of Illinois Urbana-Champaign
Urbana, IL, USA
yl101@illinois.edu

Yisong Yue
California Institute of Technology
Pasadena, CA, USA
yyue@caltech.edu

Adam Wierman
California Institute of Technology
Pasadena, CA, USA
adamw@caltech.edu

Abstract

We study online adaptive policy selection in systems with time-varying costs and dynamics. We develop the Gradient-based Adaptive Policy Selection (GAPS) algorithm together with a general analytical framework for online policy selection via online optimization. Under our proposed notion of contractive policy classes, we show that GAPS approximates the behavior of an ideal online gradient descent algorithm on the policy parameters while requiring less information and computation. When convexity holds, our algorithm is the first to achieve optimal policy regret. When convexity does not hold, we provide the first local regret bound for online policy selection. Our numerical experiments show that GAPS can adapt to changing environments more quickly than existing benchmarks.

1 Introduction

We study the problem of online adaptive policy selection for nonlinear time-varying discrete-time dynamical systems. The dynamics are given by $x_{t+1} = g_t(x_t, u_t)$, where x_t is the state and u_t is the control input at time t. The policy class is a time-varying mapping π_t from the state x_t and a policy parameter θ_t to a control input u_t. At every time step t, the online policy incurs a stage cost $c_t = f_t(x_t, u_t)$ that depends on the current state and control input. The goal of policy selection is to pick the parameter θ_t online to minimize the total stage costs over a finite horizon T.

Online adaptive policy selection and general online control have received significant attention recently [1–8] because many control tasks require running the policy on a single trajectory, as opposed to restarting the episode to evaluate a different policy from the same initial state. Adaptivity is also important when the dynamics and cost functions are time-varying. For example, in robotics, time-varying dynamics arise when we control an aircraft under changing wind conditions [9].

[†]This work is supported by NSF Grants CNS-2146814, CPS-2136197, CNS-2106403, NGSDI-2105648, CCF-1918865, and Gift from Latitude AI, with additional support for Yiheng Lin provided by Amazon AI4Science Fellowship and PIMCO Graduate Fellowship in Data Science.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

In this paper, we are interested in developing a unified framework that can leverage a broad suite of theoretical results from online optimization and efficiently translate them to online policy selection, where efficiency includes both preserving the tightness of the guarantees and computational considerations. A central issue is that, in online policy selection, the stage cost c_t depends on all previously selected parameters $(\theta_0, \ldots, \theta_{t-1})$ via the state x_t. Many prior works along this direction have addressed this issue by finite-memory reductions. This approach led to the first regret bound on online policy selection, but the bounds are not tight, the computational cost can be large, and the dynamics and policy classes studied are restrictive [1, 3, 6–8].

Contributions. We propose and analyze the algorithm Gradient-based Adaptive Policy Selection (GAPS, Algorithm 1) to address three limitations of existing results on online policy selection. First, under the assumption that c_t is a convex function of $(\theta_0, \ldots, \theta_t)$, prior work left a $\log T$ regret gap between OCO and online policy selection. We close this gap by showing that GAPS achieves the optimal regret of $O(\sqrt{T})$ (Theorem 3.3) [1]. Second, many previous approaches require oracle access to the dynamics/costs and expensive resimulation from imaginary previous states. In contrast, GAPS only requires partial derivatives of the dynamics and costs along the visited trajectory, and computes $O(\log T)$ matrix multiplications at each step. Third, the application of existing regret analysis frameworks is limited to specific policy classes and systems because they require c_t to be convex in $(\theta_0, \ldots, \theta_t)$. We address this limitation by showing the first local regret bound for online policy selection when the convexity does not hold. Specifically, GAPS achieves the local regret of $O(\sqrt{(1 + V)T})$, where V is a measure of how much (g_t, f_t, π_t) changes over the entire horizon.

To derive these performance guarantees, we develop a novel proof framework based on a general exponentially decaying, or "contractive", perturbation property (Definition 2.6) on the policy-induced closed-loop dynamics. This generalizes a key property of disturbance-action controllers [e.g. 1, 8] and includes other important policy classes such as model predictive control (MPC) [e.g. 11] and linear feedback controllers [e.g. 12]. Under this property, we prove an approximation error bound (Theorem 3.2), which shows that GAPS can mimic the update of an ideal online gradient descent (OGD) algorithm [13] that has oracle knowledge of how the current policy parameter θ_t would have performed if used exclusively over the whole trajectory. This error bound bridges online policy selection and online optimization, which means regret guarantees on OGD for online optimization can be transferred to GAPS for online policy selection.

In numerical experiments, we demonstrate that GAPS can adapt faster than an existing follow-the-leader-type baseline in MPC with imperfect disturbance predictions, and outperforms a strong optimal control baseline in a nonlinear system with non-i.i.d. disturbances. The source code for all experiments is published at `https://www.github.com/jpreiss/adaptive_policy_selection`.

Related Work. Our work is related to online control and adaptive-learning-based control [2, 14–19], especially online control with adversarial disturbances and regret guarantees [1, 6–8, 20, 21]. For example, there is a rich literature on policy regret bounds for time-invariant dynamics [1, 6, 8, 15, 20, 22]. There is also a growing interest in algorithms for time-varying systems with small adaptive regret [7, 21], dynamic regret [23–25], and competitive ratio [26–29]. Many prior works study a specific policy class called disturbance-action controller (DAC) [1, 3, 6–8]. When applied to linear dynamics g_t with convex cost functions f_t, DAC renders the stage cost c_t a convex function in past policy parameters $(\theta_0, \ldots, \theta_t)$. Our work contributes to the literature by proposing a general contractive perturbation property that includes DAC as a special case, and showing local regret bounds that do not require c_t to be convex in $(\theta_0, \ldots, \theta_t)$. A recent work also handles nonconvex c_t, but it studies an episodic setting and requires c_t to be "nearly convex", which holds under its policy class [30].

In addition to online control, this work is also related to online learning/optimization [13, 31, 32], especially online optimization with memory and/or switching costs, where the cost at each time step depends on past decisions. Specifically, our online adaptive policy selection problem is related to online optimization with memory [29, 33–39]. Our analysis for GAPS provides insight on how to handle indefinite memory when the impact of a past decision decays exponentially with time.

Our contractive perturbation property and the analytical framework based on this property are closely related to prior works on discrete-time incremental stability and contraction theory in nonlinear systems [40–46], as well as works that leverage such properties to derive guarantees for (online) controllers [47–49]. In complicated systems, it may be hard to design policies that provably satisfy

---

[1]A concurrent work [10] also closes this gap in a less general setting. See Section 3.2 for a discussion.

2

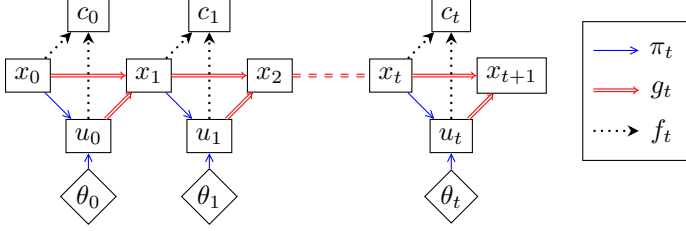



Figure 1: Diagram of the causal relationships between states, policy parameters, control inputs, and costs.

these properties. This motivates some recent works to study neural-based approaches that can learn a controller together with its certificate for contraction properties simultaneously [50, 51]. Our work contributes to this field by showing that, when the system satisfies the contractive perturbation property, one can leverage this property to bridge online policy selection with online optimization.

Notation. We use $[t_1 : t_2]$ to denote the sequence $(t_1, \ldots, t_2)$, $a_{t_1:t_2}$ to denote $(a_{t_1}, a_{t_1+1}, \ldots, a_{t_2})$ for $t_1 \leq t_2$, and $a_{\times \tau}$ for $(a, \ldots, a)$ with a repeated $\tau \geq 0$ times. We define $q(x, Q) = x^\top Q x$. Symbols $\mathbf{1}$ and $\mathbf{0}$ denote the all-one and all-zero vectors/matrices respectively, with dimension implied by context. The Euclidean ball with center $\mathbf{0}$ and radius R in $\mathbb{R}^n$ is denoted by $B_n(0, R)$. We let $\|\cdot\|$ denote the (induced) Euclidean norm for vectors (matrices). The diameter of a set Θ is $\mathrm{diam}(\Theta) := \sup_{x,y \in \Theta} \|x - y\|$. The projection onto the set Θ is $\Pi_\Theta(x) = \arg\min_{y \in \Theta} \|y - x\|$.

2 Preliminaries

We consider online policy selection on a single trajectory. The setting is a discrete-time dynamical system with state $x_t \in \mathbb{R}^n$ for time index $t \in \mathcal{T} := [0 : T - 1]$. At time step $t \in \mathcal{T}$, the policy picks a control action $u_t \in \mathbb{R}^m$, and the next state and the incurred cost are given by:

$$\text{Dynamics: } x_{t+1} = g_t(x_t, u_t), \qquad \text{Cost: } c_t := f_t(x_t, u_t),$$

respectively, where $g_t(\cdot, \cdot)$ is a time-varying dynamics function and $f_t(\cdot, \cdot)$ is a time-varying stage cost. The goal is to minimize the total cost $\sum_{t=0}^{T-1} c_t$.

We consider parameterized time-varying policies of the form of $u_t = \pi_t(x_t, \theta_t)$, where x_t is the current state at time step t and $\theta_t \in \Theta$ is the current policy parameter. Θ is a closed convex subset of $\mathbb{R}^d$. We assume the dynamics, cost, and policy functions $\{g_t, f_t, \pi_t\}_{t \in \mathcal{T}}$ are oblivious, meaning they are fixed before the game begins. The online policy selection algorithm optimizes the total cost by selecting θ_t sequentially. We illustrate how the policy parameter sequence $\theta_{0:T-1}$ affects the trajectory $\{x_t, u_t\}_{t \in \mathcal{T}}$ and per-step costs $c_{0:T-1}$ in Figure 1. The online algorithm has access to the partial derivatives of the dynamics f_t and cost g_t *along the visited trajectory*, but does not have oracle access to the f_t, g_t for arbitrary states and actions.

We provide two motivating examples for our setting. Appendix H contains more details and a third example. The first example is MPC with confidence coefficients, a generalization of [38].

Example 2.1 (MPC with Confidence Coefficients). *Consider a linear time-varying (LTV) system* $g_t(x_t, u_t) = A_t x_t + B_t u_t + w_t$, *with time-varying costs* $f_t(x_t, u_t) = q(x_t, Q_t) + q(u_t, R_t)$. *At time t, the policy observes* $\{A_{t:t+k-1}, B_{t:t+k-1}, Q_{t:t+k-1}, R_{t:t+k-1}, w_{t:t+k-1|t}\}$, *where $w_{\tau|t}$ is a (noisy) prediction of the future disturbance w_τ. Then, $\pi_t(x_t, \theta_t)$ commits the first entry of*

$$\underset{u_{t:t+k-1|t}}{\arg\min} \sum_{\tau=t}^{t+k-1} f_\tau(x_{\tau|t}, u_{\tau|t}) + q(x_{t+k|t}, \tilde{Q}) \tag{1}$$

$$\text{s.t. } x_{t|t} = x_t, \quad x_{\tau+1|t} = A_\tau x_{\tau|t} + B_\tau u_{\tau|t} + \lambda_t^{[\tau-t]} w_{\tau|t} : \ t \leq \tau < t+k,$$

where $\theta_t = \left(\lambda_t^{[0]}, \lambda_t^{[1]}, \ldots, \lambda_t^{[k-1]}\right)$, $\Theta = [0, 1]^k$ *and $\tilde{Q}$ is a fixed positive-definite matrix. Intuitively,* $\lambda_t^{[i]}$ *represents our level of confidence in the disturbance prediction i steps into the future at time step t, with entry 1 being fully confident and 0 being not confident at all.*

The second example studies a nonlinear control model motivated by [12, 47].

Example 2.2 (Linear Feedback Control in Nonlinear Systems). *Consider a time-varying nonlinear control problem with dynamics* $g_t(x_t, u_t) = A x_t + B u_t + \delta_t(x_t, u_t)$ *and costs* $f_t(x_t, u_t) = q(x_t, Q) +$

3

$q(u_t, R)$. *Here, the nonlinear residual δ_t comes from linearization and is assumed to be sufficiently small and Lipschitz. Inspired by [12], we construct an online policy based on the optimal controller $u_t = -\bar{K}x_t$ for the linear-quadratic regulator $\mathrm{LQR}(A, B, Q, R)$. Specifically, we let $\pi_t(x_t, \theta_t) = -K(\theta_t)x_t$ where K is a mapping from Θ to $\mathbb{R}^{n \times m}$ such that $\left\| K(\theta_t) - \bar{K} \right\|$ is uniformly bounded.*

2.1 Policy Class and Performance Metrics

In our setting, the state x_t at time t is uniquely determined by the combination of 1) a state x_τ at a previous time $\tau < t$, and 2) the parameter sequence $\theta_{\tau:t-1}$. Similarly, the cost at time t is uniquely determined by x_τ and $\theta_{\tau:t}$. Since we use these properties often, we introduce the following notation.

Definition 2.3 (Multi-Step Dynamics and Cost). *The multi-step dynamics $g_{t|\tau}$ between two time steps $\tau \leq t$ specifies the state x_t as a function of the previous state x_τ and previous policy parameters $\theta_{\tau:t-1}$. It is defined recursively, with the base case $g_{\tau|\tau}(x_\tau) := x_\tau$ and the recursive case*

$$g_{t+1|\tau}(x_\tau, \theta_{\tau:t}) = g_t(z_t, \pi_t(z_t, \theta_t)), \ \forall\, t \geq \tau,$$

in which $z_t := g_{t|\tau}(x_\tau, \theta_{\tau:t-1})$.[2] The multi-step cost $f_{t|\tau}$ specifies the cost c_t as function of x_τ and $\theta_{\tau:t}$. It is defined as $f_{t|\tau}(x_\tau, \theta_{\tau:t}) := f_t(z_t, \pi_t(z_t, \theta_t))$.

In this paper, we frequently compare the trajectory of our algorithm against the trajectory that would arise from applying a fixed parameter θ since time step 0, which we denote as $\hat{x}_t(\theta) := g_{t|0}(x_0, \theta_{\times t})$ and $\hat{u}_t(\theta) := \pi_t(\hat{x}_t(\theta), \theta)$. A related concept that is heavily used is the *surrogate cost F_t*, which maps a single policy parameter to a real number.

Definition 2.4 (Surrogate Cost). *The surrogate cost function is defined as $F_t(\theta) := f_t(\hat{x}_t(\theta), \hat{u}_t(\theta))$.*

Figure 1 shows the overall causal structure, from which these concepts follow.

To measure the performance of an online algorithm, we adopt the objective of **_adaptive policy regret_**, which has been used by [7, 52]. It is a stronger benchmark than the static policy regret [1, 6] and is more suited to time-varying environments. We use $\{x_t, u_t, \theta_t\}_{t \in \mathcal{T}}$ to denote the trajectory of the online algorithm throughout the paper. The adaptive policy regret $R^A(T)$ is defined as the maximum difference between the cost of the online policy and the cost of the optimal fixed-parameter policy over any sub-interval of the whole horizon $\mathcal{T}$, i.e.,

$$R^A(T) := \max_{I=[t_1:t_2]\subseteq\mathcal{T}}\left(\sum_{t\in I} f_t(x_t, u_t) - \inf_{\theta\in\Theta}\sum_{t\in I} F_t(\theta)\right). \tag{2}$$

In contrast, the (static) policy regret defined in [1, 6] restricts the time interval I to be the whole horizon $\mathcal{T}$. Thus, a bound on adaptive regret is strictly stronger than the same bound on static regret. Adaptive regret is particularly useful in time-varying environments like Examples 2.1 and 2.2 because an online algorithm must adapt quickly to compete against a comparator policy parameter that can change indefinitely with every time interval [32, Section 10.2].

In the general case when surrogate costs $F_{0:T-1}$ are nonconvex, it is difficult (if not impossible) for online algorithms to achieve meaningful guarantees on classic regret metrics like $R^A(T)$ or static policy regret because they do not have oracle optimization solvers or even the exact knowledge of the surrogate costs. Therefore, we introduce the metric of **_local regret_**, which bounds the sum of squared gradient norms over the whole horizon:

$$R^L(T) := \sum_{t=0}^{T-1}\left\|\nabla F_t(\theta_t)\right\|^2. \tag{3}$$

Similar metrics have been adopted by previous works on online nonconvex optimization [53]. Intuitively, $R^L(T)$ measures how well the online agent chases the (changing) stationary point of the surrogate cost sequence $F_{0:T-1}$. Since the surrogate cost functions are changing over time, the bound on $R^L(T)$ will depend on how much the system $\{g_t, f_t, \pi_t\}_{t\in\mathcal{T}}$ changes over the whole horizon $\mathcal{T}$. We defer the details to Section 3.3.

2.2 Contractive Perturbation and Stability

In this section, we introduce two key properties needed for our sub-linear regret guarantees in adaptive online policy selection. We define both with respect to trajectories generated by "slowly" time-varying parameters, which are easier to analyze than arbitrary parameter sequences.

---

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

4

Definition 2.5. *We denote the set of policy parameter sequences with ε-constrained step size by*

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

The first property we require is an exponentially decaying, or "contractive", perturbation property of the closed-loop dynamics of the system with the policy class. We now formalize this property.

Definition 2.6 (ε-Time-varying Contractive Perturbation). *The ε-time-varying contractive perturbation property holds for $R_C > 0, C > 0$, $\rho \in (0, 1)$, and $\varepsilon \geq 0$ if, for any $\theta_{\tau:t-1} \in S_\varepsilon(\tau : t - 1)$,*

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

holds for arbitrary $x_\tau, x'_\tau \in B_n(0, R_C)$ and time steps $\tau \leq t$.

Intuitively, ε-time-varying contractive perturbation requires two trajectories starting from different states (in a bounded ball) to converge towards each other if they adopt the same slowly time-varying policy parameter sequence. We call the special case of $\varepsilon = 0$ *time-invariant contractive perturbation*, meaning the policy parameter is fixed. Although it may be difficult to verify the time-varying property directly since it allows the policy parameters to change, we show in Lemma 2.8 that time-invariant contractive perturbation implies that the time-varying version also holds for some small $\varepsilon > 0$.

The time-invariant contractive perturbation property is closely related to discrete-time incremental stability [e.g. 45] and contraction theory [e.g. 46], which have been studied in control theory. While some specific policies including DAC and MPC satisfy ε-time-varying contractive perturbation globally in linear systems, in other cases it is hard to verify. Our property is local and thus is easier to establish for broader applications in nonlinear systems (e.g., Example 2.2).

Besides contractive perturbation, another important property we need is the stability of the policy class, which requires $\pi_{0:T-1}$ can stabilize the system starting from the zero state as long as the policy parameter varies slowly. This property is stated formally below:

Definition 2.7 (ε-Time-varying Stability). *The ε-time-varying stability property holds for $R_S > 0$ and $\varepsilon \geq 0$ if, for any $\theta_{\tau:t-1} \in S_\varepsilon(\tau : t - 1)$, $\left\| g_{t|\tau}(0, \theta_{\tau:t-1}) \right\| \leq R_S$ holds for any time steps $t \geq \tau$.*

Intuitively, ε-time-varying stability guarantees that the policy class $\pi_{0:T-1}$ can achieve stability if the policy parameters $\theta_{0:T-1}$ vary slowly.[3] Similarly to contractive perturbation, one only needs to verify time-invariant stability (i.e., $\varepsilon = 0$ and the policy parameter is fixed) to claim time-varying stability holds for some strictly positive ε (see Lemma 2.8). The reason we still use the time-varying contractive perturbation and stability in our assumptions is that they hold for $\varepsilon = +\infty$ in some cases, including DAC and MPC with confidence coefficients. Applying Lemma 2.8 for those systems will lead to a small, overly pessimistic ε.

2.3 Key Assumptions

We make two assumptions about the online policy selection problem to achieve regret guarantees.

Assumption 2.1. *The dynamics $g_{0:T-1}$, policies $\pi_{0:T-1}$, and costs $f_{0:T-1}$ are differentiable at every time step and satisfy that, for any convex compact sets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{U} \subseteq \mathcal{R}^m$, one can find Lipschitzness/smoothness constants (can depend on $\mathcal{X}$ and $\mathcal{U}$) such that:*

1. The dynamics $g_t(x, u)$ is $(L_{g,x}, L_{g,u})$-Lipschitz and $(\ell_{g,x}, \ell_{g,u})$-smooth in (x, u) on $\mathcal{X} \times \mathcal{U}$.
2. The policy function $\pi_t(x, \theta)$ is $(L_{\pi,x}, L_{\pi,\theta})$-Lipschitz and $(\ell_{\pi,x}, \ell_{\pi,\theta})$-smooth in (x, θ) on $\mathcal{X} \times \Theta$.
3. The stage cost function $f_t(x, u)$ is (L_f, L_f)-Lipschitz and $(\ell_{f,x}, \ell_{f,u})$-smooth in (x, u) on $\mathcal{X} \times \mathcal{U}$.

Assumption 2.1 is general because we only require the Lipschitzness/smoothness of g_t and f_t to hold for bounded states/actions within $\mathcal{X}$ and $\mathcal{U}$, where the coefficients may depend on $\mathcal{X}$ and $\mathcal{U}$. Similar assumptions are common in the literature of online control/optimization [25, 29, 47].

Our second assumption is on the contractive perturbation and the stability of the closed-loop dynamics induced by a slowly time-varying policy parameter sequence.

Assumption 2.2. *Let $\mathcal{G}$ denote the set of all possible dynamics/policy sequences $\{g_t, \pi_t\}_{t \in \mathcal{T}}$ the environment/policy class may provide. For a fixed $\varepsilon \in \mathbb{R}_{\geq 0}$, the ε-time-varying contractive perturbation (Definition 2.6) holds with (R_C, C, ρ) for any sequence in $\mathcal{G}$. The ε-time-varying stability*

---

[3]This property is standard in online control and is satisfied by DAC [1, 3, 6–8] as well as Examples 2.1 & 2.2.

5

(Definition 2.7) holds with $R_S < R_C$ for any sequence in $\mathcal{G}$. We assume that the initial state satisfies $\|x_0\| < (R_C - R_S)/C$. Further, we assume that if $\{g, \pi\}$ is the dynamics/policy at an intermediate time step of a sequence in $\mathcal{G}$, then the time-invariant sequence $\{g, \pi\}_{\times T}$ is also in $\mathcal{G}$.[4]

Note that Assumption 2.2 is on the joint properties of both the dynamical system and the policy class when composed together in a closed loop. The motivation is to generalize two key properties of linear systems under typical reasonable controllers: 1) the effect of past decisions on the current state decays exponentially fast, and 2) if the system is initialized near the origin, it remains near the origin. We generalize these properties via ε-time-varying contractive perturbation (Definition 2.6) and ε-time-varying stability (Definition 2.7) respectively. Although Assumption 2.2 may seem complicated to understand, it is less restrictive than the assumptions in the most closely related work (e.g., [1, 3, 7]) that focus on linear dynamics.

Compared to other settings where contractive perturbation holds globally [1, 8, 54], Assumption 2.2 only requires it to hold locally in a bounded ball $B(0, R_C)$, which becomes important in nonlinear settings. This brings a new challenge because we need to guarantee that the starting state stays within $B(0, R_C)$ whenever we apply this property in the proof. Therefore, in Assumption 2.2, we assume $R_C > R_S + C\|x_0\|$. Similarly, to leverage the Lipschitzness/smoothness property, we require $\mathcal{X} \supseteq B(0, R_x)$ where $R_x \geq C(R_S + C\|x_0\|) + R_S$ and $\mathcal{U} = \{\pi(x, \theta) \mid x \in \mathcal{X}, \theta \in \Theta, \pi \in \mathcal{G}\}$. Since the coefficients in Assumption 2.1 depend on $\mathcal{X}$ and $\mathcal{U}$, we will set $\mathcal{X} = B(0, R_x)$ and $R_x = C(R_S + C\|x_0\|) + R_S$ by default when presenting these constants. The goal is to ensure that the controller never leaves the region where contractive perturbation applies, which is critical for our analysis and again generalizes properties found in the literature (e.g., Examples 2.1, 2.2, and H.1).

For some systems, verifying Assumption 2.2 is straightforward (e.g., Example 2.1). In other cases, we can rely on the following lemma, which can convert a time-invariant version of the property to general time-varying one. We defer its proof to Appendix C.

Lemma 2.8. *Suppose Assumption 2.2 holds for $\varepsilon = 0$ and (R_C, C, ρ, R_S), which satisfies $R_C > (C + 1)R_S$. Suppose Assumption 2.1 also holds and let $\mathcal{X} := B(0, R_x)$, where $R_x = (C + 1)^2 R_S$. Then, Assumption 2.2 also holds for $\hat{\varepsilon} > 0$, $(\hat{R}_C, \hat{C}, \hat{\rho}, \hat{R}_S)$, and x_0 that satisfies $(\hat{R}_C - \hat{R}_S)/C$. Here, $\hat{R}_S, \hat{R}_C, \hat{\rho}$ are arbitrary constants that satisfies $R_S < \hat{R}_S < \hat{R}_C < R_C/(C + 1)$ and $\rho < \hat{\rho} < 1$. The positive constants $\hat{\varepsilon}$ and $\hat{C}$ are given detailed expressions in Appendix C.*

Remark 2.9. *Lemma 2.8 can also be useful when applied to some parameterized controllers for time-invariant nonlinear systems. For example, the well-known "computed torque control" feedback linearization controllers for robotic manipulators (see, e.g., [55]) renders the closed-loop dynamics exponentially stable about an equilibrium, and the feedback gains can be parameterized. Thus, it satisfies Assumption 2.2 in a neighborhood about the equilibrium, via Lemma 2.8. Even with time-invariant dynamics, the time-varying costs (such as tracking a trajectory determined online) provide a setting where selecting the policy parameters online can be beneficial.*

3 Method and Theoretical Results

Our algorithm, Gradient-Based Adaptive Policy Selection (GAPS), is inspired by the classic online gradient descent (OGD) algorithm [32, 56], with a novel approach for approximating the gradient of the surrogate stage cost F_t. In the context of online optimization, OGD works as follows. At each time t, the current stage cost describes how good the learner's current decision θ_t is. The learner updates its decision by taking a gradient step with respect to this cost. Mapping this intuition to online policy selection, the *ideal* OGD update rule would be the following.

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

This is because the surrogate cost F_t (Definition 2.4) characterizes how good θ_t is for time t if we had applied θ_t from the start, i.e., without the impact of other historical policy parameters $\theta_{0:t-1}$. However, since the complexity of computing ∇F_t exactly grows proportionally to t, the ideal OGD becomes intractable when the horizon T is large.

---

[4]For $\{g, \pi\}_{\times T}$ to be in $\mathcal{G}$, it must satisfy other assumptions about contractive perturbation and stability that we impose on $\mathcal{G}$ but does not need to occur in real problem instances. We only use this assumption in the proof of Theorem 3.6, and it can be made without the loss of generality for time-invariant dynamics and policy classes.

6

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

$$\nabla F_t(\theta_t) = \sum_{b=0}^{t} \left. \frac{\partial f_{t|0}}{\partial \theta_{t-b}} \right|_{x_0,(\theta_t)_{\times(t+1)}} \quad \text{and} \quad G_t = \sum_{b=0}^{\min\{B-1,t\}} \left. \frac{\partial f_{t|0}}{\partial \theta_{t-b}} \right|_{x_0,\theta_{0:t}}. \tag{4}$$

GAPS uses two techniques to efficiently approximate $\nabla F_t(\theta_t)$. First, we *replace the ideal sequence* $(\theta_t)_{\times(t+1)}$ *by the actual sequence* $\theta_{0:t}$. This enables computing gradients along the actual trajectory experienced by the online policy without re-simulating the trajectory under θ_t. Second, we *truncate the whole historical dependence to at most B steps*. This bounds the memory used by GAPS. ε-time-varying contractive perturbation is the key to bound the bias of G_t: Intuitively, in the first step, although θ_τ becomes more different with θ_t as τ decreases, its impact on $f_{t|0}$ decays more quickly (exponentially); In the second step, the terms that we discard are exponentially small with respect to B. We provide a rigorous bound of the bias in Theorem 3.2 and a proof outline in Appendix D.

Algorithm 1 Gradient-based Adaptive Policy Selection (GAPS)

Require: Learning rate η, buffer length B, initial θ_0.
1: **for** $t = 0, \dots, T-1$ **do**
2:     Observe the current state x_t.
3:     Pick the control action $u_t = \pi_t(x_t, \theta_t)$.
4:     Incur the stage cost $c_t = f_t(x_t, u_t)$.
5:     Compute the approximated gradient: $G_t = \sum_{b=0}^{\min\{B-1,t\}} \left. \frac{\partial f_{t|0}}{\partial \theta_{t-b}} \right|_{x_0,\theta_{0:t}}.$
6:     Perform the update $\theta_{t+1} = \prod_\Theta (\theta_t - \eta G_t)$.
7: **end for**

Algorithm 1 presents GAPS in its simplest form. Although the expression of the partial derivatives contains $\theta_{0:t}$, the time- and space-efficient implementation of GAPS only requires to store B partial derivatives for B previous time steps. Details are given in Algorithm 2 in Appendix B.

Compared to many previous online control algorithms that take a reduction approach based on OCO with Memory, our algorithm can be much more computationally efficient (see Appendix I.4 for an empirical comparison). Specifically, these works [1, 3, 6] take a different *finite-memory reduction* approach toward reducing the online control problem to OCO with Memory [33] by completely removing the dependence on policy parameters before time step $t - B$ for a fixed memory length B. In the finite-memory reduction, one must "imaginarily" reset the state at time $t - B$ to be $\mathbf{0}$ and then use the B-step truncated multi-step cost function $f_{t|t-B}(\mathbf{0}, \theta_{t-B:t})$ in the OGD with Memory algorithm [1]. When applied to our setting, this is equivalent to replacing G_t in line 1 of Algorithm 1 by $G'_t = \sum_{b=0}^{B-1} \frac{\partial f_{t|t-B}}{\partial \theta_{t-b}} |_{0,(\theta_t)_{\times(B+1)}}$. However, the estimator G'_t has limitations compared with G_t in GAPS. First, computing G'_t requires oracle access to the partial derivatives of the dynamics and cost functions for arbitrary state and actions. Second, even if those are available, G'_t is less computationally efficient than G_t in GAPS, especially when the policy is expensive to execute. Taking MPC (Example 2.1) as an example, computing G'_t at every time step requires solving B MPC optimization problems when re-simulating the system, where $B = \Omega(\log T)$. In contrast, computing G_t in GAPS only requires solving one MPC optimization problem and $O(B)$ matrix multiplications to update the partial derivatives. One may wonder how significant this improvement is, and if it affects the regret. To address this concern, we compare GAPS with Ideal OGD and OGD with Memory in the setting of MPC with confidence coefficients for a 2D double integrator. The simulation results show that GAPS achieve very similar regret with the two benchmarks, while the improvement on computation efficiency is significant (see Appendix I.4).

3.1 Bounds on Truncation Error

We now present the first part of our main result, which states 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 $\nabla F_t(\theta_t)$. In other words, GAPS mimics the ideal OGD update (Definition 3.1).

7

Theorem 3.2. *Suppose Assumptions 2.1 and 2.2 hold. Let $\{(x_t, u_t, \theta_t)\}_{t \in \mathcal{T}}$ denote the trajectory of GAPS (Algorithm 1) with buffer size B and learning rate $\eta \leq \Omega((1-\rho)\varepsilon)$. Then, we have*

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

where $\Omega(\cdot)$ and $O(\cdot)$ hide the dependence on the Lipschitz/smoothness constants defined in Assumption 2.1 and C in contractive perturbation — see details in Appendix D.1.

We defer the proof of Theorem 3.2 to Appendix D.1. Note that this result does not require any convexity assumptions on the surrogate cost F_t.

3.2 Regret Bounds for GAPS: Convex Surrogate Cost

The second part of our main result studies the case when the surrogate cost F_t is a convex function. This assumption is explicitly required or satisfied by the policy classes and dynamical systems in many prior works on online control and online policy selection [1, 3, 6, 54].

The error bounds in Theorem 3.2 can reduce the problem of GAPS' regret bound in control to the problem of OGD's regret bound in online optimization, where the following result is well known: When the surrogate cost functions F_t are convex, the ideal OGD update (Definition 3.1) achieves the regret bound $\sum_{t=0}^{T-1} F_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=0}^{T-1} F_t(\theta) = O(\sqrt{T})$, when the step size η is of the order $1/\sqrt{T}$ [32]. By taking the biases on the stage costs and the gradients into consideration, we derive the adaptive regret bound in Theorem 3.3. Besides the adaptive regret, one can use a similar reduction approach to "transfer" other regret guarantees for OGD in online optimization to GAPS in control. We include the derivation of a dynamic regret bound as an example in Appendix E.

Theorem 3.3. *Under the same assumptions as Theorem 3.2, if we additionally assume F_t is convex for every time t and $\mathrm{diam}(\Theta)$ is bounded by a constant D, then GAPS achieves adaptive regret*

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

where $O(\cdot)$ hides the same constants as in Theorem 3.2 and D — see details in Appendix D.1.

We discuss how to choose the learning rate and the regret it achieves in the following corollary.

Corollary 3.4. *Under the same assumptions as Theorem 3.3, suppose the horizon length $T \gg \frac{1}{1-\rho}$ and the buffer length $B \geq \frac{1}{2}\log(T)/\log(1/\rho)$. If we set $\eta = (1-\rho)^{\frac{5}{2}}T^{-\frac{1}{2}}$, then GAPS achieves adaptive regret $R^A(T) = O((1-\rho)^{-\frac{5}{2}}T^{\frac{1}{2}})$.*

We defer the proof of Theorem 3.3 to Appendix D.1. Compared to the (static) policy regret bounds of [1, 3], our bound is tighter by a factor of $\log T$. The key observation is that the impact of a past policy parameter θ_{t-b} on the current stage cost c_t decays exponentially with respect to b (see Appendix D for details). In comparison, the reduction-based approach first approximates c_t with $\hat{c}_t$ that depends on $\theta_{t-B+1:t}$, and then applies general OCO with memory results on $\hat{c}_t$ [1, 3]. General OCO with memory cannot distinguish the different magnitudes of the contributions that $\theta_{t-B+1:t}$ make to $\hat{c}_t$, which leads to the regret gap of $B = O(\log T)$.

In the more restrictive setting of linear time-invariant dynamics with the DAC policy class, the results of a concurrent work [10] can also be used to close the $\log T$ gap on static regret of online policy selection. In comparison, Theorem 3.3 considers more general time-varying dynamics and adopts the stronger metric of adaptive regret. As a practical matter, the follow-the-regularized-leader type of algorithm used by [10] is often (much) less computationally efficient than a gradient-based algorithm like GAPS. Nevertheless, [10] made distinct contributions by allowing the state space to be a general Banach space and providing a lower bound for OCO with unbounded memory.

3.3 Regret Bounds for GAPS: Nonconvex Surrogate Cost

The third part of our main result studies the case when the surrogate cost F_t is nonconvex. Before presenting the result, we formally define the variation intensity that measures how much the system changes over the whole horizon.

Definition 3.5 (Variation Intensity). *Let $\{g_t, \pi_t, f_t\}_{t \in \mathcal{T}}$ be a sequence of dynamics/policy/cost functions that the environment provides. The variation intensity V of this sequence is defined as*

$$\sum_{t=1}^{T-1} \sup_{x \in \mathcal{X}, u \in \mathcal{U}} \|g_t(x, u) - g_{t-1}(x, u)\| + \sup_{x \in \mathcal{X}, \theta \in \Theta} \|\pi_t(x, \theta) - \pi_{t-1}(x, \theta)\| + \sup_{x \in \mathcal{X}, u \in \mathcal{U}} |f_t(x, u) - f_{t-1}(x, u)|.$$

8

Variation intensity is used as a measure of hardness for changing environments in the literature of online optimization that often appear in regret upper bounds (see [57] for an overview). Definition 3.5 generalizes one of the standard definitions to online policy selection. Using this definition, we present our main result for GAPS applied to nonconvex surrogate costs using the metric of local regret (3).

Theorem 3.6. *Under the same assumptions as Theorem 3.2, if we additionally assume that $\Theta = \mathbb{R}^d$ for some integer d, then GAPS satisfies local regret*

$$R^L(T) = O\left(\frac{1+V}{(1-\rho)^3\eta} + \frac{\eta T}{(1-\rho)^6} + \frac{\rho^B T}{(1-\rho)^2} + \frac{\eta^3 T}{(1-\rho)^{13}} + \frac{\rho^{2B}\eta T}{(1-\rho)^5}\right),$$

where $O(\cdot)$ hides the same constants as in Theorem 3.2 — see details in Appendix F.

We discuss how to choose the learning rate and the regret it achieves in the following corollary.

Corollary 3.7. *Under the same assumptions as Theorem 3.6, suppose the horizon length $T \gg \frac{1}{1-\rho}$ and the buffer length $B \geq \frac{1}{2}\log(T)/\log(1/\rho)$. If we set $\eta = (1-\rho)^{\frac{3}{2}}(1+V)^{\frac{1}{2}}T^{-\frac{1}{2}}$, GAPS achieves local regret $R^L(T) = O((1-\rho)^{-\frac{9}{2}}(1+V)^{\frac{1}{2}}T^{\frac{1}{2}})$.*

We defer the proof of Theorem 3.6 to Appendix F. Note that the local regret will be sublinear in T if the variation intensity $V = o(T)$. To derive the local regret guarantee in Theorem 3.6, we address additional challenges compared to the convex case. First, we derive a local regret guarantee for OGD in online nonconvex optimization. We cannot directly apply results from the literature because they do not use ordinary OGD, and it is difficult to apply algorithms like Follow-the-Perturbed-Leader [e.g. 58] to online policy selection due to constraints on information and step size. Then, to transfer the regret bound from online optimization to online policy selection, we show how to convert the measure of variation defined on $F_{0:T-1}$ to our variation intensity V defined on $\{g_t, \pi_t, f_t\}_{t\in\mathcal{T}}$.

A limitation of Theorem 3.6 is that we need to assume Θ is a whole Euclidean space so that GAPS will not converge to a point at the boundary of Θ that is not a stationary point. Example 2.2 and Appendix H.3 show that one can re-parameterize the policy class to satisfy this assumption in some cases. Relaxing this assumption is our future work.

4 Numerical Experiments

In this section we compare GAPS to strong baseline algorithms in settings based on Examples 2.1 and 2.2. Details are deferred to Appendix I due to space limitations. Appendix I also includes a third experiment comparing GAPS to a bandit-based algorithm for selecting the planning horizon in MPC, and a computation time comparison between GAPS and the alternative gradient approximation of [1].

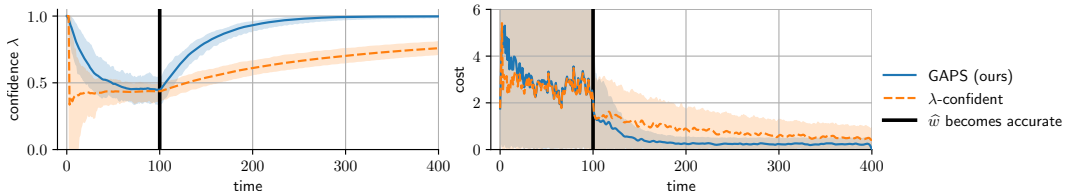



Figure 2: Comparing GAPS and baseline [11] for online adaptation of a confidence parameter for MPC with disturbance predictions. *Left:* Confidence parameter. *Right:* Per-step cost. Shaded bands show 10%-90% quantile range over randomized disturbance properties. See body for details.

MPC confidence parameter. We compare GAPS to the follow-the-leader-type method of [11] for tuning a scalar confidence parameter in model-predictive control with noisy disturbance predictions. The setting is close to Example 2.1 but restricted to satisfy the conditions of the theoretical guarantees in [11]. We consider the scalar system $x_{t+1} = 2x_t + u_t + w_t$ under non-stochastic disturbances w_t with the cost $f_t(x_t, u_t) = x_t^2 + u_t^2$. For $t = 0$ to 100, the predictions of w_t are corrupted by a large amount of noise. After $t > 100$, the prediction noise is instantly reduced by a factor of 100. In this setup, an ideal algorithm should learn to decrease confidence level at first to account for the noise, but then increase to $\lambda \approx 1$ when the predictions become accurate.

9

Figure 2 shows the values of the confidence parameter λ and the per-timestep cost generated by each algorithm. Both methods are initialized to $\lambda = 1$. The method of [11] rapidly adjusts to an appropriate confidence level at first, while GAPS adjusts more slowly but eventually reaches the same value. However, when the accuracy changes, GAPS adapts more quickly and obtains lower costs towards the end of the simulation. In other words, we see that GAPS behaves essentially like an instance of Ideal OGD with constant step size, which is consistent with our theoretical results (Theorem 3.2).

(a) Linear gains k_p, k_d tuned by GAPS compared to LQR optimal.

(b) Cumulative cost difference.

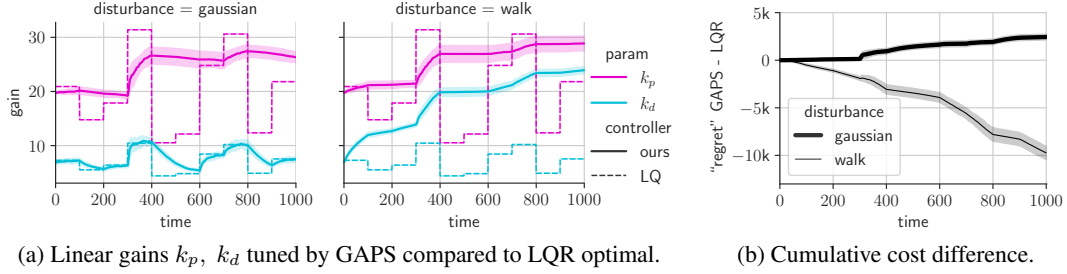



Figure 3: Comparing GAPS and LQR baseline in nonlinear inverted pendulum system. Shaded bands show ± 1 standard deviation over the randomness of the disturbances. See body for details.

Linear controller of nonlinear time-varying system. We apply GAPS to tune the gain parameters of a linear feedback controller in a nonlinear inverted pendulum system. Every 100 seconds, the pendulum mass changes. The system reflects the smooth nonlinear dynamics and nonconvex surrogate costs in Example 2.2, although it differs in other details (see Appendices H.3 and I.2). We compare GAPS to a strong and dynamic baseline that deploys the infinite-horizon linear-quadratic regulator (LQR) optimal controller for the linearized dynamics at each mass. We simulate two disturbances: 1) i.i.d. Gaussian, and 2) Ornstein-Uhlenbeck random walk.

Figure 3a shows the controller parameters tuned by GAPS, along with the baseline LQR-optimal gains, for each disturbance type. The derivative gain k_d closely follows LQR for i.i.d. disturbances but diverges for random-walk disturbances, where LQR is no longer optimal. This is reflected in the cumulative cost difference between GAPS and LQR, shown in Figure 3b. GAPS nearly matches LQR under i.i.d. disturbances, but significantly outperforms it when the disturbance is a random walk. The results show that GAPS can both 1) adapt to step changes in dynamics on a single trajectory almost as quickly as the comparator that benefits from knowledge of the near-optimal analytic solution, and 2) outperform the comparator in more general settings where the analytic solution no longer applies.

5 Conclusion and Future Directions

In this paper, we study the problem of online adaptive policy selection under a general contractive perturbation property. We propose GAPS, which can be implemented more efficiently and with less information than existing algorithms. Under convexity assumptions, we show that GAPS achieves adaptive policy regret of $O(\sqrt{T})$, which closes the $\log T$ gap between online control and OCO left open by previous results. When convexity does not hold, we show that GAPS achieves local regret of $O(\sqrt{(1+V)T})$, where V is the variation intensity of the time-varying system. This is the first local regret bound on online policy selection attained without any convexity assumptions on the surrogate cost functions. Our numerical simulations demonstrate the effectiveness of GAPS, especially for fast adaptation in time-varying settings.

Our work motivates interesting future research directions. For example, a limitation is that GAPS assumes *all* policy parameters can stabilize the system and satisfy contractive perturbation. A recent work on online policy selection relaxed this assumption by a bandit-based algorithm but requires Θ to be a finite set [59]. An interesting future direction to study is what regret guarantees can be achieved when Θ is a continuous parameter set and not all of the candidate policies satisfy these assumptions.

10

References

[1] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

[2] Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[3] Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Algorithmic Learning Theory*, pages 408–421. PMLR, 2020.

[4] Dimitar Ho, Hoang Le, John Doyle, and Yisong Yue. Online robust control of nonlinear systems with large uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 3475–3483. PMLR, 2021.

[5] Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in LQR control. In *Advances in Neural Information Processing Systems*, 2022.

[6] Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. In *Conference on Learning Theory*, pages 1114–1143. PMLR, 2021.

[7] Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. *CoRR*, abs/2007.04393, 2020.

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

[9] Michael O'Connell, Guanya Shi, Xichen Shi, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural-fly enables rapid learning for agile flight in strong winds. *Science Robotics*, 7(66):eabm6597, 2022.

[10] Raunak Kumar, Sarah Dean, and Robert Kleinberg. Online convex optimization with unbounded memory. *arXiv preprint arXiv:2210.09903*, 2022.

[11] Tongxin Li, Ruixiao Yang, Guannan Qu, Guanya Shi, Chenkai Yu, Adam Wierman, and Steven Low. Robustness and consistency in linear quadratic control with predictions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6:1–35, 2021.

[12] Guannan Qu, Chenkai Yu, Steven Low, and Adam Wierman. Exploiting linear models for model-free nonlinear control: A provably convergent policy gradient approach. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6539–6546. IEEE, 2021.

[13] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.

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

[15] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, volume 33, pages 15312–15325, 2020.

[16] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[17] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{t}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

[18] Yingying Li, Subhro Das, Jeff S. Shamma, and Na Li. Safe adaptive learning-based control for constrained linear quadratic regulators with regret guarantees. *CoRR*, abs/2111.00411, 2021.

11

[19] Jing Yu, Dimitar Ho, and Adam Wierman. Online stabilization of unknown networked systems with communication constraints. *arXiv preprint arXiv:2203.02630*, 2022.

[20] Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*, volume 119, pages 3211–3221. PMLR, 2020.

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

[22] Kim Peter Wabersich and Melanie N. Zeilinger. Performance and safety of bayesian model predictive control: Scalable model-based RL with guarantees. *CoRR*, abs/2006.03483, 2020.

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

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

[25] Yiheng Lin, Yang Hu, Guanya Shi, Haoyuan Sun, Guannan Qu, and Adam Wierman. Perturbation-based regret analysis of predictive control in linear time varying systems. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[26] Oron Sabag, Sahin Lale, and Babak Hassibi. Optimal competitive-ratio control. *CoRR*, abs/2206.01782, 2022.

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

[28] Gautam Goel and Babak Hassibi. Competitive control. *CoRR*, abs/2107.13657, 2021.

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

[30] Xinyi Chen, Edgar Minasyan, Jason D Lee, and Elad Hazan. Regret guarantees for online deep control. *Proceedings of Machine Learning Research vol XX*, 1:34, 2023.

[31] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[32] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[33] Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

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

[35] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

[37] Niangjun Chen, Gautam Goel, and Adam Wierman. Smoothed online convex optimization in high dimensions via online balanced descent. In *Conference On Learning Theory*, pages 1574–1594. PMLR, 2018.

12

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

[39] Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *International Conference on Machine Learning*, pages 1747–1754, 2012.

[40] Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, pages 471–483. PMLR, 2021.

[41] Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, and Yisong Yue. Meta-adaptive nonlinear control: Theory and algorithms. In *Advances in Neural Information Processing Systems*, volume 34, pages 10013–10025, 2021.

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

[43] David Angeli. A Lyapunov approach to incremental stability properties. *IEEE Transactions on Automatic Control*, 47(3):410–421, 2002.

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

[45] Florian Bayer, Mathias Bürger, and Frank Allgöwer. Discrete-time incremental iss: A framework for robust nmpc. In *2013 European Control Conference (ECC)*, pages 2068–2073. IEEE, 2013.

[46] Hiroyasu Tsukamoto, Soon-Jo Chung, and Jean-Jaques E Slotine. Contraction theory for nonlinear stability analysis and learning-based control: A tutorial overview. *Annual Reviews in Control*, 52:135–169, 2021.

[47] Tongxin Li, Ruixiao Yang, Guannan Qu, Yiheng Lin, Adam Wierman, and Steven H. Low. Certifying black-box policies with stability for nonlinear control. *IEEE Open Journal of Control Systems*, 2:49–62, 2023.

[48] Hiroyasu Tsukamoto, Soon-Jo Chung, and Jean-Jacques Slotine. Learning-based adaptive control using contraction theory. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2533–2538. IEEE, 2021.

[49] Stephen Tu, Alexander Robey, Tingnan Zhang, and Nikolai Matni. On the sample complexity of stability constrained imitation learning. In *Learning for Dynamics and Control Conference*, pages 180–191. PMLR, 2022.

[50] Charles Dawson, Sicun Gao, and Chuchu Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, pages 1–19, 2023.

[51] Dawei Sun, Susmit Jha, and Chuchu Fan. Learning certified control using contraction metric. In *Conference on Robot Learning*, pages 1519–1539. PMLR, 2021.

[52] Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 14, 2007.

[53] Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pages 1433–1441. PMLR, 2017.

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

[55] Jean-Jacques E Slotine, Weiping Li, et al. *Applied nonlinear control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.

[56] Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019.

13

[57] Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7195–7201. IEEE, 2016.

[58] Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pages 845–861. PMLR, 2020.

[59] Yingying Li, James A Preiss, Na Li, Yiheng Lin, Adam Wierman, and Jeff S Shamma. Online switching control with stability and regret guarantees. In *Learning for Dynamics and Control Conference*, pages 1138–1151. PMLR, 2023.

[60] Rolf Schneider. *Convex Bodies: the Brunn–Minkowski Theory*. Cambridge University Press, 2014.

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

[62] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[63] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *ACM Symposium on the Theory of Computing*, pages 459–467, 2014.

[64] Yiheng Lin, Yang Hu, Guannan Qu, Tongxin Li, and Adam Wierman. Bounded-regret MPC via perturbation analysis: Prediction error, constraints, and nonlinearity. In *Advances in Neural Information Processing Systems*, 2022.

14

Outline of the appendices.
- Appendix A presents additional definitions and remarks that support the main body of the paper.

- Appendix B provides an efficient implementation of our algorithm GAPS for the continuous-parameter setting.

- Appendices C to F contain the detailed proofs of our theoretical results.

- Appendix G details a bandit-based algorithm for finite policy classes and its regret bound.

- Appendix H provides additional details on our MPC and nonlinear control example problem settings, as well as a discussion of the disturbance-action controller popular in online control literature.

- Appendix I contains 1) details on the numerical experiments in Section 4, and 2) a third numerical experiment whose results do not appear in Section 4.

A Mathematical background and additional remarks

Definition A.1 (Lipschitzness and Smoothness). *We say a function $h : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$ is (L_1, L_2)-Lipschitz on a set $\mathcal{D} \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ if for any $(y, z), (y', z') \in \mathcal{D}$, we have*

$$\|h(y, z) - h(y', z')\| \le L_1 \|y - y'\| + L_2 \|z - z'\|.$$

We say h is (ℓ_1, ℓ_2)-smooth on $\mathcal{D}$ if h is differentiable and for any $(y, z), (y', z') \in \mathcal{D}$, we have

$$\|\nabla h(y, z) - \nabla h(y', z')\| \le \ell_1 \|y - y'\| + \ell_2 \|z - z'\|.$$

Remark A.2 (Relation to other notions of contraction/stability). *Contractive perturbation is closely related to incremental exponential stability in nonlinear control theory [42–44], which considers a time-varying dynamical system $x_{t+1} = h_t(x_t)$ and requires the following convergent property:*

$$\|x_t - x_t'\| \le C\rho^{t-\tau}\|x_\tau - x_\tau'\| \quad \text{for any} \quad x_\tau, x_\tau' \in \mathbb{R}^n, \ x_{t+1}' = h_t(x_t'), \ t \ge \tau.$$

Basically, the time-invariant contractive perturbation property requires that any closed-loop dynamics $x_{t+1} = g_t(x_t, \pi_t(x_t, \theta))$ generated by policies with a fixed parameter $\theta \in \Theta$ guarantees incremental exponential stability. A sufficient condition to verify the incremental exponential stability is the contraction analysis. See [42] for more details.

B Practical Algorithm Implementation

In Algorithm 2, we present an equivalent version of the GAPS (Algorithm 1) for practical implementation that specifies how to compute the partial derivative $\left.\frac{\partial f_{t|0}}{\partial \theta_{t-b}}\right|_{x_0, \theta_{0:t}}$ efficiently using the buffered information.

C Proof of Lemma 2.8

We first restate Lemma 2.8 with detailed coefficients in Lemma C.1:

Lemma C.1. *Suppose Assumption 2.2 holds for $\varepsilon = 0$ and (R_C, C, ρ, R_S), which satisfies $R_C > (C + 1)R_S$. Suppose Assumption 2.1 also holds and let $\mathcal{X} := B_n(0, R_x)$, where $R_x = (C + 1)^2 R_S$. Then, Assumption 2.2 also holds for $\hat{\varepsilon} > 0$, $(\hat{R}_C, \hat{C}, \hat{\rho}, \hat{R}_S)$, and x_0 that satisfies the inequality $\|x_0\| \le (\hat{R}_C - \hat{R}_S)/C$. Here, $\hat{R}_S, \hat{R}_C, \hat{\rho}$ are arbitrary constants that satisfies $R_S < \hat{R}_S < \hat{R}_C < R_C/(C + 1)$ and $\rho < \hat{\rho} < 1$. Other coefficients are given by*

$$\hat{C} = ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x}) \cdot (1 + L_{g,x} + L_{g,u}L_{\pi,x})^{2h}\hat{\rho}^{-h},$$

$$\hat{\varepsilon} = \min\left\{\frac{(\hat{\rho}^h - C\rho^h)(1-\rho)^2}{C \cdot C'\rho(1 + L_{g,x} + L_{g,u}L_{\pi,x})^{2h}}, \frac{(1-\rho)^2(\hat{R}_S - R_S)}{CL_{g,u}L_{\pi,\theta}}\right\}, \ where$$

$$C' = \left((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u} \cdot (L_{\pi,x} + L_{\pi,\theta})) + L_{g,u} \cdot (\ell_{\pi,x} + \ell_{\pi,\theta})\right)(L_{g,u}L_{\pi,\theta} + 1),$$

where h is a constant integer that satisfies $C\rho^h < \min\{\hat{\rho}^h, 1 - \hat{R}_S/\hat{R}_C\}$.

15

Algorithm 2 Gradient-based Adaptive Policy Selection (For practical implementation)

Require: Learning rate η, buffer length B, initial parameter θ_0.

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

$$G_t = \left(\frac{\partial f_t}{\partial x_t}\Big|_{x_t, u_t} + \frac{\partial f_t}{\partial u_t}\Big|_{x_t, u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x_t, \theta_t} \right) \cdot \sum_{b=1}^{B-1} \frac{\partial x_t}{\partial \theta_{t-b}} + \frac{\partial f_t}{\partial u_t}\Big|_{x_t, u_t} \cdot \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x_t, \theta_t}$$

11:    Perform the projected gradient update $\theta_{t+1} = \prod_\Theta(\theta_t - \eta G_t)$.
12:    Empty the buffer, and save $\left[\frac{\partial u_t}{\partial \theta_t}, \frac{\partial x_t}{\partial \theta_{t-1}}, \ldots, \frac{\partial x_t}{\partial \theta_{t-B+1}} \right]$ (if the term exists) into the buffer.
13: **end for**

Before showing Lemma C.1, we first show that the composition of Lipschitz and smooth functions is still Lipschitz and smooth in the following technical lemma:

Lemma C.2. *Suppose the sequence of functions $\iota_{1:t}$ satisfies that $\iota_i : D_i \to D_{i+1}$ is L-Lipschitz and ℓ-smooth for all $i \in \{1, 2, \cdots, t\}$. Then, their composition $(\iota_t \circ \iota_{t-1} \circ \cdots \circ \iota_1)$ is Lt-Lipschitz and $\ell(1+L)^{2t}$-smooth.*

Proof of Lemma C.2. We show the conclusion by induction. For $t = 1$, ι_1 is L-Lipschitz and $\ell(1+L)$-smooth.

Suppose we have shown that $(\iota_t \circ \iota_{t-1} \circ \cdots \circ \iota_1)$ is Lt-Lipschitz and $\ell(1+L)^{2t}$-smooth for any t functions that satisfies the assumptions of Lemma C.2. For $t+1$, we simplify the notation by defining $\hat{\iota} := (\iota_{t+1} \circ \iota_t \circ \cdots \circ \iota_2)$. Our goal is to show $(\hat{\iota} \circ \iota_1)$ is L^{t+1}-Lipschitz and $\ell(1+L)^{2(t+1)}$-smooth. Note that

$$\|(\hat{\iota} \circ \iota_1)(x) - (\hat{\iota} \circ \iota_1)(x')\| \leq L^t \|\iota_1(x) - \iota_1(x')\| \leq L^{t+1}\|x - x'\|,$$

where we use the induction assumption in the first inequality and the assumption of Lemma C.2 in the second inequality. We also see that

$$\left\| \frac{\partial(\hat{\iota} \circ \iota_1)}{\partial x}\Big|_x - \frac{\partial(\hat{\iota} \circ \iota_1)}{\partial x}\Big|_{x'} \right\|$$

$$= \left\| \frac{\partial \hat{\iota}}{\partial y}\Big|_{\iota_1(x)} \cdot \frac{\partial \iota_1}{\partial x}\Big|_x - \frac{\partial \hat{\iota}}{\partial y}\Big|_{\iota_1(x')} \cdot \frac{\partial \iota_1}{\partial x}\Big|_{x'} \right\| \tag{5a}$$

$$\leq \left\| \frac{\partial \hat{\iota}}{\partial y}\Big|_{\iota_1(x)} - \frac{\partial \hat{\iota}}{\partial y}\Big|_{\iota_1(x')} \right\| \cdot \left\| \frac{\partial \iota_1}{\partial x}\Big|_x \right\| + \left\| \frac{\partial \hat{\iota}}{\partial y}\Big|_{\iota_1(x')} \right\| \cdot \left\| \frac{\partial \iota_1}{\partial x}\Big|_x - \frac{\partial \iota_1}{\partial x}\Big|_{x'} \right\|$$

$$\leq \ell(1+L)^{2t}\|\iota_1(x) - \iota_1(x')\| \cdot L + L^t \cdot \ell\|x - x'\| \tag{5b}$$

$$\leq \ell\big(L^2(1+L)^{2t} + L^t\big)\|x - x'\| \tag{5c}$$

$$\leq \ell(1+L)^{2(t+1)}\|x - x'\|,$$

where we use the chain rule decomposition in (5a); we use the induction assumption in (5b); we use the assumption that ι_1 is L-Lipschitz in (5c).

Therefore, we have shown Lemma C.2 by induction. $\qquad\qquad\square$

16

Now we are ready to show Lemma C.1.

We first discuss the intuition behind the proof. Since the multi-step dynamics is differentiable under Assumption 2.1, we only need to show an upper bound of $\left\|\frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}}\right\|$ that is exponentially decaying. Intuitively, we use the chain rule to decompose the partial derivative $\frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}}$ as the product of multiple partial derivatives $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i},\theta_{t_i:t_{i+1}-1}}$, where each time interval $[t_i : t_{i+1} - 1]$ has (approximately) length h. By the time-invariant contractive property, we know the norm of $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}$ can be upper bounded by $C\rho^h$ once it is realized on the trajectory $\{x_{t_i}, (\theta_{t_i})_{\times h}\}$, where the policy parameter is repeating. By the smoothness guarantee derived in Lemma C.2, we can show the difference between $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i},\theta_{t_i:t_{i+1}-1}}$ and $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i},(\theta_{t_i})_{\times h}}$ is in the order of $O(\varepsilon)$. This implies that the norm of $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i},\theta_{t_i:t_{i+1}-1}}$ can be bounded by $\hat\rho^h$ once ε is sufficiently small with respect to the gap $\left(\hat\rho^h - C\rho^h\right)$.

We present the formal proof of Lemma C.1 below.

Proof of Lemma C.1. We first show that, under a time-invariant policy parameter θ, starting from an arbitrary x_τ that satisfies $\|x_\tau\| \le R_C$, we have $\|x_t\| \le R_S + C\rho^{t-\tau}\|x_\tau\|$ for all $t \ge \tau$.

To see this, note that by time-invariant contractive perturbation, we have

$$\left\|g_{t|\tau}(x_\tau,\theta_{\times(t-\tau)})\right\| \le \left\|g_{t|\tau}(x_\tau,\theta_{\times(t-\tau)}) - g_{t|\tau}(0,\theta_{\times(t-\tau)})\right\| + \left\|g_{t|\tau}(0,\theta_{\times(t-\tau)})\right\|$$
$$\le C\rho^{t-\tau}\|x_\tau\| + R_S.$$

Now, we show that if starting from x_τ that satisfies $\|x_\tau\| \le \hat R_C$, the trajectory induced by an $\hat\varepsilon$-time-varying parameter sequence satisfies that $\|x_t\| \le \hat R_S + C\rho^{t-\tau}\|x_\tau\|$ for all $t \ge \tau$. Since this ball is contained in $B(0, R_C)$, we know the time-invariant contractive perturbation always apply.

We show this statement by induction. Suppose the statement holds for all time steps $\tau, \ldots, t-1$. We see that

$$\left\|g_{t|\tau}(x_\tau,\theta_{\tau:t-1}) - g_{t|\tau}(x_\tau,\theta_{\times(t-\tau)})\right\|$$

$$\le \sum_{j=\tau}^{t}\left\|g_{t|\tau}(x_\tau,\theta_{\tau:j},(\theta_t)_{\times(t-j-1)}) - g_{t|\tau}(x_\tau,\theta_{\tau:j-1},(\theta_t)_{\times(t-j)})\right\|$$

$$\le CL_{g,u}L_{\pi,\theta}\sum_{j=\tau}^{t}\rho^{t-j-1}\|\theta_t - \theta_j\| \le \frac{CL_{g,u}L_{\pi,\theta}\hat\varepsilon}{(1-\rho)^2}.$$

Therefore, we see that

$$\left\|g_{t|\tau}(x_\tau,\theta_{\tau:t-1})\right\| \le \left\|g_{t|\tau}(x_\tau,\theta_{\times(t-\tau)})\right\| + \frac{CL_{g,u}L_{\pi,\theta}\hat\varepsilon}{(1-\rho)^2}$$
$$\le C\rho^{t-\tau}\|x_\tau\| + R_S + \frac{CL_{g,u}L_{\pi,\theta}\hat\varepsilon}{(1-\rho)^2}$$
$$= C\rho^{t-\tau}\|x_\tau\| + \hat R_S.$$

This finishes the proof of ε-time-varying stability with $\hat\varepsilon$ and $\hat R_S$.

Note that we can decompose the partial derivative $\frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}}$ as

$$\frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}}$$
$$= \frac{\partial g_{t_p|t_{p-1}}}{\partial x_{t_{p-1}}}\Big|_{x_{t_{p-1}},\theta_{t_{p-1}:t_p-1}} \cdot \frac{\partial g_{t_{p-1}|t_{p-2}}}{\partial x_{t_{p-2}}}\Big|_{x_{t_{p-2}},\theta_{t_{p-2}:t_{p-1}-1}} \cdots \frac{\partial g_{t_1|t_0}}{\partial x_{t_0}}\Big|_{x_{t_0},\theta_{t_0:t_1-1}}, \quad (6)$$

17

where $t_0 = \tau, t_p = t$; $t_i = t_{i-1} + h$ holds for $i = 1, \ldots, p-1$, and $t_{p-1} < t_p \leq t_{p-1} + h$.
For $i \in [0, p-2]$, we see that

$$\left\| \frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}} \Bigg|_{x_{t_i}, \theta_{t_i:t_{i+1}-1}} - \frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}} \Bigg|_{x_{t_i}, (\theta_{t_i}) \times h} \right\|$$

$$\leq \sum_{j=0}^{h-1} \left\| \frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}} \Bigg|_{x_{t_i}, (\theta_{t_i}) \times j, \theta_{t_i+j;t_{i+1}-1}} - \frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}} \Bigg|_{x_{t_i}, (\theta_{t_i}) \times (j+1), \theta_{t_i+j+1;t_{i+1}-1}} \right\| \quad \text{(7a)}$$

$$\leq \sum_{j=0}^{h-1} \left\| \frac{\partial g_{t_i+j}}{\partial x_{t_i}} \Bigg|_{x_{t_i}, (\theta_{t_i}) \times j} \right\| \cdot$$

$$\left\| \frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i+j;t_{i+1}-1}} - \frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i}, \theta_{t_i+j+1;t_{i+1}-1}} \right\| \quad \text{(7b)}$$

$$\leq C \sum_{j=0}^{h-1} \rho^j \left\| \frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i+j;t_{i+1}-1}} - \frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i}, \theta_{t_i+j+1;t_{i+1}-1}} \right\|, \quad \text{(7c)}$$

where we use the shorthand notation $\bar{x}_{t_i+j} = g_{t_i+j|t_i}(x_{t_i}, (\theta_{t_i}) \times j)$. We use the triangle inequality in (7a), the chain rule decomposition in (7b). In (7c), we can apply the time-invariant contractive perturbation property because $\|x_\tau\| \leq \hat{R}_C$, which implies that $\|x_{t_i}\| \leq R_C$. Note that

$$\left\| \frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i+j;t_{i+1}-1}} - \frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i}, \theta_{t_i+j+1;t_{i+1}-1}} \right\|$$

$$\leq \left\| \frac{\partial g_{t_{i+1}|(t_i+j+1)}}{\partial x_{t_i+j+1}} \Bigg|_{\bar{x}_{t_i+j+1}, \theta_{t_i+j+1;t_{i+1}-1}} \right\| \cdot$$

$$\left\| \frac{\partial g_{(t_i+j+1)|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i+j}} - \frac{\partial g_{(t_i+j+1)|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i}} \right\|$$

$$+ \left\| \frac{\partial g_{t_{i+1}|(t_i+j+1)}}{\partial x_{t_i+j+1}} \Bigg|_{\bar{x}_{t_i+j+1}, \theta_{t_i+j+1;t_{i+1}-1}} - \frac{\partial g_{t_{i+1}|(t_i+j+1)}}{\partial x_{t_i+j+1}} \Bigg|_{\bar{x}'_{t_i+j+1}, \theta_{t_i+j+1;t_{i+1}-1}} \right\| \cdot$$

$$\left\| \frac{\partial g_{(t_i+j+1)|(t_i+j)}}{\partial x_{t_i+j}} \Bigg|_{\bar{x}_{t_i+j}, \theta_{t_i+j}} \right\|$$

$$\leq (L_{g,x} + L_{g,u} L_{\pi,x})^{h-j-1} \cdot ((1 + L_{\pi,x}) \ell_{g,u} L_{\pi,\theta} + L_{g,u} \ell_{\pi,\theta}) \|\theta_{t_i+j} - \theta_{t_i}\|$$

$$+ ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x}) + L_{g,u} \cdot \ell_{\pi,x}) \cdot (1 + L_{g,x} + L_{g,u} L_{\pi,x})^{2(h-j)} \cdot$$

$$L_{g,u} L_{\pi,\theta} \|\theta_{t_i+j} - \theta_{t_i}\|$$

$$\leq C'(1 + L_{g,x} + L_{g,u} L_{\pi,x})^{2h} \cdot j\varepsilon,$$

where we adopt the shorthand $\bar{x}'_{t_i+j+1} := g_{t_i+j+1|t_i+j}(\bar{x}_{t_i+j}, \theta_{t_i+j})$, and define

$$C' = ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u} \cdot (L_{\pi,x} + L_{\pi,\theta})) + L_{g,u} \cdot (\ell_{\pi,x} + \ell_{\pi,\theta}))(L_{g,u} L_{\pi,\theta} + 1).$$

Here, we use the chain rule and the triangle inequality in the first inequality. We can apply Lemma C.2 and

$$\left\| \bar{x}'_{t_i+j+1} - \bar{x}_{t_i+j+1} \right\| \leq L_{g,u} L_{\pi,\theta} \|\theta_{t_i+j} - \theta_{t_i}\|$$

in the second inequality because the trajectory induced by $(\theta_{t_i}) \times j, \theta_{t_i+j:t_{i+1}-1}$ always stay within the ball $B(0, R_x)$ where the Lipschitzness/smoothness of dynamics/policies hold. Substituting this into (7) gives

$$\left\| \frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}} \Bigg|_{x_{t_i}, \theta_{t_i:t_{i+1}-1}} - \frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}} \Bigg|_{x_{t_i}, (\theta_{t_i}) \times h} \right\| \leq CC'(1 + L_{g,x} + L_{g,u} L_{\pi,x})^{2h} \varepsilon \sum_{j=0}^{h-1} \rho^j \cdot j$$

18

$$\leq \frac{CC'\rho(1 + L_{g,x} + L_{g,u}L_{\pi,x})^{2h}}{(1-\rho)^2} \cdot \varepsilon$$

$$\leq \hat{\rho}^h - C\rho^h. \tag{8}$$

Therefore, by (6), we see that

$$\left\| \frac{\partial g_{t|\tau}}{\partial x_\tau} \bigg|_{x_\tau, \theta_{\tau:t-1}} \right\|$$

$$\leq \left\| \frac{\partial g_{t_p|t_{p-1}}}{\partial x_{t_{p-1}}} \bigg|_{x_{t_{p-1}}, \theta_{t_{p-1}:t_p-1}} \right\| \cdot \left\| \frac{\partial g_{t_{p-1}|t_{p-2}}}{\partial x_{t_{p-2}}} \bigg|_{x_{t_{p-2}}, \theta_{t_{p-2}:t_{p-1}-1}} \right\| \cdots \left\| \frac{\partial g_{t_1|t_0}}{\partial x_{t_0}} \bigg|_{x_{t_0}, \theta_{t_0:t_1-1}} \right\|$$

$$\leq ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x}) + L_{g,u} \cdot \ell_{\pi,x}) \cdot (1 + L_{g,x} + L_{g,u}L_{\pi,x})^{2h} \cdot (\hat{\rho}^h)^{p-1} \tag{9a}$$

$$\leq \hat{C}(\hat{\rho})^{t-\tau}, \tag{9b}$$

where we use (8) in (9a); we use the definition of $\hat{C}$ in (9b). This finishes the proof of ε-time-varying contractive perturbation with $\hat{\epsilon}, \hat{R}_C, \hat{C}, \hat{\rho}$. $\square$

D Regret of GAPS under convex surrogate costs

In this section, we provide a proof outline of the adaptive regret bound for GAPS. As we discussed in Section 3, 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 size and computational complexity. While the existing literature of OCO guarantees that the ideal OGD update with constant step size η of the order $1/\sqrt{T}$ achieves a policy regret of $O\left(\sqrt{T}\right)$, GAPS incurs an approximation error at every time step since it uses G_t (Algorithm 1, Line 1) 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 D.1. The proof is deferred to Appendix D.2.

Theorem D.1. *Consider the update rule $\theta_{t+1} = \prod_\Theta(\theta_t - \eta G_t)$. Suppose Θ is a convex compact set with diameter D. If F_t is convex and $\|\nabla F_t(\theta)\| \leq W$ for all $\theta \in \Theta$, and $\|\nabla F_t(\theta_t) - G_t\| \leq \alpha$ holds for all time steps t, then, for arbitrary $I = [r:s] \subseteq \mathcal{T}$,*

$$\sum_{t=r}^{s} F_t(\theta_t) - \min_{\theta_I \in \Theta} \sum_{t=r}^{s} F_t(\theta_I) \leq \alpha DT + (W^2 + \alpha^2)\eta T + \frac{D^2}{2\eta}.$$

With Theorem D.1, obtaining the policy regret bounds for GAPS reduces to showing both $|f_t(x_t, u_t) - F_t(\theta_t)|$ and $\|\nabla F_t(\theta_t) - G_t\|$ are in the order of $O(1/\sqrt{T})$. Here, we only consider the order of magnitude with respect to the horizon T for clarity. As we will show in Theorem D.5 and Theorem D.6, both of these quantities are in the order of $O(\eta)$ when GAPS adopts the learning rate η.

To obtain these results, we first show a lemma about the stability of the trajectory achieved by an ε-time-varying policy parameter sequence.

Lemma D.2. *Suppose Assumptions 2.1 and 2.2 hold. For any starting state $x_\tau \in B_n(0, R_S + C\|x_0\|)$ and $\theta_{\tau:t-1} \in S_\varepsilon(\tau:t-1)$, the final state $x_t := g_{t|\tau}(x_\tau, \theta_{\tau:t-1})$ satisfies $\|x_t\| \leq C\rho^{t-\tau}\|x_\tau\| + R_S$.*

Proof of Lemma D.2. By ε-time-varying contractive perturbation, we see that

$$\left\| x_t - g_{t|\tau}(0, \theta_{\tau:t-1}) \right\| \leq C\rho^{t-\tau}\|x_\tau\|.$$

Thus, by the triangle inequality, we see that

$$\|x_t\| \leq \left\| x_t - g_{t|\tau}(0, \theta_{\tau:t-1}) \right\| + \left\| g_{t|\tau}(0, \theta_{\tau:t-1}) \right\| \leq C\rho^{t-\tau}\|x_\tau\| + R_S,$$

where we use ε-time-varying stability in the last inequality. $\square$

Next, we show a lemma about the contractive property of the partial derivatives of the multi-step dynamics.

19

Lemma D.3 (Lipschitzness/Smoothness of the Multi-Step Dynamics). *Suppose Assumptions 2.1 and 2.2 hold. Given two time steps $t > \tau$, for any $x_\tau, x'_\tau \in B_n(0, R_S + C\|x_0\|)$ and $\theta_\tau, \theta'_\tau \in \Theta$, $\theta_{\tau+1:t-1} \in S_\varepsilon(\tau+1 : t-1)$, if $x'_{\tau+1} := g_{\tau+1|\tau}(x'_\tau, \theta'_\tau)$ is also in $B_n(0, R_S + C\|x_0\|)$, the multi-step dynamical function $g_{t|\tau}$ satisfies that*

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

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

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

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

$$C_{\ell,g,(x,x)} = ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,x}\ell_{\pi,x})C^3\rho^{-1}(1-\rho)^{-1},$$

$$C_{\ell,g,(x,\theta)} = ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,x}\ell_{\pi,x})C^3 L_{g,u}L_{\pi,\theta}\rho^{-1}(1-\rho)^{-1}$$
$$+ ((1 + L_{\pi,x})\ell_{g,u}L_{\pi,\theta} + L_{g,u}\ell_{\pi,\theta})C\rho^{-1}(1-\rho)^{-1},$$

$$C_{\ell,g,(\theta,x)} = ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,x}\ell_{\pi,x})(L_{g,x} + L_{g,u}L_{\pi,x})C^3 L_{g,u}L_{\pi,\theta}\rho^{-2}(1-\rho)^{-1}$$
$$+ C(L_{\pi,\theta}(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x})\rho^{-1},$$

$$C_{\ell,g,(\theta,\theta)} = ((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x}) + L_{g,x} \cdot \ell_{\pi,x})L_{g,u}^2 L_{\pi,\theta}^2 C^3 \rho^{-2}(1-\rho)^{-1}$$
$$+ (L_{g,u}\ell_{\pi,\theta} + \ell_{g,u}L_{\pi,\theta}^2)C\rho^{-1}.$$

Intuitively, Lemma D.3 shows that the dependence of the state x_t on the previous state x_τ and θ_τ 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_τ and $\theta_{\tau:t-1}$. When other variables are fixed, the Lipschitzness and smoothness constants with respect to x_τ and θ_τ are both $O(\rho^{t-\tau})$. While the contractive Lipschitzness on x_τ is automatically guaranteed by ε-time-varying contractive perturbation (Definition 2.6), we use this property and the chain rule decomposition to show the Lipschitzness on θ_τ and the smoothness.

The first inequality in Lemma D.3 directly follows from ε-time-varying contractive perturbation. To reflect the main technical difficulty, we show the third inequality here with the assumption that the first two inequalities hold. The proof of other inequalities are deferred to Appendix D.3.

Proof of the 3rd inequality in Lemma D.3. Note that we have the chain rule decomposition

$$\frac{\partial g_{t|\tau}}{\partial x_\tau} \Big|_{x_\tau, \theta_\tau, \theta_{\tau+1:t-1}} = \frac{\partial g_{t|t-1}}{\partial x_{t-1}} \Big|_{x_{t-1}, \theta_{t-1}} \cdot \frac{\partial g_{t-1|t-2}}{\partial x_{t-2}} \Big|_{x_{t-2}, \theta_{t-2}} \cdots \frac{\partial g_{\tau+1|\tau}}{\partial x_\tau} \Big|_{x_\tau, \theta_\tau},$$

$$\frac{\partial g_{t|\tau}}{\partial x_\tau} \Big|_{x'_\tau, \theta'_\tau, \theta_{\tau+1:t-1}} = \frac{\partial g_{t|t-1}}{\partial x_{t-1}} \Big|_{x'_{t-1}, \theta_{t-1}} \cdot \frac{\partial g_{t-1|t-2}}{\partial x_{t-2}} \Big|_{x'_{t-2}, \theta_{t-2}} \cdots \frac{\partial g_{\tau+1|\tau}}{\partial x_\tau} \Big|_{x'_\tau, \theta'_\tau}, \quad (10)$$

where we use the notation $x_{\tau'} = g_{\tau'|\tau}(x_\tau, \theta_{\tau:\tau'-1})$ and $x'_{\tau'} = g_{\tau'|\tau}(x'_\tau, \theta'_\tau, \theta_{\tau+1:\tau'-1})$ for $\tau' \in [\tau+1 : t-1]$.

Note that for any $i \in [1 : t - \tau]$ and any $\theta'_{t-i} \in \Theta$, we have the decomposition

$$\frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}} \Big|_{x_{t-i}, \theta_{t-i}} - \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}} \Big|_{x'_{t-i}, \theta'_{t-i}}$$

$$= \frac{\partial g_{t-i}}{\partial x_{t-i}} \Big|_{x_{t-i}, u_{t-i}} - \frac{\partial g_{t-i}}{\partial x_{t-i}} \Big|_{x'_{t-i}, u'_{t-i}} + \frac{\partial g_{t-i}}{\partial u_{t-i}} \Big|_{x_{t-i}, u_{t-i}} \frac{\partial \pi_{t-i}}{\partial x_{t-i}} \Big|_{x_{t-i}, \theta_{t-i}}$$

$$- \frac{\partial g_{t-i}}{\partial x_{t-i}} \Big|_{x'_{t-i}, u'_{t-i}} \frac{\partial \pi_{t-i}}{\partial x_{t-i}} \Big|_{x'_{t-i}, \theta'_{t-i}}$$

20

$$
= \frac{\partial g_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},u_{t-i}} - \frac{\partial g_{t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}} + \left(\frac{\partial g_{t-i}}{\partial u_{t-i}}\bigg|_{x_{t-i},u_{t-i}} - \frac{\partial g_{t-i}}{\partial u_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}} \right) \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}}
$$

$$
+ \frac{\partial g_{t-i}}{\partial u_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}} \left(\frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}} - \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},\theta'_{t-i}} \right),
$$

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

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

$$
\leq \ell_{g,x}\|x_{t-i} - x'_{t-i}\| + \ell_{g,u}\|\pi_{t-i}(x_{t-i},\theta_{t-i}) - \pi_{t-i}(x'_{t-i},\theta'_{t-i})\|
$$

$$
+ L_{\pi,x}\left(\ell_{g,x}\|x_{t-i} - x'_{t-i}\| + \ell_{g,u}\|\pi_{t-i}(x_{t-i},\theta_{t-i}) - \pi_{t-i}(x'_{t-i},\theta'_{t-i})\|\right)
$$

$$
+ L_{g,u} \cdot \left(\ell_{\pi,x}\|x_{t-i} - x'_{t-i}\| + \ell_{\pi,\theta}\|\theta_{t-i} - \theta'_{t-i}\|\right) \tag{11a}
$$

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

$$
+ \left((1 + L_{\pi,x})\ell_{g,u}L_{\pi,\theta} + L_{g,u}\ell_{\pi,\theta}\right)\|\theta_{t-i} - \theta'_{t-i}\|, \tag{11b}
$$

where we use Assumption 2.1 and the definition of u_{t-i}, u'_{t-i} in (11a); and Assumption 2.1 in (11b). Therefore, by (10) and (11), we see that

$$
\left\| \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t-1}} - \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta_{\tau:t-1}} \right\|
$$

$$
\leq \sum_{i=1}^{t-\tau-1} \left(\left\| \prod_{\tau'=1}^{i-1} \frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\bigg|_{x'_{t-\tau'},\theta_{t-\tau'}} \right\| \cdot \left\| \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}} - \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},\theta_{t-i}} \right\| \cdot \right.
$$

$$
\left. \left\| \prod_{\tau'=i+1}^{t-\tau} \frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\bigg|_{x_{t-\tau'},\theta_{t-\tau'}} \right\| \right)
$$

$$
+ \left\| \prod_{\tau'=1}^{t-\tau-1} \frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\bigg|_{x'_{t-\tau'},\theta_{t-\tau'}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau} \right\|
$$

$$
\leq \sum_{i=1}^{t-\tau} (C\rho^{i-1}) \cdot \left((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}\cdot L_{\pi,x}) + L_{g,x}\cdot\ell_{\pi,x}\right)\|x_{t-i} - x'_{t-i}\| \cdot (C\rho^{t-\tau-i})
$$

$$
+ C\rho^{t-\tau-1} \cdot \left((1 + L_{\pi,x})\ell_{g,u}L_{\pi,\theta} + L_{g,u}\ell_{\pi,\theta}\right)\|\theta_\tau - \theta'_\tau\| \tag{12a}
$$

$$
= \left((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}\cdot L_{\pi,x}) + L_{g,x}\cdot\ell_{\pi,x}\right)C^2 \cdot \rho^{t-\tau-1} \sum_{i=1}^{t-\tau}\|x_{t-i} - x'_{t-i}\|
$$

$$
+ C\rho^{t-\tau-1} \cdot \left((1 + L_{\pi,x})\ell_{g,u}L_{\pi,\theta} + L_{g,u}\ell_{\pi,\theta}\right)\|\theta_\tau - \theta'_\tau\|
$$

$$
\leq C_{\ell,g,(x,x)} \cdot \rho^{t-\tau}\|x_\tau - x'_\tau\| + C_{\ell,g,(x,\theta)} \cdot \rho^{t-\tau}\|\theta_\tau - \theta'_\tau\|, \tag{12b}
$$

where we use the ε-time-varying contractive perturbation property and (11) in (12a); we use the first two inequalities to bound $\|x_{t-i} - x'_{t-i}\| \leq C\rho^{t-i-\tau}\|x_\tau - x'_\tau\| + \frac{CL_{g,u}L_{\pi,\theta}}{\rho} \cdot \rho^{t-i-\tau}\|\theta_\tau - \theta'_\tau\|$ in (12b). $\qquad\square$

Since we will need more general forms of policy sequence later to bound $\|G_t - \nabla F_t\|$ than the sequence with small step sizes, we state the contractive Lipschitzness/smoothness of the multi-step cost function $f_{t|\tau}$. This is an implication of Lemma D.3 because for any $\tau < t$, the previous state x_τ and previous policy parameter θ_τ can only affect the current stage cost c_t by affecting the current state x_t. We formalize this result in Corollary D.4 and defer the detailed proof to Appendix D.4.

Corollary D.4 (Lipschitzness/Smoothness of the Multi-Step Costs). *Under the same assumptions as Lemma D.3, let*

$$
x_t := g_{t|\tau}(x_\tau,\theta_{\tau:t-1}), u_t := \pi_t(x_t,\theta_t); \text{ and } x'_t := g_{t|\tau}(x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}), u'_t := \pi_t(x'_t,\theta_t).
$$

21

Then, the multi-step cost function $f_{t|\tau}$ satisfies that

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

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

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

where $C_{L,f,x} = L_f C(1 + L_{\pi,x}), C_{L,f,\theta} = L_f \max\{C_{L,g,\theta}(1 + L_{\pi,x}), L_{\pi,\theta}\}$, and

$$C_{\ell,f,(x,x)} = L_f(1 + L_{\pi,x})C_{\ell,g,(x,x)} + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) + L_f\ell_{\pi,x})C_{L,g,x}^2,$$

$$C_{\ell,f,(x,\theta)} = L_f(1 + L_{\pi,x})C_{\ell,g,(x,\theta)} + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) + L_f\ell_{\pi,x})C_{L,g,x}C_{L,g,\theta},$$

$$C_{\ell,f,(\theta,x)} = L_f(1 + L_{\pi,x})C_{\ell,g,(\theta,x)} + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) + L_f\ell_{\pi,x})C_{L,g,x}C_{L,g,\theta},$$

$$C_{\ell,f,(\theta,\theta)} = L_f(1 + L_{\pi,x})C_{\ell,g,(\theta,\theta)} + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) + L_f\ell_{\pi,x})C_{L,g,\theta}^2.$$

With the help of Lemma D.3 and Corollary D.4, we first bound the cost difference $|f_t(x_t, u_t) - F_t(\theta_t)|$ in Theorem D.5. 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 D.5, we also bound the distance between GAPS' trajectory and the imaginary trajectory if the same policy parameter θ_t had been used from time 0 to time t, which will be useful for showing Theorem D.6 later in this section.

Theorem D.5. *Suppose Assumptions 2.1 and 2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of GAPS (Algorithm 1) with buffer size B and a constant learning rate $\eta \le \frac{(1-\rho)\varepsilon}{C_{L,f,\theta}}$. Then, both $\|G_t\|$ and $\|\nabla F_t(\theta_t)\|$ are upper bounded by $\frac{C_{L,f,\theta}}{1-\rho}$, and the following inequalities holds for any two time steps τ, t ($\tau \le t$):*

$$\|\theta_t - \theta_\tau\| \le \frac{C_{L,f,\theta}}{1-\rho} \cdot (t-\tau)\eta, \text{ and } \|x_\tau - \hat{x}_\tau(\theta_t)\| \le \frac{C_{L,f,\theta}C_{L,g,\theta}\rho}{(1-\rho)^2}\left((t-\tau) + \frac{1}{1-\rho}\right) \cdot \eta,$$

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

$$|f_t(x_t, u_t) - F_t(\theta_t)| \le \frac{C_{L,f,\theta}C_{L,g,\theta}L_f(1 + L_{\pi,x})\rho}{(1-\rho)^3} \cdot \eta.$$

In addition, for any parameter sequence $\tilde{\theta}_{0:t} \in \Theta^{t+1}$, let $\tilde{x}_t$ and $\tilde{u}_t$ be the state/control action achieved by this sequence $\tilde{x}_t := g_{t|0}(x_0, \tilde{\theta}_{0:t-1})$ and $\tilde{u}_t := \pi_t(\tilde{x}_t, \tilde{\theta}_t)$. If $\|\tilde{x}_t\| \le \min\{R_C, R_x\}$ holds for all t, then the following inequality holds for all time t:

$$\left|f_t(\tilde{x}_t, \tilde{u}_t) - F_t(\tilde{\theta}_t)\right| \le \frac{CL_{\pi,\theta}L_{g,x}L_f(1 + L_{\pi,x})}{1-\rho}\sum_{\tau=0}^{t-1}\rho^{t-\tau-1}\left\|\tilde{\theta}_{\tau+1} - \tilde{\theta}_\tau\right\|.$$

To show Theorem D.5, 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 η. We then leverage strong contractive perturbation to bound $\|x_\tau - \hat{x}_\tau(\theta_t)\|$ and use it to bound $|f_t(x_t, u_t) - F_t(\theta_t)|$ by the Lipschitzness of f_t. We defer the detailed proof to Appendix D.5.

In Theorem D.6 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 D.6 (Gradient Bias). *Suppose Assumptions 2.1 and 2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of GAPS (Algorithm 1) with buffer size B and learning rate $\eta \le \frac{(1-\rho)\varepsilon}{C_{L,f,\theta}}$. Then, the following holds for all $\tau \le t$:*

$$\left\|\left.\frac{\partial f_{t|0}}{\partial \theta_\tau}\right|_{x_0,\theta_{0:t}} - \left.\frac{\partial f_{t|0}}{\partial \theta_\tau}\right|_{x_0,(\theta_t)_{\times(t+1)}}\right\| = O\left(\left(\frac{1}{(1-\rho)^4} + \frac{t-\tau}{(1-\rho)^3} + \frac{(t-\tau)^2}{(1-\rho)^2}\right)\rho^{t-\tau} \cdot \eta\right),$$

22

Further, we see that

$$\|G_t - \nabla F_t(\theta_t)\| \le O\left(\frac{\eta}{(1-\rho)^5} + \frac{\rho^B}{1-\rho}\right).$$

(See Appendix D.6 for the detailed expressions.)

The key technique we used to show Theorem D.6 is a sequential decomposition of the error based on the triangle inequality. Specifically, note that Corollary D.4 only allow us to compare the partial derivatives when $\theta_{\tau+1:t}$ are fixed and the only perturbations are on x_τ and θ_τ. To compare the partial derivatives realized on two trajectory instances $(\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)})$ and $(x_\tau, \theta_{\tau:t})$, we change the parameters sequentially one by one, following the path

$$(\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)}) \to (x_\tau, \theta_\tau, (\theta_t)_{\times(t-\tau)}) \to (x_\tau, \theta_{\tau:\tau+1}, (\theta_t)_{\times(t-\tau-1)}) \to \cdots \to (x_\tau, \theta_{\tau:t}).$$

A complete proof of Theorem D.6 can be found in Appendix D.6.

The bounds in Theorems D.5 and D.6 show that we can achieve our desired bounds $|f_t(x_t, u_t) - F_t(\theta_t)| = O(1/\sqrt{T})$ and $\|\nabla F_t(\theta_t) - G_t\| = O(1/\sqrt{T})$ if we set the learning rate and the buffer length to be $O(1/\sqrt{T})$ and $O(\log T)$ respectively. Substituting these bounds into Theorem D.1 with a more careful analysis on the order of the factor $1/(1-\rho)$ will finish the proof of Theorem 3.3. The detailed proof can be found in Appendix D.6.

D.1 Detailed Statement and Proofs of Theorems 3.2 and 3.3

We restate Theorem 3.2 with detailed expressions in Theorem D.7.

Theorem D.7. *Suppose Assumptions 2.1 and 2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t \in \mathcal{T}}$ denote the trajectory of GAPS (Algorithm 1) with buffer size B and learning rate $\eta_t = \eta \le \frac{(1-\rho)\varepsilon}{C_{L,f,\theta}}$, where $C_{L,f,\theta}$ is defined in Corollary D.4. Then, we have*

$$|f_t(x_t, u_t) - F_t(\theta_t)| \le \frac{C_{L,f,\theta}C_{L,g,\theta}L_f(1 + L_{\pi,x})\rho}{(1-\rho)^3} \cdot \eta, \text{ and}$$

$$\|G_t - \nabla F_t(\theta_t)\| \le \left(\hat{C}_0(1-\rho)^{-1} + (\hat{C}_1 + \hat{C}_2)(1-\rho)^{-2} + \hat{C}_2(1-\rho)^{-3}\right)\eta$$
$$+ C_{L,f,\theta}(1-\rho)^{-1} \cdot \rho^B, \tag{13}$$

where $\hat{C}_0, \hat{C}_1, \hat{C}_2$ are defined in Theorem D.9.

Proof of Theorem D.7. Theorem D.7 directly followed from Theorems D.5 and D.6. □

We restate Theorem 3.3 with detailed expressions in Theorem D.8.

Theorem D.8. *Under the same assumptions as Theorem 3.2, if we additionally assume the surrogate stage cost F_t is convex for every time step t, then GAPS achieves the adaptive regret bound*

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

Proof of Theorem D.8. The first two inequalities are shown in Theorem D.5 and Theorem D.6. Thus, we focus on the adaptive regret part here in the proof.

Fix a time interval $I = [r : s] \subseteq \mathcal{T}$ and let θ_I be an arbitrary policy parameter in Θ. By Theorem D.1 and Theorem D.6, we see that the sequence of policy parameters of the online policy satisfies that

$$\sum_{t=r}^{s} F_t(\theta_t) - \sum_{t=r}^{s} F_t(\theta_I) \le (W^2 + \alpha^2)\eta T + \frac{D^2}{2\eta} + \alpha DT, \tag{14}$$

23

where $W = \frac{C_{L,f,\theta}}{1-\rho}$ by Theorem D.5 and $\alpha = \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3} \right) \eta + \frac{C_{L,f,\theta}}{1-\rho} \cdot \rho^B$.

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 D.5, we have

$$|f_t(x_t, u_t) - F_t(\theta_t)| \leq \frac{C_{L,f,\theta} C_{L,g,\theta} L_f (1 + L_{\pi,x}) \rho}{(1-\rho)^3} \cdot \eta. \tag{15}$$

Substituting these into (14) gives that

$$\sum_{t=r}^{s} f_t(x_t, u_t) - \sum_{t=r}^{s} f_t(\hat{x}_t(\theta_I), \hat{u}_t(\theta_I))$$

$$\leq \left(\sum_{t=r}^{s} F_t(\theta_t) - \sum_{t=r}^{s} F_t(\theta_I) \right) + \sum_{t=r}^{s} |f_t(x_t, u_t) - F_t(\theta_t)|$$

$$\leq \frac{D^2}{2\eta} + \alpha D T + \left(W^2 + \alpha^2 + \frac{C_{L,f,\theta} C_{L,g,\theta} L_f (1 + L_{\pi,x}) \rho}{(1-\rho)^3} \right) \cdot \eta T$$

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

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

where we used (14) and (15) in the second inequality. $\qquad\square$

D.2 Proof of Theorem D.1

Our proof is inspired by the proof of Theorem 2.1 in [56]. For a fixed time interval $I = [r : s] \subseteq \mathcal{T}$ and $\theta_I \in \Theta$, we consider the potential function $\Phi_t = \frac{1}{2\eta} \|\theta_t - \theta_I\|^2$. Note that θ_I satisfies $\|\theta_r - \theta_I\| \leq D$ because we assume $\mathrm{diam}(\Theta) \leq D$. To simplify the notation, we define $\theta'_{t+1} = \theta_t - \eta G_t$.

By proposition 2.2 in [56], we see that

$$\frac{1}{2} (\|\theta_{t+1} - \theta_I\|^2 - \|\theta_t - \theta_I\|^2) \leq \frac{1}{2} (\|\theta'_{t+1} - \theta_I\|^2 - \|\theta_t - \theta_I\|^2)$$

$$= \langle \theta'_{t+1} - \theta_t, \theta_t - \theta_I \rangle + \frac{1}{2} \|\theta'_{t+1} - \theta_t\|^2$$

$$= \eta \langle G_t, \theta_I - \theta_t \rangle + \frac{\eta^2}{2} \|G_t\|^2.$$

Using this inequality, we see that

$$F_t(\theta_t) - F_t(\theta_I) + \Phi_{t+1} - \Phi_t$$

$$= F_t(\theta_t) - F_t(\theta_I) + \langle G_t, \theta_I - \theta_t \rangle + \frac{\eta}{2} \|G_t\|^2$$

$$= F_t(\theta_t) - F_t(\theta_I) + \langle \nabla F_t(\theta_t) + (G_t - \nabla F_t(\theta_t)), \theta_I - \theta_t \rangle + \frac{\eta}{2} \|\nabla F_t(\theta_t) + (G_t - F_t(\theta_t))\|^2$$

$$\leq F_t(\theta_t) - F_t(\theta_I) + \langle \nabla F_t(\theta_t), \theta_I - \theta_t \rangle + \langle G_t - \nabla F_t(\theta_t), \theta_I - \theta_t \rangle + \eta \|\nabla F_t(\theta_t)\|^2$$

$$+ \eta \|G_t - \nabla F_t(\theta_t)\|^2 \tag{16a}$$

$$\leq 0 + \|G_t - \nabla F_t(\theta_t)\| \cdot \|\theta_I - \theta_t\| + \eta \|\nabla F_t(\theta_t)\|^2 + \eta \alpha^2 \tag{16b}$$

$$\leq \alpha D + W^2 \eta + \eta \alpha^2, \tag{16c}$$

where we used the triangle inequality and the AM-GM inequality in (16a); we used the assumption that F_t is convex, $\|G_t - \nabla F_t(\theta_t)\| \leq \alpha$, and the Cauchy-Schwarz inequality in (16b); and we used the assumptions $\|G_t - \nabla F_t(\theta_t)\| \leq \alpha$, $\mathrm{diam}(\Theta) \leq D$, $\|\nabla F_t(\theta_t)\| \leq W$ in (16c).

Summing (16) over the time interval $[r : s]$ gives that

$$\sum_{t=r}^{s} (F_t(\theta_t) - F_t(\theta_I)) \leq (s - r) \cdot \left(\alpha D + W^2 \eta + \eta \alpha^2 \right) + (\Phi_r - \Phi_{s+1})$$

24

$$\leq \left(\alpha D + W^2\eta + \eta\alpha^2\right)T + \frac{D^2}{2\eta},$$

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

D.3 Proof of Inequalities 1,2, and 4 in Lemma D.3

The first inequality directly follows from ε-time-varying contractive perturbation (Definition 2.6). For the second inequality, when $t = \tau + 1$, note that $\left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} = \left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x_\tau,u_\tau} \cdot \left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau}$, where $u_\tau = \pi_\tau(x_\tau, \theta_\tau)$. Taking norms of both sides of the equation gives

$$\left\| \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\| = \left\| \left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x_\tau,u_\tau} \cdot \left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\| \leq \left\| \left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x_\tau,u_\tau} \right\| \cdot \left\| \left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\| \leq L_{g,u}L_{\pi,\theta}.$$

When $t > \tau + 1$, we see that

$$\left\| \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t-1}} \right\| = \left\| \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x_{\tau+1},\theta_{\tau+1:t-1}} \cdot \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\|$$

$$\leq \left\| \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\| \leq \frac{C_0 L_{g,u}L_{\pi,\theta}}{\rho} \cdot \rho^{t-\tau},$$

where $x_{\tau+1} = g_{\tau+1|\tau}(x_\tau, \theta_\tau)$.

For the last inequality of Lemma D.3, when $t = \tau + 1$, we see that

$$\left\| \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} - \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau} \right\|$$

$$= \left\| \left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x_\tau,u_\tau} \cdot \left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} - \left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x'_\tau,u'_\tau} \cdot \left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau} \right\|$$

$$\leq \left\| \left(\left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x_\tau,u_\tau} - \left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x'_\tau,u'_\tau} \right) \cdot \left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\| + \left\| \left.\frac{\partial g_\tau}{\partial u_\tau}\right|_{x'_\tau,u'_\tau} \left(\left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} - \left.\frac{\partial \pi_\tau}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau} \right) \right\|$$

$$\tag{17a}$$

$$\leq L_{\pi,\theta}(\ell_{g,x}\|x_\tau - x'_\tau\| + \ell_{g,u}\|u_\tau - u'_\tau\|) + L_{g,u}(\ell_{\pi,x}\|x_\tau - x'_\tau\| + \ell_{\pi,\theta}\|\theta_\tau - \theta'_\tau\|) \tag{17b}$$

$$\leq (L_{\pi,\theta}(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x})\|x_\tau - x'_\tau\| + (L^2_{\pi,\theta}\ell_{g,u} + L_{g,u}\ell_{\pi,\theta})\|\theta_\tau - \theta'_\tau\|, \tag{17c}$$

where we use the notations $u_\tau = \pi_\tau(x_\tau, \theta_\tau), u'_\tau = \pi_\tau(x_\tau, \theta'_\tau)$. We use the triangle inequality in (17a); we use Assumption 2.1 in both (17b) and (17c).

When $t > \tau + 1$, we see that

$$\left\| \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\|$$

$$\leq \left\| \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x_{\tau+1},\theta_{\tau+1:t-1}} \cdot \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} - \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \cdot \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau} \right\| \tag{18a}$$

$$\leq \left\| \left(\left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x_{\tau+1},\theta_{\tau+1:t-1}} - \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right) \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\|$$

$$+ \left\| \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \cdot \left(\left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} - \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta'_\tau} \right) \right\| \tag{18b}$$

$$\leq \left\| \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x_{\tau+1},\theta_{\tau+1:t-1}} - \left.\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\right|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau} \right\|$$

25

$$+ \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}} \right|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau} \right|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau} \right|_{x'_\tau,\theta'_\tau} \right\|$$

$$\leq \frac{((1+L_{\pi,x})(\ell_{g,x}+\ell_{g,u}\cdot L_{\pi,x}) + L_{g,x}\cdot\ell_{\pi,x})C_0^3}{\rho(1-\rho)} \cdot \rho^{t-\tau-1} \cdot \left\| x_{\tau+1} - x'_{\tau+1} \right\| \cdot L_{g,u}L_{\pi,\theta}$$

$$+ C_0 \cdot \rho^{t-\tau-1} \cdot (L_{\pi,\theta}(\ell_{g,x}+\ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x})\|x_\tau - x'_\tau\|$$

$$+ C_0 \cdot \rho^{t-\tau-1} \cdot (L_{\pi,\theta}^2\ell_{g,u} + L_{g,u}\ell_{\pi,\theta})\|\theta_\tau - \theta'_\tau\| \tag{18c}$$

$$\leq C_{\ell,g,(\theta,x)}\rho^{t-\tau}\|x_\tau - x'_\tau\| + C_{\ell,g,(\theta,\theta)}\rho^{t-\tau}\|\theta_\tau - \theta'_\tau\|, \tag{18d}$$

where we use the notations $x_{\tau+1} = g_{\tau+1|\tau}(x_\tau,\theta_\tau), x'_{\tau+1} = g_{\tau+1|\tau}(x_\tau,\theta'_\tau)$. We use the chain rule decomposition in (18a); we use the triangle inequality in (18b); we use the first and the third inequality of Lemma D.3 as well as (17) in (18c); we use the first two inequalities of Lemma D.3 in (18d).

D.4 Proof of Corollary D.4

To show the first inequality, note that

$$\frac{\partial f_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t}} = \left(\frac{\partial f_t}{\partial x_t}\bigg|_{x_t,u_t} + \frac{\partial f_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t-1}}, \tag{19}$$

where $x_t = g_{t|\tau}(x_\tau,\theta_{\tau:t-1}), u_t = \pi_t(x_t,\theta_t)$. Thus, by ε-time-varying contractive perturbation, we see that

$$\left\| \frac{\partial f_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t}} \right\| \leq \left(\left\| \frac{\partial f_t}{\partial x_t}\bigg|_{x_t,u_t} \right\| + \left\| \frac{\partial f_t}{\partial u_t}\bigg|_{x_t,u_t} \right\| \cdot \left\| \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \right\| \right) \cdot \left\| \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t-1}} \right\|$$

$$\leq L_f(1+L_{\pi,x}) \cdot C\rho^{t-\tau}.$$

For the second inequality, when $\tau = t$, since $x_t \in B_n(0,R'_x)$ and $u_t \in B_m(0,R'_u)$, we see that

$$\left\| \frac{\partial f_{t|t}}{\partial \theta_t}\bigg|_{x_t,\theta_t} \right\| = \left\| \frac{\partial f_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} \right\| \leq \left\| \frac{\partial f_t}{\partial u_t}\bigg|_{x_t,u_t} \right\| \cdot \left\| \frac{\partial \pi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} \right\| \leq L_f L_{\pi,\theta}.$$

When $\tau < t$, the second inequality can be shown similarly with the first inequality in Corollary D.4 because we have the chain-rule decomposition

$$\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau,\theta_{\tau:t}} = \left(\frac{\partial f_t}{\partial x_t}\bigg|_{x_t,u_t} + \frac{\partial f_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau,\theta_{\tau:t-1}}. \tag{20}$$

Applying Lemma D.3 gives that $C_{L,f,\theta} = L_f C_{L,g,\theta}(1+L_{\pi,x})$.

For the third inequality, using (19), we see that

$$\left\| \frac{\partial f_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \frac{\partial f_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t}} \right\|$$

$$\leq \left\| \frac{\partial f_t}{\partial x_t}\bigg|_{x_t,u_t} \cdot \left(\frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\|$$

$$+ \left\| \left(\frac{\partial f_t}{\partial x_t}\bigg|_{x_t,u_t} - \frac{\partial f_t}{\partial x_t}\bigg|_{x'_t,u'_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\|$$

$$+ \left\| \frac{\partial f_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \cdot \left(\frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\|$$

$$+ \left\| \frac{\partial f_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \left(\frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} - \frac{\partial \pi_t}{\partial x_t}\bigg|_{x'_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\|$$

26

$$+ \left\| \left(\left. \frac{\partial f_t}{\partial u_t} \right|_{x_t,u_t} - \left. \frac{\partial f_t}{\partial u_t} \right|_{x'_t,u'_t} \right) \cdot \left. \frac{\partial \pi_t}{\partial x_t} \right|_{x'_t,\theta_t} \cdot \left. \frac{\partial g_{t|\tau}}{\partial x_\tau} \right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \tag{21a}$$

$$\leq L_f \rho^{t-\tau} \left(C_{\ell,g,(x,x)} \| x_\tau - x'_\tau \| + C_{\ell,g,(x,\theta)} \| \theta_\tau - \theta'_\tau \| \right)$$
$$+ (\ell_{f,x} \| x_t - x'_t \| + \ell_{f,u} \| u_t - u'_t \|) \cdot C_{L,g,x} \rho^{t-\tau}$$
$$+ L_f L_{\pi,x} \rho^{t-\tau} \left(C_{\ell,g,(x,x)} \| x_\tau - x'_\tau \| + C_{\ell,g,(x,\theta)} \| \theta_\tau - \theta'_\tau \| \right)$$
$$+ L_f \ell_{\pi,x} \| x_t - x'_t \| \cdot C_{L,g,x} \rho^{t-\tau} + (\ell_{f,x} \| x_t - x'_t \| + \ell_{f,u} \| u_t - u'_t \|) \cdot L_{\pi,x} C_{L,g,x} \rho^{t-\tau}, \tag{21b}$$

where we use the notations $x_t = g_{t|\tau}(x_\tau, \theta_{\tau:t-1}), x'_t = g_{t|\tau}(x'_\tau, \theta_{\tau:t-1}), u_t = \pi_t(x_t, \theta_t), u'_t = \pi_t(x'_t, \theta_t)$. We use (19) and the triangle inequality in (21a); we use Lemma D.3 in (21b). Note that by the first two inequalities in Lemma D.3, we have

$$\| x_t - x'_t \| \leq \rho^{t-\tau} (C_{L,g,x} \| x_\tau - x'_\tau \| + C_{L,g,\theta} \| \theta_\tau - \theta'_\tau \|),$$
$$\| u_t - u'_t \| \leq L_{\pi,x} \rho^{t-\tau} (C_{L,g,x} \| x_\tau - x'_\tau \| + C_{L,g,\theta} \| \theta_\tau - \theta'_\tau \|).$$

Substituting these two inequalities into (21) finishes the proof of the third inequality.

For the last inequality, when $\tau = t$, we have that

$$\left\| \left. \frac{\partial f_{t,t}}{\partial \theta_t} \right|_{x_t,\theta_t} - \left. \frac{\partial f_{t,t}}{\partial \theta_t} \right|_{x'_t,\theta'_t} \right\|$$

$$= \left\| \left. \frac{\partial f_t}{\partial u_t} \right|_{x_t,u_t} \cdot \left. \frac{\partial \pi_t}{\partial \theta_t} \right|_{x_t,\theta_t} - \left. \frac{\partial f_t}{\partial u_t} \right|_{x'_t,u'_t} \cdot \left. \frac{\partial \pi_t}{\partial \theta_t} \right|_{x'_t,\theta'_t} \right\| \tag{22a}$$

$$\leq \left\| \left. \frac{\partial f_t}{\partial u_t} \right|_{x_t,u_t} \cdot \left(\left. \frac{\partial \pi_t}{\partial \theta_t} \right|_{x_t,\theta_t} - \left. \frac{\partial \pi_t}{\partial \theta_t} \right|_{x'_t,\theta'_t} \right) \right\| + \left\| \left(\left. \frac{\partial f_t}{\partial u_t} \right|_{x_t,u_t} - \left. \frac{\partial f_t}{\partial u_t} \right|_{x'_t,u'_t} \right) \cdot \left. \frac{\partial \pi_t}{\partial \theta_t} \right|_{x'_t,\theta'_t} \right\| \tag{22b}$$

$$\leq L_f (\ell_{\pi,\theta} \| \theta_t - \theta'_t \| + \ell_{\pi,x} \| x_t - x'_t \|) + (\ell_{f,x} \| x_t - x'_t \| + \ell_{f,u} \| u_t - u'_t \|) \cdot L_{\pi,\theta} \tag{22c}$$

$$\leq (L_f \ell_{\pi,\theta} + (\ell_{f,x} + \ell_{f,u} L_{\pi,x}) L_{\pi,\theta}) \| x_t - x'_t \| + (L_f \ell_{\pi,\theta} + \ell_{f,u} L_{\pi,\theta}^2) \| \theta_t - \theta'_t \|, \tag{22d}$$

where we use the chain rule decomposition in (22a); we use the triangle inequality in (22b); we use Assumption 2.1 in both (22c) and (22d).

When $\tau < t$, by (20), we have that

$$\left\| \left. \frac{\partial f_{t|\tau}}{\partial \theta_\tau} \right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \left. \frac{\partial f_{t|\tau}}{\partial \theta_\tau} \right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t}} \right\|$$

$$\leq \left\| \left. \frac{\partial f_t}{\partial x_t} \right|_{x_t,u_t} \cdot \left(\left. \frac{\partial g_{t|\tau}}{\partial \theta_\tau} \right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \left. \frac{\partial g_{t|\tau}}{\partial \theta_\tau} \right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\|$$

$$+ \left\| \left(\left. \frac{\partial f_t}{\partial x_t} \right|_{x_t,u_t} - \left. \frac{\partial f_t}{\partial x_t} \right|_{x'_t,u'_t} \right) \cdot \left. \frac{\partial g_{t|\tau}}{\partial \theta_\tau} \right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\|$$

$$+ \left\| \left. \frac{\partial f_t}{\partial u_t} \right|_{x_t,u_t} \cdot \left. \frac{\partial \pi_t}{\partial x_t} \right|_{x_t,\theta_t} \cdot \left(\left. \frac{\partial g_{t|\tau}}{\partial \theta_\tau} \right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \left. \frac{\partial g_{t|\tau}}{\partial \theta_\tau} \right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\|$$

$$+ \left\| \left. \frac{\partial f_t}{\partial u_t} \right|_{x_t,u_t} \cdot \left(\left. \frac{\partial \pi_t}{\partial x_t} \right|_{x_t,\theta_t} - \left. \frac{\partial \pi_t}{\partial x_t} \right|_{x'_t,\theta_t} \right) \cdot \left. \frac{\partial g_{t|\tau}}{\partial \theta_\tau} \right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\|$$

$$+ \left\| \left(\left. \frac{\partial f_t}{\partial u_t} \right|_{x_t,u_t} - \left. \frac{\partial f_t}{\partial u_t} \right|_{x'_t,u'_t} \right) \cdot \left. \frac{\partial \pi_t}{\partial x_t} \right|_{x'_t,\theta_t} \cdot \left. \frac{\partial g_{t|\tau}}{\partial \theta_\tau} \right|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \tag{23a}$$

$$\leq L_f \rho^{t-\tau} \left(C_{\ell,g,(\theta,x)} \| x_\tau - x'_\tau \| + C_{\ell,g,(\theta,\theta)} \| \theta_\tau - \theta'_\tau \| \right)$$
$$+ (\ell_{f,x} \| x_t - x'_t \| + \ell_{f,u} \| u_t - u'_t \|) \cdot C_{L,g,\theta} \rho^{t-\tau}$$
$$+ L_f L_{\pi,x} \rho^{t-\tau} \left(C_{\ell,g,(\theta,x)} \| x_\tau - x'_\tau \| + C_{\ell,g,(\theta,\theta)} \| \theta_\tau - \theta'_\tau \| \right)$$
$$+ L_f \ell_{\pi,x} \| x_t - x'_t \| \cdot C_{L,g,\theta} \rho^{t-\tau} + (\ell_{f,x} \| x_t - x'_t \| + \ell_{f,u} \| u_t - u'_t \|) \cdot L_{\pi,x} C_{L,g,\theta} \rho^{t-\tau}, \tag{23b}$$

27

where we use (20) and the triangle inequality in (23a); we use Assumption 2.1 in (23b). Note that by the first two inequalities in Lemma D.3, we have

$$\|x_t - x_t'\| \leq \rho^{t-\tau}(C_{L,g,x}\|x_\tau - x_\tau'\| + C_{L,g,\theta}\|\theta_\tau - \theta_\tau'\|),$$
$$\|u_t - u_t'\| \leq L_{\pi,x}\rho^{t-\tau}(C_{L,g,x}\|x_\tau - x_\tau'\| + C_{L,g,\theta}\|\theta_\tau - \theta_\tau'\|).$$

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

D.5 Proof of Theorem D.5

To simplify the notation, we use the shorthand $\hat{x}_\tau(\theta) := g_{\tau|0}(x_0, \theta_{\times\tau})$ and $\hat{u}_\tau(\theta) = \pi_\tau(\hat{x}_\tau(\theta), \theta)$ for any time τ and policy parameter θ.

We first derive an upper bound of G_t in order to bound the difference between θ_t and θ_{t+1}. Recall that

$$G_t := \sum_{\tau=0}^{\min\{t, B-1\}} \left.\frac{\partial f_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0, \theta_{0:t}} = \sum_{\tau=0}^{\min\{t, B-1\}} \left.\frac{\partial f_{t|t-\tau}}{\partial \theta_{t-\tau}}\right|_{x_{t-\tau}, \theta_{t-\tau:t}}. \tag{24}$$

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

$$\|G_t\| \leq \frac{C_{L,f,\theta}}{1-\rho}, x_t \in B_n(0, R_S + C\|x_0\|), u_t \in \mathcal{U}, \text{ and } \|\theta_{t+1} - \theta_t\| \leq \varepsilon. \tag{25}$$

Note that $\|G_0\| \leq C_{L,f,\theta} \leq \frac{C_{L,f,\theta}}{1-\rho}$ by Corollary D.4. We also have $x_0 \in B_n(0, R_S + C\|x_0\|)$ and $u_0 \in \mathcal{U}$.

Suppose $\|G_{t-1}\| \leq \frac{C_{L,f,\theta}}{1-\rho}$ for some $t \geq 1$. Then, since $\eta \leq \frac{(1-\rho)\varepsilon}{C_{L,f,\theta}}$ and the projection onto Θ is a contraction (see Theorem 1.2.1 in [60]), we see that

$$\|\theta_t - \theta_{t-1}\| \leq \|\eta G_{t-1}\| \leq \varepsilon.$$

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

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

Taking norm on both sides of (24), we see that

$$\|G_t\| = \left\|\sum_{\tau=0}^{\min\{t, B-1\}} \left.\frac{\partial f_{t|t-\tau}}{\partial \theta_{t-\tau}}\right|_{x_{t-\tau}, \theta_{t-\tau:t}}\right\|$$

$$\leq \sum_{\tau=0}^{\min\{t, B-1\}} \left\|\left.\frac{\partial f_{t|t-\tau}}{\partial \theta_{t-\tau}}\right|_{x_{t-\tau}, \theta_{t-\tau:t}}\right\| \tag{26a}$$

$$\leq \sum_{\tau=0}^{\min\{t, B-1\}} C_{L,f,\theta}\rho^\tau \tag{26b}$$

$$\leq \frac{C_{L,f,\theta}}{1-\rho},$$

where we use the triangle inequality in (26a) and Corollary D.4 in (26b). Note that we can apply Corollary D.4 because $x_t \in B_n(0, R_S + C\|x_0\|)$. Therefore, we have shown (25) by induction. One can use the same technique as (26) to show $\|\nabla F_t(\theta_t)\| \leq \frac{C_{L,f,\theta}}{1-\rho}$.

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

$$\|\theta_t - \theta_\tau\| \leq \frac{C_{L,f,\theta}\eta(t-\tau)}{1-\rho}. \tag{27}$$

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

$$\|x_\tau - \hat{x}_\tau(\theta_t)\| = \left\|g_{\tau|0}(x_0, \theta_{0:\tau-1}) - g_{\tau|0}(x_0, (\theta_t)_{\times\tau})\right\|$$

28

$$\leq \sum_{\tau'=0}^{\tau-1} \left\| g_{\tau|0}(x_0, \theta_{0:\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g_{\tau|0}(x_0, \theta_{0:\tau'-1}, (\theta_t)_{\times(\tau-\tau')}) \right\| \quad (28a)$$

$$\leq \sum_{\tau'=0}^{\tau-1} \left\| g_{\tau|\tau'}(x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g_{\tau|\tau'}(x_{\tau'}, (\theta_t)_{\times(\tau-\tau')}) \right\| \quad (28b)$$

$$\leq \sum_{\tau'=0}^{\tau-1} C_{L,g,\theta}\rho^{\tau-\tau'} \|\theta_t - \theta_{\tau'}\| \quad (28c)$$

$$\leq \frac{C_{L,f,\theta}C_{L,g,\theta}\eta}{1-\rho} \sum_{\tau'=0}^{\tau-1} (t-\tau')\rho^{\tau-\tau'} \quad (28d)$$

$$\leq \frac{C_{L,f,\theta}C_{L,g,\theta}\rho}{(1-\rho)^2}\left((t-\tau) + \frac{1}{1-\rho}\right)\cdot\eta,$$

where we use the triangle inequality in (28a); we use the definition of multi-step dynamics in (28b); we use Lemma D.3 in (28c); we use (27) in (28d).

Similarly, since (24) guarantees that $x_t \in B_n(0, R_S + C\|x_0\|)$ and we also see that $\hat{x}_t(\theta_t) \in B_n(0, R_S + C\|x_0\|)$, we obtain that

$$|f_t(x_t, u_t) - F_t(\theta_t)| = |f_t(x_t, u_t) - f_t(\hat{x}_t(\theta_t), \hat{u}_t(\theta_t))|$$

$$\leq L_f(\|x_t - \hat{x}_t(\theta_t)\| + \|u_t - \hat{u}_t(\theta_t)\|) \quad (29a)$$

$$= L_f(\|x_t - \hat{x}_t(\theta_t)\| + \|\pi_t(x_t, \theta_t) - \pi_t(\hat{x}_t(\theta_t), \theta_t)\|)$$

$$\leq L_f(1 + L_{\pi,x})\|x_t - \hat{x}_t(\theta_t)\| \quad (29b)$$

$$\leq \frac{C_{L,f,\theta}C_{L,g,\theta}L_f(1 + L_{\pi,x})\rho}{(1-\rho)^3}\cdot\eta, \quad (29c)$$

where we use Assumption 2.1 in (29a) and (29b); we use (28) in (29c).

To show the last inequality in Theorem D.5, note that we have

$$\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| \leq \sum_{\tau=0}^{t-1}\left\| g_{t|0}\Big(x_0, \tilde{\theta}_{0:\tau-1}, (\tilde{\theta}_t)_{\times(t-\tau)}\Big) - g_{t|0}\Big(x_0, \tilde{\theta}_{0:\tau}, (\tilde{\theta}_t)_{\times(t-\tau-1)}\Big)\right\| \quad (30a)$$

$$\leq C L_{\pi,\theta}L_{g,x} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1}\left\|\tilde{\theta}_t - \tilde{\theta}_\tau\right\| \quad (30b)$$

$$\leq C L_{\pi,\theta}L_{g,x} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \sum_{\tau'=\tau}^{t-1}\left\|\tilde{\theta}_{\tau'+1} - \tilde{\theta}_{\tau'}\right\| \quad (30c)$$

$$\leq \frac{C L_{\pi,\theta}L_{g,x}}{1-\rho} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1}\left\|\tilde{\theta}_{\tau+1} - \tilde{\theta}_\tau\right\|, \quad (30d)$$

where we use the triangle inequality in (30a) and (30c); we use the assumption that $\|\tilde{x}_t\| \leq \min\{R_C, R_x\}$ and the time-invariant contractive perturbation property in (30b); we rearrange the terms and use $\sum_{\tau=0}^{\infty} \rho^\tau \leq \frac{1}{1-\rho}$ in (30d).

Therefore, since $\tilde{x}_t, \hat{x}_t(\tilde{\theta}_t) \in \mathcal{X}$ and $\tilde{u}_t, \hat{u}_t(\tilde{\theta}_t) \in \mathcal{U}$, we see that

$$\left| f_t(\tilde{x}_t, \tilde{u}_t) - F_t(\tilde{\theta}_t)\right| = \left| f_t(\tilde{x}_t, \tilde{u}_t) - f_t(\hat{x}_t(\tilde{\theta}_t), \hat{u}_t(\tilde{\theta}_t))\right|$$

$$\leq L_f\left(\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| + \left\|\tilde{u}_t - \hat{u}_t(\tilde{\theta}_t)\right\|\right) \quad (31a)$$

$$= L_f\left(\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| + \left\|\pi_t(\tilde{x}_t, \tilde{\theta}_t) - \pi_t(\hat{x}_t(\tilde{\theta}_t), \tilde{\theta}_t)\right\|\right)$$

$$\leq L_f(1 + L_{\pi,x})\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| \quad (31b)$$

$$\leq \frac{C L_{\pi,\theta}L_{g,x}L_f(1 + L_{\pi,x})}{1-\rho} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1}\left\|\tilde{\theta}_{\tau+1} - \tilde{\theta}_\tau\right\|, \quad (31c)$$

29

where we use Assumption 2.1 in (31a) and (31b); we use (30) in (31c).

D.6 Proof of Theorem D.6

Theorem D.9 (Gradient Bias). *Suppose Assumptions 2.1 and 2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t \in \mathcal{T}}$ denote the trajectory of GAPS (Algorithm 1) with buffer size B and learning rate $\eta \leq \frac{(1-\rho)\varepsilon}{C_{L,f,\theta}}$. Then, the following holds for all $\tau \leq t$:*

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

for

$$\hat{C}_0 = \frac{\rho C_{L,f,\theta} C_{L,g,\theta} C_{\ell,f,(\theta,x)}}{(1-\rho)^3}, \hat{C}_1 = \frac{(1-\rho) C_{L,f,\theta} C_{\ell,f,(\theta,x)} + \rho C_{L,f,\theta} C_{L,g,\theta} C_{\ell,f,(\theta,\theta)}}{(1-\rho)^2},$$

$$\hat{C}_2 = \frac{C_{\ell,f,(x,\theta)} C_{L,g,\theta} C_{L,f,\theta}}{1-\rho}.$$

Next,

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

Proof of Theorem D.9. To simplify the notation, we adopt the shorthand notations $\hat{x}_\tau(\theta) := g_{\tau|0}(x_0, \theta_{\times \tau})$ and $\hat{u}_\tau(\theta) := \pi_\tau(\hat{x}_\tau(\theta), \theta)$ throughout the proof.

As we discussed below Theorem D.6 in the proof outline, we use the triangle inequality to do the decomposition

$$
\begin{aligned}
&\left\| \frac{\partial f_{t|0}}{\partial \theta_\tau}\bigg|_{x_0, \theta_{0:t}} - \frac{\partial f_{t|0}}{\partial \theta_\tau}\bigg|_{x_0, (\theta_t) \times (t+1)} \right\| \\
&= \left\| \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_{\tau:t}} - \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{\hat{x}_\tau(\theta_t), (\theta_t) \times (t-\tau+1)} \right\| \\
&\leq \left\| \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_\tau, (\theta_t) \times (t-\tau)} - \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{\hat{x}_\tau(\theta_t), (\theta_t) \times (t-\tau+1)} \right\| \\
&\quad + \sum_{\tau'=\tau+1}^{t-1} \left\| \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_{\tau:\tau'}, (\theta_t) \times (t-\tau')} - \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_{\tau:\tau'-1}, (\theta_t) \times (t-\tau'+1)} \right\|.
\end{aligned}
\tag{32}
$$

Note that we can apply Corollary D.4 to bound each term in (32). For the first term in (32), since $x_\tau, \hat{x}_\tau(\theta_t), x_{\tau+1} \in B_n(0, R_S + C\|x_0\|)$, we see that

$$
\begin{aligned}
&\left\| \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_\tau, (\theta_t) \times (t-\tau)} - \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{\hat{x}_\tau(\theta_t), (\theta_t) \times (t-\tau+1)} \right\| \\
&\leq \rho^{t-\tau} \left(C_{\ell,f,(\theta,x)} \|x_\tau - \hat{x}_\tau(\theta_t)\| + C_{\ell,f,(\theta,\theta)} \|\theta_t - \theta_\tau\| \right) \\
&\leq \frac{(1-\rho) C_{L,f,\theta} C_{\ell,f,(\theta,x)} + \rho C_{L,f,\theta} C_{L,g,\theta} C_{\ell,f,(\theta,\theta)}}{(1-\rho)^2} \cdot (t-\tau) \rho^{t-\tau} \cdot \eta \\
&\quad + \frac{\rho C_{L,f,\theta} C_{L,g,\theta} C_{\ell,f,(\theta,x)}}{(1-\rho)^3} \cdot \rho^{t-\tau} \cdot \eta,
\end{aligned}
$$

(33a)

(33b)

where we use Corollary D.4 in (33a) and Theorem D.5 in (33b).

For any $\tau' \in [\tau + 1 : t - 1]$, since $x_{\tau'}, x_{\tau'+1} \in B_n(0, R_S + C\|x_0\|)$, we see that

$$\left\| \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_{\tau:\tau'}, (\theta_t) \times (t-\tau')} - \frac{\partial f_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_{\tau:\tau'-1}, (\theta_t) \times (t-\tau'+1)} \right\|$$

30

$$= \left\| \left(\frac{\partial f_{t|\tau'}}{\partial x_{\tau'}} \bigg|_{x_{\tau'}, \theta_{\tau'}, (\theta_t) \times (t-\tau')} - \frac{\partial f_{t|\tau'}}{\partial x_{\tau'}} \bigg|_{x_{\tau'}, (\theta_t) \times (t-\tau'+1)} \right) \frac{\partial g_{\tau'|\tau}}{\partial \theta_\tau} \bigg|_{x_\tau, \theta_{\tau:\tau'-1}} \right\|$$

$$\leq \left\| \frac{\partial f_{t|\tau'}}{\partial x_{\tau'}} \bigg|_{x_{\tau'}, \theta_{\tau'}, (\theta_t) \times (t-\tau')} - \frac{\partial f_{t|\tau'}}{\partial x_{\tau'}} \bigg|_{x_{\tau'}, (\theta_t) \times (t-\tau'+1)} \right\| \cdot \left\| \frac{\partial g_{\tau'|\tau}}{\partial \theta_\tau} \bigg|_{x_\tau, \theta_{\tau:\tau'-1}} \right\|$$

$$\leq C_{\ell, f, (x,\theta)} \rho^{t-\tau'} \|\theta_t - \theta_{\tau'}\| \cdot C_{L, g, \theta} \rho^{\tau'-\tau} \tag{34a}$$

$$\leq \frac{C_{\ell, f, (x,\theta)} C_{L, g, \theta} C_{L, f, \theta}}{1 - \rho} \cdot (t - \tau) \rho^{t-\tau} \cdot \eta, \tag{34b}$$

where we use Lemma D.3 and Corollary D.4 in (34a); we use Theorem D.5 in (34b). Substituting (33) and (34) into (32) finishes the proof of the first inequality.

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

$$G_t := \sum_{\tau=0}^{\min\{t, B-1\}} \frac{\partial f_{t|0}}{\partial \theta_{t-\tau}} \bigg|_{x_0, \theta_{0:t}}, \nabla \ell_t(\theta_t) = \sum_{\tau=0}^{t} \frac{\partial f_{t|0}}{\partial \theta_{t-\tau}} \bigg|_{x_0, (\theta_t) \times (t+1)}.$$

Therefore, we see that

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

$$\leq \sum_{\tau=0}^{\min\{t, B-1\}} \left\| \frac{\partial f_{t|0}}{\partial \theta_{t-\tau}} \bigg|_{x_0, \theta_{0:t}} - \frac{\partial f_{t|0}}{\partial \theta_{t-\tau}} \bigg|_{x_0, (\theta_t) \times (t+1)} \right\|$$

$$+ \mathbf{1}(t \geq B) \sum_{\tau=B}^{t} \left\| \frac{\partial f_{t|t-\tau}}{\partial \theta_{t-\tau}} \bigg|_{\hat{x}_{t-\tau}(\theta_t), (\theta_t) \times (\tau+1)} \right\| \tag{35a}$$

$$\leq \sum_{\tau=0}^{B-1} \left(\hat{C}_0 + \hat{C}_1 \tau + \hat{C}_2 \tau^2 \right) \rho^\tau \eta + \sum_{\tau=B}^{t} C_{L, f, \theta} \rho^\tau \tag{35b}$$

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

where we use the triangle inequality in (35a); we use the first inequality in Theorem D.6 that we have shown and Corollary D.4 in (35b). $\qquad\square$

E Dynamic regret of GAPS

When the surrogate cost F_t is convex, dynamic policy regret is another benchmark designed for time-varying environments [32, 61]. To define the dynamic regret, we first need to define the optimal total cost under the parameter path length constraint (denoted by $\hat{J}_{\mathcal{T}}(\Upsilon)$):

$$\min_{\tilde{\theta}_{0:T-1} \in \Theta^T} J_{\mathcal{T}}(\tilde{\theta}_{0:T-1}) \text{ s.t. } \sum_{\tau=1}^{T-1} \left\| \tilde{\theta}_\tau - \tilde{\theta}_{\tau-1} \right\| \leq \Upsilon - 1, \text{ and } \|\tilde{x}_t\| \leq \min\{R_C, R_x\}, \forall t \in \mathcal{T},$$

where $\tilde{x}_t := g_{t|0}(x_0, \tilde{\theta}_{0:t-1})$ and $\Upsilon - 1 \geq 0$ is the upper bound on the total path length of the parameter sequence. The dynamic regret is defined as

$$R_\Upsilon^D(T) := J_{\mathcal{T}}(\theta_{0:T-1}) - \hat{J}_{\mathcal{T}}(\Upsilon). \tag{36}$$

Compared with the (static) policy regret $R^S(T)$ and the adaptive policy regret $R^A(T)$, the dynamic policy regret $R^D(T)$ is more flexible by allowing the offline benchmark to change policy parameters over time.

To bound the dynamic regret of GAPS, we follow a similar outline as the proof of the adaptive regret (Appendix D): We first show a dynamic regret of the Ideal OGD algorithm that is robust to gradient biases.

31

Theorem E.1. *Under the same assumptions as Theorem D.1, for any sequence $\tilde{\theta}_{0:T-1} \in \Theta^T$, we have*

$$\sum_{t=0}^{T-1} F_t(\theta_t) - \sum_{t=0}^{T-1} F_t(\tilde{\theta}_t) \leq \frac{3D^2}{\eta} + \frac{2D}{\eta} \sum_{t=1}^{T-1} \left\| \tilde{\theta}_t - \tilde{\theta}_{t-1} \right\| + \eta(W + \alpha)^2 T + 2\alpha DT.$$

The proof of Theorem E.1 is deferred to Appendix E.1.

Recall that we have estabilished the similarity between GAPS and Ideal OGD in Theorem 3.2. Using these bounds on costs/gradients differences, we can "transfer" the dynamic regret guarantee on OGD (for online optimization) to GAPS (for online policy selection). We formally state this result Theorem E.2 and defer the proof to Appendix E.2.

Theorem E.2. *Under the same assumptions as Theorem 3.3, GAPS achieves the dynamic regret bound*

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

$$\tag{37}$$

(See Appendix E.2 for the detailed expressions.)

By choosing the appropriate learning rate η, one can achieve sublinear dynamic regret when the parameter path length constraint Υ of the benchmark is $o(T)$. We discuss how to set the learning rate and the dynamic regret bound it achieves in Corollary E.3

Corollary E.3. *Under the same assumptions as Theorem 3.3, suppose $T \gg \frac{1}{1-\rho}$ and $B \geq \frac{1}{2}\log(T)/\log(1/\rho)$: If we set $\eta = (1-\rho)^{\frac{5}{2}}(\Upsilon/T)^{\frac{1}{2}}$, we get the dynamic regret of $R_\Upsilon^D(T) = O\left((1-\rho)^{-\frac{5}{2}}(\Upsilon T)^{\frac{1}{2}}\right)$.*

It is worth noticing that the approach to derive the dynamic regret for GAPS can generalize to other performance metrics: If we can derive some performance bounds for OGD in online optimization, we can use Theorem 3.2 to show a corresponding guarantee for GAPS in online policy selection. An interesting future direction is using this approach to derive results for GAPS under nonconvex surrogate costs F_t.

E.1 Proof of Theorem E.1

This proof is inspired by the proof of Theorem 10.1 in [32].

Without loss of generality, we can assume $\theta_0 = 0$, because otherwise we change the variable by subtracting all other parameters by θ_0. We see that

$$\begin{aligned}
\left\| \theta_{t+1} - \tilde{\theta}_t \right\|^2 &\leq \left\| \theta'_{t+1} - \tilde{\theta}_t \right\|^2 \\
&= \left\| \theta_t - \eta G_t - \tilde{\theta}_t \right\|^2 \\
&= \left\| \theta_t - \tilde{\theta}_t \right\|^2 + \eta^2 \| G_t \|^2 - 2\eta \langle G_t, \theta_t - \tilde{\theta}_t \rangle.
\end{aligned}$$

Thus, we see that

$$\begin{aligned}
&2\langle \nabla F_t(\theta_t), \theta_t - \tilde{\theta}_t \rangle \\
&\leq 2\langle \nabla F_t(\theta_t) - G_t, \theta_t - \tilde{\theta}_t \rangle + \frac{1}{\eta}\left(\left\| \theta_t - \tilde{\theta}_t \right\|^2 - \left\| \theta_{t+1} - \tilde{\theta}_t \right\|^2 \right) + \eta(W+\alpha)^2 \\
&\leq \frac{1}{\eta}\left(\left\| \theta_t - \tilde{\theta}_t \right\|^2 - \left\| \theta_{t+1} - \tilde{\theta}_t \right\|^2 \right) + \eta(W+\alpha)^2 + 2\alpha D.
\end{aligned}$$

Therefore, we obtain that

$$2\left(\sum_{t=0}^{T-1} F_t(\theta_t) - F_t(\tilde{\theta}_t) \right)$$

32

$$\leq 2 \sum_{t=0}^{T-1} \langle \nabla F_t(\theta_t), \theta_t - \tilde{\theta}_t \rangle$$

$$\leq \frac{1}{\eta} \sum_{t=0}^{T-1} \left(\left\| \theta_t - \tilde{\theta}_t \right\|^2 - \left\| \theta_{t+1} - \tilde{\theta}_t \right\|^2 \right) + \eta(W + \alpha)^2 T + 2\alpha DT$$

$$\leq \frac{1}{\eta} \sum_{t=0}^{T-1} \left(\|\theta_t\|^2 - \|\theta_{t+1}\|^2 + 2\langle \tilde{\theta}_t, \theta_{t+1} - \theta_t \rangle \right) + \eta(W + \alpha)^2 T + 2\alpha DT$$

$$\leq \frac{1}{\eta} \left(D^2 + 2\sum_{t=1}^{T-1} \langle \theta_t, \tilde{\theta}_{t-1} - \tilde{\theta}_t \rangle + 2D^2 \right) + \eta(W + \alpha)^2 T + 2\alpha DT$$

$$\leq \frac{3D^2}{\eta} + \frac{2D}{\eta} \sum_{t=1}^{T-1} \left\| \tilde{\theta}_t - \tilde{\theta}_{t-1} \right\| + \eta(W + \alpha)^2 T + 2\alpha DT.$$

E.2 Complete Version of Theorem E.2 and Its Proof

Theorem E.4. *Under the same assumptions as Theorem 3.3, GAPS achieves the dynamic regret bound*

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

$$+ 4\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3} \right)^2 D^2 \eta^3 T + \frac{4C_{L,f,\theta}^2}{(1-\rho)^2} \rho^{2B} \eta T$$

$$+ \left(\frac{2D}{\eta} + \frac{C_{L,f,\theta}}{(1-\rho)^2} \right) \sum_{t=1}^{T-1} \left\| \tilde{\theta}_t - \tilde{\theta}_{t-1} \right\|.$$

Proof of Theorem E.4. By Theorem D.1 and Theorem D.6, we see that the sequence of policy parameters of the online policy satisfies that

$$\sum_{t=0}^{T-1} F_t(\theta_t) - \sum_{t=0}^{T-1} F_t(\tilde{\theta}_t) \leq \frac{3D^2}{\eta} + \frac{2D}{\eta} \sum_{t=1}^{T-1} \left\| \tilde{\theta}_t - \tilde{\theta}_{t-1} \right\| + \eta(W + \alpha)^2 T + 2\alpha DT, \qquad (38)$$

where $W = \frac{C_{L,f,\theta}}{1-\rho}$ by Theorem D.5 and $\alpha = \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3} \right) \eta + \frac{C_{L,f,\theta}}{1-\rho} \cdot \rho^B$.

Recall that by Theorem D.5, we have that

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

$$\left| f_t(\tilde{x}_t, \tilde{u}_t) - F_t(\tilde{\theta}_t) \right| \leq \frac{C_{L,f,\theta}}{1-\rho} \sum_{\tau=0}^{t-1} \rho^\tau \left\| \tilde{\theta}_{t-\tau} - \tilde{\theta}_{t-\tau-1} \right\|.$$

Substituting these into (38) gives that

$$f_t(x_t, u_t) - f_t(\tilde{x}_t, \tilde{u}_t)$$

$$\leq \left(\sum_{t=0}^{T-1} F_t(\theta_t) - \sum_{t=0}^{T-1} F_t(\tilde{\theta}_t) \right) + \sum_{t=0}^{T-1} |f_t(x_t, u_t) - F_t(\theta_t)| + \sum_{t=0}^{T-1} \left| f_t(\tilde{x}_t, \tilde{u}_t) - F_t(\tilde{\theta}_t) \right|$$

$$\leq \frac{3D^2}{\eta} + \frac{2D}{\eta} \sum_{t=1}^{T-1} \left\| \tilde{\theta}_t - \tilde{\theta}_{t-1} \right\| + \eta(W + \alpha)^2 T + 2\alpha DT + \frac{C_{L,f,\theta}^2 \rho}{(1-\rho)^3} \cdot \eta T$$

$$+ \frac{C_{L,f,\theta}}{1-\rho} \sum_{t=0}^{T-1} \sum_{\tau=0}^{t-1} \rho^\tau \left\| \tilde{\theta}_{t-\tau} - \tilde{\theta}_{t-\tau-1} \right\|$$

33

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

$$+ 4\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)^2 D^2\eta^3 T + \frac{4C_{L,f,\theta}^2}{(1-\rho)^2}\rho^{2B}\eta T$$

$$+ \left(\frac{2D}{\eta} + \frac{C_{L,f,\theta}}{(1-\rho)^2}\right)\sum_{t=1}^{T-1}\left\|\tilde{\theta}_t - \tilde{\theta}_{t-1}\right\|.$$

$\square$

F Regret of GAPS under nonconvex surrogate costs

To derive the local regret bound for GAPS in online policy selection, we first bound the local regret for OGD (with biased gradients) in online nonconvex optimization.

Theorem F.1. *Suppose $\Theta = \mathbb{R}^d$. Consider the biased OGD update rule $\theta_{t+1} = \theta_t - \eta G_t$, where G_t satisfies $\|G_t - \nabla F_t(\theta_t)\| \leq \beta$. Suppose at every time t, F_t is L_F-Lipschitz and ℓ_F-smooth. If the learning rate $\eta < \frac{1}{\ell_F}$, we have that*

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

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

The proof of Theorem F.1 is deferred to Appendix F.1. Our proof is inspired by the analysis for (stochastic) gradient descent in offline nonconvex optimization (see [e.g. 62]) with the additional step to handle the time-varying function sequence $F_{0:T-1}$ via the measure of variation $\mathsf{dist}_s(F_t, F_{t-1})$.

Our approximation error bound (Theorem 3.2) guarantees that the bias $\beta = O(1/\sqrt{T})$ if we set the learning rate $\eta = O(1/\sqrt{T})$. Therefore, the remaining task is to bound the measure of variation in online nonconvex optimization $\sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1})$ by the variation intensity in online policy selection V (Definition 3.5). To derive this bound, we need to show a convergence result on applying a fixed policy parameter in a time-invariant system. We begin with a definition that characterize this (imaginary) dynamical process.

Definition F.2. *For fixed dynamics function g, policy function π, and policy parameter θ, we define $\tilde{x}_\tau^{(g,\pi)}(\theta)$ recursively by the equation*

$$\tilde{x}_{\tau+1}^{(g,\pi)}(\theta) = g\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \pi\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \theta\right)\right), \forall\tau \geq 0, \text{ where } \tilde{x}_0^{(g,\pi)}(\theta) = x_0.$$

Compared with $\hat{x}_t(\theta)$ we defined before Definition 2.4, the state $\tilde{x}_{\tau+1}^{(g,\pi)}(\theta)$ is produced by a time-invariant dynamical system induced by g and π, while $\hat{x}_t(\theta)$ is produced by the actual time-varying dynamics induced by $g_{0:t-1}$ and $\pi_{0:t-1}$.

We show the time-invariant evolution $\tilde{x}_\tau^{(g,\pi)}$ has a unique limitation point as τ tends to infinity. This limit is also a fixed point, and the states will converge to the limit exponentially fast with respect to τ. We state this result formally in Lemma F.3.

Lemma F.3. *Suppose Assumptions 2.1 and 2.2 hold, and $(g,\pi) \in \mathcal{G}$. The limit $\lim_{\tau\to\infty}\tilde{x}_\tau^{(g,\pi)}(\theta)$ exists. Let $\tilde{x}_\infty^{(g,\pi)}(\theta) := \lim_{\tau\to\infty}\tilde{x}_\tau^{(g,\pi)}(\theta)$. Further, we also have that*

$$\left\|\tilde{x}_\tau^{(g,\pi)}(\theta) - \tilde{x}_\infty^{(g,\pi)}(\theta)\right\| \leq C\rho^\tau\left\|x_0 - \tilde{x}_\infty^{(g,\pi)}(\theta)\right\| \leq C\rho^\tau\cdot\mathsf{diam}(\mathcal{X}),$$

where $\mathsf{diam}(\mathcal{X}) = 2C(R_S + C\|x_0\|) + 2R_S$.

The proof of Lemma F.3 is deferred to Appendix F.2. With the fixed point and convergence result in Lemma F.3, we bound the measure of variation based on $F_{0:T-1}$ by the variation intensity V based on $g_{0:T-1}, \pi_{0:T-1}$, and $f_{0:T-1}$ in Lemma F.4.

34

Lemma F.4. *Suppose Assumptions 2.1 and 2.2 hold. Then, we have*

$$\sum_{t=1}^{T-1} \text{dist}_s(F_t, F_{t-1}) \leq \frac{2CL_f(1+L_{\pi,x})(1+L_{g,u})}{(1-\rho)^2\rho} \cdot V + \frac{2CL_f(1+L_{\pi,x})}{1-\rho} \cdot \text{diam}(\mathcal{X}),$$

where $\text{diam}(\mathcal{X}) = 2C(R_S + C\|x_0\|) + 2R_S$.

With these auxiliary results, we restate Theorem 3.6 with complete expressions and present the proof.

Theorem F.5. *Under the same assumptions as Theorem 3.2, if we additionally assume that $\Theta = \mathbb{R}^d$ for some integer d, then GAPS satisfies local regret*

$$R^L(T) \leq \frac{2}{\eta}\left(c_0 + \frac{2CL_FL_f(1+L_{\pi,x})}{1-\rho}\left(\frac{(1+L_{g,u})V}{(1-\rho)\rho} + 2C(R_S + C\|x_0\|) + 2R_S\right)\right)$$

$$+ 2(1-\ell_F\eta)L_F\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta T$$

$$+ 2(1-\ell_F\eta)L_F\frac{C_{L,f,\theta}}{1-\rho} \cdot \rho^B T$$

$$+ 2\ell_F\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)^2 \eta^3 T + 2\ell_F\frac{C_{L,f,\theta}^2}{(1-\rho)^2} \cdot \rho^{2B}\eta T, \tag{39}$$

where $\hat{C}_0, \hat{C}_1, \hat{C}_2$ *are defined in Theorem D.9,* $c_0 = f_0(x_0, \pi_0(x_0, \theta_0))$, *and*

$$L_F = \frac{C_{L,f,\theta}}{1-\rho}, \quad \ell_F = \frac{CL_{g,u}L_{\pi,\theta}C_{\ell,f,(\theta,x)} + \rho C_{\ell,f,(x,\theta)}C_{L,g,\theta}}{(1-\rho)^2}.$$

Proof of Theorem F.5. By Theorem F.1 and Theorem 3.2, we know the parameter sequence of GAPS satisfies that

$$\sum_{t=0}^{T-1} \|\nabla F_t(\theta_t)\|^2 \leq \frac{2}{\eta}\left(F_0(\theta_0) + L_F\sum_{t=1}^{T-1}\text{dist}_s(F_t, F_{t-1})\right) + \left(2(1-\ell_F\eta)L_F\beta + \ell_F\eta \cdot \beta^2\right)T, \tag{40}$$

where $\beta = \left(\hat{C}_0(1-\rho)^{-1} + (\hat{C}_1 + \hat{C}_2)(1-\rho)^{-2} + \hat{C}_2(1-\rho)^{-3}\right)\eta + C_{L,f,\theta}(1-\rho)^{-1} \cdot \rho^B$. Here, $\hat{C}_0, \hat{C}_1, \hat{C}_2$ are defined in Theorem D.9.

Note that $L_F = \frac{C_{L,f,\theta}}{1-\rho}$ by Theorem D.5. Now we show that we can set

$$\ell_F = \frac{CL_{g,u}L_{\pi,\theta}C_{\ell,f,(\theta,x)} + \rho C_{\ell,f,(x,\theta)}C_{L,g,\theta}}{(1-\rho)^2}.$$

To see this, by Lemma D.3, we obtain that the following inequality holds for every time step t,

$$\|\nabla F_t(\theta) - \nabla F_t(\theta')\| = \left\|\sum_{\tau=0}^{t}\frac{\partial f_{t|\tau}}{\partial\theta_\tau}\bigg|_{\hat{x}_\tau(\theta),\theta_{\times(t-\tau+1)}} - \sum_{\tau=0}^{t}\frac{\partial f_{t|\tau}}{\partial\theta_\tau}\bigg|_{\hat{x}_\tau(\theta'),\theta'_{\times(t-\tau+1)}}\right\| \tag{41a}$$

$$\leq \sum_{\tau=0}^{t}\left\|\frac{\partial f_{t|\tau}}{\partial\theta_\tau}\bigg|_{\hat{x}_\tau(\theta),\theta_{\times(t-\tau+1)}} - \frac{\partial f_{t|\tau}}{\partial\theta_\tau}\bigg|_{\hat{x}_\tau(\theta'),\theta'_{\times(t-\tau+1)}}\right\|, \tag{41b}$$

where we use the definition of surrogate cost functions in (41a); we use the triangle inequality in (41b).

Note that for each term in (41), we can decompose it as

$$\left\|\frac{\partial f_{t|\tau}}{\partial\theta_\tau}\bigg|_{\hat{x}_\tau(\theta),\theta_{\times(t-\tau+1)}} - \frac{\partial f_{t|\tau}}{\partial\theta_\tau}\bigg|_{\hat{x}_\tau(\theta'),\theta'_{\times(t-\tau+1)}}\right\|$$

35

$$\leq \left\| \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta),\theta_{\times(t-\tau+1)}} - \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta_{\times(t-\tau+1)}} \right\|$$

$$+ \sum_{j=\tau}^{t} \left\| \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta'_{\times(j-\tau)},\theta_{\times(t-j+1)}} - \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta'_{\times(j-\tau+1)},\theta_{\times(t-j)}} \right\|. \tag{42}$$

Note that for any time step τ, by the triangle inequality, we have

$$\|\hat{x}_\tau(\theta) - \hat{x}_\tau(\theta')\| \leq \sum_{j=0}^{\tau-1} \left\| g_{\tau|0}(x_0, \theta_{\times j}, \theta'_{\times(\tau-j)}) - g_{\tau|0}(x_0, \theta_{\times(j+1)}, \theta'_{\times(\tau-j-1)}) \right\|$$

$$\leq \sum_{j=0}^{\tau-1} C\rho^{\tau-j-1} \cdot L_{g,u}L_{\pi,\theta}\|\theta - \theta'\| \leq \frac{CL_{g,u}L_{\pi,\theta}}{1-\rho} \cdot \|\theta - \theta'\|,$$

where we apply the time-invariant contractive perturbation in the last inequality. Therefore, by Corollary D.4, we obtain that

$$\left\| \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta),\theta_{\times(t-\tau+1)}} - \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta_{\times(t-\tau+1)}} \right\|$$

$$\leq C_{\ell,f,(\theta,x)}\rho^{t-\tau}\|\hat{x}_\tau(\theta) - \hat{x}_\tau(\theta')\| \leq \frac{CL_{g,u}L_{\pi,\theta}C_{\ell,f,(\theta,x)}}{1-\rho} \cdot \rho^{t-\tau} \cdot \|\theta - \theta'\|. \tag{43}$$

We also see that

$$\left\| \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta'_{\times(j-\tau)},\theta_{\times(t-j+1)}} - \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta'_{\times(j-\tau+1)},\theta_{\times(t-j)}} \right\|$$

$$= \left\| \left(\left.\frac{\partial f_{t|j}}{\partial x_j}\right|_{\hat{x}_j(\theta'),\theta_{\times(t-j+1)}} - \left.\frac{\partial f_{t|j}}{\partial \theta_j}\right|_{\hat{x}_j(\theta'),\theta',\theta_{\times(t-j)}} \right) \left.\frac{\partial g_{j|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta'_{\times(j-\tau)}} \right\|$$

$$\leq \left\| \left.\frac{\partial f_{t|j}}{\partial x_j}\right|_{\hat{x}_j(\theta'),\theta_{\times(t-j+1)}} - \left.\frac{\partial f_{t|j}}{\partial \theta_j}\right|_{\hat{x}_j(\theta'),\theta',\theta_{\times(t-j)}} \right\| \cdot \left\| \left.\frac{\partial g_{j|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta'_{\times(j-\tau)}} \right\|$$

$$\leq C_{\ell,f,(x,\theta)}\rho^{t-j}\|\theta - \theta'\| \cdot C_{L,g,\theta}\rho^{j-\tau} = C_{\ell,f,(x,\theta)}C_{L,g,\theta}\rho^{t-\tau}\|\theta - \theta'\|. \tag{44}$$

Substituting (43) and (44) into (42) gives that

$$\left\| \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta),\theta_{\times(t-\tau+1)}} - \left.\frac{\partial f_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'),\theta'_{\times(t-\tau+1)}} \right\|$$

$$\leq \left(\frac{CL_{g,u}L_{\pi,\theta}C_{\ell,f,(\theta,x)}}{1-\rho} + (t-\tau)C_{\ell,f,(x,\theta)}C_{L,g,\theta} \right) \cdot \rho^{t-\tau}\|\theta - \theta'\|.$$

Substituting this inequality into (41) gives that

$$\|\nabla F_t(\theta) - \nabla F_t(\theta')\| \leq \sum_{\tau=0}^{t} \left(\frac{CL_{g,u}L_{\pi,\theta}C_{\ell,f,(\theta,x)}}{1-\rho} + (t-\tau)C_{\ell,f,(x,\theta)}C_{L,g,\theta} \right) \cdot \rho^{t-\tau}\|\theta - \theta'\|$$

$$\leq \frac{CL_{g,u}L_{\pi,\theta}C_{\ell,f,(\theta,x)} + \rho C_{\ell,f,(x,\theta)}C_{L,g,\theta}}{(1-\rho)^2} \cdot \|\theta - \theta'\|.$$

Therefore, we can set $\ell_F = \frac{CL_{g,u}L_{\pi,\theta}C_{\ell,f,(\theta,x)}+\rho C_{\ell,f,(x,\theta)}C_{L,g,\theta}}{(1-\rho)^2}$.

Recall that the notation dist_s is defined in Theorem F.1. By Lemma F.4, we know that

$$\sum_{t=1}^{T-1} \text{dist}_s(F_t, F_{t-1})$$

$$\leq \frac{2CL_f(1+L_{\pi,x})(1+L_{g,u})}{(1-\rho)^2\rho} \cdot V + \frac{4CL_f(1+L_{\pi,x})(C(R_S + C\|x_0\|) + R_S)}{1-\rho}.$$

Substituting this inequality and the expressions of L_F, ℓ_F into (40) finishes the proof. $\qquad\square$

36

F.1 Proof of Theorem F.1

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

$$
\begin{aligned}
F_t(\theta_{t+1}) &\leq F_t(\theta_t) + \langle \nabla F_t(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\ell_F}{2} \|\theta_{t+1} - \theta_t\|^2 \\
&= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), G_t \rangle + \frac{\ell_F \eta^2}{2} \|G_t\|^2 \\
&= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), \nabla F_t(\theta_t) \rangle + \frac{\ell_F \eta^2}{2} \|\nabla F_t(\theta_t)\|^2 \\
&\quad - \eta \langle \nabla F_t(\theta_t), G_t - \nabla F_t(\theta_t) \rangle + \ell_F \eta^2 \langle \nabla F_t(\theta_t), G_t - \nabla F(\theta_t) \rangle \\
&\quad + \frac{\ell_F \eta^2}{2} \|\nabla F_t(\theta_t) - G_t\|^2 \\
&\leq F_t(\theta_t) - \eta \left(1 - \frac{\ell_F \eta}{2}\right) \|\nabla F_t(\theta_t)\|^2 + \eta(1 - \ell_F \eta) L\beta + \frac{\ell_F \eta^2}{2} \cdot \beta^2. \quad (45)
\end{aligned}
$$

Summing (45) over $t = 0, 1, \ldots, T - 1$ and rearranging the terms gives that

$$
\begin{aligned}
&\eta \left(1 - \frac{\ell_F \eta}{2}\right) \sum_{t=0}^{T-1} \|\nabla F_t(\theta_t)\|^2 \\
&\leq \sum_{t=0}^{T-1} (F_t(\theta_t) - F_t(\theta_{t+1})) + \left(\eta(1 - \ell_F \eta) L\beta + \frac{\ell_F \eta^2}{2} \cdot \beta^2\right) T \\
&\leq F_0(\theta_0) + \sum_{t=1}^{T-1} (F_t(\theta_t) - F_{t-1}(\theta_t)) + \left(\eta(1 - \ell_F \eta) L\beta + \frac{\ell_F \eta^2}{2} \cdot \beta^2\right) T \\
&\leq F_0(\theta_0) + L_F \sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1}) + \left(\eta(1 - \ell_F \eta) L_F \beta + \frac{\ell_F \eta^2}{2} \cdot \beta^2\right) T. \quad (46)
\end{aligned}
$$

F.2 Proof of Lemma F.3

We first show the limit $\lim_{\tau \to \infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$ exists. It suffices to show that $\{\tilde{x}_\tau^{(g,\pi)}(\theta)\}$ is a Cauchy sequence. Note that for $\tau' > \tau \geq 0$, we have

$$
\left\| \tilde{x}_{\tau'}^{(g,\pi)}(\theta) - \tilde{x}_\tau^{(g,\pi)}(\theta) \right\| \leq \sum_{j=\tau}^{\tau'-1} \left\| \tilde{x}_{j+1}^{(g,\pi)}(\theta) - \tilde{x}_j^{(g,\pi)}(\theta) \right\| \quad (47a)
$$

$$
\leq \sum_{j=\tau}^{\tau'-1} C \rho^j \left\| \tilde{x}_1^{(g,\pi)}(\theta) - x_0 \right\| \quad (47b)
$$

$$
\leq \frac{C \rho^\tau}{1 - \rho} \cdot \left\| \tilde{x}_1^{(g,\pi)}(\theta) - x_0 \right\|,
$$

where we use the triangle inequality in (47a) and Assumption 2.2 in (47b). Therefore, we see the limit $\lim_{\tau \to \infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$ exists because $\{\tilde{x}_\tau^{(g,\pi)}(\theta)\}$ is a Cauchy sequence and we denote $\tilde{x}_\infty^{(g,\pi)}(\theta) := \lim_{\tau \to \infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$.

Now, we show that $\tilde{x}_\infty^{(g,\pi)}(\theta)$ is a fixed point of the time-invariant closed-loop dynamics induced by (g, π, θ). To see this, note that

$$
\begin{aligned}
g\left(\tilde{x}_\infty^{(g,\pi)}(\theta), \pi(\tilde{x}_\infty^{(g,\pi)}(\theta), \theta)\right) &= g\left(\lim_{\tau \to \infty} \tilde{x}_\tau^{(g,\pi)}(\theta), \pi(\lim_{\tau \to \infty} \tilde{x}_\tau^{(g,\pi)}(\theta), \theta)\right) \\
&= \lim_{\tau \to \infty} g\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \pi(\tilde{x}_\tau^{(g,\pi)}(\theta), \theta)\right) \\
&= \lim_{\tau \to \infty} \tilde{x}_{\tau+1}^{(g,\pi)}(\theta) = \tilde{x}_\infty^{(g,\pi)}(\theta),
\end{aligned}
$$

37

where we can pull out $\lim_{\tau \to \infty}$ in the second equation because the right hand side is a continuous function of $\tilde{x}_\tau^{(g,\pi)}(\theta)$ at the point $\tilde{x}_\infty^{(g,\pi)}(\theta)$ by Assumption 2.1.

Therefore, applying the contractive perturbation property in Assumption 2.2 gives that

$$\left\| \tilde{x}_\tau^{(g,\pi)}(\theta) - \tilde{x}_\infty^{(g,\pi)}(\theta) \right\| \leq C\rho^\tau \left\| x_0 - \tilde{x}_\infty^{(g,\pi)}(\theta) \right\| \leq C\rho^\tau \mathsf{diam}(\mathcal{X}).$$

F.3 Proof of Lemma F.4

To simplify the notation, we introduce the notations

$$\mathsf{dist}_d(g, g') := \sup_{x \in \mathcal{X}, u \in \mathcal{U}} \|g(x, u) - g'(x, u)\|, \quad \mathsf{dist}_p(\pi, \pi') := \sup_{x \in \mathcal{X}, \theta \in \Theta} \|\pi(x, \theta) - \pi'(x, \theta)\|,$$

$$\mathsf{dist}_c(f, f') := \sup_{x \in \mathcal{X}, u \in \mathcal{U}} |f(x, u) - f'(x, u)|.$$

For any $\theta \in \Theta$, we see that

$$
\begin{aligned}
&|F_t(\theta) - F_{t-1}(\theta)| \\
&= |f_t(\hat{x}_t(\theta), \pi_t(\hat{x}_t(\theta), \theta)) - f_{t-1}(\hat{x}_{t-1}(\theta), \pi_{t-1}(\hat{x}_{t-1}(\theta), \theta))| && \text{(48a)} \\
&\leq |f_t(\hat{x}_t(\theta), \pi_t(\hat{x}_t(\theta), \theta)) - f_t(\hat{x}_{t-1}(\theta), \pi_{t-1}(\hat{x}_{t-1}(\theta), \theta))| + \mathsf{dist}_c(f_t, f_{t-1}) && \text{(48b)} \\
&\leq L_f(\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| + \|\pi_t(\hat{x}_t(\theta), \theta) - \pi_{t-1}(\hat{x}_{t-1}(\theta), \theta)\|) + \mathsf{dist}_c(f_t, f_{t-1}) && \text{(48c)} \\
&\leq L_f(\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| + \|\pi_t(\hat{x}_t(\theta), \theta) - \pi_t(\hat{x}_{t-1}(\theta), \theta)\|) \\
&\quad + (L_f \mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(f_t, f_{t-1})) && \text{(48d)} \\
&\leq L_f(1 + L_{\pi,x})\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| + (L_f \mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(f_t, f_{t-1})), && \text{(48e)}
\end{aligned}
$$

where we use the definition of the surrogate cost in (48a); we use the definition of dist_c in (48b); we use the assumption that f_t is Lipschitz in (48c); we use the definition of dist_π on (48d); we use the assumption that π_t is Lipschitz in (48e).

To bound $\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\|$, we first bound the difference between $\hat{x}_t(\theta)$ and $\tilde{x}_t^{(g,\pi)}(\theta)$ for arbitrary $(g, \pi) \in \mathcal{G}$. Note that $\hat{x}_t(\theta) = g_{t|0}\left(\tilde{x}_0^{(g,\pi)}, \theta_{\times t} \right)$, thus

$$
\begin{aligned}
\left\| \hat{x}_t(\theta) - \tilde{x}_t^{(g,\pi)}(\theta) \right\| &\leq \sum_{\tau=0}^{t-1} \left\| g_{t|\tau+1}\left(\tilde{x}_{\tau+1}^{(g,\pi)}(\theta), \theta_{\times(t-\tau-1)} \right) - g_{t|\tau}\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \theta_{\times(t-\tau)} \right) \right\| && \text{(49a)} \\
&\leq C \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \left\| \tilde{x}_{\tau+1}^{(g,\pi)}(\theta) - g_{\tau+1|\tau}\left(\tilde{x}_{\tau+1}^{(g,\pi)}(\theta), \theta \right) \right\| && \text{(49b)} \\
&\leq C \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} (\mathsf{dist}_d(g_\tau, g) + L_{g,u} \mathsf{dist}_p(\pi_\tau, \pi)), && \text{(49c)}
\end{aligned}
$$

where we use the triangle inequality in (49a); we use the contractive perturbation property in (49b); we use the definitions of dist_d and dist_p in (49c).

Therefore, we see that

$$
\begin{aligned}
\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| &\leq \left\| \hat{x}_t(\theta) - \tilde{x}_t^{(g_t,\pi_t)}(\theta) \right\| + \left\| \hat{x}_{t-1}(\theta) - \tilde{x}_{t-1}^{(g_t,\pi_t)}(\theta) \right\| \\
&\quad + \left\| \tilde{x}_t^{(g_t,\pi_t)}(\theta) - \tilde{x}_\infty^{(g_t,\pi_t)}(\theta) \right\| + \left\| \tilde{x}_{t-1}^{(g_t,\pi_t)}(\theta) - \tilde{x}_\infty^{(g_t,\pi_t)}(\theta) \right\| && \text{(50a)} \\
&\leq C \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} (\mathsf{dist}_d(g_\tau, g_t) + L_{g,u} \mathsf{dist}_p(\pi_\tau, \pi_t)) \\
&\quad + C \sum_{\tau=0}^{t-2} \rho^{t-\tau-2} (\mathsf{dist}_d(g_\tau, g_t) + L_{g,u} \mathsf{dist}_p(\pi_\tau, \pi_t)) \\
&\quad + C\rho^t \mathsf{diam}(\mathcal{X}) + C\rho^{t-1} \mathsf{diam}(\mathcal{X}) && \text{(50b)}
\end{aligned}
$$

38

$$\leq \frac{2C}{1-\rho} \cdot \sum_{\tau=0}^{t-1} \rho^{t-\tau-2}(\mathsf{dist}_d(g_\tau, g_{\tau+1}) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_{\tau+1}))$$
$$+ 2C\rho^{t-1}\mathsf{diam}(\mathcal{X}), \tag{50c}$$

where we use the triangle inequality in (50a); in (50b), we use the bound we derived in (49); in (50c), we use the triangle inequality decomposition

$$\mathsf{dist}_d(g_\tau, g_t) \leq \sum_{j=\tau}^{t-1} \mathsf{dist}_d(g_{j+1}, g_j), \text{ and } \mathsf{dist}_p(\pi_\tau, \pi_t) \leq \sum_{j=\tau}^{t-1} \mathsf{dist}_p(\pi_{j+1}, \pi_j).$$

Substituting this into (48) gives that

$$\mathsf{dist}_s(F_t, F_{t-1}) \leq \frac{2CL_f(1 + L_{\pi,x})}{1-\rho} \cdot \sum_{\tau=0}^{t-1} \rho^{t-\tau-2}(\mathsf{dist}_d(g_\tau, g_{\tau+1}) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_{\tau+1}))$$
$$+ 2CL_f(1 + L_{\pi,x})\rho^{t-1}\mathsf{diam}(\mathcal{X}) + (L_f\mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(f_t, f_{t-1})) \tag{51}$$

because (50) holds for arbitrary $\theta \in \Theta$.

Summing (51) over $t = 1, \ldots, T - 1$ and rearranging the terms give that

$$\sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1})$$
$$\leq \frac{2CL_f(1 + L_{\pi,x})}{1-\rho} \cdot \sum_{t=1}^{T-1}\sum_{\tau=0}^{t-1} \rho^{t-\tau-2}(\mathsf{dist}_d(g_\tau, g_{\tau+1}) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_{\tau+1}))$$
$$+ 2CL_f(1 + L_{\pi,x})\sum_{t=1}^{T-1} \rho^{t-1}\mathsf{diam}(\mathcal{X}) + \sum_{t=1}^{T-1}(L_f\mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(f_t, f_{t-1}))$$
$$\leq \frac{2CL_f(1 + L_{\pi,x})(1 + L_{g,u})}{(1-\rho)^2\rho} \cdot P_{\mathcal{T}} + \frac{2CL_f(1 + L_{\pi,x})}{1-\rho} \cdot \mathsf{diam}(\mathcal{X}).$$

G Learning Finite Policy Parameters

In this appendix, we consider a different setting of online policy selection where the policy parameters come from a finite set $\Theta = \{1, 2, \ldots, k\}$ rather than a convex subset of $\mathbb{R}^d$. In this setting, we can also adopt a proof approach based on the contractive perturbation property to achieve a sublinear (static) policy regret bound. Specifically, we establish a regret bound in expectation under the same assumptions as the continuous parameter case [5], though it requires a different algorithm. Our results for the finite setting also requires less information: here we do not assume knowledge of the partial derivatives of the dynamics and cost.

G.1 Algorithm

We propose an algorithm, Bandit-based Adaptive Policy Selection (BAPS, Algorithm 3), that deploys the same policy parameter for a batch of b time steps before switching. Intuitively, when we switch the policy parameter, we wait until the state "forgets" the impact of the parameter used in the previous batch before we evaluate the current parameter's performance. With a sufficiently large batch size b, we can guarantee that the difference between the mean actual cost over a batch and the mean surrogate cost is negligible.

G.2 Analysis

We present our main result for BAPS in Theorem G.1.

---

[5]Since step size constraints on policy parameters are not well-defined in the finite setting, by "under the same assumptions", we mean Assumption 2.2 holds with $\varepsilon = 0$.

39

Algorithm 3 Bandit-based Adaptive Policy Selection (BAPS)

Require: policy count k, batch size b, learning rate η.
 1: Set the initial distribution to uniform: $s_0 = (1/k)_{\times k}$.
 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 $u_t = \pi_t(x_t, j_m)$.
 4:   Compute the batched loss vector $\hat{\ell}_m^b \in \mathbb{R}^k$:

$$\hat{\ell}_m^b(j) = \begin{cases} \frac{1}{s_m(j_m)} \sum_{t=mb}^{(m+1)b-1} f_t(x_t, u_t), & j = j_m \\ 0, & j \neq j_m. \end{cases}$$

 5:   Set $s_{m+1}(j) = s_m(j)e^{-\eta\hat{\ell}_m^b(j)}, \forall j \in \{1, \ldots, k\}$ and do the normalization $s_{m+1} \leftarrow s_{m+1}/\|s_{m+1}\|_1$.
 6: **end for**

Theorem G.1. *For the finite parameter setting, suppose Assumption 2.1 holds and Assumption 2.2 holds with $\varepsilon = 0$. Then, BAPS (Algorithm 3) with batch size $b \geq \tau_0$ and learning step size $\eta > 0$ attains the following expected policy regret, where the expectation is taken over the randomness of the sampling mechanism in line 3 of Algorithm 3:*

$$\mathbb{E}\big[R^S(T)\big] \leq CD_0 \cdot (1-\rho)^{-1}b^{-1}T + kD_0\eta bT + \log k \cdot \eta^{-1},$$

where $R^S(T) = \sum_{t=0}^{T-1} f_t(x_t, u_t) - \min_{\theta \in \Theta} \sum_{t=0}^{T-1} F_t(\theta)$. Here, we define

$$R_0 = \min\{\frac{C\|x_0\| + R_S}{\rho}, R_C\}, \tau_0 = \log\left(\frac{CR_0}{R_0 - R_S}\right)/\log\left(\frac{1}{\rho}\right), \text{ and}$$

$$D_0 = \max_{t \in \mathcal{T}} \sup_{x \in \mathcal{X}, u \in \mathcal{U}} |f_t(x, u)|.$$

We discuss how to choose the learning rate and the regret it achieves in the following corollary.

Corollary G.2. *Under the same assumptions as Theorem G.1, suppose $T \gg \tau_0$, if we set the parameters as $b = \left(\frac{C_0^2 D_0 T}{(1-\rho)^2 k \log k}\right)^{\frac{1}{3}}$ and $\eta = \left(\frac{(1-\rho)(\log k)^2}{C_0 D_0^2 k T^2}\right)^{\frac{1}{3}}$, we have $\mathbb{E}\big[R^S(T)\big] = 3\left(\frac{CD_0^2 \cdot k \log k \cdot T^2}{1-\rho}\right)^{\frac{1}{3}}$.*

Note that the $O(T^{2/3})$ static policy regret bound in expectation is worse than the $O(\sqrt{T})$ bound for GAPS in Theorem 3.3 in the dependence on the horizon T. However, it matches the regret lower bound for MAB with switching costs [63], which we conjecture to be the parallel online learning setting for online policy 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 θ can be modeled as an unknown switching cost with a constant upper bound. An interesting open question is to verify this intuition and see whether the $O(T^{2/3})$ static policy regret is tight for finite policy selection under the assumptions of BAPS.

Now we present the proof of Theorem G.1.

Proof of Theorem G.1. We first show by induction that the trajectory achieved by a "batched" policy parameter with batch size $b \geq \tau_0$ will always stay within $\mathcal{X} \times \mathcal{U}$, so that the cost is uniformly bounded by D_0.

To see this, we show the state x_{mb} at first time step in each batch satisfies $\|x_{mb}\| \leq R_0$. Note that the statement holds for $m = 0$. Suppose it holds for $m - 1$. By Assumption 2.2, we see that

$$\|x_{mb}\| \leq C\rho^b\|x_{(m-1)b}\| + R_S \leq C\rho^{\tau_0} \cdot R_0 + R_S \leq R_0,$$

where we use the definition of τ_0 in the last inequality. So $\|x_{mb}\| \leq R_0$ holds for all m by induction. Then, for any time step t that satisfies $mb < t < (m+1)b$, we have

$$\|x_t\| \leq C\rho^{t-mb}\|x_{mb}\| + R_S \leq C(C\|x_0\| + R_S) + R_S.$$

Thus, we see that $(x_t, u_t) \in \mathcal{X} \times \mathcal{U}$ and $|f_t(x_t, u_t)| \leq D_0$.

40

The rest of the proof is inspired by the proof of Lemma 6.3 in [32]. By Theorem 1.5 in [32], the following inequality holds for any $j \in \Theta$:

$$\sum_{m=0}^{T/b-1} s_m^\top \hat{\ell}_m^b \leq \sum_{m=0}^{T/b-1} \hat{\ell}_m^b(j) + \eta \sum_{m=0}^{T/b-1} s_m^\top (\hat{\ell}_m^b)^2 + \frac{\log k}{\eta}. \tag{52}$$

Taking the expectation on both sides gives that

$$
\sum_{m=0}^{T/b-1} \mathbb{E}\left[s_m^\top \hat{\ell}_m^b \right]
$$

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

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

$$
= \sum_{m=0}^{T/b-1} \mathbb{E}\left[\sum_{t=mb}^{(m+1)b-1} f_{t|mb}(x_{mb}, j_{\times(t-mb)}) \right]
$$

$$
+ \eta \sum_{m=0}^{T/b-1} \mathbb{E}\left[\sum_{q\in\Theta} \left(\sum_{t=mb}^{(m+1)b-1} f_{t|mb}(x_{mb}, q_{\times(t-mb)}) \right)^2 \right] + \frac{\log k}{\eta}
$$

$$
\leq \left(\sum_{t=0}^{T-1} F_t(j) + \frac{2C_0 L_0 R_x}{1-\rho} \cdot \frac{T}{b} \right) + \eta b D_0^2 kT + \frac{\log k}{\eta}. \tag{53}
$$

Note that the left hand side of (53) is equal to the expected cost incurred by Algorithm 3, i.e., $\mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t, u_t) \right]$. Therefore, we see that for any $j \in \Theta$, we have

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

Since the cost functions and dynamics are oblivious, this finishes the proof. $\square$

H Discussion of Examples

In this section, we discuss why Examples 2.1 and 2.2 satisfy the time-varying contractive perturbation and stability in Assumption 2.2 as well as Assumption 2.1 about Lipschitzness and smoothness. Before presenting these two examples, we first explain why contractive perturbation generalizes the notion of the disturbance-action controller (DAC), which is a policy class that has been studied intensively in the literature of online control [1, 3, 6].

H.1 Disturbance-action controller

The example about disturbance-action controller (DAC) considers 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 bounded adversarial disturbance at time t. The DAC class is given by

$$\pi_t(x_t, \theta_t) = -K_t^{stb} x_t + \sum_{i=1}^{h} M_t^{[i]} w_{t-i},$$

where K_t^{stb} is a given matrix sequence such that for some positive constants C and $\rho < 1$

$$\left\| \prod_{t=t_1}^{t_2-1} (A_t - B_t K_t^{stb}) \right\| \leq C \rho^{t_2 - t_1}, \forall t_2 > t_1.$$

In this example, $\Theta := \{ (M^{[i]})_{i\in[1:h]} \mid \sum_{i=1}^{h} \| M^{[i]} \| \leq R_M \}$ for some $R_M > 0$.

41

DAC has been widely used in recent online control literature [1, 3, 6]. When applied to linear dynamics and convex cost functions, the DAC class renders the multi-step cost functions $f_{t|\tau}$ convex in the policy parameters, which enables the reduction to OCO with Memory [1, 3]. It is also shown to be sufficient to approximate any strongly stable state feedback policy when the memory length h is large enough.

Note that the parameter set Θ is a convex compact subset of $\mathbb{R}^{hnm}$ with diameter $D \leq nR_M$. For this example to satisfy Assumption 2.1, we additionally assume that there exists positive constants $a, b, \bar{k}, \bar{w}$ such that $\|A_t\| \leq a$, $\|B_t\| \leq b$, $\|K_t^{stb}\| \leq \bar{k}$, $\|w_t\| \leq \bar{w}$ holds for all t. The cost functions f_t is $(\ell_{f,x}, \ell_{f,u})$-smooth in (x, u) and $\arg \min f_t = (0, 0)$.

DAC is special because it satisfies Assumption 2.2 with arbitrarily large ε. To see this, we define the state transition matrix as

$$\Psi_{t|\tau} := \begin{cases} \prod_{\tau'=1}^{t-\tau}(A_{t-\tau'} - B_{t-\tau'}K_{t-\tau'}^{stb}) & \text{if } t > \tau, \\ I & \text{otherwise.} \end{cases}$$

By assumption, we know that $\|\psi_{t|\tau}\| \leq C\rho^{t-\tau}$.

We first show the ε-time-varying contractive perturbation property (Definition 2.6) holds for $R_C = +\infty, \varepsilon = +\infty$. To see this, note that for any two time steps $t \geq \tau$, states $x_\tau, x_\tau' \in \mathbb{R}^n$, and $\theta_{\tau:t-1} \in \Theta^{t-\tau}$, we have that

$$g_{t|\tau}(x_\tau, \theta_{\tau:t-1}) - g_{t|\tau}(x_\tau', \theta_{\tau:t-1}) = \Psi_{t|\tau}(x_\tau - x_\tau'),$$

which implies that

$$\left\|g_{t|\tau}(x_\tau, \theta_{\tau:t-1}) - g_{t|\tau}(x_\tau', \theta_{\tau:t-1})\right\| \leq \left\|\Psi_{t|\tau}\right\| \cdot \|x_\tau - x_\tau'\| \leq C\rho^{t-\tau}\|x_\tau - x_\tau'\|.$$

To see that the ε-time-varying stability (Definition 2.7) holds with $R_S = \frac{CR_M\bar{w}}{1-\rho}$ and $\varepsilon = +\infty$, note that for arbitrary $\theta_{\tau:t-1} \in \Theta^{t-\tau}$, we have

$$x_t = g_{t|\tau}(x_\tau, \theta_{\tau:t-1}) = \psi_{t|\tau}x_\tau + \sum_{\tau'=\tau}^{t} \psi_{t|\tau'}\left(\sum_{i=1}^{h} M_{\tau'}^{[i]}w_{\tau'-i}\right). \tag{54}$$

Taking norms on both side of the equation and applying the triangle inequality gives that

$$\|x_t\| \leq \|\psi_{t|\tau}\| \cdot \|x_\tau\| + \sum_{\tau'=\tau}^{t} \|\psi_{t|\tau'}\|\left(\sum_{i=1}^{h}\left\|M_{\tau'}^{[i]}\right\| \cdot \|w_{\tau'-i}\|\right)$$

$$\leq C\rho^{t-\tau}\|x_\tau\| + \sum_{\tau'=\tau}^{t} C\rho^{t-\tau'}R_M\bar{w}$$

$$\leq C\rho^{t-\tau}\|x_0\| + \frac{CR_M\bar{w}}{1-\rho},$$

where we use the assumption that $\|\psi_{t|\tau}\| \leq C\rho^{t-\tau}$, the definition of the policy class, and the bound on disturbances in the second inequality. Therefore, ε-time-varying stability holds with $R_S = \frac{CR_M\bar{w}}{1-\rho}$.

H.2 MPC with confidence coefficients

Our example about Model Predictive Control (MPC) with Confidence Coefficients generalizes the λ-confident policy proposed by [11]. In this setting, some policy parameter $\theta^{(c)} \in \Theta$ 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). Minimizing regret in this setting implies that the online policy is both robust and consistent.

Recall that in this example, MPC selects the current control action by solving the optimization problem

$$\arg\min_{u_{t:t+k-1|t}} \sum_{\tau=t}^{t+k-1} f_\tau(x_{\tau|t}, u_{\tau|t}) + q(x_{t+k|t}, \tilde{Q})$$

$$\text{s. t. } x_{t|t} = x_t, \tag{55}$$

$$x_{\tau+1|t} = A_\tau x_{\tau|t} + B_\tau u_{\tau|t} + \lambda_t^{[\tau-t]}w_{\tau|t} : t \leq \tau < t+k,$$

42

where $\theta_t = \left(\lambda_t^{[0]}, \lambda_t^{[1]}, \ldots, \lambda_t^{[k-1]}\right)$, $\Theta \subseteq [0,1]^k$. Thus, Θ is a convex compact set with diameter $\sqrt{k}$.

For this example to satisfy Assumptions 2.1 and 2.2, we make two standard assumptions that are also required by prior works on online control with MPC [25, 64]. The first assumption is about the uniform bounds on the dynamical matrices, cost function matrices, and disturbances.

Assumption H.1. *For any time step $t \in \mathcal{T}$, we have $\|A_t\| \leq a, \|B_t\| \leq b, \|w_t\| \leq \bar{w}$, and*

$$\mu I_n \preceq Q_t \preceq \ell I_n, \mu I_m \preceq R_t \preceq \ell I_m, \mu I_n \preceq \tilde{Q} \preceq \ell I_n.$$

We also assume that $\left\|w_{\tau|t}\right\| \leq \bar{w}$ for predicted disturbances.

The second assumption is about the uniform controllability of the LTV system:

Assumption H.2. *We define the transition matrix in this LTV system as*

$$\Phi(t_2, t_1) := \begin{cases} A_{t_2-1} \cdots A_{t_1}, & \text{if } t_2 > t_1, \\ I, & \text{otherwise.} \end{cases}$$

For any two time steps $t' > t$, we define

$$\Xi_{t,t'} = [\Phi(t', t+1)B_t, \Phi(t', t+2)B_{t+1}, \ldots, \Phi(t', t')B_{t'}].$$

We assume that there exists a positive integer d_0 such that the smallest singular value of the matrix $\Xi_{t,t'}$ is uniformly lower bounded by some positive constant σ when $t' \geq t + d_0$, i.e., $\sigma_{min}(\Xi_{t,t'}) \geq \sigma$ holds for any $t' \geq t + d_0$, where $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix.

Before we proceed to show that Assumptions 2.1 and 2.2 hold in this example, we first define an auxiliary parameterized optimal problem control problem that will be used in our analysis: For any two time steps $t < t'$, let $\psi_t^{t'}(y_t, v_{t:t'}; P_{t'})$ denote the optimal trajectory planned according to initial state y_t, disturbances $v_{t:t'}$, and terminal cost matrix $P_{t'}$, i.e.,

$$\psi_t^{t'}(y_t, v_{t:t'}; P_{t'}) = \underset{x_{t:t'}, u_{t:(t'-1)}}{\arg\min} \sum_{\tau=t}^{t'-1} f_\tau(x_\tau, u_\tau) + \frac{1}{2}x_{t'}^\top P_{t'} x_{t'}$$

$$\text{s.t. } x_{\tau+1} = A_\tau x_\tau + B_\tau u_\tau + v_\tau, \forall \tau \in [t : t'-1];$$

$$x_t = y_t.$$

Using this notation, we can express MPC with confidence coefficients as

$$\pi_t(x_t, \theta_t) = \psi_t^{t+k}(x_t, \{\lambda_t^{[\tau-t]}w_{\tau|t}\}_{\tau \in [t:t+k-1]}; \tilde{Q})_{u_t},$$

where the index u_t denotes the corresponding entry in the predictive optimal solution. The perturbation bound in [25, Theorem 3.3] states that for any $\tau \in [t, t']$, we have

$$\left\| \psi_t^{t'}(y_t, v_{t:t'}; \tilde{Q})_{u_\tau} - \psi_t^{t'}(y_t', v_{t:t'}'; \tilde{Q})_{u_\tau} \right\| \leq C_0 \left(\rho_0^{\tau-t}\|y_t - y_t'\| + \sum_{j=t}^{t'} \rho_0^{|\tau-j|}\|v_j - v_j'\| \right), \quad (56)$$

where $C_0 > 0, \rho_0 \in (0, 1)$ are constants that depends on the system parameters including a, b, μ, ℓ, σ, and d_0. And the inequality (56) still holds if we replace the index u_τ on the left hand side by x_τ. Therefore, we know that $\psi_t^{t+k}(y_t, v_{t:t+k}; \tilde{Q})_{u_t}$ is bounded Lipschitz in y_t and $v_{t:t+k}$. Note that $\pi_t(x_t, \theta_t)$ is an affine function in its inputs (x_t, θ_t), i.e., $\pi_t(x_t, \theta_t)$ can be expressed equivalently as

$$\pi_t(x_t, \theta_t) = -\bar{K}_t^{(k)}x_t - \sum_{\tau=t}^{t+k-1} \lambda_t^{[\tau-t]}\bar{K}_t^{(k,\tau)}w_{\tau|t}, \quad (57)$$

where the matrices $\bar{K}_t^{(k)}, \bar{K}_t^{(k,\tau)}$ only depends on $\{(A_t, B_t, Q_t, R_t)\}_{t\in\mathcal{T}}$ and $\tilde{Q}$ [23, 54]. The superscript k denotes the prediction horizon of the MPC we adopt, thus $k = T$ will give MPC future predictions all the way to the end of the online policy selection game. So the smoothness constants $\ell_{\pi,x} = \ell_{\pi,\theta} = 0$. Thus, we see that Assumption 2.1 holds. We also see that the surrogate cost function F_t is convex.

Next, we show a lemma about contractive perturbation and stability.

43

Lemma H.1. *Suppose Assumptions H.1 and H.2 hold. Recall that C_0 and ρ_0 are given in (56). Then, for any $\rho \in (\rho_0, 1)$, if the prediction horizon satisfies*

$$k \geq \frac{1}{2} \log\big(C_0^3 ab\rho_0/(\rho - \rho_0)\big)/\log(1/\rho_0),$$

the MPC with confidence coefficients policy class satisfies ε-time-varying contractive perturbation with $\varepsilon = +\infty$, $R_C = +\infty$, $C = C_0$ and decay factor ρ. It also satisfies ε-time-varying stability with $\varepsilon = +\infty$ and $R_S = \frac{C_0(1 - \rho_0 + C_0)\bar{w}}{(1 - \rho_0)(1 - \rho)}$.

Proof of Lemma H.1. By the perturbation bound in (56), we see that for any $t' > t$,

$$\left\| \psi_t^T(y_t, 0_{\times(T-t)}; \tilde{Q})_{x_{t'}} - \psi_t^T(y_t', 0_{\times(T-t)}; \tilde{Q})_{x_{t'}} \right\| \leq C_0 \rho_0^{t'-t} \| y_t - y_t' \|.$$

Therefore, we obtain that for any $t' > t$,

$$\left\| (A_{t'-1} - B_{t'-1} \bar{K}_{t'-1}^{(T)})(A_{t'-2} - B_{t'-2} \bar{K}_{t'-2}^{(T)}) \cdots (A_t - B_t \bar{K}_t^{(T)}) \right\| \leq C_0 \rho_0^{t'-t}. \tag{58}$$

Now we show that

$$\left\| (A_{t'-1} - B_{t'-1} \bar{K}_{t'-1}^{(k)})(A_{t'-2} - B_{t'-2} \bar{K}_{t'-2}^{(k)}) \cdots (A_t - B_t \bar{K}_t^{(k)}) \right\| \leq C_0 \rho^{t'-t}. \tag{59}$$

To see this, we construct a sequence $\hat{v}_{t:T-1}$ such that $\hat{v}_{t+k} = -A_{t+k} \psi_t^{t+k}(x_t, 0_{\times k}; \tilde{Q})_{x_{t+k}}$ and $\hat{v}_\tau = 0$ for $\tau \neq t+k, \tau \in [t : T-1]$. We observe that

$$\left\| \psi_t^{t+k}(x_t, 0_{\times k}; \tilde{Q}) - \psi_t^T(x_t, 0_{\times(T-t)}; \tilde{Q}) \right\|$$
$$= \left\| \psi_t^T(x_t, \hat{v}_{t:T-1}; \tilde{Q}) - \psi_t^T(x_t, 0_{\times(T-t)}; \tilde{Q}) \right\|$$
$$\leq C_0 \rho_0^k \| \hat{v}_{t+k} \| \leq C_0^2 a \rho_0^{2k} \| x_t \|,$$

where we use (56) in the last line. To simplify the notation, we define $M_\tau^{(p)} := A_\tau - B_\tau \bar{K}_\tau^{(p)}$ and $\alpha := C_0^2 \rho_0^{2k} ab$. By the above inequality, we see that

$$\left\| M_\tau^{(k)} - M_\tau^{(T)} \right\| \leq \alpha, \forall \tau \in \mathcal{T}. \tag{60}$$

Therefore, we obtain that

$$\left\| M_{t'-1}^{(k)} M_{t'-2}^{(k)} \cdots M_t^{(k)} \right\| \leq \sum_{j=0}^{t'-t} \binom{t'-t}{j} C_0^{j+1} \rho_0^{t'-t} \alpha^j \tag{61a}$$

$$\leq C_0 \rho_0^{t'-t} (1 + C_0 \alpha)^{t'-t} \leq C_0 \rho^{t'-t}, \tag{61b}$$

where we use the decomposition $M_\tau^{(k)} = M_\tau^{(T)} + (M_\tau^{(k)} - M_\tau^{(T)})$, the triangle inequality, and (60) in (61a); we use the condition that $k \geq \frac{1}{2} \log\big(C_0^3 ab\rho_0/(\rho - \rho_0)\big)/\log(1/\rho_0)$ in (61b). This finishes the proof of (59).

Now we consider two trajectories that apply the same policy parameter sequence but start from different states x_τ and x_τ'. For arbitrary $\theta_{\tau:t-1} = \Theta^{t-\tau}$, we see that

$$\left\| g_{t|\tau}(x_\tau, \theta_{\tau:t-1}) - g_{t|\tau}(x_\tau', \theta_{\tau:t-1}) \right\|$$
$$= \left\| (A_{t-1} - B_{t-1} \bar{K}_{t-1}^{(k)})(A_{t-2} - B_{t-2} \bar{K}_{t-2}^{(k)}) \cdots (A_\tau - B_\tau \bar{K}_\tau^{(k)})(x_\tau - x_\tau') \right\| \tag{62a}$$
$$\leq C_0 \rho^{t-\tau} \| x_\tau - x_\tau' \|, \tag{62b}$$

where we use the affine expression of π_t (57) and the fact that these two trajectories experience the same sequence of disturbances and predictions. This finishes the proof of ε-time-varying contractive perturbation with $\varepsilon = +\infty$ and $R_C = +\infty$.

By the perturbation bound in (56), we also see that

$$\left\| \psi_t^{t+k}(0, v_{t:t+k-1}; \tilde{Q})_{u_t} - \psi_t^{t+k}(0, v_{t:t+k-1}'; \tilde{Q})_{u_t} \right\| \leq C_0 \sum_{\tau=t}^{t+k-1} \rho_0^{\tau-t} \| v_\tau - v_\tau' \|.$$

44

Combining this inequality with the affine relationship in (58), we see that

$$\left\|\bar{K}_t^{(k,\tau)}\right\| \le C_0 \rho_0^{\tau - t}. \tag{63}$$

Therefore, we obtain that

$$
\begin{aligned}
&\left\|g_{t|\tau}(0, \theta_{\tau:t-1})\right\| \\
&= \left\|\sum_{i=\tau}^{t-1}(A_{t-1} - B_{t-1}\bar{K}_{t-1}^{(k)}) \cdots (A_{i+1} - B_{i+1}\bar{K}_{i+1}^{(k)})\left(w_i - \sum_{j=i}^{i+k-1}\lambda_i^{[j-i]}\bar{K}_i^{(k,j)}w_{j|i}\right)\right\| \\
&\le \sum_{i=\tau}^{t-1}C_0\rho^{t-1-i}\left\|w_i - \sum_{j=i}^{i+k-1}\lambda_i^{[j-i]}\bar{K}_i^{(k,j)}w_{j|i}\right\| \tag{64a} \\
&\le \sum_{i=\tau}^{t-1}C_0\rho^{t-1-i}\left(\bar{w} + \sum_{j=i}^{i+k-1}C_0\rho_0^{j-i}\bar{w}\right) \tag{64b} \\
&\le \frac{C_0(1 - \rho_0 + C_0)\bar{w}}{(1 - \rho_0)(1 - \rho)},
\end{aligned}
$$

where we use the triangle inequality and (59) in (64a); we use the triangle inequality and (63) in (64b). This finishes the proof of ε-time-varying stability. $\square$

H.3 Linear Feedback Control in Nonlinear System

The example we present below is an extension of the nonlinear control model in [47].

We consider the dynamical system $g_t(x_t, u_t) = Ax_t + Bu_t + \delta_t(x_t, u_t)$, where (A, B) is stabilizable, i.e., there exists $K \in \mathbb{R}^{n \times m}$ such that $\rho(A - BK) < 1$. The cost functions are given by $f_t(x_t, u_t) = q(x_t, Q) + q(u_t, R)$, where $Q \succ 0$ and $R \succeq 0$. Under these assumptions, we know the discrete algebraic Riccati equation (DARE) has a unique solution $P \in \mathbb{S}^n$ such that $P \succeq 0$ and

$$P = Q + A^\top PA - A^\top PB(R + B^\top PB)^{-1}B^\top PA.$$

We let $\bar{K} := (R + B^\top PB)^{-1}B^\top PA$ and introduce the notation $M := A - B\bar{K}$. We know that $\rho(M) < 1$. Thus, there exists constants $C_0 > 0, \rho_0 \in (0, 1)$ such that the inequality $\|M^t\| \le C_0 \rho_0^t$ holds. To simplify the presentation, we assume $x_0 = 0$, $\mathcal{X} = B(0, R_x)$, and $\mathcal{U} = B(0, R_u)$ with $R_u > \|\bar{K}\|R_x$. Our main assumption bounds the magnitude and Lipschitzness of the nonlinear residual δ_t:

Assumption H.3. *Function δ_t is Lipschitz continuous on $\mathcal{X} \times \mathcal{U}$ and satisfies that*

$$\|\delta_t(0, 0)\| \le \bar{\delta}, \text{ and } \|\delta_t(x, u) - \delta_t(x', u')\| \le L_{\delta,x}\|x - x'\| + L_{\delta,u}\|u - u'\|$$

for any $x, x' \in \mathcal{X}$ and $u, u' \in \mathcal{U}$.

Compared with Assumption 1 in [47], Assumption H.3 does not require $\delta_t(0, 0) = 0$ or assume the Lipschitzness holds globally.

Define the set $\mathcal{K}_\sigma := \{\hat{K} \in \mathbb{R}^{n \times m} \mid \|\hat{K} - \bar{K}\| \le \sigma\}$. The policy class in this example is $\pi_t(x_t, \theta_t) = K(\theta_t)x_t$, where $K : \Theta \to \mathcal{K}_\sigma$ is a L_K-Lipschitz function. Here, we do not directly use $\mathcal{K}_\sigma$ as our parameter set because it may be difficult to do projection and we want to leave the freedom to set Θ to the whole Euclidean space.

Now we provide an example where $\Theta = \mathbb{R}^d$. Let $K_1, K_2, \ldots, K_d$ be d different linear feedback controllers in the set $\mathcal{K}_\sigma$. For any $\theta \in \Theta$, we define

$$K(\theta) := \sum_{i=1}^{d} \mathsf{softmax}(\theta)_i K_i, \text{ where } \mathsf{softmax}(\theta)_i = \frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}}, \text{ for } i = 1, \ldots, d.$$

Note that K is a Lipschitz and differentiable function in this example.

45

Lemma H.2. *Under Assumption H.3, suppose σ and the Lipschitzness constants satisfies that*

$$0 < \sigma \le \frac{R_u - \|\bar{K}\|R_x}{R_x}, \text{ and } \rho_0 + C_0\big(\sigma\|B\| + L_{\delta,x} + L_{\delta,u}(\|\bar{K}\| + \sigma)\big) < 1.$$

Let $\rho := \rho_0 + C_0\big(\sigma\|B\| + L_{\delta,x} + L_{\delta,u}(\|\bar{K}\| + \sigma)\big)$. We also assume that $\bar{\delta} < (1-\rho)R_x/C_0$. Then, the joint system of (g_t, f_t, π_t) satisfies time-invariant contractive perturbation with

$$C = C_0, \rho = \rho, R_C = \frac{(1-\rho)R_x - C_0\bar{\delta}}{(1-\rho)C_0}.$$

It also satisfies time-invariant stability with $R_S = \frac{C_0\bar{\delta}}{1-\rho}$.

The assumptions of Lemma H.2 requires the nonlinear residual to be sufficiently small and Lipschitz, and the radius σ of the set $\mathcal{K}_\sigma$ should also be sufficiently small. Such assumptions are also required by prior works [12, 47].

One can use Lemma 2.8 to extend the time-invariant contractive perturbation/stability to the ε-time-varying version. And to satisfy the assumptions of Lemma 2.8, we need to additionally assume $\bar{\delta}$ is sufficiently small such that

$$\bar{\delta} \le \frac{(1-\rho)R_x}{C_0(C_0+1)^2}.$$

Proof of Lemma H.2. We first study stability. When the policy parameter is fixed to be θ, the control action is $u_t = -K(\theta)x_t$ for all time step t. We see that

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t + \delta_t(x_t, u_t) = (A - BK(\theta))x_t + \delta_t(x_t, -K(\theta)x_t) \\
&= Mx_t + \underbrace{\big(B(\bar{K} - K(\theta))x_t + \delta_t(x_t, -K(\theta)x_t)\big)}_{:=r_t(x_t)}.
\end{aligned} \tag{65}$$

Using Assumption H.3, we know that if $x_t \in \mathcal{X}$,

$$\|r_t(x_t)\| \le \bar{\delta} + \big(\sigma\|B\| + L_{\delta,x} + L_{\delta,u}(\|\bar{K}\| + \sigma)\big)\|x_t\|. \tag{66}$$

Therefore, for any $0 \le \tau < t$, if $x_\tau, \ldots, x_{t-1} \in \mathcal{X}$, we have

$$\|x_t\| \le \|M^{t-\tau}x_\tau\| + \left\|\sum_{j=\tau}^{t-1} M^{t-1-j}r_j(x_j)\right\| \tag{67a}$$

$$\le C_0\rho_0^{t-\tau}\left(\|x_\tau\| + \sum_{j=\tau}^{t-1} \rho_0^{\tau-1-j}\|r_j(x_j)\|\right) \tag{67b}$$

$$\le C_0\rho_0^{t-\tau}\left(\|x_\tau\| + \big(\sigma\|B\| + L_{\delta,x} + L_{\delta,u}(\|\bar{K}\| + \sigma)\big)\sum_{j=\tau}^{t-1} \rho_0^{\tau-1-j}\|x_j\|\right) + \frac{C_0\bar{\delta}}{1-\rho_0}, \tag{67c}$$

where we use (65) and the triangle inequality in (67a); we use the property of matrix M in (67b); we use (66) in (67c).

Let $\rho := \rho_0 + C_0\big(\sigma\|B\| + L_{\delta,x} + L_{\delta,u}(\|\bar{K}\| + \sigma)\big)$ and $\beta := \frac{C_0\bar{\delta}}{1-\rho}$. Now we use induction to show that if $\|x_\tau\| \le \frac{R_x - \beta}{C_0}$, then for every $t \ge \tau$, $x_t \in \mathcal{X}$ and $\|x_t\| \le C_0\rho^{t-\tau}\|x_\tau\| + \beta$.

Note that the statement holds for τ. Suppose it holds for $\tau, \ldots, t-1$, then for t, by (67), we see that

$$\|x_t\| \le C_0\rho_0^{t-\tau}\|x_\tau\| + (\rho - \rho_0)\rho_0^{t-\tau}\sum_{j=\tau}^{t-1} \rho_0^{\tau-1-j}\|x_j\| + \frac{C_0\bar{\delta}}{1-\rho_0} \tag{68a}$$

$$\le C_0\rho_0^{t-\tau}\|x_\tau\| + (\rho - \rho_0)\rho_0^{t-\tau}\sum_{j=\tau}^{t-1} \rho_0^{\tau-1-j}\big(C_0\rho^{j-\tau}\|x_\tau\| + \beta\big) + \frac{C_0\bar{\delta}}{1-\rho_0} \tag{68b}$$

46

$$\leq C_0 \rho_0^{t-\tau} \|x_\tau\| + C_0(\rho^{t-\tau} - \rho_0^{t-\tau})\|x_\tau\| + \frac{(\rho - \rho_0)\beta}{1 - \rho_0} + \frac{C_0\bar{\delta}}{1 - \rho_0} \tag{68c}$$

$$= C_0 \rho^{t-\tau} \|x_\tau\| + \beta,$$

where we use (67) in (68a); we use the induction assumption in (68b) and rearrange the terms in (68c). Therefore, x_t is also in $\mathcal{X}$, and we have shown that $x_t \in \mathcal{X}$ and $\|x_t\| \leq C_0\rho^{t-\tau}\|x_\tau\| + \beta$ hold for all t by inductions. This finishes the proof for stability.

Now, we show that contractive perturbation also holds. Suppose $x'_\tau, \ldots, x'_t$ is another trajectory that starts from a different state x'_τ, which also satisfies $\|x'_\tau\| \leq \frac{R_x - \beta}{C_0}$. Note that this trajectory evolves according to $x'_{j+1} = Mx'_j + r_j(x'_j), \forall j \geq \tau$. By the Lipschitzness of the nonlinear residual δ_j, we see that
$$\|r_t(x_t) - r_t(x'_t)\| \leq (\sigma\|B\| + L_{\delta,x} + L_{\delta,u}(\|K\| + \sigma))\|x_t - x'_t\|.$$

Therefore, we obtain that

$$\|x_t - x'_t\| \leq \|M^{t-\tau}(x_\tau - x'_\tau)\| + \left\|\sum_{j=\tau}^{t-1} M^{t-1-j}\big(r_j(x_j) - r_j(x'_j)\big)\right\|$$

$$\leq C_0\rho_0^{t-\tau}\left(\|x_\tau - x'_\tau\| + \sum_{j=\tau}^{t-1} \rho_0^{\tau-1-j}\|r_j(x_j) - r_j(x'_j)\|\right)$$

$$\leq C_0\rho_0^{t-\tau}\left(\|x_\tau - x'_\tau\| + (\sigma\|B\| + L_{\delta,x} + L_{\delta,u}(\|K\| + \sigma))\sum_{j=\tau}^{t-1} \rho_0^{\tau-1-j}\|x_j - x'_j\|\right).$$

Using the same technique as (68), we can show that
$$\|x_t - x'_t\| \leq C_0\rho^{t-\tau}\|x_\tau - x'_\tau\|$$

holds if $\|x_\tau\| \leq \frac{R_x - \beta}{C_0}$ and $\|x'_\tau\| \leq \frac{R_x - \beta}{C_0}$. $\qquad\qquad\square$

I Numerical experiment details

In this appendix we provide details on the numerical experiments of Section 4. We also present a third experiment that compares GAPS to our bandit-based algorithm (BAPS, Algorithm 3) in a setting where the policy class for GAPS is a strict superset of the policy class for BAPS. Finally, we present an experiment demonstrating the improved computational efficiency of GAPS relative to prior work and the ideal OGD Definition 3.1.

I.1 Adapting a scalar confidence parameter for MPC

This section provides details on the baseline algorithm [11] to which GAPS was compared in Section 4. The main result was shown in Figure 2.

The method of Li et al. [11] tunes a scalar confidence parameter for model-predictive control in a LTI system with imperfect disturbance predictions. The setting is similar to Example 2.1 but has several restrictions: 1) the dynamics matrices A, B and cost matrices Q, R are time-invariant, 2) the disturbance predictions $\{\widehat{w}_t\}_{t=0}^{T-1}$ are static and known before the game begins, and 3) the policy is parameterized by a single confidence parameter $\lambda \in [0, 1]$ instead of separate confidence parameters $(\lambda^{[0]}, \ldots, \lambda^{[k-1]}) \in [0, 1]^k$ for each prediction in the planning horizon.

Li et al. [11] propose a "follow-the-leader"-style algorithm where the entire trajectory history is used to select λ at each time step. Their method commits the $u_{t|t}$ entry of the optimal solution

$$\underset{u_{t:t+k-1|t}}{\arg\min} \quad \sum_{\tau=t}^{t+k-1} \big(q\big(x_{\tau|t}, Q\big) + q\big(u_{\tau|t}, R\big)\big) + q\big(x_{t+k|t}, P\big)$$

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

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

47

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

$$\lambda_t := \frac{\sum_{s=0}^{t-1} (\eta(w; s, t-1))^\top H(\eta(\widehat{w}; s, t-1))}{\sum_{s=0}^{t-1} (\eta(\widehat{w}; s, t-1))^\top H(\eta(\widehat{w}; s, t-1))}, \quad \text{where} \quad \eta(w; s, t) := \sum_{\tau=s}^{t} \left(F^\top\right)^{\tau-s} P w_\tau,$$

where $H = B(R + B^\top P B)^{-1} B^\top$ and

$$F = A - BK^\star = A - B(R + B^\top P B)^{-1} B^\top P A$$

is the closed-loop linear dynamics matrix induced by the LQR-optimal infinite-horizon linear policy.

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 $\widehat{w}_t$ following

$$w_t = \sin(2\pi f t + p), \qquad \widehat{w}_t = w_t + n_t, \qquad n_t \sim \begin{cases} \text{Uniform}([-2, 2]) & : t \leq T/4 \\ \text{Uniform}([-0.02, 0.02]) & : \text{otherwise,} \end{cases}$$

where $f > 0$ and $p \in [0, 2\pi)$ are frequency and phase constants. In other words, the predictions are initially noisy but become accurate after $t = T/4$. We set $T = 400$ and perform 100 random trials. In each trial, the frequency f is sampled from a log-uniform distribution on $[0.01, 0.1]$ and the phase p is sampled uniformly from $[0, 2\pi)$. Results and discussion are found in Section 4.

I.2 Adapting a linear controller for a nonlinear time-varying system

This section provides details on the experimental setting and baseline controller for the second experiment in Section 4, in which GAPS is compared to a LQR baseline for tuning feedback control gains in the nonlinear inverted pendulum. The main result was shown in Figure 3.

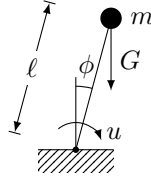



Figure 4: The inverted pendulum system.

We consider the inverted pendulum illustrated in Figure 4. The system is a point mass on a massless rod with a stationary and frictionless pivot. The input is a torque about the pivot. The continuous-time nonlinear dynamics are given by

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

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

$$\frac{d}{dt} x \approx \begin{pmatrix} 0 & 1 \\ G/\ell & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1/m\ell^2 \end{pmatrix} u. \tag{69}$$

A linear feedback policy 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) policy. We apply GAPS to tune the parameters $\theta = (k_p, k_d)$. We simulate time-varying pendulum mass m_t with regularly-spaced step changes. We measure policy performance by the LQR cost

$$f_t(x_t, u_t) = \Delta_t \cdot (\phi_t^2 + \dot{\phi}_t^2 + 0.1u_t^2), \tag{70}$$

where Δ_t is the discretization interval. We compare GAPS to a natural baseline that deploys the policy $u_t = K^\star(m_t)x_t$, where $K^\star$ is the optimal PD policy for the infinite-horizon summation of the LQR cost (70) when the pendulum dynamics for mass m_t are linearized as in (69) and discretized with

48

a zero-order hold. 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 a Ornstein-Uhlenbeck-type random walk with state $s_t \in \mathbb{R}$ and dynamics $s_{t+1} = \gamma s_t + \delta_t$, where $\gamma \in [0, 1)$ and δ_t is Gaussian-distributed with zero mean and variance σ^2. When $\gamma = 0$, the process becomes i.i.d. uniform noises. The disturbance enters as a perturbation of the velocity, i.e. the system is driven by the true input $u_t + m\ell^2 s_t$. We test two scenarios: the i.i.d. case ($\gamma = 0, \sigma = 8$) and the random walk ($\gamma = 0.95, \sigma = 0.5$). These values lead to similar closed-loop state magnitudes under optimal LQR control. We deploy GAPS with buffer size $B = 400$ and learning rate $\eta = 0.03$.

This setting is intended to match the spirit of the nonlinear system in Example 2.2 and the local regret bound for nonconvex surrogate cost in Theorem 3.6. A simple randomized search yields two controller parameters that certify the nonconvexity of the surrogate cost F_1. However, we relax some requirements: 1) we have unbounded (but concentrated) noise, 2) we do not parameterize the controller to allow Θ to be a full Euclidean space, and 3) we do not restrict the controller set or noise magnitude to ensure that the conditions of Lemma H.2 are satisfied. Nonetheless, we observe strong performance from GAPS.

I.3 Selecting the horizon in MPC: discrete and continuous approaches

In this section, we compare GAPS to the bandit-based algorithm (BAPS, Algorithm 3) in a setting where both finite and infinite policy classes are well motivated. We consider MPC with disturbance predictions that are accurate in the near future but inaccurate in the far future. The inaccurate far-future predictions mean it is counterproductive to plan over a long horizon. One natural solution is to optimize the MPC planning horizon, which yields a small finite policy class. On the other hand, we can also apply MPC with confidence coefficients (Example 2.1) and optimize the confidence coefficients using GAPS.

The continuous policy class is a strict superset of the finite class. To see this, first let k denote the maximum MPC horizon we wish to consider. Let $\mathsf{MPC}_j(\theta)$ denote the MPC algorithm (1) with horizon $j \leq k$ and confidences $\theta \in [0, 1]^j$. When the MPC terminal cost matrix $\tilde{Q}$ is the solution of the discrete-time algebraic Riccati equation for the associated infinite-horizon LQR problem, the Bellman optimality condition and the convexity of the MPC optimization problem imply that selecting from $\{\mathsf{MPC}_0(\mathbf{1}), \ldots, \mathsf{MPC}_k(\mathbf{1})\}$ is equivalent to selecting from the parameter set

$$\Theta_{\mathrm{dis}} = \{[0 \ \cdots \ 0], [1 \ 0 \ \cdots \ 0], [1 \ 1 \ 0 \ \cdots \ 0], \ldots, [1 \ \cdots \ 1]\}$$

for MPC_k. For the continuous parameter set $\Theta_{\mathrm{cts}} = [0, 1]^k$ we have $\Theta_{\mathrm{dis}} \subsetneq \Theta_{\mathrm{cts}}$, Therefore, in the long term GAPS should outperform BAPS, but it is not clear which will adapt faster.

We consider a two-dimensional discretized double integrator system subject to disturbances in both velocity and position, with quadratic cost. The continuous-time system is discretized with forward Euler integration over the interval Δ_t, resulting in the dynamics and cost matrices

$$A = \begin{pmatrix} I_2 & \Delta_t I_2 \\ \mathbf{0} & I_2 \end{pmatrix}, \qquad B = \begin{pmatrix} \mathbf{0} \\ \Delta_t I_2 \end{pmatrix}, \qquad Q = s_n \begin{pmatrix} I_2 & \mathbf{0} \\ \mathbf{0} & s_Q I_2 \end{pmatrix}, \qquad R = s_n s_R I_2,$$

with positive cost scale factors s_Q, s_R, s_n. The cost scale factors s_Q, s_R balance the costs of position error, velocity error, and control effort. The overall scale factor s_n is set to ensure compatibility with the cost bound assumptions of Algorithm 3. Note that the optimal policy is invariant to changes in s_n.

The true disturbances $\{w_t\}_{t=0}^{H-1}$ are sampled i.i.d. from a uniform distribution on $[-\Delta_t, \Delta_t]^4$. 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

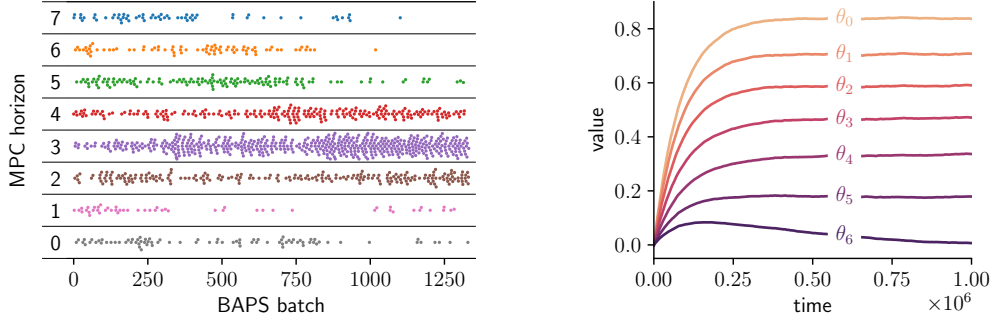
$$w_{t+\tau|t} := w_{t+\tau} + s_\epsilon \Delta_t \sum_{i=t}^{t+\tau} \epsilon_i,$$

where s_ϵ is a scale factor and $\{\epsilon_t\}_{t=0}^{H-1}$ are sampled i.i.d. from a uniform distribution on $[-1, 1]^4$.

In this experiment, we select the discretization interval $\Delta_t = 0.1$ and the scale factors $s_Q = 0.1$, $s_R = 0.01$, and $s_\epsilon = 0.25$. These choices ensure that the optimal cost is strongly affected by the MPC

49

horizon (or confidence weights). We use the GAPS/BAPS learning rates and batch sizes suggested by our regret bounds.

Our regret bounds are realized by particular choices of learning rate and EXP3 batch size that depend on the decay constants (C, ρ) induced by the MPC policies. The conservative bounds on (C, ρ) yielded by Lemma H.1 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 in this setting, and use the constants of the optimal linear policy $K^\star$. We compute ρ empirically from the spectrum of $A - BK^\star$ and let $C = \max\{\rho^{-n}\|(A - BK)^n\|\}_{n=0}^{N}$ for sufficiently large N.

(a) BAPS: Each dot represents the MPC horizon selected for a single batch.

(b) GAPS: Evolution of prediction confidence parameters θ_i.

Figure 5: Application of GAPS and BAPS to model-predictive control with noisy disturbance predictions. GAPS tunes per-step confidence weights $\theta_i \in [0, 1]$, while BAPS selects between different horizon lengths. Both methods learn to disregard far-future predictions.

Figure 5a visualizes the behavior of BAPS on the policies $\{\mathsf{MPC}_0(\mathbf{1}), \ldots, \mathsf{MPC}_k(\mathbf{1})\}$. A dot at the coordinate (β, j) indicates that BAPS selected $\mathsf{MPC}_j(\mathbf{1})$ during the βth batch. The selections eventually concentrate on the optimal $\mathsf{MPC}_3(\mathbf{1})$, with some continued exploration of the nearly-optimal $\mathsf{MPC}_2(\mathbf{1})$ and $\mathsf{MPC}_4(\mathbf{1})$.

For GAPS, the policy is MPC_k with initial trust parameters $\theta_0 = \mathbf{0}$. Figure 5b shows the evolution of the trust parameter θ_t under GAPS. It quickly converges to values that are (roughly) inversely proportional to the prediction errors.

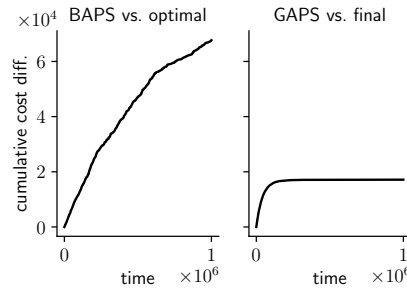



Figure 6: True regret of BAPS; "Regret" of GAPS compared to the final θ_{T-1} GAPS selects.

Due to the gap between Θ_{cts} and Θ_{dis}, we cannot evaluate the regret of each algorithm with respect to the same optimum. In this example, we have $J(\theta_{T-1}) = 0.89 \min_{\theta \in \Theta_{\text{dis}}} J(\theta)$, where θ_{T-1} is the final parameter value selected by GAPS. Therefore, we plot separate regret curves in Figure 6. Thanks to the LTI dynamics, convex costs, and time-invariant noise properties, we have $\theta_{T-1} \approx \min_{\theta \in \Theta_{\text{cts}}} J(\theta)$. Therefore, we use θ_{T-1} as the regret baseline for GAPS. 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.

50

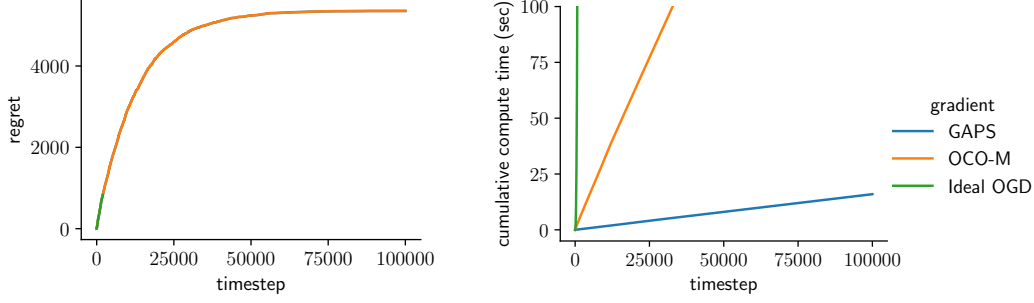



Figure 7: Regret (left) and computation time (right) comparison between: 1) Our algorithm (*GAPS*), 2) The method of [1] (*OCO-M*), and 3) the ideal online gradient descent on the surrogate cost (*Ideal OGD*, Definition 3.1). The algorithms are applied to tune MPC confidence coefficients in a two-dimensional double-integrator system, as in Appendix I.3. Ideal OGD was terminated after 2000 time steps due to intractable (linearly growing per timestep) computational cost. The regret curves of GAPS and OCO-M are indistinguishable ($< 0.02\%$ difference at completion). Both have a constant computational cost per timestep, but GAPS is over 15 times faster.

I.4 Comparison with Ideal OGD and alternative truncation

In this experiment we demonstrate the computational efficiency of GAPS. We compare GAPS to the Ideal OGD update (Definition 3.1) as well as the gradient approximation of [1].

Recall from the discussion in Section 3 that the method of [1] approximates the Ideal OGD update by counterfactually resimulating the policy and dynamics from an "imaginary" zero state starting B steps in the past. In our notation, this is equivalent to replacing G_t in line 1 of Algorithm 1 by

$$G'_t = \sum_{b=0}^{B-1} \frac{\partial f_{t|t-B}}{\partial \theta_{t-b}} \bigg|_{\mathbf{0}, (\theta_t) \times (B+1)},$$

while, recall for comparison, GAPS's gradient approximation is

$$G_t = \sum_{b=0}^{B-1} \frac{\partial f_{t|0}}{\partial \theta_{t-b}} \bigg|_{x_0, \theta_{0:t}}.$$

Note that we have written both approximations ignoring the initial $t < B$ cases for simplicity. As discussed in Section 3, computing G'_t requires evaluating the policy B times, where $B = \Omega(\log T)$, whereas computing G_t requires only one policy evaluation and B matrix multiplications.

To illustrate the impact of this difference, we compare each algorithm in the setting of MPC with confidence coefficients for a 2D double integrator from Appendix I.3. In this setting, computing G'_t at every time step requires solving B MPC optimization problems, while computing G_t in GAPS only requires solving one MPC optimization problem. We implement all algorithms in Python using NumPy and the explicit solution of the MPC optimization problem as discussed in Appendix H.2.

Comparisons of regret and computation time are shown in Figure 7. We see that the Ideal OGD computation time grows quadratically, and we must terminate it early to keep the experiment runtime reasonable. Ideal OGD is not tractable in practice. GAPS and the method of [1] both use constant time per step, but GAPS's constant is smaller. On the regret plots, the three methods are indistinguishable. The final regret of GAPS and the method of [1] differ by less than 0.02%, while the computation time of GAPS is over 15 times faster.

51