Online Policy Optimization in Unknown Nonlinear Systems

Shengnan Liu SHENGNAN_LIU176@UCSB.EDU
University of California, Santa Barbara

James A. Preiss PREISS@UCSB.EDU
University of California, Santa Barbara

Abstract

1. Introduction

Online policy optimization concerns controllers that learn while operating a system. Decisions are made sequentially along a single, non-episodic trajectory: at each step the agent observes the current state and outcome of past actions, selects an action using a parameterized policy, and immediately incurs a stage cost before moving on. There is no reset or repeated episode in which to re-try alternatives; improvements must be squeezed out of the one trajectory the system actually follows. This viewpoint is attractive in applications such as robotic manipulation, aerial vehicles, and process control, where data arrive during operation and the dynamics and objectives can change over time.

Two themes make this problem challenging in practice. First, the mechanism that maps actions to future outcomes may be partially known, time-varying, or outright unknown, which complicates choosing a reliable update direction. Second, algorithms must be both general and lightweight: they should handle nonlinear, time-varying systems and flexible policy classes, yet run in real-time without large replay buffers or repeated resimulation of counterfactual trajectories. Much of the early literature addressed linear or time-invariant settings; later work relaxed either the modeling assumptions or the generality of policies, but rarely both at once.

To navigate these constraints, the GAPS family adopts a surrogate-objective lens. Instead of evaluating a policy parameter by hypothetically rewinding and re-executing the entire past, the surrogate judges how a single parameter would have performed if it had been used from the start, and a gradient-based update is computed from information gathered along the realized trajectory. The original GAPS algorithm implements this idea with a short, logarithmically growing memory that captures the most influential recent sensitivities. Under a contractive perturbation view of the closed loop, this yields sublinear local regret guarantees. It also avoids heavy resimulation. M-GAPS refines the mechanism with an internal sensitivity recursion, collapsing the history into a constant-size memory and keeping per-step computation constant, which makes the approach suitable for embedded and realtime deployments.

In prior work on unknown dynamics, M-GAPS has been embedded in a meta-framework that couples an online policy optimizer (ALG) with an online estimator (EST). There, theoretical guarantees track how prediction errors in values and state derivatives translate into bounded perturbations of the policy update. M-GAPS standalone removes the estimator layer and studies the unperturbed regime in which the update has direct access to the model components it needs. This isolates the algorithmic phenomena underlying stability and regret, independent of identification quality, and provides a clean baseline against which estimation strategies can later be reintroduced.

Design principles. Two structural ingredients recur in our analysis and design. The first is a contraction- stability requirement for the joint evolution of the system and the algorithm's internal memory under slowly varying parameters: small parameter changes should induce state deviations that shrink over time, ensuring boundedness and rapid forgetting of past perturbations. The second is a slow- variation condition on the parameter path, achieved by standard stepsize choices on a smooth surrogate objective; this keeps the update stable while allowing adaptation to nonstationarity.

Contributions. This project contributes a clarified and tutorialized standalone presentation of M-GAPS, with emphasis on exposition rather than introducing new algorithmic or theoretical results:

- **Standalone regime formalization and correspondence.** We isolate the "standalone" case as a clean baseline for studying the policy-update mechanism without estimation, residual terms, exploration perturbations, or exogenous disturbances, via restating the online, non-episodic single-trajectory setting in terms of known components (g_t, π_t, h_t), and we make the correspondence to the original M-GAPS formulation explicit (what is removed/kept, and why).
- **Motivating worked example (early in the paper).** We develop a concrete nonlinear online control example (e.g., obstacle-avoiding tracking with saturating actuators) that explicitly instantiates the key ingredients (g_t, π_t, h_t), illustrates why online adaptation of policy parameters is useful in practice, and makes the operational role of the surrogate objective and the sensitivity recursion transparent.
- **Tutorial rewrite with proof scaffolding.** We reorganize the standalone analysis into a tutorial format with additional explanations and intermediate derivations (e.g., explicit chain-rule decompositions and term-by-term bounds), expanded/annotated proof details for key lemmas and theorems, consolidated notation and constant tables, and a dependency flowchart that visually connects all the formal statements of the paper, making the logical dependencies and proof pathways between statements immediately clear.

Relation to prior work. Our standpoint connects three threads. From online optimization, we inherit surrogate- based progress measures and projected gradient updates tailored to nonconvex, time- varying objectives. From learning- based control and adaptive control, we adopt the insight that stabilization and performance can be achieved by fitting along the realized trajectory rather than identifying a globally accurate model. From recent online control algorithms, we retain the focus on scalability and implementability, replacing history truncation with an internal recursion that preserves constant memory. In the companion unknown- dynamics setting, the same update can be combined with on- trajectory regression; here we deliberately abstract away estimation to expose the core algorithmic mechanisms.

Paper roadmap. We begin by detailing the standalone setting and assumptions, then present the standalone M-GAPS update and its computational structure. We follow with a comparison between the standalone M-GAPS and the original M-GAPS, and conclude with stability and regret analyses under contraction and slow variation.

2. Problem Setting

We first present the M-GAPS standalone setting. The main focus is on the policy update of θ from M-GAPS but with no perturbation, nonlinear residual, and disturbance.

Thus, in a discrete-time dynamical system, our simplified dynamics become

$$x_{t+1} = g_t(x_t, u_t), \tag{1}$$

where $x_t \in \mathbb{R}^n$ denotes the system state, $u_t \in \mathbb{R}^m$ denotes the control input, and g_t is the dynamical function.

To control this system, the online agent adopts a time-varying control policy π_t that is parameterized by a policy parameter $\theta_t \in \Theta \subseteq \mathbb{R}^d$. Specifically, the online agent picks the control input from the policy class

$$u_t = \pi_t(x_t, \theta_t). \tag{2}$$

The objective of the online agent is to minimize the total cost

$$\sum_{t=0}^{T-1} c_t \tag{3}$$

incurred over a finite horizon, where the stage cost at time step t is given by

$$c_t = h_t(x_t, u_t, \theta_t)^{1} \tag{4}$$

As a minor technical detail, we require that $h_{T-1} \geq 0$ in the proof of Theorem 17 at (22c) in Appendix I when upper bounding expressions.

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

Example 1 (Obstacle-avoiding tracking with saturating actuators) *Consider the problem setting with known dynamics g_t and a time–varying policy $\pi_t(\cdot, \theta_t)$. Let $x_t \in \mathbb{R}^2$ be the state and $u_t \in \mathbb{R}^2$ the control. The dynamics include smooth actuator saturation and a known exogenous drift (disturbance) $d_t \in \mathbb{R}^2$:*

$$x_{t+1} = g_t(x_t, u_t) = x_t + \Delta\, \sigma(u_t) \;+\; d_t, \qquad \sigma(z) = \tanh(z) \text{ (applied elementwise)}, \;\; \Delta > 0.$$

At time t we are given a reference $r_t \in \mathbb{R}^2$ and the center of a moving obstacle $o_t \in \mathbb{R}^2$. Define the repulsive potential/proximity penalty and its gradient

$$\phi_t(x) = \frac{1}{\|x - o_t\|^2 + \varepsilon}, \qquad p_t(x) = \nabla \phi_t(x) = -\frac{2(x - o_t)}{(\|x - o_t\|^2 + \varepsilon)^2}, \quad \varepsilon > 0.$$

The online agent chooses inputs from the policy class

$$u_t = \pi_t(x_t, \theta_t) = \alpha_t\,(r_t - x_t) + \beta_t\, p_t(x_t), \qquad \theta_t = (\alpha_t, \beta_t) \in \Theta := [-\kappa, \kappa]^2,$$

where $\kappa > 0$ is a fixed design bound so Θ is convex and compact.

The stage cost used in the problem setting ($c_t = h_t(x_t, u_t, \theta_t)$) balances tracking, effort, and safety:

$$c_t = h_t(x_t, u_t, \theta_t) = \|x_t - r_t\|_2^2 \;+\; \lambda_u \|u_t\|_2^2 \;+\; \lambda_o\, \phi_t(x_t).$$

This models a ground robot (or planar UAV) that must track a moving waypoint r_t while keeping distance from a pedestrian/vehicle at o_t; the tanh *nonlinearity captures actuator limits, the term d_t captures a known drift (disturbance)(e.g., wind or conveyor motion), and the gains $\theta_t = (\alpha_t, \beta_t)$ adapt online to trade off tracking and obstacle clearance.*

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

2.1. Performance Metrics

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

Definition 1 (Surrogate Cost) *The surrogate cost function is*

$$F_t(\theta) \coloneqq h_t(\tilde{x}_t(\theta), \tilde{u}_t(\theta), \theta),$$

where $\tilde{x}_t(\theta)$ and $\tilde{u}_t(\theta)$ are the system state and control input at time t if the agent is at the state recursively defined as

$$\tilde{x}_\tau(\theta) \coloneqq g_\tau(\tilde{x}_{\tau-1}(\theta), \tilde{u}_{\tau-1}(\theta))$$

and applies the control inputs

$$\tilde{u}_\tau(\theta) \coloneqq \pi_\tau(\tilde{x}_\tau(\theta), \theta)$$

at all previous time steps $\tau = 0, 1, \ldots, t$. [Ryan: Shouldn't $\tilde{x}_\tau(\theta)$ also be defined here?

$$\tilde{x}_\tau(\theta) \coloneqq g_\tau(\tilde{x}_{\tau-1}(\theta), \tilde{u}_{\tau-1}(\theta))$$

]

[Shengnan: Yes, I agree. Added that!] Intuitively, the surrogate cost $F_t(\theta)$ evaluates how good a policy parameter θ is at an intermediate time step t by eliminating the influence of any past policy parameters $\theta_{0:t-1}$. This concept is useful for defining different regret metrics. For example, the *static regret* is a widely-used metric in the literature of online policy optimization (Cesa-Bianchi and Lugosi (2006); Hazan and Seshadhri (2009)) that compares the total cost of an online agent with the best static policy parameter in hindsight can be written as

$$R^S(T) \coloneqq \sum_{t=0}^{T-1} c_t - \min_{\theta \in \Theta} \sum_{t=0}^{T-1} F_t(\theta).$$

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

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

Here, the projected gradient $\nabla_{\eta,\Theta} F_t(\theta_t)$ (parameterized by η) is a surrogate of the original gradient $\nabla F_t(\theta_t)$ that also considers the constraint set Θ. Specifically, an update step with the projected gradient is equivalent to projecting the output of the original gradient descent step back onto Θ, i.e.

$$\text{for any } \theta \in \Theta, \quad \theta - \eta \nabla_{\eta,\Theta} F_t(\theta) = \Pi_\Theta(\theta - \eta \nabla F_t(\theta)),$$

where Π_Θ is the Euclidean projection to Θ. This notion of local regret is first introduced by Hazan et al. (2017), and we provide a formal definition in Definition 6 in Appendix B. Intuitively, the local regret measures how well the policy parameter sequence $\theta_{0:T-1}$ tracks the changing stationary points of the surrogate cost functions $F_{0:T-1}$ (when $\Theta = \mathbb{R}^d$). In the time-invariant setting, sublinear local regret implies convergence to a stationary point.

Although local regret is useful for measuring the performance of an online policy optimization algorithm under nonconvex surrogate costs, a limitation of applying it alone to our setting with time-varying dynamics is that the surrogate cost F_t is defined in terms of ALG's behavior under known true dynamics.

To address this limitation, in addition to bounding the local regret of the policy parameters $\theta_{0:T-1}$, we also bound the distance between the actual trajectory of the online agent and the trajectory it would achieve with the same policy parameters $\theta_{0:T-1}$ under the known dynamics.

3. Main Results

This section introduces a concrete, simplified online policy optimization algorithm, Memoryless Gradient-based Adaptive Policy Selection (M-GAPS, Algorithm 1).

We model the controller together with the plant as the following joint recursion:

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \\ \theta_{t+1} \end{pmatrix} = q_t(x_t, y_t, \theta_t) = \begin{pmatrix} q_t^x(x_t, y_t, \theta_t) \\ q_t^y(x_t, y_t, \theta_t) \\ q_t^\theta(x_t, y_t, \theta_t) \end{pmatrix}, \quad x_t \in \mathbb{R}^n,\ y_t \in \mathbb{R}^p,\ \theta_t \in \Theta \subset \mathbb{R}^d. \tag{5}$$

Here, $y_t \in \mathbb{R}^p$ is an auxiliary state that M-GAPS can use to store something besides the system state x_t and the policy parameter θ_t to help it perform the update. For example, y_t can be a finite memory buffer that stores information from the past. It can also be the integral of past tracking error in an integral controller. To understand (5) intuitively, it is helpful to draw connections with the process of using a gradient-based optimizer to update θ_t iteratively where the exact gradients are available. In more practical scenarios though, the optimizer can only use biased gradient estimations, which still perform well in general.

To make explicit how θ_t affects both the next state and the stage cost, we collect the closed-loop maps:

$$u_t \coloneqq \pi_t(x_t, \theta_t), \tag{6a}$$

$$g_{t+1|t}(x_t, \theta_t) \coloneqq g_t(x_t,\ u_t) = g_t(x_t, \pi_t(x_t, \theta_t)), \tag{6b}$$

$$h_{t|t}(x_t, \theta_t) \coloneqq h_t(x_t,\ u_t,\ \theta_t) = h_t(x_t, \pi_t(x_t, \theta_t), \theta_t). \tag{6c}$$

These operational steps (applying u_t, plant evolution to x_{t+1}, and incurring c_t) are shown inline in Algorithm 1. We then analyze the properties of the induced joint recursion (x_t, y_t, θ_t).

The standalone M-GAPS is modified from the M-GAPS algorithm but only with simpler dynamics and settings. Thus, in the setting where the online agent has exact knowledge of the time-varying dynamical models, simplified M-GAPS can achieve better regret guarantees than M-GAPS, and the same memory/computational complexities $O(1)$. This significantly improves computational efficiency as the Gradient-based Adaptive Policy Selection (GAPS) algorithm (Lin et al., 2023), which M-GAPS takes inspiration from, has memory/computational complexities of $O(\log T)$. To understand why this improvement is possible, note that, on the one hand, the M-GAPS algorithm considers the perturbation/biases which contributes additively to the local regret bound; on the other hand,

Algorithm 1: Memoryless Gradient-based Adaptive Policy Selection

Parameters: Learning rate η, initial parameter θ_0.
Initialize: Policy parameter θ_0; Internal state $y_0 = \mathbf{0}_{n\times d}$.
for $t = 0, 1, \ldots, T-1$ **do**
  Take inputs x_t, θ_t, g_t, h_t, and π_t.
  Set $u_t \leftarrow \pi_t(x_t, \theta_t)$; $x_{t+1} \leftarrow g_t(x_t, u_t)$; $c_t \leftarrow h_t(x_t, u_t, \theta_t)$. */* Controller action and plant evolution (context, not part of gradient calc) */*
  Use these to obtain $g_{t+1|t}$ and $h_{t|t}$. */* $g_{t+1|t}$ and $h_{t|t}$ are defined in (??). */*
  Update $y_{t+1} \leftarrow \left.\frac{\partial g_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}$. */* Update partial derivatives accumulator. */*
  Let $G_t \leftarrow \left.\frac{\partial h_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}$.
  Update and output $\theta_{t+1} \leftarrow \prod_\Theta(\theta_t - \eta G_t)$. */* Π_Θ is the Euclidean projection to Θ. */*
end

to design an efficient estimation of $\nabla F_t(\theta_t)$, GAPS does two steps of approximations: first replacing the imaginary trajectory with the actual trajectory and then doing a bounded-memory truncation, but the (standalone) M-GAPS only keeps the first approximation step of GAPS by introducing the auxiliary internal state y_t that accumulates past partial derivatives. This greatly simplifies the computation. A more detailed comparison between standalone M-GAPS and GAPS can be found in Appendix J.

We state two important properties of the joint dynamics induced by M-GAPS. The first property is about contraction stability of x_t and y_t under slow time-varying constraint.

Property 2 *[Contraction Stability] For any sequence $\theta_{0:T-1}$ that satisfies the slowly time-varying constraint $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$, the partial dynamical system*

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t), \qquad y_{t+1} = q_t^y(x_t, y_t, \theta_t) \tag{7}$$

satisfying that $\|x_t\| \le R_x^* < R_x$ and $\|y_t\| \le R_y^* < R_y$ always hold if the system starts from $(x_\tau, y_\tau) = (0, 0)$. Further, there exists a function $\gamma : \mathbb{Z}_{\ge 0} \to \mathbb{R}_{\ge 0}$ that satisfies $\sum_{t=0}^\infty \gamma(t) \le \Gamma$, such that from any initial states $(x_\tau, y_\tau), (x'_\tau, y'_\tau)$ where $\|x_\tau\|, \|x'_\tau\| \le R_x$ and $\|y_\tau\|, \|y'_\tau\| \le R_y$ [Ryan: Should these be $\|x_\tau\|, \|x'_\tau\| \le R_x^*$ and $\|y_\tau\|, \|y'_\tau\| \le R_y^*$] [Shengnan: I don't think those should be R_x^*, R_y^*. In Property 2 / Lemma 15, R_x^*, R_y^* are reachable bounds for the special trajectory initialized at $(0,0)$, while R_x, R_y in the "from any initial states" clause specify the working region where the contraction/Lipschitz assumptions and the proof hold for arbitrary initializations. In particular, the proof needs a uniform bound on $\|y_\tau\|$ to control terms like $\left|\frac{\partial g}{\partial x}(\cdot) - \frac{\partial g}{\partial x}(\cdot)\right| \cdot \|y_\tau\|$ when bounding $\|y_t - y'_t\|$, which cannot be guaranteed by R_y^* unless we restrict to the $(0,0)$-initialized trajectory. Since $R_x^* < R_x$ and $R_y^* < R_y$, that special trajectory is automatically contained in the R_x, R_y domain anyway.], the trajectory satisfies $\left\|(x_{\tau+t}, y_{\tau+t}) - (x'_{\tau+t}, y'_{\tau+t})\right\| \le \gamma(t) \cdot \|(x_\tau, y_\tau) - (x'_\tau, y'_\tau)\|$.

Note that Property 2 is different with the contraction assumption of Lin et al. (2023) because it also considers the internal state y_t of ALG besides the system state x_t. The requirement that $\sum_{t=0}^\infty \gamma(t) \le \Gamma$ is also weaker than the exponential decay rate in Lin et al. (2023).

The second property we need is the robustness of the update rule that M-GAPS uses to update the policy parameter θ_t. In the unperturbed setting (no additive disturbances on the update), the statement specializes as follows.

Property 3 *[Regret Bound] Consider the joint dynamics in (5). Under Assumptions 13 and 14, the resulting $\{\theta_t\}$ satisfies the slowly-time-varying constraint $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ for all time t. Further, when $\eta \le \bar{\eta}$ for some positive constant $\bar{\eta}$, M-GAPS with (5) can achieve a local regret guarantee $R_\eta^L(T) = O\Big(\frac{1}{\eta}(1 + V_{sys}) + \eta T + \eta^3 T\Big)$, where the variation intensity is defined as*

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

Remark. If $\|\nabla F_t(\theta)\| \le G$ on Θ, then $\|\theta_{t+1} - \theta_t\| \le \eta G$, so one can take $\epsilon_\theta = \eta G$.

To understand Property 3, think about online gradient descent (OGD) in online optimization without state or dynamics. With unbiased gradients, each step is just a small move of size η, so the parameter path is naturally slowly varying (i.e., $\|\theta_{t+1} - \theta_t\|$ stays small). And because no disturbance is injected, the regret reduces to the standard OGD guarantee—precisely the $R(T, 0)$ special case of Theorem 17.

Taken together, these two properties align with the OGD comparison above: small per-step moves keep the parameter path slowly varying, and the baseline rate $R_\eta^L(T)$ emerges under a carefully balanced stepsize within the admissible range $\eta \le \bar{\eta}$. We record this as the following lemma.

Lemma 4 *Under Assumptions 13 and 14, M-GAPS (Algorithm 1) satisfy Properties 2, and 3 when applied to dynamics (5) and policy class (6a).*

We present the specific constants and the formal proof of Lemma 4 in Appendix F. Note that for the standalone M-GAPS and the nonconvex local regret result in Lin et al. (2023), we have $\Theta = \mathbb{R}^d$. Since the projected gradients are identical with the original gradients when $\Theta = \mathbb{R}^d$, the local regret bound $R_\eta^L(T)$ of standalone M-GAPS given by Lemma 4 matches the local regret bound of GAPS in Lin et al. (2023).

4. Comparison with the M-GAPS

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

Standalone correspondence. The standalone variant corresponds to the special case of the original formulation in which

- the exact a_t^* are always known;

- the nonlinear residual is absent (equivalently, $f_t(x_t, a_t^*) \equiv 0$), so the known model is $g_t(x_t, u_t)$ and no residual estimation is required;
- no exploration/algorithmic perturbation is used (i.e., $\zeta_t \equiv 0$);
- exogenous disturbances are ignored (i.e., w_t is set to zero for analysis);

Under these conditions, the working dynamics are reduced to $x_{t+1} = g_t(x_t, u_t)$ with policy $u_t = \pi_t(x_t, \theta_t)$.

5. Conclusion

Conclusion. We studied online policy optimization in a simplified "standalone" regime where the controller has direct access to the maps it needs and no estimator or exogenous disturbance is present. In this setting we formalized the M-GAPS update with constant memory and per-step computation, and analyzed the induced joint recursion of (x_t, y_t, θ_t). Under standard smoothness and a time-varying contraction condition, the update is stable, the parameter path remains slowly varying with an appropriate stepsize, and the algorithm achieves sublinear local regret with an explicit dependence on system variation. These guarantees mirror those known for GAPS Lin et al. (2023) while removing the estimation layer, thereby isolating which effects are intrinsic to the policy update mechanism.

Outlook. Manifold–aware M-GAPS: A natural next step is to place the policy parameter on a Riemannian manifold (Θ, g) and replace the Euclidean update by a manifold step.

Acknowledgements

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

References

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

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

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

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

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

Table 1: Important notations in this paper.

Notation	Meaning
$t_1 : t_2$	The integer sequence $\{t_1, \dots, t_2\}$;
$a_{t_1:t_2}$	A sequence of variables $\{a_t\}_{t=t_1,\dots,t_2}$;
$\|\cdot\|$	ℓ_2 (Euclidean) norm;
$\|\cdot\|_F$	Frobenius norm;
$\|\cdot\|_P$	Norm induced by matrix P;
$\mathbb{Z}_{\geq 0}$	The set of non-negative integers;
$\mathbb{R}_{\geq 0}$	The set of non-negative reals;
$\sigma(z_{1:t}, z'_{1:t})$	Product sigma-algebra generated by sequences $z_{1:t}$ and $z'_{1:t}$;
x_t	$x_t \in \mathbb{R}^n$ is the system state;
u_t	$u_t \in \mathbb{R}^m$ is the control input;
R_η^L	Local regret with learning rate η;
$q_t(x_t, y_t, \theta_t)$	The joint dynamics of the system at time t;
$g_{t\|\tau}$	Multi-step dynamics between two time steps $\tau \leq t$;
$h_{t\|\tau}$	Multi-step cost between two time steps $\tau \leq t$;
$\Pi_\Theta(y)$	Euclidean projection of y to set Θ;
$\nabla_{\Theta,\eta} F(\theta)$	The (Θ, η)-projected gradient of F;
$S_{\epsilon_\theta}(t_1 : t_2)$	The set of policy parameter sequences with ϵ_θ-constrained step size;
$B_n(\cdot, \cdot)$	Closed Euclidean ball in $\mathbb{R}^n$ (center, radius);

Outline of the appendices.

- In Appendix B, we provide a notation table and list important definitions used in the proofs.
- In Appendix C, we present the details about the application of matched-disturbance dynamics.
- In Appendix F, we show M-GAPS satisfies the properties for `ALG` in the application of matched-disturbance dynamics.
- In Appendix I, we show online gradient descent with inexact updates can achieve local regret bounds in online nonconvex optimization, which is used in the proof of M-GAPS.
- In Appendix J, we make a detailed comparison between our M-GAPS algorithm with the GAPS algorithm proposed by Lin et al. (2023). We summarize some results from Lin et al. (2023) that are useful for us to analyze M-GAPS.

Appendix A. Simulation Results

Appendix B. Notations and Definitions

We provide a notation table (Table 1) and a constant table (Table 2) that summarize the important notations in this paper.

A key concept in this paper is the ideal trajectory obtained by rolling out a given policy–parameter sequence $\theta_{0:T-1}$ on the true dynamics. To formalize this, we introduce the multi-step dynamics/cost,

Table 2: Important constants in this paper.

Notation	Meaning
Γ	The Contraction Stability Constant Γ associated with γ, defined in Property 2;
$L_{g,x}$	State-wise Lipschitz Constant of Dynamics, defined in Assumption 13;
$L_{g,u}$	Input-wise Lipschitz Constant of Dynamics, defined in Assumption 13;
$\ell_{g,x}$	State-Jacobian Smoothness of Dynamics, defined in Assumption 13;
$\ell_{g,u}$	Input-Jacobian Smoothness of dynamics, defined in Assumption 13;
$L_{\pi,x}$	State-wise Lipschitz Constant of Policy, defined in Assumption 13;
$L_{\pi,\theta}$	Parameter-wise Lipschitz Constant of Policy, defined in Assumption 13;
$\ell_{\pi,x}$	State-Jacobian Smoothness of Policy, defined in Assumption 13;
$\ell_{\pi,\theta}$	Parameter-Jacobian Smoothness of Policy, defined in Assumption 13;
$L_{h,x}$	Cost Lipschitz Constant in x, defined in Assumption 13;
$L_{h,u}$	Cost Lipschitz Constant in u, defined in Assumption 13;
$L_{h,\theta}$	Cost Lipschitz Constant in θ, defined in Assumption 13;
$\ell_{h,x}$	Gradient Smoothness Constant of cost in x, defined in Assumption 13;
$\ell_{h,u}$	Gradient Smoothness Constant of cost in u, defined in Assumption 13;
$\ell_{h,\theta}$	Gradient Smoothness Constant of cost in θ, defined in Assumption 13;
R_C	Contraction Region Radius, defined in Assumption 14;
R_S	Uniform Stability Bound, defined in Assumption 14;
S_0	Baseline Variation Constant, defined in Lemma 16;
S_1	Linear Perturbation Gain, defined in Lemma 16;
S_2	Quadratic Perturbation Gain, defined in Lemma 16;
$\bar{C} = C_{L,g,x}$	Multi-Step State-Sensitivity Lipschitz Constant, defined in Lemma 19;
$C_{L,g,\theta}$	Multi-Step Parameter-Sensitivity Lipschitz Constant, defined in Lemma 19;
$C_{\ell,g,(x,x)}$	Self-Smoothness of the State-Jacobian in x, defined in Lemma 19;
$C_{\ell,g,(x,\theta)}$	Cross-Smoothness of the State-Jacobian w.r.t. Parameter Change, defined in Lemma 19;
$C_{\ell,g,(\theta,x)}$	Cross-Smoothness of the Parameter-Jacobian w.r.t. State Change, defined in Lemma 19;
$C_{\ell,g,(\theta,\theta)}$	Self-Smoothness of the Parameter-Jacobian in θ, defined in Lemma 19;
$C_{L,h,x}$	Multi-Step Cost Gradient Lipschitz in State, defined in Corollary 20;
$C_{L,h,\theta}$	Multi-Step Cost Gradient Lipschitz in Parameter, defined in Corollary 20;
$C_{\ell,h,(x,x)}$	Self-smoothness of the Cost State-Gradient in x, defined in Corollary 20;
$C_{\ell,h,(x,\theta)}$	Cross-Smoothness of the Cost State-Gradient w.r.t. Parameter Change, defined in Corollary 20;
$C_{\ell,h,(\theta,x)}$	Cross-Smoothness of the Cost Parameter-Gradient w.r.t. State Change, defined in Corollary 20;
$C_{\ell,h,(\theta,\theta)}$	Self-smoothness of the Cost Parameter-Gradient in θ, defined in Corollary 20;
$\hat{C}_0$	Constant η-Bias Coefficient, defined in Theorem 22;
$\hat{C}_1$	Linear η-Bias Coefficient, defined in Theorem 22;
$\hat{C}_2$	Quadratic η-Bias Coefficient, defined in Theorem 22;

which characterize how the system evolves under $\theta_{0:T-1}$ without any estimation module. This mirrors the construction in Lin et al. (2023) (which studies online policy selection with known dynamical systems) and differs only in notation. Concretely, throughout this work we use the multi-step mappings in Definition 5 to define the state and cost at time t resulting from applying the control inputs $u_\tau^*(\theta) := \pi_\tau(\tilde{x}_\tau^*(\theta), \theta)$ recursively from $\tau = 0$ to t on the true system. No estimated quantities or "actual" trajectories are needed in our setting; the ideal rollout under the true dynamics is the sole reference path used in both analysis and evaluation.

Definition 5 (Multi-Step Dynamics and Cost) *The multi-step dynamics $g_{t|\tau}$ between two time steps $\tau \le t$ specifies the state x_t as a function of the previous state x_τ and previous policy parameters $\theta_{\tau:t-1}$. It is defined recursively, with the base case $g_{\tau|\tau}(x_\tau) := x_\tau$ and the recursive case*

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

in which $z_t := g_{t|\tau}(x_\tau, \theta_{\tau:t-1})$.[2]

The multi-step cost $h_{t|\tau}$ specifies the cost c_t as function of x_τ and $\theta_{\tau:t}$. It is defined as

$$h_{t|\tau}(x_\tau, \theta_{\tau:t}) := h_t(z_t, \pi_t(z_t, \theta_t), \theta_t).$$

It is worth emphasizing that, in our work, the concepts of multi-step dynamics/cost are only used for the theoretical analysis, because their definitions involve true model parameters that are unknown to the online agent. [Shengnan: Could you please go over the explanation part as you said in the meeting you have some ideas on how to formulate this one]

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

Definition 6 (Projected gradient) *Let $F : \Theta \to \mathbb{R}$ be a differentiable function on a closed convex set $\Theta \subseteq \mathbb{R}^d$. For $\eta > 0$, the (Θ, η)-projected gradient of F is defined as*

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

This concept of projected gradient is used to define the local regret in Section 2.1, and Appendix I that studies online gradient descent for online nonconvex optimization with constraints.

Projected gradient and stationarity. Let $\Theta \subset \mathbb{R}^d$ be closed and convex, $\eta > 0$. Define the projected gradient

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

Then the projected step can be written equivalently as

$$\theta^+ \ = \ \theta - \eta\, \nabla_{\Theta,\eta} F(\theta) \ = \ \Pi_\Theta(\theta - \eta \nabla F(\theta)).$$

By the metric projection optimality condition, $y - \Pi_\Theta(y) \in N_\Theta(\Pi_\Theta(y))$, where

$$N_\Theta(p) \ := \ \{\, w : \ \langle w,\, v - p \rangle \le 0 \ \ \forall v \in \Theta \,\}$$

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

is the normal cone. Taking $y = \theta - \eta \nabla F(\theta)$ yields the KKT-style characterization

$$\nabla_{\Theta,\eta} F(\theta) = 0 \iff \theta = \Pi_\Theta(\theta - \eta \nabla F(\theta)) \iff 0 \in \nabla F(\theta) + N_\Theta(\theta).$$

Two immediate facts: (i) if $\Theta = \mathbb{R}^d$, then $\nabla_{\Theta,\eta} F(\theta) = \nabla F(\theta)$; (ii) zeros of the projected gradient coincide with first–order stationarity under the constraint Θ.

Appendix C. Dynamics for Application

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

Definition 7 *We denote the set of policy parameter sequences with ϵ_θ-constrained step size by*

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

Definition 8 (ϵ_θ-Time-varying Contractive Perturbation) *For $\epsilon_\theta \ge 0$, we say the ϵ_θ-time-varying contractive perturbation property holds for $R_C > 0, C \ge 1$, and $\rho \in (0,1)$ if, for any parameter sequence $\theta_{\tau:t-1} \in S_{\epsilon_\theta}(\tau : t-1)$, the following inequality holds for arbitrary $x_\tau, x'_\tau \in B_n(0, R_C)$ and time steps $\tau \le t$:*

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

Definition 9 (ϵ_θ-Time-varying Stability) *For $\epsilon_\theta \ge 0$, the ϵ_θ-time-varying stability property holds for $R_S > 0$ if, for any $\theta_{\tau:t-1} \in S_{\epsilon_\theta}(\tau : t-1)$, $\left\| g_{t|\tau}(0, \theta_{\tau:t-1}) \right\| \le R_S$ holds for any $t \ge \tau$.*

Definition 10 (Block Lipschitzness) *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ and let $z = (z^{(1)}, \ldots, z^{(k)})$ with $z^{(i)} \in \mathbb{R}^{d_i}$, $\sum_i d_i = d$. A (possibly vector–valued) map $\phi : \mathcal{Z} \to \mathbb{R}^p$ is $(L_1, \ldots, L_k)$–Lipschitz on $\mathcal{Z}$ if there exist nonnegative constants $L_1, \ldots, L_k$ such that, for all $z, z' \in \mathcal{Z}$,*

$$\left\| \phi(z) - \phi(z') \right\| \;\le\; \sum_{i=1}^{k} L_i \left\| z^{(i)} - z'^{(i)} \right\|.$$

When $k = 1$, this reduces to the usual L–Lipschitz condition. The constants L_i may depend on $\mathcal{Z}$.

Definition 11 (Block smoothness (Lipschitz Jacobian / gradient)) *With the same block decomposition, a differentiable map $\phi : \mathcal{Z} \to \mathbb{R}^p$ is $(\ell_1, \ldots, \ell_k)$–smooth on $\mathcal{Z}$ if its Jacobian $J\phi(z)$ exists for all $z \in \mathcal{Z}$ and is block–Lipschitz: for all $z, z' \in \mathcal{Z}$,*

$$\left\| J\phi(z) - J\phi(z') \right\| \;\le\; \sum_{i=1}^{k} \ell_i \left\| z^{(i)} - z'^{(i)} \right\|.$$

In the scalar case $f : \mathcal{Z} \to \mathbb{R}$, this is equivalent to the gradient being block–Lipschitz:

$$\left\| \nabla f(z) - \nabla f(z') \right\| \;\le\; \sum_{i=1}^{k} \ell_i \left\| z^{(i)} - z'^{(i)} \right\| \quad \text{for all } z, z' \in \mathcal{Z}.$$

Remark (block vs. single constant). In finite dimensions, the block and single-constant definitions are equivalent up to a fixed multiplicative factor; we keep the block form to track per-block contributions for tighter, cleaner sensitivity/chain-rule bounds.

With the definitions of time-varying contractive perturbation and time-varying stability, we state our key assumptions below: Assumption 12 is about regulating our domain and the choice of radius.

Assumption 12 (Auxiliary radii and compact domains) *Like Lin et al. (2023), assume there exists a real number $\bar{R}_C > 0$ such that*

$$R_C \;>\; \bar{R}_C \;>\; R_S + \bar{C}\,\|x_0\|.$$

When the ϵ_θ-time-varying contractive perturbation (Definition 8) holds globally (so that $R_C = \infty$), we still fix any $\bar{R}_C$ satisfying the strict inequality above. Similarly, to leverage the Lipschitzness/smoothness property, we require $\mathcal{X} \supseteq B_n(0, R_x)$ where $R_x \geq \bar{C}\bar{R}_C + R_S$ and $\mathcal{U} = \{\pi(x,\theta) \mid x \in \mathcal{X}, \theta \in \Theta, \pi \in \mathcal{G}\}$.

But to pin down the compact domains on which the Lipschitz/smooth constants in Assumption 13 are evaluated, we choose

$$\mathcal{X} \;:=\; B_n(0, R_x), \qquad R_x \;:=\; \bar{C}\,\bar{R}_C + R_S,$$

and define the control-value set

$$\mathcal{U} \;:=\; \{\, \pi_t(x,\theta) \mid x \in \mathcal{X},\ \theta \in \Theta,\ (g,\pi) \in \mathcal{G},\ t \in \mathcal{T} \,\}.$$

as the coefficients in Assumption 13 depend on $\mathcal{X}$ and $\mathcal{U}$. We also fix

$$\mathcal{Y} \;:=\; B_p(0, R_y), \qquad R_y \;:=\; \frac{\bar{C}\,L_{g,u}\,L_{\pi,\theta}}{\rho(1-\rho)},$$

so that the internal state y_t remains in $\mathcal{Y}$ along the trajectories considered.

This assumption is purely notational/technical: it selects compact radii for evaluating Lipschitz/smoothness and implies the initial-state condition $\|x_0\| < (R_C - R_S)/\bar{C}$ in Assumption 13 and Assumption 14.

The second assumption 13 is about the Lipschitzness/smoothness properties of the dynamics, policy, and the cost functions.

Assumption 13 *[James: We should remind the reader the definitions of Lipschitz and Smooth.] The dynamics $g_{0:T-1}$, policies $\pi_{0:T-1}$, and costs $h_{0:T-1}$ are differentiable at every time step and satisfy that, for any convex compact sets $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{U} \subseteq \mathbb{R}^m$, one can find Lipschitzness/smoothness constants (can depend on $\mathcal{X}$ and $\mathcal{U}$) such that:*

1. *$g_t(x,u)$ is $(L_{g,x}, L_{g,u})$-Lipschitz and $(\ell_{g,x}, \ell_{g,u})$-smooth in (x,u) on $\mathcal{X} \times \mathcal{U}$.*
2. *$\pi_t(x,\theta)$ 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. *$h_t(x,u,\theta)$ is $(L_{h,x}, L_{h,u}, L_{h,\theta})$-Lipschitz and $(\ell_{h,x}, \ell_{h,u}, \ell_{h,\theta})$-smooth in (x,u,θ) on $\mathcal{X} \times \mathcal{U} \times \Theta$.*

The third assumption (Assumption 14) is on the contractive perturbation and the stability of the multi-step dynamics $g_{t|\tau}$.

Assumption 14 *Let $\mathcal{G}$ denote the set of all possible sequences $\{g_t, \pi_t\}_{t\in\mathcal{T}}$ the environment may provide. For a fixed $\epsilon_\theta \in \mathbb{R}_{\geq 0}$, the ϵ_θ-time-varying contractive perturbation holds with $(R_C, \bar{C}, \rho)$ for any sequence in $\mathcal{G}$. The ϵ_θ-time-varying stability holds with $R_S < R_C$ for any sequence in $\mathcal{G}$. We assume that the initial state satisfies $\|x_0\| < (R_C - R_S)/\bar{C}$. Further, we assume that if $\{g, \pi\}$ is the dynamics/disturbance/policy at an intermediate time step of a sequence in $\mathcal{G}$, then the time-invariant sequence $\{g, \pi\}_{\times T}$ is also in $\mathcal{G}$.*[3]

[James: The paragraph below contains further assumptions, yes? Should be a labeled assumption (merge or create a new one).] [Shengnan: Could you please double check the paragraph below and Assumption 12 to see if everything aligns] Like Lin et al. (2023), in Assumption 14, we assume there exists a positive real number $\bar{R}_C$ such that $R_C > \bar{R}_C > R_S + \bar{C}\|x_0\|$. Here, we introduce the real constant $\bar{R}_C$ because R_C can be $+\infty$ when time-varying contractive perturbation (Definition 8) holds globally. Similarly, to leverage the Lipschitzness/smoothness property, we require $\mathcal{X} \supseteq B(0, R_x)$ where $R_x \geq \bar{C}\bar{R}_C + R_S$ and $\mathcal{U} = \{\pi(x, \theta) \mid x \in \mathcal{X}, \theta \in \Theta, \pi \in \mathcal{G}\}$. Since the coefficients in Assumption 13 depend on $\mathcal{X}$ and $\mathcal{U}$, we will set $\mathcal{X} = B_n(0, R_x)$ and $R_x = \bar{C}\bar{R}_C + R_S$ by default when presenting these constants. We also set $\mathcal{Y} = B_p(0, R_y)$ with $R_y = \frac{\bar{C}L_{g,u}L_{\pi,\theta}}{\rho(1-\rho)}$, so that the internal state y_t will stay in $\mathcal{Y}$.

Appendix D. Proof of Theorem ??

Appendix E. Proof of Theorem ??

Appendix F. Proof of Lemma 4

When applied to the dynamical system (5) and the policy class (6a), the joint dynamics induced by applying standalone M-GAPS are given by

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t) = g_t(x_t, \pi_t(x_t, \theta_t)), \tag{8a}$$

$$y_{t+1} = q_t^y(x_t, y_t, \theta_t) = \left.\frac{\partial g_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}, \tag{8b}$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t) = \Pi_\Theta\left(\theta_t - \eta\left(\left.\frac{\partial h_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right). \tag{8c}$$

Since Lemma 4 consists two properties, we show them separately in Lemmas 15-16.

Lemma 15 *Suppose the sequence $\theta_{0:T-1}$ is given and it satisfies the constraint that $\|\theta_t - \theta_{t-1}\| \leq \epsilon_\theta$ for all time step t. Consider the dynamical system*

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t) = g_t(x_t, \pi_t(x_t, \theta_t)),$$

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

3. 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. This assumption can be made without the loss of generality for time-invariant dynamics and policy classes.

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

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

where recall that ρ *is the decay factor defined in Assumption 14. Further, from any initial states* $(x_\tau, y_\tau), (x'_\tau, y'_\tau)$ *that satisfy* $\|x_\tau\|, \|x'_\tau\| \le R_x$ *and* $\|y_\tau\|, \|y'_\tau\| \le R_y$, *the trajectory satisfies*

$$\big\|(x_t, y_t) - (x'_t, y'_t)\big\| \le \gamma(t-\tau) \cdot \big\|(x_\tau, y_\tau) - (x'_\tau, y'_\tau)\big\|,$$

where

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

Note that γ *satisfies*

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

The definitions of the coefficients $C_{L,g,\theta}, C_{\ell,g,(x,x)}, C_{\ell,g,(\theta,x)}$ can be found in Lemma 19 in Appendix J. And the proof of Lemma 15 can be found in Appendix F.2.

Lemma 16 *Consider the dynamical system*

$$\begin{aligned}
x_{t+1} &= q_t^x(x_t, y_t, \theta_t) = g_t(x_t, \pi_t(x_t, \theta_t)), \\
y_{t+1} &= q_t^y(x_t, y_t, \theta_t) = \left.\frac{\partial g_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}, \\
\theta_{t+1} &= q_t^\theta(x_t, y_t, \theta_t) = \Pi_\Theta\left(\theta_t - \eta\left(\left.\frac{\partial h_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right).
\end{aligned} \tag{9}$$

Suppose the learning rate η *satisfies* $\eta < \min\left\{\frac{(1-\rho)\epsilon_\theta}{C_{L,h,\theta}}, \frac{1-\rho}{2C_{\ell,h,(\theta,\theta)}}\right\}$. *Then, the parameter increments obey* $\|\theta_t - \theta_{t-1}\| \le \frac{C_{L,h,\theta}}{1-\rho}\eta \le \epsilon_\theta$ *for all* t, *so* $\{\theta_t\}$ *is slowly-time–varying. Further, the trajectory* $\{\theta_t\}$ *achieves the local regret guarantee*

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

$$\begin{aligned}
S_0 &\coloneqq \frac{2\bar{C}L_h(1 + L_{\pi,x} + L_{f,x})(1 + L_{g,u})}{(1-\rho)^2\rho} \cdot V_{sys} \\
&\quad + \frac{2\bar{C}L_h(1 + L_{\pi,x} + L_{f,x})}{1-\rho} \cdot \big(2\bar{C}\bar{R}_C + 2R_S\big), \\
S_1 &\coloneqq \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta T, \\
S_2 &\coloneqq \left(1 + \frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2 + \hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_2 + \hat{C}_4}{(1-\rho)^3}\right).
\end{aligned}$$

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

Here, the variation intensity is defined as

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

The bound can be simplified to

$$R_\eta^L(T)=O\left(\frac{1}{\eta}(1+V_{sys})+\eta T+\eta^3 T\right),$$

where the $O(\cdot)$ notation hides dependence on $\frac{1}{1-\rho}, R_x, R_y, \bar{C}$, and the Lipchitzness/smoothness coefficients defined in Assumption 13.

The definition of the coefficient $C_{L,h,\theta}$ can be found in Corollary 20 in Appendix J. The proof of Lemma 16 can be found in Appendix F.3.

F.1. Stepsize requirements for slow variation and smoothness

Let the one–step surrogate be

$$F_t(\theta) \;=\; h_t(\tilde{x}_t(\theta),\ \pi_t(\tilde{x}_t(\theta),\theta),\ \theta),$$

[Ryan: Should be

$$F_t(\theta) \;=\; h_t(\tilde{x}_t(\theta),\ \pi_t(\tilde{x}_t(\theta),\theta),\ \theta),$$

?] [Shengnan: Yes, that's a typo from me. Modified that!] and denote the (exact) sensitivity of the state to the policy parameter by

$$y_t(\theta) \;:=\; \frac{\partial x_t(\theta)}{\partial\theta}.$$

(For M-GAPS in the exact case, the internal variable updated by the Jacobian recursion coincides with this sensitivity.)

F.1.1. Slow–variation requirement

The projected update is $\theta_{t+1}=\Pi_\Theta(\theta_t-\eta\nabla F_t(\theta_t))$. By nonexpansiveness of Euclidean projection,

$$\|\theta_{t+1}-\theta_t\| \;=\; \|\Pi_\Theta(\theta_t-\eta\nabla F_t(\theta_t))-\Pi_\Theta(\theta_t)\| \;\le\; \eta\,\|\nabla F_t(\theta_t)\|.$$

By the chain rule,

$$\nabla F_t(\theta) \;=\; \underbrace{\partial_\theta h_t(\cdot)}_{\text{direct }\theta\text{-channel}} \;+\; \underbrace{\partial_x h_t(\cdot)\,y_t(\theta)}_{\text{through state}} \;+\; \underbrace{\partial_u h_t(\cdot)\,\partial_\theta \pi_t(\cdot)}_{\text{through policy}}.$$

Property 2 (contraction) implies that perturbing θ affects the state with a gain bounded by the geometric series of decay factors, hence $\|y_t(\theta)\| \le \frac{\bar{C}_y}{1-\rho}$ for a time–uniform constant $\bar{C}_y$. Bounding $\|\partial_x h_t\|$, $\|\partial_u h_t\|$, $\|\partial_\theta h_t\|$ and $\|\partial_\theta \pi_t\|$ by their Lipschitz constants and folding $\bar{C}_y$ into a single coefficient gives

$$\|\nabla F_t(\theta)\| \;\le\; \frac{C_{L,h,\theta}}{1-\rho}.$$

Therefore

$$\|\theta_{t+1} - \theta_t\| \;\le\; \eta\, \frac{C_{L,h,\theta}}{1-\rho}.$$

To enforce the slowly–time–varying constraint $\|\theta_{t+1} - \theta_t\| \le \epsilon_\theta$, it suffices to choose

$$\eta \;<\; \frac{(1-\rho)\,\epsilon_\theta}{C_{L,h,\theta}}.$$

F.1.2. Smoothness requirement (stability of the gradient step)

We also need a stepsize that is small enough for the projected gradient analysis on a smooth objective. We seek a Lipschitz bound of the form

$$\big\|\nabla F_t(\theta) - \nabla F_t(\theta')\big\| \;\le\; \ell_F \,\big\|\theta - \theta'\big\|.$$

Split the difference using the chain rule and bound each term:

$$\begin{aligned}
\big\|\nabla F_t(\theta) - \nabla F_t(\theta')\big\| &= \big\|\partial_\theta h_t(\theta) - \partial_\theta h_t(\theta') + \partial_x h_t(\theta)\, y_t(\theta) - \partial_x h_t(\theta')\, y_t(\theta')\big\| \\
&\le \big\|\partial_\theta h_t(\theta) - \partial_\theta h_t(\theta')\big\| + \big\|\partial_x h_t(\theta)\, y_t(\theta) - \partial_x h_t(\theta')\, y_t(\theta')\big\| \\
&\le \underbrace{C_{\ell,h,(\theta,\theta)}\,\|\theta-\theta'\|}_{\text{direct }\theta\text{-channel}} + \underbrace{\big\|\partial_x h_t(\theta) - \partial_x h_t(\theta')\big\|}_{\le C_{\ell,h,(\theta,x)}\|\theta-\theta'\|}\;\underbrace{\|y_t(\theta)\|}_{\le C_y/(1-\rho)} \\
&\quad + \underbrace{\|\partial_x h_t(\theta')\|}_{\le L_{h,x}}\;\underbrace{\|y_t(\theta) - y_t(\theta')\|}_{\le \tilde{C}_y/(1-\rho)\,\|\theta-\theta'\|} \\
&\le \Big(C_{\ell,h,(\theta,\theta)} + \frac{C_{\ell,h,(\theta,x)}\,C_y + L_{h,x}\,\tilde{C}_y}{1-\rho}\Big)\,\|\theta-\theta'\|.
\end{aligned}$$

The sensitivity bounds

$$\|y_t(\theta)\| \;\le\; \frac{C_y}{1-\rho}, \qquad \|y_t(\theta) - y_t(\theta')\| \;\le\; \frac{\tilde{C}_y}{1-\rho}\,\|\theta-\theta'\|$$

[Ryan: Shouldnt this be

$$\|y_t(\theta)\| \;\le\; \frac{\tilde{C}_y}{1-\rho},$$

?] [Shengnan: I think it should be C_y: $\|y_t(\theta)\|$ is a magnitude bound from the geometric-series/Jacobian-sum argument when $y_\tau = 0$, while $\|y_t(\theta) - y_t(\theta')\|$ is a Lipschitz-in-θ bound that involves Jacobian differences and smoothness constants, so it gets a different coefficient $\tilde{C}_y$.] follow from the contraction property (Property 2): $\theta - \theta'$ changes the state by at most $\sum_\tau \gamma(\tau) \le \frac{1}{1-\rho}$, which directly

yields the same factor for the state sensitivity. Collecting terms and folding $C_{\ell,h,(\theta,x)}C_y + L_{h,x}\tilde{C}_y$ into the same symbol as the direct θ–channel for a convenient single coefficient, we obtain

$$\big\|\nabla F_t(\theta) - \nabla F_t(\theta')\big\| \;\le\; \left(C_{\ell,h,(\theta,\theta)} + \frac{C_{\ell,h,(\theta,\theta)}}{1-\rho}\right)\|\theta-\theta'\| \;\le\; \frac{2\,C_{\ell,h,(\theta,\theta)}}{1-\rho}\,\|\theta-\theta'\|.$$

Hence F_t is ℓ_F–smooth with $\ell_F \le \dfrac{2\,C_{\ell,h,(\theta,\theta)}}{1-\rho}$, as claimed. Then, for the projected update $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta\,\nabla F_t(\theta_t))$, the standard descent lemma on an ℓ_F-smooth objective gives

$$F_t(\theta_{t+1}) \;\le\; F_t(\theta_t) - \eta\,\big\langle \nabla F_t(\theta_t),\, \nabla_{\Theta,\eta} F_t(\theta_t)\big\rangle + \frac{\ell_F}{2}\,\eta^2\,\big\|\nabla_{\Theta,\eta} F_t(\theta_t)\big\|^2.$$

To keep this stable (nonexpansive up to the inner product term), it suffices to require $\eta \le 1/\ell_F$. Using the bound on ℓ_F above, a convenient sufficient condition is

$$\eta \;\le\; \frac{1-\rho}{2\,C_{\ell,h,(\theta,\theta)}}\,.$$

Combining the two requirements yields the final stepsize condition

$$\boxed{\eta \;<\; \min\left\{\frac{(1-\rho)\,\epsilon_\theta}{C_{L,h,\theta}}\;,\;\; \frac{1-\rho}{2\,C_{\ell,h,(\theta,\theta)}}\right\}.}$$

This choice simultaneously (i) guarantees the slowly time–varying update $\|\theta_{t+1}-\theta_t\| \le \epsilon_\theta$ and (ii) keeps the projected gradient step stable under the ℓ_F-smooth surrogate.

F.2. Proof of Lemma 15

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

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

and

$$\begin{aligned} x'_{\tau+1} &= g_\tau(x'_\tau, \pi_t(x'_\tau,\theta_\tau)),\\ y'_{\tau+1} &= \left.\frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\right|_{x'_\tau,\theta_\tau}\cdot y'_\tau + \left.\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_\tau}, \end{aligned}$$

where $\tau = t_1, t_1+1, \ldots, t_2$.

Note that by Assumption 14, we have that $\|x_{t_2}\| \le R_S$ and for any x_{t_1}, x'_{t_1} whose norms are upper bounded by R_C

$$\big\|x_{t_2} - x'_{t_2}\big\| \le \bar{C}\rho^{t_2-t_1}\big\|x_{t_1} - x'_{t_1}\big\|. \tag{10}$$

where ρ is the decay factor of the contractive perturbation property defined in Assumption 14. For the y sequence, note that y_{t_2} and y'_{t_2} can be expressed equivalently as

$$y_{t_2} = \left.\frac{\partial g_{t_2|t_1}}{\partial x_{t_1}}\right|_{x_{t_1},\theta_{t_1:t_2-1}} \cdot y_{t_1} + \sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}, \tag{11a}$$

$$y'_{t_2} = \left.\frac{\partial g_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}} \cdot y'_{t_1} + \sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g_{t_2|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_{\tau:t_2-1}}. \tag{11b}$$

By Lemma 19, we see that if $y_{t_1} = 0$, then

$$\|y_{t_2}\| = \left\|\sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}\right\| \le \sum_{\tau=t_1}^{t_2-1} \left\|\left.\frac{\partial g_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}\right\| \le \sum_{\tau=t_1}^{t_2-1} C_{L,g,\theta}\rho^{t_2-\tau} = \frac{C_{L,g,\theta}}{1-\rho}. \tag{12}$$

We also see that

$$\begin{aligned}
&\left\|y_{t_2} - y'_{t_2}\right\| \\
&= \left\|\left(\left.\frac{\partial g_{t_2|t_1}}{\partial x_{t_1}}\right|_{x_{t_1},\theta_{t_1:t_2-1}} - \left.\frac{\partial g_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}}\right) \cdot y_{t_1}\right\| + \left\|\left.\frac{\partial g_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}} \cdot (y_{t_1} - y'_{t_1})\right\| \\
&\quad + \sum_{\tau=t_1}^{t_2-1} \left\|\left.\frac{\partial g_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}} - \left.\frac{\partial g_{t_2|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_{\tau:t_2-1}}\right\| && \text{(13a)} \\
&\le C_{\ell,g,(x,x)}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| \cdot R_y + C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\| \\
&\quad + C_{\ell,g,(\theta,x)} \sum_{\tau=t_1}^{t_2-1} \rho^{t_2-\tau}\left\|x_\tau - x'_\tau\right\| && \text{(13b)} \\
&\le C_{\ell,g,(x,x)}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| \cdot R_y + C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\| \\
&\quad + C_{\ell,g,(\theta,x)} \sum_{\tau=t_1}^{t_2-1} \rho^{t_2-\tau} \cdot \bar{C}\rho^{\tau-t_1}\left\|x_{t_1} - x'_{t_1}\right\| && \text{(13c)} \\
&\le \left(C_{\ell,g,(x,x)}R_y + C_{\ell,g,(\theta,x)}\bar{C}(t_2-t_1)\right)\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| + C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\|. && \text{(13d)}
\end{aligned}$$

Therefore, we see that

$$\begin{aligned}
&\left\|(x_{t_2}, y_{t_2}) - (x'_{t_2}, y'_{t_2})\right\| \\
&\le \left\|x_{t_2} - x'_{t_2}\right\| + \left\|y_{t_2} - y'_{t_2}\right\| && \text{(14a)} \\
&\le \bar{C}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| + \left(C_{\ell,g,(x,x)}R_y + C_{\ell,g,(\theta,x)}\bar{C}(t_2-t_1)\right)\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\| \\
&\quad + \bar{C}\rho^{t_2-t_1}\left\|y_{t_1} - y'_{t_1}\right\| && \text{(14b)} \\
&\le \gamma(t_2-t_1)\left\|(x_{t_1}, y_{t_1}) - (x'_{t_1}, y'_{t_1})\right\|, && \text{(14c)}
\end{aligned}$$

where we use the triangle inequality in (14a); we use (10) and (13) and $\bar{C} = C_{L,g,x}$ in (14b); we use the inequality that

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

and the definition of $\gamma(\cdot)$ in (14c).

F.3. Proof of Lemma 16

We compare the dynamical system (9) with the Ideal OGD update rule:

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta \nabla F_t(\theta_t)). \tag{15}$$

Note that the update on θ_t that the dynamical system (9) performs can be written equivalently as

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta G_t), \tag{16}$$

where

$$G_t \coloneqq \sum_{\tau=0}^{t} \frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\bigg|_{x_0, \theta_{0:t}}. \tag{17}$$

By Theorem 22, we know that

$$\|G_t - \nabla F_t(\theta_t)\| \le \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta,$$

where the constants $\hat{C}_{0:5}$ are given in Theorem 22. Let θ_{t+1} be the actual next policy parameter (following the update rule (16)). By Lemma 18, we see that

$$\begin{aligned}\|\theta_{t+1} - \Pi_\Theta(\theta_t - \eta \nabla F_t(\theta_t))\| &\le \eta \|G_t - \nabla F_t(\theta_t)\| \\ &\le \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta^2.\end{aligned}$$

Then, we can apply Theorem 17 to obtain that

$$\sum_{t=0}^{T-1} \|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2 \le \frac{1}{\eta(1-\eta\ell_F)}\left(F_0(\theta_0) + \sum_{t=1}^{T-1} \|F_t - F_{t-1}\|_\infty\right) + \frac{L_F S_1 + \ell_F \eta S_2}{1-\eta\ell_F}, \tag{18}$$

where $\|\cdot\|_\infty$ is a metric that measures the distance between two surrogate cost functions (see Theorem 17 for definition), and S_1 and S_2 are given by

$$\begin{aligned}S_1 &\coloneqq (\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta T, \\ S_2 &\coloneqq \left(1 + \frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2 + \hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_2 + \hat{C}_4}{(1-\rho)^3}\right)\cdot \\ &\quad \left[\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta^2 T\right].\end{aligned}$$

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

$$\sum_{t=1}^{T-1} \|F_t - F_{t-1}\|_\infty \leq \frac{2\bar{C}L_h(1 + L_{\pi,x} + L_{f,x})(1 + L_{g,u})}{(1-\rho)^2\rho} \cdot V_{sys} + \frac{2\bar{C}L_h(1 + L_{\pi,x} + L_{f,x})}{1-\rho} \cdot \big(2\bar{C}\bar{R}_C + 2R_S\big).$$

Substituting the above inequality and $L_F = \frac{C_{L,f,\theta}}{1-\rho}, \ell_F = \frac{C_{\ell,h,(\theta,\theta)}}{1-\rho}$ into (18) finishes the proof.

Appendix G. Proof of Lemma ??

Appendix H. Proof of Theorem ??

Appendix I. Local Regret of Online Gradient Descent

[James: The following result controls the local regret of parameter sequences whose local steps have a bounded Euclidean error compared to online gradient descent.]

Theorem 17 *Consider a parameter sequence (θ_t) that satisfies*

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

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

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

where $\|F - F'\|_\infty \coloneqq \sup_{\theta\in\Theta}|F(\theta) - F'(\theta)|$.

Proof To begin, we assert the nonexpansiveness of Euclidean projection onto the compact convex set $\Theta \in \mathbb{R}^d$. This is a classic result in convex optimization (see, for example, Theorem 1.2.1 in Schneider (2014)):

Lemma 18 *Let q and q' be arbitrary points in $\mathbb{R}^d$. Let $p = \Pi_\Theta(q)$ and $p' = \Pi_\Theta(q')$. Then, the following inequality holds:*

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

Now we return to the proof of Theorem 17. We define the quantity

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

to name the deviation of the steps in our parameter sequence from those of ideal OGD:

$$\theta_{t+1} - \theta_t = -\eta\nabla_{\Theta,\eta}F_t(\theta_t) + \eta\epsilon_t. \tag{19}$$

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

$$\begin{aligned} F_t(\theta_{t+1}) &\le F_t(\theta_t) + \langle \nabla F_t(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\ell_F}{2}\|\theta_{t+1} - \theta_t\|^2 \\ &= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) - \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\|\nabla_{\Theta,\eta} F_t(\theta_t) - \epsilon_t\|^2 \qquad (20a) \\ &= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) \rangle + \frac{\ell_F \eta^2}{2}\|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2 \\ &\quad + \eta \langle \nabla F_t(\theta_t), \epsilon_t \rangle - \ell_F \eta^2 \langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\|\epsilon_t\|^2, \qquad (20b) \end{aligned}$$

where we use (19) in (20a). Recall that Θ is a closed convex subset of $\mathbb{R}^d$.

Since $\theta_t - \eta \nabla_{\Theta,\eta} F_t(\theta_t)$ is the projection of $\theta_t - \eta \nabla F_t(\theta_t)$ onto Θ and $\theta_t \in \Theta$, we have [James: Why?] [Shengnan: Could you please double check this also?]

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

Rearranging terms gives that

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

Let

$$y := \theta_t - \eta \nabla F_t(\theta_t), \qquad p := \Pi_\Theta(y) = \theta_t - \eta \nabla_{\Theta,\eta} F_t(\theta_t).$$

Note that $\theta_t - \eta \nabla_{\Theta,\eta} F_t(\theta_t)$ is the projection of $\theta_t - \eta \nabla F_t(\theta_t)$ onto Θ and $\theta_t \in \Theta$. By the projection variational inequality for metric projection onto a closed convex set,

$$\langle y - p,\ v - p \rangle \le 0 \qquad \forall\, v \in \Theta.$$

Taking $v = \theta_t \in \Theta$ gives

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

Expanding $y - p = -\eta(\nabla F_t(\theta_t) - \nabla_{\Theta,\eta} F_t(\theta_t))$ and $\theta_t - p = \eta\, \nabla_{\Theta,\eta} F_t(\theta_t)$ and dividing by $-\eta^2$ (with $\eta > 0$) yields

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

Substituting this inequality into (20) gives that

$$\begin{aligned} F_t(\theta_{t+1}) &\le F_t(\theta_t) - \eta \big\|\nabla_{\Theta,\eta} F_t(\theta_t)\big\|^2 + \frac{\ell_F \eta^2}{2}\big\|\nabla_{\Theta,\eta} F_t(\theta_t)\big\|^2 \\ &\quad + \eta \big\langle \nabla F_t(\theta_t), \epsilon_t \big\rangle - \ell_F \eta^2 \big\langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \big\rangle + \frac{\ell_F \eta^2}{2}\big\|\epsilon_t\big\|^2 \\ &\le\ F_t(\theta_t) - \eta \big\|\nabla_{\Theta,\eta} F_t(\theta_t)\big\|^2 + \frac{\ell_F \eta^2}{2}\big\|\nabla_{\Theta,\eta} F_t(\theta_t)\big\|^2 + \eta \big\|\nabla F_t(\theta_t)\big\|\,\big\|\epsilon_t\big\| \\ &\quad - \ell_F \eta^2 \big\langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \big\rangle + \frac{\ell_F \eta^2}{2}\big\|\epsilon_t\big\|^2 \qquad (21a) \\ &=\ F_t(\theta_t) - \eta \big\|\nabla_{\Theta,\eta} F_t(\theta_t)\big\|^2 - \frac{\ell_F \eta^2}{2}\big\|\nabla_{\Theta,\eta} F_t(\theta_t) + \epsilon_t\big\|^2 \end{aligned}$$

$$+ \ell_F \eta^2 \left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 + \ell_F \eta^2 \left\|\epsilon_t\right\|^2 + \eta \left\|\nabla F_t(\theta_t)\right\| \left\|\epsilon_t\right\| \tag{21b}$$

$$= F_t(\theta_t) - \eta(1 - \ell_F \eta) \left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 - \frac{\ell_F \eta^2}{2} \left\|\nabla_{\Theta,\eta} F_t(\theta_t) + \epsilon_t\right\|^2 + \ell_F \eta^2 \left\|\epsilon_t\right\|^2 + \eta \left\|\nabla F_t(\theta_t)\right\| \left\|\epsilon_t\right\| \tag{21c}$$

$$\leq F_t(\theta_t) - \eta(1 - \ell_F \eta) \left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 + \eta L_F \beta_t + \ell_F \eta^2 \beta_t^2, \tag{21d}$$

where in (21a) we apply Cauchy–Schwarz to the linear term $\eta \langle \nabla F_t(\theta_t), \epsilon_t \rangle \leq \eta \|\nabla F_t(\theta_t)\| \|\epsilon_t\|$, then in (21b) use the completion-of-squares identity to rewrite $\frac{\ell_F \eta^2}{2} \|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2$, $\frac{\ell_F \eta^2}{2} \|\epsilon_t\|^2$, and $-\ell_F \eta^2 \langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \rangle$ in (21a); (21c) factors out the common constant for $-\eta \|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2$ and $+\ell_F \eta^2 \|\nabla_{\Theta,\eta} F_t(\theta_t)\|^2$; lastly, (21d) uses the Lipschitz constant L_F of F_t, the assumption bound $\|\epsilon_t\| \leq \beta_t$, and drops the negative term $-\frac{\ell_F \eta^2}{2} \|\nabla_{\Theta,\eta} F_t(\theta_t) + \epsilon_t\|^2$.

[James: Expand the steps to reach (21c).]

Rearranging and summing (21) over $t = 0, 1, \ldots, T-1$ gives

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

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

$$= F_0(\theta_0) + \sum_{t=1}^{T-1} (F_t(\theta_t) - F_{t-1}(\theta_t)) - F_{T-1}(\theta_T) + \eta L_F \sum_{t=0}^{T-1} \beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1} \beta_t^2 \tag{22a}$$

$$\leq F_0(\theta_0) + \sum_{t=1}^{T-1} \|F_t - F_{t-1}\|_\infty - F_{T-1}(\theta_T) + \eta L_F \sum_{t=0}^{T-1} \beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1} \beta_t^2 \tag{22b}$$

$$\leq F_0(\theta_0) + \sum_{t=1}^{T-1} \|F_t - F_{t-1}\|_\infty + \eta L_F \sum_{t=0}^{T-1} \beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1} \beta_t^2. \tag{22c}$$

where in (22a) we write out $t = 0$ case for the first sum and leave the rest in the summation format; in (22b) we apply the definition of $\|\cdot\|_\infty$ to $\sum_{t=1}^{T-1} (F_t(\theta_t) - F_{t-1}(\theta_t))$; in (22c) we use $F_{T-1}(\theta_T) \geq 0$ and thus drop this negative term.[4] ■

Fact 1 (Projection variational inequality / normal-cone characterization) *Let $\Theta \subset \mathbb{R}^d$ be nonempty, closed, and convex. For any $y \in \mathbb{R}^d$, let $p = \Pi_\Theta(y)$ be the (Euclidean) metric projection. Then*

$$\langle y - p,\ v - p \rangle \leq 0 \qquad \forall\, v \in \Theta,$$

equivalently, $y - p \in N_\Theta(p)$ where $N_\Theta(p) = \{w : \langle w, v - p \rangle \leq 0,\ \forall v \in \Theta\}$.

4. This follows from our assumptions, which make the inputs of h_t uniformly bounded: $\theta \in \Theta$ with Θ compact, $\tilde{x}_t(\theta)$ remains in a compact ball by ϵ_θ–time-varying stability (hence $\tilde{x}_t(\theta) \in B_n(0, R_x)$), and $\tilde{u}_t(\theta) = \pi_t(\tilde{x}_t(\theta), \theta)$ belongs to a compact set $\mathcal{U}$ by continuity of π_t on a compact domain. Therefore $\mathcal{D}_t := \{(\tilde{x}_t(\theta), \tilde{u}_t(\theta), \theta) : \theta \in \Theta\}$ is compact. Since h_t is continuous, it attains a finite infimum $I_t := \inf_{(x,u,\theta) \in \mathcal{D}_t} h_t(x, u, \theta)$ on $\mathcal{D}_t$. Define $\bar{h}_t(x, u, \theta) := h_t(x, u, \theta) - I_t$. Then $\bar{h}_t \geq 0$ on $\mathcal{D}_t$, and replacing h_t by $\bar{h}_t$ leaves all regret expressions unchanged (constants cancel) and does not affect gradients or projected updates.

Proof Consider $\phi(v) = \frac{1}{2}\|v - y\|^2$. Since Θ is convex and ϕ is differentiable convex with $\nabla\phi(v) = v - y$, the first-order necessary optimality condition for the constrained minimizer $p = \arg\min_{v\in\Theta} \phi(v)$ gives

$$\langle \nabla\phi(p),\, v - p\rangle \;=\; \langle p - y,\, v - p\rangle \;\geq\; 0 \quad \forall\, v \in \Theta,$$

which is equivalent to $\langle y - p,\, v - p\rangle \leq 0$ for all $v \in \Theta$. ■

Appendix J. Useful Lemmas

In this section, we summarize some useful existing results in Lin et al. (2023) that can help us in the proof of M-GAPS (Algorithm 1). We can build our proof upon some results shown in Lin et al. (2023) because of the similarity between our M-GAPS algorithm and the GAPS algorithm proposed by Lin et al. (2023) when applied to known dynamical systems: Both algorithms are designed to efficiently approximate the gradient $\nabla F_t(\theta_t)$ of the surrogate cost. Note that $\nabla F_t(\theta_t)$ can be expressed as

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

M-GAPS adopts the approximation G_t that is given by

$$G_t = \sum_{\tau=0}^{t} \frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\bigg|_{x_0,\theta_{0:t}}, \tag{23}$$

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

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

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

In this section, we translate some results from Lin et al. (2023) into the settings of matched-disturbance dynamics application discussed in Section 2. Lemma 19 is Lemma D.3 in Lin et al. (2023). We changed the condition $x_\tau, x'_\tau \in B_n(0, R_S + C\|x_0\|)$ to $x_\tau, x'_\tau \in B_n(0, \bar{R}_C)$, where $\bar{R}_C$ can be any positive number that satisfies $\bar{R}_C < R_C$ and $C\bar{R}_C + R_S \leq R_x$. This minor change will not affect the proof provided in Lin et al. (2023).

Lemma 19 (Lipschitzness/Smoothness of the Multi-Step Dynamics) *Suppose Assumptions 13 and 14 hold. Given two time steps $t > \tau$, for any $x_\tau, x'_\tau \in B_n(0, \bar{R}_C)$ and $\theta_\tau, \theta'_\tau \in \Theta$, $\theta_{\tau+1:t-1} \in S_\varepsilon(\tau+1 : t-1)$, if $x'_{\tau+1} := g_{\tau+1|\tau}(x'_\tau, \theta'_\tau)$ is also in $B_n(0, \bar{R}_C)$, the multi-step dynamical function $g_{t|\tau}$ satisfies that*

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

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

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

where $C_{L,g,x} = \bar{C}$ *as defined in Assumption 12* $C_{L,g,\theta} = \dfrac{\bar{C}L_{g,u}L_{\pi,\theta}}{\rho}$, *and*

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

[James: Possible to factor out any common constant from the list of $C_{\ell,g,*}$ *above?] [Shengnan: could you please compare the two versions of* Γ*'s]*

$$A \coloneqq (1+L_{\pi,x})(\ell_{g,x}+\ell_{g,u}L_{\pi,x})+L_{g,x}\,\ell_{\pi,x}.$$

$$\begin{aligned}
C_{\ell,g,(x,x)} &= \Gamma\,\rho^{-1}\big[\,A\,\Gamma^2\,(1-\rho)^{-1}\big],\\
C_{\ell,g,(x,\theta)} &= \Gamma\,\rho^{-1}\big[\,A\,\Gamma^2 L_{g,u}L_{\pi,\theta}\,(1-\rho)^{-1}+((1+L_{\pi,x})\,\ell_{g,u}\,L_{\pi,\theta}+L_{g,u}\,\ell_{\pi,\theta})\,(1-\rho)^{-1}\big],\\
C_{\ell,g,(\theta,x)} &= \Gamma\,\rho^{-1}\big[\,A\,(L_{g,x}+L_{g,u}L_{\pi,x})\,\Gamma^2 L_{g,u}L_{\pi,\theta}\,\rho^{-1}(1-\rho)^{-1}+(L_{\pi,\theta}\,(\ell_{g,x}+\ell_{g,u}L_{\pi,x})+L_{g,u}\,\ell_{\pi,x})\big],\\
C_{\ell,g,(\theta,\theta)} &= \Gamma\,\rho^{-1}\big[\,A\,L_{g,u}^2L_{\pi,\theta}^2\,\Gamma^2\,\rho^{-1}(1-\rho)^{-1}+\big(L_{g,u}\,\ell_{\pi,\theta}+\ell_{g,u}\,L_{\pi,\theta}^2\big)\big].
\end{aligned}$$

[James: Copy the proof from the original GAPS paper and adapt as needed.] Intuitively, Lemma 19 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 8), we use this property and the chain rule decomposition to show the Lipschitzness on θ_τ and the smoothness.

[Shengnan: could you please check the proof for Lemma 19 below?]

Proof The first inequality in Lemma 19 directly follows from ϵ-time-varying contractive perturbation (Definition 8).

For the second inequality, when $t = \tau + 1$, note that

$$\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} = \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau}, \text{where } u_\tau = \pi_\tau(x_\tau,\theta_\tau).$$

Taking norms on both sides gives

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

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

$$\begin{aligned}\left\|\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}}\right\| &= \left\|\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\,\theta_{\tau+1:t-1}} \cdot \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\,\theta_\tau}\right\| \\ &\le \left\|\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}}\right\| \cdot \left\|\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau}\right\| \le \frac{C_0}{\rho}\, L_{g,u}L_{\pi,\theta}\,\rho^{\,t-\tau},\end{aligned}$$

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

For the third inequality, note that we have the chain rule decompositions

$$\begin{aligned}\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} &= \frac{\partial g_{t|t-1}}{\partial x_{t-1}}\Big|_{x_{t-1},\theta_{t-1}} \cdot \frac{\partial g_{t-1|t-2}}{\partial x_{t-2}}\Big|_{x_{t-2},\theta_{t-2}} \cdot \ldots \cdot \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau}, \\ \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} &= \frac{\partial g_{t|t-1}}{\partial x_{t-1}}\Big|_{x'_{t-1},\theta'_{t-1}} \cdot \frac{\partial g_{t-1|t-2}}{\partial x_{t-2}}\Big|_{x'_{t-2},\theta'_{t-2}} \cdot \ldots \cdot \frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau},\end{aligned} \tag{24}$$

where we use the notation

$$x_{\tau'} \coloneqq g_{\tau'|\tau}(x_\tau,\theta_{\tau:\tau'-1}) \quad \text{and} \quad x'_{\tau'} \coloneqq g_{\tau'|\tau}(x'_\tau,\theta'_\tau,\theta_{\tau+1:\tau'-1}) \quad \text{for } \tau' \in [\tau+1:t-1].$$

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

$$\begin{aligned}&\frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\theta_{t-i}} - \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\Big|_{x'_{t-i},\theta'_{t-i}} \\ &= \underbrace{\left(\frac{\partial g_{t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},u_{t-i}} - \frac{\partial g_{t-i}}{\partial x_{t-i}}\Big|_{x'_{t-i},u'_{t-i}}\right)}_{\text{change in } \partial g/\partial x \text{ (both } x \text{and } u \text{ vary)}} + \underbrace{\left(\frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x_{t-i},u_{t-i}} - \frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x_{t-i},u'_{t-i}}\right)\frac{\partial \pi_{t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\theta_{t-i}}}_{\text{change in } u \text{ at fixed } x_{t-i}} \\ &\quad + \underbrace{\left(\frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x_{t-i},u'_{t-i}} - \frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x'_{t-i},u'_{t-i}}\right)\frac{\partial \pi_{t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\theta_{t-i}}}_{\text{change in } x \text{ propagated through } u} + \underbrace{\frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x'_{t-i},u'_{t-i}}\left(\frac{\partial \pi_{t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\theta_{t-i}} - \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\Big|_{x'_{t-i},\theta'_{t-i}}\right)}_{\text{change in the policy Jacobian}} \\ &= \frac{\partial g_{t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\,u_{t-i}} - \frac{\partial g_{t-i}}{\partial x_{t-i}}\Big|_{x'_{t-i},\,u'_{t-i}} + \left(\frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x_{t-i},\,u_{t-i}} - \frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x'_{t-i},\,u'_{t-i}}\right)\frac{\partial \pi_{t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\,\theta_{t-i}} \\ &\quad + \frac{\partial g_{t-i}}{\partial u_{t-i}}\Big|_{x'_{t-i},\,u'_{t-i}}\left(\frac{\partial \pi_{t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\,\theta_{t-i}} - \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\Big|_{x'_{t-i},\,\theta'_{t-i}}\right).\end{aligned}$$

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

$$\begin{aligned}
&\left\| \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\theta_{t-i}} - \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\Big|_{x'_{t-i},\theta'_{t-i}} \right\| \\
&\le \ell_{g,x} \left\| x_{t-i} - x'_{t-i} \right\| + \ell_{g,u} \left\| \pi_{t-i}(x_{t-i}, \theta_{t-i}) - \pi_{t-i}(x'_{t-i}, \theta'_{t-i}) \right\| \\
&\quad + L_{\pi,x}\big(\ell_{g,x} \left\| x_{t-i} - x'_{t-i} \right\| + \ell_{g,u} \left\| \pi_{t-i}(x_{t-i}, \theta_{t-i}) - \pi_{t-i}(x'_{t-i}, \theta'_{t-i}) \right\|\big) \\
&\quad + L_{g,u}\big(\ell_{\pi,x} \left\| x_{t-i} - x'_{t-i} \right\| + \ell_{\pi,\theta} \left\| \theta_{t-i} - \theta'_{t-i} \right\|\big) \qquad (25) \\
&\le \underbrace{\big((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u} L_{\pi,x}) + L_{g,u}\, \ell_{\pi,x}\big)}_{\text{coeff. on } \left\| x_{t-i} - x'_{t-i} \right\|} \left\| x_{t-i} - x'_{t-i} \right\| \\
&\quad + \underbrace{\big((1 + L_{\pi,x})\, \ell_{g,u}\, L_{\pi,\theta} + L_{g,u}\, \ell_{\pi,\theta}\big)}_{\text{coeff. on} \left\| \theta_{t-i} - \theta'_{t-i} \right\|} \left\| \theta_{t-i} - \theta'_{t-i} \right\|, \qquad (26)
\end{aligned}$$

where we use Assumption 13 and definition of u_{t-i}, u'_{t-i} in (25), and Assumption 13 in (26). Therefore, by (24) and (26), we obtain

$$\begin{aligned}
&\left\| \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}} - \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta_{\tau:t-1}} \right\| \\
&\le \sum_{i=1}^{t-\tau-1} \left(\left\| \prod_{\tau'=1}^{i-1} \frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\Big|_{x'_{t-\tau'},\theta_{t-\tau'}} \right\| \cdot \left\| \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\Big|_{x_{t-i},\theta_{t-i}} - \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\Big|_{x'_{t-i},\theta_{t-i}} \right\| \right. \\
&\qquad \left. \cdot \left\| \prod_{\tau'=i+1}^{t-\tau} \frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\Big|_{x_{t-\tau'},\theta_{t-\tau'}} \right\| \right) \\
&\quad + \left\| \prod_{\tau'=1}^{t-\tau-1} \frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\Big|_{x'_{t-\tau'},\theta_{t-\tau'}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right\| \\
&\le \sum_{i=1}^{t-\tau} \Gamma\, \rho^{\,i-1} \cdot \Big((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}\, L_{\pi,x}) + L_{g,x}\, \ell_{\pi,x}\Big) \cdot \|x_{t-i} - x'_{t-i}\| \cdot \Gamma\, \rho^{\,t-\tau-i} \\
&\qquad + \Gamma\, \rho^{\,t-\tau-1} \cdot \Big((1 + L_{\pi,x})\, \ell_{g,u}\, L_{\pi,\theta} + L_{g,u}\, \ell_{\pi,\theta}\Big) \|\theta_\tau - \theta'_\tau\| \qquad (27) \\
&= \Big((1 + L_{\pi,x}(\ell_{g,x} + \ell_{g,u}\, L_{\pi,x}) + L_{g,x}\, \ell_{\pi,x}\Big)\, \Gamma^2 \cdot \rho^{\,t-\tau-1} \sum_{i=1}^{t-\tau} \|x_{t-i} - x'_{t-i}\| \\
&\qquad + \Gamma\, \rho^{\,t-\tau-1} \cdot \Big((1 + L_{\pi,x})\, \ell_{g,u}\, L_{\pi,\theta} + L_{g,u}\, \ell_{\pi,\theta}\Big) \|\theta_\tau - \theta'_\tau\| \\
&\le C_{l,g,(x,x)} \cdot \rho^{\,t-\tau} \|x_\tau - x'_\tau\| \;+\; C_{l,g,(x,\theta)} \cdot \rho^{\,t-\tau} \|\theta_\tau - \theta'_\tau\| \qquad (28)
\end{aligned}$$

where we use the ϵ-time-varying contractive perturbation property and (26) in (27); we use the first two inequalities to bound $\|x_{t-i} - x'_{t-i}\| \le \Gamma\, \rho^{\,t-i-\tau} \|x_\tau - x'_\tau\| + \frac{C L_{g,u} L_{\pi,\theta}}{\rho}\, \rho^{\,t-i-\tau} \|\theta_\tau - \theta'_\tau\|$ in (28).

For the last inequality, when $t = \tau + 1$, we see that

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

$$
\begin{aligned}
&= \left\| \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x'_\tau,u'_\tau} \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right\| \\
&\le \left\| \left(\frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} - \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x'_\tau,u'_\tau} \right) \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| + \left\| \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x'_\tau,u'_\tau} \cdot \left(\frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right) \right\| \quad (29) \\
&\le L_{\pi,\theta}\big(\ell_{g,x}\|x_\tau - x'_\tau\| + \ell_{g,u}\|u_\tau - u'_\tau\|\big) + L_{g,u}\big(\ell_{\pi,x}\|x_\tau - x'_\tau\| + \ell_{\pi,\theta}\|\theta_\tau - \theta'_\tau\|\big) \quad (30) \\
&\le \big(L_{\pi,\theta}(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x}\big)\|x_\tau - x'_\tau\| + \big(L^2_{\pi,\theta}\ell_{g,u} + L_{g,u}\ell_{\pi,\theta}\big)\|\theta_\tau - \theta'_\tau\|, \quad (31)
\end{aligned}
$$

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 (29); we use Assumption 13 in both (30) and (31).
Meanwhile, when $t > \tau + 1$, we also see that

$$
\begin{aligned}
&\left\| \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\
&\le \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}} \cdot \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \cdot \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right\| \quad (32) \\
&\le \left\| \left(\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right) \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| \\
&\quad + \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \left(\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right) \right\| \quad (33) \\
&\le \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| \\
&\quad + \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right\| \\
&\le \frac{((1 + L_{\pi,x})(\ell_{g,x} + \ell_{g,u}\, L_{\pi,x}) + L_{g,x}\, \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} \\
&\quad + C_0 \cdot \rho^{\,t-\tau-1} \cdot \left(L_{\pi,\theta}(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x}\right) \left\| x_\tau - x'_\tau \right\| \\
&\quad + C_0 \cdot \rho^{\,t-\tau-1} \cdot \left(L^2_{\pi,\theta}\ell_{g,u} + L_{g,u}\ell_{\pi,\theta}\right) \left\| \theta_\tau - \theta'_\tau \right\| \quad (34) \\
&\le C_{\ell,g,(\theta,x)}\, \rho^{\,t-\tau} \left\| x_\tau - x'_\tau \right\| \;+\; C_{\ell,g,(\theta,\theta)}\, \rho^{\,t-\tau} \left\| \theta_\tau - \theta'_\tau \right\|, \quad (35)
\end{aligned}
$$

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 (32); we use the triangle inequality in (33); we use the first and the third inequality of Lemma 19 as well as (29) - (31) in (34); we use the first two inequalities of Lemma 19 in (35). ■

Corollary 20 is implied by Lemma 19 and corresponds to Corollary D.4 in Lin et al. (2023).

Corollary 20 (Lipschitzness/Smoothness of the Multi-Step Costs) *Under the same assumptions as Lemma 19, the multi-step cost function $h_{t|\tau}$ satisfies that*

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

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

where $C_{L,h,x} = L_h\Gamma(1+L_{\pi,x})$, $C_{L,h,\theta} = L_h \max\{C_{L,g,\theta}(1+L_{\pi,x}), L_{\pi,\theta}\}$, *and*

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

[James: Copy proof from GAPS paper – or if it is nothing more than a product of Lipschitz constants for function composition, then say that.]

[Shengnan: could you please check the proof for Corollary 20 below?]

Proof For the first inequality in Corollary 20, note that

$$\frac{\partial h_{t|\tau}}{\partial x_\tau} = \left(\frac{\partial h_t}{\partial x_t}\Big|_{x_t,u_t} + \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t}\right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}}, \tag{36}$$

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

$$\begin{aligned}
\left\| \frac{\partial h_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_{\tau:t}} \right\| &\le \left(\left\| \frac{\partial h_t}{\partial x}\Big|_{x_t,u_t} \right\| + \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \right\| \cdot \left\| \frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t} \right\|\right) \left\| \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}} \right\| &(37)\\
&\le L_h(1+L_{\pi,x})\,\Gamma\,\rho^{t-\tau}. &(38)
\end{aligned}$$

For the second inequality, when $\tau = t$, since $x_t \in B_n(0,\bar{R}_C)$ and $u_t \in \mathcal{U}$, we see that

$$\left\| \frac{\partial h_{t|t}}{\partial \theta_t}\Big|_{x_t,\theta_t} \right\| = \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x_t,\theta_t} \right\| \le \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \right\| \left\| \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x_t,\theta_t} \right\| \le L_h L_{\pi,\theta}.$$

When $\tau < t$, the second inequality can be shown similarly with the first inequality in Corollary 20 because we have the chain-rule decomposition

$$\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:t}} = \left(\frac{\partial h_t}{\partial x_t}\Big|_{x_t,u_t} + \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t}\right) \cdot \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}}. \tag{39}$$

Applying Lemma 19 gives that $C_{L,h,\theta} = L_h\, C_{L,g,\theta}\,(1+L_{\pi,x})$.
For the third inequality, using (36), we see that

$$\begin{aligned}
&\left\| \frac{\partial h_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \frac{\partial h_{t|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t}} \right\|\\
&\le \left\| \frac{\partial h_t}{\partial x_t}\Big|_{x_t,u_t} \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial h_t}{\partial x_t}\Big|_{x_t,u_t} \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\|
\end{aligned}$$

$$
\begin{aligned}
&+ \left\| \left(\frac{\partial h_t}{\partial x_t}\Big|_{x_t,u_t} - \frac{\partial h_t}{\partial x_t}\Big|_{x'_t,u'_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\
&+ \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t} \cdot \left(\frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\| \\
&+ \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \left(\frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t} - \frac{\partial \pi_t}{\partial x_t}\Big|_{x'_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\
&+ \left\| \left(\frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} - \frac{\partial h_t}{\partial u_t}\Big|_{x'_t,u'_t} \right) \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x'_t,\theta_t} \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \qquad (40) \\
&\le L_h \rho^{t-\tau} \big(C_{\ell,g,(x,x)} \|x_\tau - x'_\tau\| + C_{\ell,g,(x,\theta)} \|\theta_\tau - \theta'_\tau\| \big) \\
&\quad + \big(\ell_{h,x} \|x_t - x'_t\| + \ell_{h,u} \|u_t - u'_t\| \big) C_{L,g,x} \rho^{t-\tau} \\
&\quad + L_h L_{\pi,x} \rho^{t-\tau} \big(C_{\ell,g,(x,x)} \|x_\tau - x'_\tau\| + C_{\ell,g,(x,\theta)} \|\theta_\tau - \theta'_\tau\| \big) \\
&\quad + L_h \ell_{\pi,x} \|x_t - x'_t\| C_{L,g,x} \rho^{t-\tau} + \big(\ell_{h,x} \|x_t - x'_t\| + \ell_{h,u} \|u_t - u'_t\| \big) L_{\pi,x} C_{L,g,x} \rho^{t-\tau}, \qquad (41)
\end{aligned}
$$

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 (36) and the triangle inequality in (40); we use Lemma 19 in (41). Note that by the first two inequalities in Lemma 19, we have

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

Substituting these two inequalities into (41) finishes the proof of the third inequality.
For the last inequality, note that when $\tau = t$, we have that

$$
\begin{aligned}
&\left\| \frac{\partial h_{t,t}}{\partial \theta_t}\Big|_{x_t,\theta_t} - \frac{\partial h_{t,t}}{\partial \theta_t}\Big|_{x'_t,\theta'_t} \right\| \qquad (42) \\
&= \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x_t,\theta_t} - \frac{\partial h_t}{\partial u_t}\Big|_{x'_t,u'_t} \cdot \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x'_t,\theta'_t} \right\| \qquad (43) \\
&\le \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \left(\frac{\partial \pi_t}{\partial \theta_t}\Big|_{x_t,\theta_t} - \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x'_t,\theta'_t} \right) \right\| + \left\| \left(\frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} - \frac{\partial h_t}{\partial u_t}\Big|_{x'_t,u'_t} \right) \cdot \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x'_t,\theta'_t} \right\| \qquad (44) \\
&\le L_h \big(\ell_{\pi,\theta} \|\theta_t - \theta'_t\| + \ell_{\pi,x} \|x_t - x'_t\| \big) + \big(\ell_{h,x} \|x_t - x'_t\| + \ell_{h,u} \|u_t - u'_t\| \big) L_{\pi,\theta} \qquad (45) \\
&\le (L_h \ell_{\pi,\theta} + (\ell_{h,x} + \ell_{h,u} L_{\pi,x}) L_{\pi,\theta}) \|x_t - x'_t\| + \big(L_h \ell_{\pi,\theta} + \ell_{h,u} L^2_{\pi,\theta} \big) \|\theta_t - \theta'_t\|, \qquad (46)
\end{aligned}
$$

where we use the chain rule decomposition in (43); we use the triangle inequality in (44); we use Assumption 13 in both (45) and (46).

When $\tau < t$, by (39), we have that

$$
\begin{aligned}
&\left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t}} \right\| \\
&\le \left\| \frac{\partial h_t}{\partial x_t}\Big|_{x_t,u_t} \cdot \left(\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\| \\
&\quad + \left\| \left(\frac{\partial h_t}{\partial x_t}\Big|_{x_t,u_t} - \frac{\partial h_t}{\partial x_t}\Big|_{x'_t,u'_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\
&\quad + \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t} \cdot \left(\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\|
\end{aligned}
$$

$$
\begin{aligned}
&+ \left\| \frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \left(\frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t} - \frac{\partial \pi_t}{\partial x_t}\Big|_{x'_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\
&+ \left\| \left(\frac{\partial h_t}{\partial u_t}\Big|_{x_t,u_t} - \frac{\partial h_t}{\partial u_t}\Big|_{x'_t,u'_t} \right) \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x'_t,\theta_t} \cdot \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \qquad (47) \\
&\le L_h \rho^{t-\tau} \left(C_{\ell,g,(\theta,x)} \|x_\tau - x'_\tau\| + C_{\ell,g,(\theta,\theta)} \|\theta_\tau - \theta'_\tau\| \right) \\
&\quad + \left(\ell_{h,x} \|x_t - x'_t\| + \ell_{h,u} \|u_t - u'_t\| \right) C_{L,g,\theta} \rho^{t-\tau} \\
&\quad + L_h L_{\pi,x} \rho^{t-\tau} \left(C_{\ell,g,(\theta,x)} \|x_\tau - x'_\tau\| + C_{\ell,g,(\theta,\theta)} \|\theta_\tau - \theta'_\tau\| \right) \\
&\quad + L_h \ell_{\pi,x} \|x_t - x'_t\| C_{L,g,\theta} \rho^{t-\tau} + \left(\ell_{h,x} \|x_t - x'_t\| + \ell_{h,u} \|u_t - u'_t\| \right) L_{\pi,x} C_{L,g,\theta} \rho^{t-\tau}, \qquad (48)
\end{aligned}
$$

where we use (39) and the triangle inequality in (47); we use Assumption 13 in (48). Note that by the first two inequalities in Lemma 19, , we have

$$
\begin{aligned}
\|x_t - x'_t\| &\le \rho^{t-\tau} \left(C_{L,g,x} \|x_\tau - x'_\tau\| + C_{L,g,\theta} \|\theta_\tau - \theta'_\tau\| \right), \\
\|u_t - u'_t\| &\le L_{\pi,x} \rho^{t-\tau} \left(C_{L,g,x} \|x_\tau - x'_\tau\| + C_{L,g,\theta} \|\theta_\tau - \theta'_\tau\| \right).
\end{aligned}
$$

Substituting these into (48) finishes the proof of the fourth inequality.

■

Theorem 21 bounds the distances between the trajectory of M-GAPS and the imaginary trajectory achieved by using θ_t repeatedly from time step 0. We include the proof of Theorem 21 in Appendix J.1 for completeness.

Theorem 21 *Suppose Assumptions 13 and 14 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of*

$$
\begin{aligned}
x_{t+1} &= q_t^x(x_t, y_t, \theta_t) = g_t(x_t, \pi_t(x_t, \theta_t)), && (49a) \\
y_{t+1} &= q_t^y(x_t, y_t, \theta_t) = \frac{\partial g_{t+1|t}}{\partial x_t}\Big|_{x_t,\theta_t} \cdot y_t + \frac{\partial g_{t+1|t}}{\partial \theta_t}\Big|_{x_t,\theta_t}, && (49b) \\
\theta_{t+1} &= q_t^\theta(x_t, y_t, \theta_t) = \Pi_\Theta \left(\theta_t - \eta \left(\frac{\partial h_{t|t}}{\partial x_t}\Big|_{x_t,\theta_t} \cdot y_t + \frac{\partial h_{t|t}}{\partial \theta_t}\Big|_{x_t,\theta_t} \right) \right). && (49c)
\end{aligned}
$$

[Shengnan: why theta t+1 not theta t. (should be a mistake.) put this into the potential mistake list, and also the $C \ge 1$ instead of 0 in contractive perturbation definition] Suppose η satisfies the constraint that $\bar{\varepsilon} := \frac{C_{L,h,\theta}\eta}{1-\rho} \le \epsilon_\theta$, where ϵ_θ is the per-step parameter-variation budget from the ϵ_θ-time-varying contractive perturbation/stability assumption.. [James: Where was ε defined? Was this supposed to be ϵ_θ?] [Shengnan: Yes, I think so, and I changed ε so that now it's ϵ_θ] Then, both $\|G_t\|$ and $\|\nabla F_t(\theta_t)\|$ are upper bounded by $\frac{C_{L,h,\theta}}{1-\rho}$ (as defined in Corollary 20), and the following inequalities holds for any two time steps $\tau \le t$: [James: the ρ in the numerator on the RHS blends in with the preceding subscripts, what about moving it next to η?] [Shengnan: moved it!]

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

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

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

[James: Move the proof directly here.] [Shengnan: the proof is now immediately below the thm]

J.1. Proof of Theorem 21

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

$$\|G_t\| \le \frac{C_{L,h,\theta}}{1-\rho},\ x_t \in B_n(0, R_S + \bar{C}\|x_0\|),\ u_t \in \mathcal{U},\ \text{and } \|\theta_{t+1} - \theta_t\| \le \epsilon_\theta, \tag{50}$$

where $\mathcal{U}$ is defined in Assumption 12.

Note that $\|G_0\| \le C_{L,h,\theta} \le \frac{C_{L,h,\theta}}{1-\rho}$ by Corollary 20.

Temporary Remark: Why do we have $x_0 \in B(0,\ R_S + \bar{C}\|x_0\|)$? This step typically relies on the fact that one may assume $\bar{C} \ge 1$ (or take it to be so without loss of generality). There are two complementary reasons.

Reason 1: The definition of contractive perturbation implies $C \ge 1$. Recall the contractive-perturbation condition, written in Definition 8 and Assumption 14: for any $\tau \le t$,

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

Taking $t = \tau$ yields

$$\|x_\tau - x'_\tau\| \le C\|x_\tau - x'_\tau\|.$$

Since this must hold for arbitrary distinct x_τ, x'_τ, we must have $C \ge 1$; hence in this paper $\bar{C} \ge 1$. Therefore

$$R_S + \bar{C}\|x_0\| \ \ge\ \bar{C}\|x_0\| \ \ge\ \|x_0\|,$$

which directly implies $x_0 \in B(0,\ R_S + \bar{C}\|x_0\|)$.

Reason 2: Even if an original bound gave $C < 1$, we can enlarge it w.l.o.g. The contractive-perturbation property is monotone in the gain: if it holds for some $\bar{C}_0$, then it also holds for any $\bar{C} \ge \bar{C}_0$. Thus one can always redefine

$$\bar{C} := \max\{1, \bar{C}_0\}$$

without violating Assumption 19. With this convention, the containment $x_0 \in B(0,\ R_S + \bar{C}\|x_0\|)$ follows automatically.

We also have $x_0 \in B_n(0, R_S + \bar{C}\|x_0\|)$ and $u_0 \in \mathcal{U}$.

Suppose $\|G_{t-1}\| \le \frac{C_{L,h,\theta}}{1-\rho}$ for some $t \ge 1$. Then, since the projection onto Θ is nonexpansive (Lemma 18) and $\eta \le \frac{(1-\rho)\epsilon_\theta}{C_{L,h,\theta}}$, we see that

$$\|\theta_t - \theta_{t-1}\| \le \|\eta G_{t-1}\| \le \epsilon_\theta.$$

Suppose $\|\theta_\tau - \theta_{\tau-1}\| \le \epsilon_\theta$ holds for all $\tau \le t$, i.e., $\theta_{0:t} \in S_{\epsilon_\theta}(0:t)$. The next lemma (a restatement of Lemma D.2 in Lin et al. (2023) with notation adapted to this paper) gives the needed state bound.

Lemma D.2 (restated; notation adapted). Suppose Assumptions 13 and 14 hold. For any starting state $x_\tau \in B_n(0, R_S + \bar{C}\|x_0\|)$ and $\theta_{\tau:t-1} \in S_{\epsilon_\theta}(\tau : t-1)$, the final state $x_t := g_{t|\tau}(x_\tau, \theta_{\tau:t-1})$ satisfies

$$\|x_t\| \ \le\ \bar{C}\,\rho^{\,t-\tau}\,\|x_\tau\| + R_S.$$

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

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

Thus, by the triangle inequality, we see that

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

[James: ref contraction stability assumption] where we use ε-time-varying stability in the last inequality. □

Applying Lemma D.2 with $x_\tau = x_0$ yields $x_t \in B_n\big(0, R_S + \bar{C}\,\|x_0\|\big)$ for any t. By the definition of $\mathcal{U}$ (constructed over a state domain that contains this ball), we also have $u_t \in \mathcal{U}$.

Therefore, from the definition of G_t in M-GAPS (23), we see that

$$\begin{aligned}
\|G_t\| &= \left\| \sum_{\tau=0}^{t} \frac{\partial h_{t|t-\tau}}{\partial \theta_{t-\tau}}\bigg|_{x_{t-\tau},\theta_{t-\tau:t}} \right\| \\
&\le \sum_{\tau=0}^{t} \left\| \frac{\partial h_{t|t-\tau}}{\partial \theta_{t-\tau}}\bigg|_{x_{t-\tau},\theta_{t-\tau:t}} \right\| && \text{(51a)} \\
&\le \sum_{\tau=0}^{t} C_{L,h,\theta}\rho^{\tau} && \text{(51b)} \\
&\le \frac{C_{L,h,\theta}}{1-\rho},
\end{aligned}$$

where we use the triangle inequality in (51a) and Corollary 20 in (51b). Note that we can apply Corollary 20 because $x_t \in B_n(0, R_S + \bar{C}\|x_0\|)$. Therefore, we have shown (50) by induction. One can use the same technique as (51) to show $\|\nabla F_t(\theta_t)\| \le \frac{C_{L,h,\theta}}{1-\rho}$. [James: Should the preceding RHS be $C_{L,h,\theta}$ instead? Might have been a mistake in the M-GAPS paper, because in the original GAPS paper f was used for costs.] [Shengnan: Yes. Modified that.]

Since the projection onto Θ is nonexpansive according to Lemma 18, we have for every k,

$$\begin{aligned}
\|\theta_{k+1} - \theta_k\| &= \|\Pi_\Theta(\theta_k - \eta G_k) - \Pi_\Theta(\theta_k)\| \\
&\le \|(\theta_k - \eta G_k) - \theta_k\| = \eta\,\|G_k\| \;\le\; \frac{C_{L,h,\theta}}{1-\rho}\eta,
\end{aligned}$$

where the last step uses the uniform bound on $\|G_k\|$ established above. Summing from $k = \tau$ to $t-1$ (triangle inequality) yields

$$\|\theta_t - \theta_\tau\| \le \sum_{k=\tau}^{t-1} \|\theta_{k+1} - \theta_k\| \le \frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau)\eta. \tag{52}$$

Now we bound the distance between x_τ and $\hat{x}_\tau(\theta_t)$ for $\tau \le t$. We see that [James: Think about ways to break this up.]

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

$$\leq \sum_{\tau'=0}^{\tau-1} \left\| g_{\tau|0}(x_0, \theta_{0:\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g_{\tau|0}(x_0, \theta_{0:\tau'-1}, (\theta_t)_{\times(\tau-\tau')}) \right\| \tag{53a}$$

$$\leq \sum_{\tau'=0}^{\tau-1} \left\| g_{\tau|\tau'}(x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g_{\tau|\tau'}(x_{\tau'}, (\theta_t)_{\times(\tau-\tau')}) \right\| \tag{53b}$$

$$\tag{53c}$$

where we use the triangle inequality in (53a) [James: Explain more] [Shengnan: Added the explanation part below] as follows: we bridge the two parameter strings $\theta_{0:\tau-1}$ and $(\theta_t)_{\times\tau}$ by a sequence that changes one coordinate at a time, from left to right. Concretely, start from $g_{\tau|0}(x_0, \theta_{0:\tau-1})$ and, for each $\tau' = 0, 1, \ldots, \tau-1$, replace the τ'-th entry $\theta_{\tau'}$ by θ_t, keeping all other entries fixed; after τ such single-coordinate replacements we reach $g_{\tau|0}(x_0, (\theta_t)_{\times\tau})$ (here $(\theta_t)_{\times m}$ means m copies of θ_t). Writing the difference between the endpoints as the sum of these τ one-step differences and applying $\|\sum_i v_i\| \leq \sum_i \|v_i\|$ yields (53a).

[James: Explain more] [Shengnan: added the explanation for (53b) below] We use the definition of the multi-step dynamics in (53b) by factoring the horizon at each τ'. It is equality by the semigroup (composition) property of the multi-step map. By definition,

$$g_{t|\tau}(x, \theta_{\tau:t-1}) := g_{t-1}(g_{t-2}(\cdots g_\tau(x, \theta_\tau) \cdots, \theta_{t-2}),\, \theta_{t-1}).$$

Hence, concatenating the parameter string as $(\theta_{0:\tau'},\, (\theta_t)_{\times(\tau-\tau'-1)}) = (\theta_{0:\tau'-1},\, \theta_{\tau'},\, (\theta_t)_{\times(\tau-\tau'-1)})$, we can factor the $0 \to \tau$ evolution at time τ':

$$\begin{aligned} g_{\tau|0}(x_0, \theta_{0:\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) &= g_{\tau-1}\big(\cdots g_{\tau'}\big(g_{\tau'|0}(x_0, \theta_{0:\tau'-1}),\, \theta_{\tau'}\big) \cdots,\, \underbrace{\theta_t, \ldots, \theta_t}_{\tau-\tau'-1 \text{ times}}\big) \\ &= g_{\tau|\tau'}\big(g_{\tau'|0}(x_0, \theta_{0:\tau'-1}),\, \theta_{\tau'},\, (\theta_t)_{\times(\tau-\tau'-1)}\big). \end{aligned}$$

Similarly, for the string $(\theta_{0:\tau'-1},\, (\theta_t)_{\times(\tau-\tau')})$ we obtain

$$g_{\tau|0}(x_0, \theta_{0:\tau'-1}, (\theta_t)_{\times(\tau-\tau')}) = g_{\tau|\tau'}\big(g_{\tau'|0}(x_0, \theta_{0:\tau'-1}),\, (\theta_t)_{\times(\tau-\tau')}\big).$$

Defining the intermediate state $x_{\tau'} := g_{\tau'|0}(x_0, \theta_{0:\tau'-1})$, both identities become

$$\begin{aligned} g_{\tau|0}(x_0, \theta_{0:\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) &= g_{\tau|\tau'}\big(x_{\tau'},\, \theta_{\tau'},\, (\theta_t)_{\times(\tau-\tau'-1)}\big), \\ g_{\tau|0}(x_0, \theta_{0:\tau'-1}, (\theta_t)_{\times(\tau-\tau')}) &= g_{\tau|\tau'}\big(x_{\tau'},\, (\theta_t)_{\times(\tau-\tau')}\big), \end{aligned}$$

so each "one–coordinate change" term in (53a) is exactly a difference of two $\tau' \to \tau$ evolutions that share the same intermediate state $x_{\tau'}$ and the same suffix length $\tau - \tau'$, which is the form in (53b).

Then, we see that

$$\left\| x_\tau - \hat{x}_\tau(\theta_t) \right\| \leq \sum_{\tau'=0}^{\tau-1} C_{L,g,\theta}\, \rho^{\tau-\tau'} \left\| \theta_t - \theta_{\tau'} \right\| \tag{54a}$$

$$\leq \sum_{\tau'=0}^{\tau-1} C_{L,g,\theta}\, \rho^{\tau-\tau'} \left(\frac{C_{L,h,\theta}}{1-\rho} (t - \tau')\, \eta \right) \tag{54b}$$

$$= \frac{C_{L,h,\theta}\, C_{L,g,\theta}}{1-\rho}\, \eta \sum_{\tau'=0}^{\tau-1} (t-\tau')\, \rho^{\tau-\tau'}$$

$$= \frac{C_{L,h,\theta}\, C_{L,g,\theta}}{1-\rho}\, \eta \left((t-\tau) \sum_{j=1}^{\tau} \rho^j \;+\; \sum_{j=1}^{\tau} j\, \rho^j \right) \tag{54c}$$

$$\tag{54d}$$

where we use Lemma 19 in (54a); we use (52) in (54b). To pass from (54b) to (54c), we rewrite and reindex the sum:

$$\sum_{\tau'=0}^{\tau-1} (t-\tau')\, \rho^{\tau-\tau'} = \sum_{j=1}^{\tau} [(t-\tau)+j]\, \rho^j = (t-\tau) \sum_{j=1}^{\tau} \rho^j \;+\; \sum_{j=1}^{\tau} j\, \rho^j,$$

where we set $j = \tau - \tau'$ so that as τ' runs $0 \to \tau - 1$, j runs $1 \to \tau$, and $t - \tau' = t - (\tau - j) = (t-\tau) + j$.

Then, the derivation follows as:

$$\left\| x_\tau - \hat{x}_\tau(\theta_t) \right\| \le \frac{C_{L,h,\theta}\, C_{L,g,\theta}}{1-\rho}\, \eta \left((t-\tau) \frac{\rho}{1-\rho} \;+\; \frac{\rho}{(1-\rho)^2} \right) \tag{55a}$$

$$= \frac{C_{L,h,\theta}\, C_{L,g,\theta}}{(1-\rho)^2} \left((t-\tau) + \frac{1}{1-\rho} \right) \rho\, \eta.$$

To pass from (54c) to (55a), we upper bound the finite sums by their infinite geometric counterparts:

$$\sum_{j=1}^{\tau} \rho^j \le \sum_{j=1}^{\infty} \rho^j = \frac{\rho}{1-\rho}, \quad \sum_{j=1}^{\tau} j\, \rho^j \le \sum_{j=1}^{\infty} j\, \rho^j = \frac{\rho}{(1-\rho)^2}.$$

Finally, to reach the last line, we simply rearrange and factor out common terms. [James: fix references or fold this further explanation into the main proof.] [Shengnan: folded with line references.]

[James: why?] Similarly, given that $\hat{x}_t(\theta_t) = g_{t|0}(x_0, (\theta_t)_{\times t})$, since $(\theta_t)_{\times t} \in S_{\epsilon_\theta}(0 : t-1)$, and consecutive differences are $0 \le \epsilon_\theta$, by Lemma D.2 with $\tau = 0$ and ϵ_θ-time-varying stability, we have $\|\hat{x}_t(\theta_t) - g_{t|0}(0, (\theta_t)_{\times t})\| \le \bar{C}\rho^t \|x_0\|$ and $\|g_{t|0}(0, (\theta_t)_{\times t})\| \le R_S$, hence $\|\hat{x}_t(\theta_t)\| \le \bar{C}\|x_0\| + R_S)$, which means $\hat{x}_t(\theta_t) \in B_n(0, R_S + \bar{C}\|x_0\|)$. Together with $x_t \in B_n(0, R_S + \bar{C}\|x_0\|)$ derived earlier, we obtain that

$$\begin{aligned} |h_t(x_t, u_t, \theta_t) - F_t(\theta_t)| &= |h_t(x_t, u_t, \theta_t) - h_t(\hat{x}_t(\theta_t), \hat{u}_t(\theta_t), \theta_t)| \\ &\le L_h(\|x_t - \hat{x}_t(\theta_t)\| + \|u_t - \hat{u}_t(\theta_t)\|) && \text{(56a)} \\ &\le L_h(1 + L_{\pi,x} + L_{f,x})\|x_t - \hat{x}_t(\theta_t)\| && \text{(56b)} \\ &\le \frac{C_{L,h,\theta} C_{L,g,\theta} L_h (1 + L_{\pi,x} + L_{h,x})\rho}{(1-\rho)^3} \cdot \eta, && \text{(56c)} \end{aligned}$$

where we use Theorem 13 in (56a) and (56b); we use (55) in (56c). [James: Where was L_h defined? I thought Theorem 13 only gives us separate Lipschitz constants for each input of h.]

Recall that we define the gradient approximation G_t for M-GAPS in (23). Using this notation, the update rule of $\theta_{0:T-1}$ in joint dynamics (49) can be simplified as

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

To compare the trajectory of M-GAPS with the trajectory achieved by the online gradient descent trajectory $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta \nabla F_t(\theta_t))$, we bound the difference between G_t and $\nabla F_t(\theta_t)$ in Theorem 22. We provide its proof in Appendix J.2 for completeness.

Theorem 22 (Gradient Bias) *Suppose Assumptions 13 and 14 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of (49). Suppose η satisfies the constraint that $\bar{\varepsilon} := \frac{C_{L,h,\theta}\eta}{1-\rho} \le \varepsilon$. Then, the following holds for all $\tau \le t$:*

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

for

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

Next,

$$\|G_t - \nabla F_t(\theta_t)\| \le \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta.$$

J.2. Proof of Theorem 22

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.

We use the triangle inequality to do the decomposition

$$\begin{aligned}
&\left\| \left.\frac{\partial h_{t|0}}{\partial \theta_\tau}\right|_{x_0,\theta_{0:t}} - \left.\frac{\partial h_{t|0}}{\partial \theta_\tau}\right|_{x_0,(\theta_t)_{\times(t+1)}} \right\| \\
&= \left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t}} - \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
&\le \left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,(\theta_t)_{\times(t-\tau)}} - \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
&\quad + \sum_{\tau'=\tau+1}^{t-1} \left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'},(\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'-1},(\theta_t)_{\times(t-\tau'+1)}} \right\|.
\end{aligned} \tag{57}$$

Note that we can apply Corollary 20 to bound each term in (57). For the first term in (57), since $x_\tau, \hat{x}_\tau(\theta_t), x_{\tau+1} \in B_n(0, \bar{R}_C)$, we see that

$$\begin{aligned}
&\left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,(\theta_t)_{\times(t-\tau)}} - \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
&\le \rho^{t-\tau}\big(C_{\ell,h,(\theta,x)}\|x_\tau - \hat{x}_\tau(\theta_t)\| + C_{\ell,h,(\theta,\theta)}\|\theta_t - \theta_\tau\|\big)
\end{aligned} \tag{58a}$$

$$\leq \frac{(1-\rho)C_{L,h,\theta}C_{\ell,h,(\theta,x)} + \rho C_{L,h,\theta}C_{L,g,\theta}C_{\ell,h,(\theta,\theta)}}{(1-\rho)^2} \cdot (t-\tau)\rho^{t-\tau} \cdot \eta$$
$$+ \frac{\rho C_{L,h,\theta}C_{L,g,\theta}C_{\ell,h,(\theta,x)}}{(1-\rho)^3} \cdot \rho^{t-\tau} \cdot \eta, \tag{58b}$$

where we use Corollary 20 in (58a) and Theorem 21 in (58b).
For any $\tau' \in [\tau+1 : t-1]$, since $x_{\tau'}, x_{\tau'+1} \in B_n(0, \bar{R}_C)$, we see that

$$\left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau, \theta_{\tau:\tau'}, (\theta_t)_{\times(t-\tau')}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau, \theta_{\tau:\tau'-1}, (\theta_t)_{\times(t-\tau'+1)}} \right\|$$
$$= \left\| \left(\frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\Big|_{x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(t-\tau')}} - \frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\Big|_{x_{\tau'}, (\theta_t)_{\times(t-\tau'+1)}} \right) \frac{\partial g_{\tau'|\tau}}{\partial \theta_\tau}\Big|_{x_\tau, \theta_{\tau:\tau'-1}} \right\|$$
$$\leq \left\| \frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\Big|_{x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(t-\tau')}} - \frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\Big|_{x_{\tau'}, (\theta_t)_{\times(t-\tau'+1)}} \right\| \cdot \left\| \frac{\partial g_{\tau'|\tau}}{\partial \theta_\tau}\Big|_{x_\tau, \theta_{\tau:\tau'-1}} \right\|$$
$$\leq C_{\ell,h,(x,\theta)}\rho^{t-\tau'}\|\theta_t - \theta_{\tau'}\| \cdot C_{L,g,\theta}\rho^{\tau'-\tau} \tag{59a}$$
$$\leq C_{\ell,h,(x,\theta)}C_{L,g,\theta} \cdot \rho^{t-\tau} \cdot \left(\frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau')\eta \right), \tag{59b}$$

where we use Lemma 19 and Corollary 20 in (59a); we use Theorem 21 in (59b). Substituting (58) and (59) into (57) finishes the proof of the first inequality.

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

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

Therefore, we see that

$$\|G_t - \nabla F_t(\theta_t)\| = \left\| \sum_{\tau=0}^{t} \frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\Big|_{x_0, \theta_{0:t}} - \sum_{\tau=0}^{t} \frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\Big|_{x_0, (\theta_t)_{\times(t+1)}} \right\|$$
$$\leq \sum_{\tau=0}^{t} \left\| \frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\Big|_{x_0, \theta_{0:t}} - \frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\Big|_{x_0, (\theta_t)_{\times(t+1)}} \right\| \tag{60a}$$
$$\leq \sum_{\tau=0}^{t} \left(\hat{C}_0 + \hat{C}_1\tau + \hat{C}_2\tau^2 \right)\rho^\tau\eta \tag{60b}$$
$$\leq \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3} \right)\eta, \tag{60c}$$

where we use the triangle inequality in (60a); we use the first inequality in Theorem 22 that we have shown and Corollary 20 in (60b). [Shengnan: folded this further explanation as well] To pass from (60b) to (60c), we bound the finite sum by its infinite counterpart and apply standard

geometric–series identities. Specifically, starting from

$$\sum_{\tau=0}^{t}\Big(\hat{C}_0+\hat{C}_1\tau+\hat{C}_2\tau^2\Big)\,\rho^\tau \;\le\; \sum_{\tau=0}^{\infty}\Big(\hat{C}_0+\hat{C}_1\tau+\hat{C}_2\tau^2\Big)\,\rho^\tau,$$

we use, for $|\rho|<1$,

$$\sum_{\tau=0}^{\infty}\rho^\tau=\frac{1}{1-\rho},\quad \sum_{\tau=0}^{\infty}\tau\,\rho^\tau=\frac{\rho}{(1-\rho)^2},\quad \sum_{\tau=0}^{\infty}\tau(\tau-1)\,\rho^\tau=\frac{2\rho^2}{(1-\rho)^3}.$$

Since $\tau^2=\tau(\tau-1)+\tau$, it follows that

$$\sum_{\tau=0}^{\infty}\tau^2\,\rho^\tau=\frac{\rho}{(1-\rho)^2}+\frac{2\rho^2}{(1-\rho)^3}.$$

Substituting into the infinite sum gives

$$\sum_{\tau=0}^{\infty}(\hat{C}_0+\hat{C}_1\tau+\hat{C}_2\tau^2)\,\rho^\tau=\frac{\hat{C}_0}{1-\rho}+\hat{C}_1\,\frac{\rho}{(1-\rho)^2}+\hat{C}_2\Big(\frac{\rho}{(1-\rho)^2}+\frac{2\rho^2}{(1-\rho)^3}\Big).$$

Finally, since $\rho\le 1$, we can loose the bound to obtain

$$\frac{\hat{C}_0}{1-\rho}+\frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2}+\frac{\hat{C}_2}{(1-\rho)^3},$$

which is exactly the coefficient stated in (60c).