Stability-Certified Reinforcement Learning: A Control-Theoretic Perspective

Ming Jin and Javad Lavaei

***Abstract*—We investigate the important problem of certifying stability of reinforcement learning policies when interconnected with nonlinear dynamical systems. We show that by regulating the partial gradients of policies, strong guarantees of robust stability can be obtained based on a proposed semidefinite programming feasibility problem. The method is able to certify a large set of stabilizing controllers by exploiting problem-specific structures; furthermore, we analyze and establish its (non)conservatism. Empirical evaluations on two decentralized control tasks, namely multi-flight formation and power system frequency regulation, demonstrate that the reinforcement learning agents can have high performance within the stability-certified parameter space, and also exhibit stable learning behaviors in the long run.**

***Index Terms*—Reinforcement learning, robust control, decentralized control synthesis, safe reinforcement learning**

I. Introduction

REINFORCEMENT learning (RL) aims at guiding an agent to perform a task as efficiently and skillfully as possible through interactions with the environment. Consider the interconnected system illustrated in Fig. 1, where $\mathcal{G}$ is the environment and π_θ is the policy. The goal of RL is to maximize the expected return:

$$\eta(\pi_\theta) = \mathop{\mathbb{E}}_{x_0, u_t \sim \pi_\theta(\cdot|x_t), x_{t+1} \sim \mathcal{G}(x_t, u_t)} \left[\sum_{t=0}^{\infty} \rho^t r(x_t, u_t) \right], \tag{1}$$

where $r(x, u)$ is the reward at state x and action u, $\rho \in (0, 1]$ is the future discount factor, and the expectation is taken over the policy π_θ as well as the initial state distribution and the world dynamics $\mathcal{G}$. While remarkable progress has been made in RL algorithms, such as policy gradient [1]–[3], Q-learning [4], [5], and actor-critic methods [6], [7], a fundamental issue that is unresolved in the literature is how to analyze or certify stability of the interconnected system, which is closely related to the safety aspect of mission-critical systems, such as autonomous cars and power grids [8]–[10].

Stability verification is challenging for two key reasons: **(i)** both the environment and the control policy (e.g., deep neural networks) are often highly nonlinear; and **(ii)** the policy changes dynamically during the learning phase. In this study, we propose a "safety set" of policies $\mathcal{P}(\xi)$ characterized by partial gradients, i.e.,[1]

$$\left\{ \pi \;\middle|\; \underline{\xi}_{ij} \le \partial_j \pi_i(y) \le \overline{\xi}_{ij}, \forall i \in [n_a], j \in [n_s], y \in \mathbb{R}^{n_s} \right\}, \tag{2}$$

where $\partial_j \pi_i$ is the partial derivative of π_i for the j-th input. The proposed framework verifies numerical bounds $\underline{\xi} \in \mathbb{R}^{n_a \times n_s}$ and $\overline{\xi} \in \mathbb{R}^{n_a \times n_s}$ such that as long as the policy stays within the "safety set", the stability of the interconnected system is guaranteed. This offers significant freedom to RL algorithms for exploring the optimal control policies while maintaining basic requirement of stability. Bounds on partial gradients are more expressive than Lipschitz constant, since they encode a wide range of neural network behaviors [11], [12].

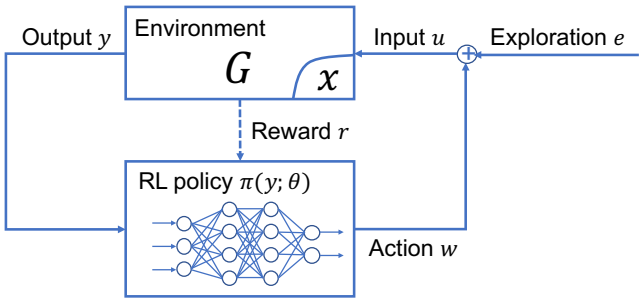


Fig. 1: Overview of the interconnected system of an RL policy and the environment. The goal of RL is to maximize expected rewards through interaction and exploration.

Our result can be also regarded as a computational method for the verification of the adversarial robustness of RL agents in the real world when the control inputs are subject to adversarial perturbations. A similar condition has been observed in computer vision [13], where bounding the Lipschitz constant is used to defend against adversarial injections in image classification. This work is an analogy to reinforcement learning in a dynamic environment, and it also incorporates the special case where controllers are known to have bounded Lipschitz constants.

The paper is organized as follows. We formulate the problem as a robust control problem in Section II. Main results on gradient bounds for a linear or nonlinear system G are presented in Section III, where we also analyze the conservatism of the certificate. Related work is discussed in Section IV. The method is evaluated in Section V on two nonlinear decentralized control tasks. Conclusions are drawn in Section VI.

This work was supported by grants from AFOSR, ONR, ARO and NSF. Parts of this work have appeared in the conference paper "Ming Jin and Javad Lavaei, Control-Theoretic Analysis of Smoothness for Stability-Certified Reinforcement Learning, Proc. of 57th IEEE Conference on Decision and Control, 2018."

M. Jin and J. Lavaei are from Department of Industrial Engineering and Operations Research, University of California, Berkeley. Emails: {jinming, lavaei}@berkeley.edu.

[1] We use $[n] = \{1, ..., n\}$ as the set notation.

II. Problem formulation

The problem under study focuses on a general continuous-time dynamical system:

$$\dot{x}(t) = f_t(x(t), u(t)), \tag{3}$$

with the state $x(t) \in \mathbb{R}^{n_s}$ and the control action $u(t) \in \mathbb{R}^{n_a}$. In general, f_t can be a time-varying and nonlinear function, but for the purpose of stability analysis, we study the important case that

$$f_t(x(t)) = Ax(t) + Bu(t) + g_t(x(t)), \tag{4}$$

where f_t comprises of a linear time-invariant (LTI) component $A \in \mathbb{R}^{n_s \times n_s}$ that is Hurwitz (i.e., every eigenvalue of A has strictly negative real part), a control matrix $B \in \mathbb{R}^{n_s \times n_a}$, and a slowly time-varying component g_t that is allowed to be nonlinear and even uncertain.[2] The condition that A is stable is a basic requirement, but the goal of reinforcement learning is to design a controller that optimizes some performance metric that is not necessarily related to the stability condition. For feedback control, we also allow the controller to obtain observations $y(t) = Cx(t) \in \mathbb{R}^{n_s}$ that are a linear function of the states, where $C \in \mathbb{R}^{n_s \times n_s}$ may have a sparsity pattern to account for partial observations in the context of decentralized controls [14], but the control action may rely on a nonlinear transformation of the state.

The control input from the RL policy is given by $u(t) = \pi_t(y(t); \theta_t) + e(t)$, where $\pi_t(y(t); \theta_t)$ is a neural network parametrized by θ_t, which can be time-varying during learning. The vector $e(t) \in \mathbb{R}^{n_a}$ captures the adversarial injections that is assumed to have a bounded energy over time ($\|e\|_2 = \sqrt{\int |e(t)|_2^2 dt} \leq \infty$). To achieve adversarial robustness, it is sufficient to certify stability, as can be seen in the classical definition of the L_2 gain [15], [16].[3]

Definition 1 (Input-output stability)**.** *The L_2 gain of the system G controlled by π is the worst-case ratio between total output energy and total input energy:*

$$\gamma(G, \pi) = \sup_{u \in L_2} \frac{\|y\|_2}{\|u\|_2}, \tag{5}$$

where L_2 is the set of all square-summable signals, $\|y\|_2 = \sqrt{\int |y(t)|_2^2 dt}$ is the total energy over time, and $u(t) = \pi_t(y(t); \theta_t) + e(t)$ is the control input with adversarial injections. If $\gamma(G, \pi)$ is finite, then the interconnected system is said to have input-output stability (or finite L_2 gain).

To this end, we develop a new quadratic constraint on gradient-bounded functions, which exploits the sparsity of the control architecture and the non-homogeneity of the output vector. Some key features of the stability-certified bounds are as follows: **(a)** the bounds are inherent to the targeted real-world control task; **(b)** they can be computed efficiently by solving some semi-definite programming (SDP) problem; **(c)** they can be used to certify stability when reinforcement learning is employed in real-world control with either off-policy or on-policy learning [17]. Furthermore, the stability certification can be regarded as an $\mathcal{S}$-procedure, and we analyze its conservatism to show that it is necessary for the robustness of a surrogate system that is closely related to the original system.

[2]This requirement is not difficult to meet in practice, because one can linearize any nonlinear systems around the equilibrium point to obtain a linear component and a nonlinear part.

[3]This stability metric is widely adopted in practice, and is closely related to bounded-input bounded-output (BIBO) stability and absolute stability (or asymptotic stability). For controllable and observable LTI systems, the equivalence can be established.

III. Main results

A. Quadratic constraints on gradient-bounded functions

To begin with, we characterize a set of policies based on Lipschitz constant for the purpose of stability analysis. A given function f is *Lipschitz continuous* if there exists a constant $\xi > 0$ (i.e., Lipschitz constant) such that

$$|f(x) - f(y)| \leq \xi |x - y|, \qquad \forall\, x, y \in \mathbb{R}^n. \tag{6}$$

A function is Lipschitz continuous with parameter ξ if and only if it satisfies the following point-wise quadratic constraint for all $x, y \in \mathbb{R}^n$:

$$\begin{bmatrix} x - y \\ f(x) - f(y) \end{bmatrix}^\top \begin{bmatrix} \xi^2 I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} x - y \\ f(x) - f(y) \end{bmatrix} \geq 0. \tag{7}$$

It can be observed that a Lipschitz continuous function with constant ξ is a member of $\mathcal{P}(\xi)$, where $\overline{\xi}_{ij} = -\underline{\xi}_{ij} = \xi$. Nevertheless, this constraint can be sometimes too conservative, because it does not explore the structure of a given problem. To elaborate on this, consider the function $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$f(x_1, x_2) = \left[\tanh(0.5x_1) - ax_1, \sin(x_2)\right]^\top, \tag{8}$$

where $x_1, x_2 \in \mathbb{R}$ and $|a| \leq 0.1$ is a deterministic but unknown parameter with a bounded magnitude. Clearly, to satisfy (6) on $\mathbb{R}^2$ for all possible tuples (a, x_1, x_2), we need to choose $\xi \geq 1$ (i.e., the function has the Lipshitz constant 1). However, this characterization is too general in this case, because it ignores the *non-homogeneity* of f_1 and f_2, as well as the *sparsity* of the problem representation. Indeed, f_1 only depends on x_1 with its slope restricted to $[-0.1, 0.6]$ for all possible $|a| \leq 0.1$, and f_2 only depends on x_2 with its slope restricted to $[-1, 1]$. In the context of controller design, the non-homogeneity of control outputs often arises from physical constraints and domain knowledge, and the sparsity of control architecture is inherent in scenarios with distributed local information. To explicitly address these requirements, we propose the following quadratic constraint.

Lemma 2 (Function with bounded partial gradients)**.** *For a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ that is differentiable with bounded partial derivatives (i.e., $\underline{\xi}_{ij} \leq \partial_j f_i(x) \leq \overline{\xi}_{ij}$ for all $x \in \mathbb{R}^n$), the following quadratic constraint is satisfied for all $\lambda_{ij} \geq 0$, $i \in [m]$, $j \in [n]$, and $x, y \in \mathbb{R}^n$:*

$$\begin{bmatrix} x - y \\ q(x, y) \end{bmatrix}^\top M(\lambda; \xi) \begin{bmatrix} x - y \\ q(x, y) \end{bmatrix} \geq 0, \tag{9}$$

where $M(\lambda;\xi)$ is given by

$$\begin{bmatrix} \operatorname{diag}\left(\left\{\sum_i \lambda_{ij}(\bar{c}_{ij}^2 - c_{ij}^2)\right\}\right) & U(\{\lambda_{ij}, c_{ij}\})^\top \\ U(\{\lambda_{ij}, c_{ij}\}) & \operatorname{diag}\left(\{-\lambda_{ij}\}\right) \end{bmatrix}, \quad (10)$$

where $\operatorname{diag}(x)$ denotes a diagonal matrix with diagonal entries specified by x, and $q(x,y) = \left[q_{11},\ldots,q_{1n},\ldots,q_{m1},\ldots,q_{mn}\right]^\top$ is determined by x and y, $\{-\lambda_{ij}\}$ is a set of non-negative multipliers that follow the same index order as q, $U(\{\lambda_{ij}, c_{ij}\}) = \begin{bmatrix} \operatorname{diag}\left(\{-\lambda_{1j}c_{1j}\}\right) & \cdots & \operatorname{diag}\left(\{-\lambda_{mj}c_{mj}\}\right) \end{bmatrix} \in \mathbb{R}^{n\times mn}$, $c_{ij} = \frac{1}{2}\left(\underline{\xi}_{ij} + \overline{\xi}_{ij}\right)$, $\bar{c}_{ij} = \overline{\xi}_{ij} - c_{ij}$, and q is related to the output of f by the constraint:

$$f(x) - f(y) = \left[I_m \otimes 1_{1\times n}\right] q = Wq, \quad (11)$$

where $\otimes$ denotes the Kronecker product.

Proof. See Appendix B. □

This above bound is a direct consequence of standard tools in real analysis [18]. To understand this result, it can be observed that (9) is equivalent to:

$$\sum_{i,j} \lambda_{ij}\left((\bar{c}_{ij}^2 - c_{ij}^2)(x_j - y_j)^2 + 2c_{ij}q_{ij}(x_j - y_j) - q_{ij}^2\right) \geq 0 \quad (12)$$

for $\lambda_{ij} \geq 0$ and $f_i(x) - f_i(y) = \sum_{j=1}^n q_{ij}$, where q_{ij} depends on x and y. Since (12) holds for all $\lambda_{ij} \geq 0$, it is equivalent to the condition that $(\bar{c}_{ij}^2 - c_{ij}^2)(x_j - y_j)^2 + 2c_{ij}q_{ij}(x_j - y_j) - q_{ij}^2 \geq 0$ for all $i \in [m]$ and $j \in [n]$, which is a direct result of the bounds imposed on the partial derivatives of f_i. If we apply it to the example function (8), we can specify that $\underline{\xi}_{11} = -0.1, \overline{\xi}_{11} = 0.6, \underline{\xi}_{22} = -1, \overline{\xi}_{22} = 1$, and all the other bounds ($\underline{\xi}_{12}, \overline{\xi}_{12}, \underline{\xi}_{21}, \overline{\xi}_{21}$) are zero. This clearly yields a more informative constraint than merely relying on the Lipschitz constraint (7). In fact, for a differentiable l-Lipschitz function, we have $\overline{\xi}_{ij} = -\underline{\xi}_{ij} = l$, and by limiting the choice of $\lambda_{ij} = \begin{cases} \lambda & \text{if } i = 1 \\ 0 & \text{if } i \neq 1 \end{cases}$, (12) is reduced to (7). However, as illustrated in this example, the quadratic constraint in Lemma 2 can incorporate richer information about the structure of the problem; therefore, it often gives rise to non-trivial stability bounds in practice.

Remarks: The constraint introduced above is not a classical integral quadratic constraint (IQC), since it involves an intermediate variable q that relates to the output f through a set of linear equalities. For stability analysis, let $y = x^* \in \mathcal{B}$ be the equilibrium point, and without loss of generality, assume that $x^* = 0$ and $f(x^*) = 0$. Then, one can define the quadratic functions

$$\phi_{ij}(x,q) = (\bar{c}_{ij}^2 - c_{ij}^2)x_j^2 + 2c_{ij}q_{ij}x_j - q_{ij}^2,$$

and the condition (9) can be written as

$$\sum_{ij} \lambda_{ij}\phi_{ij}(x,q) \geq 0, \qquad \forall \lambda_{ij} \geq 0, \quad (13)$$

which can be used to characterize the set of (x,q) associated with the function f, as we will discuss in Section III-D.

To simplify the mathematical treatment, we have focused on differentiable functions in Lemma 2; nevertheless, the analysis can be extended to non-differentiable but continuous functions (e.g., the ReLU function $\max\{0,x\}$) using the notion of generalized gradient [19, Chap. 2]. In brief, by re-assigning the bounds on partial derivatives to uniform bounds on the set of generalized partial derivatives, the constraint (9) can be directly applied.

In relation to the existing IQCs, this constraint has wider applications for the characterization of gradient-bounded functions (an overview of IQC is provided in Appendix A). The Zames-Falb IQC introduced in [20] has been widely used for single-input single-output (SISO) functions $f : \mathbb{R} \to \mathbb{R}$, but it requires the function to be monotone with the slope restricted to $[\alpha, \beta]$ with $\alpha \geq 0$, i.e., $0 \leq \alpha \leq \frac{f(x)-f(y)}{x-y} \leq \beta$ whenever $x \neq y$. The multi-input multi-output (MIMO) extension holds true only if the nonlinear function $f : \mathbb{R}^n \to \mathbb{R}^n$ is restricted to be the gradient of a convex real-valued function [21], [22]. As for the sector IQC, the scalar version can not be used (because it requires $f_i(x) = 0$ whenever there exists $j \in [n]$ such that $x_j = 0$, which is extremely restrictive), and the vector version is in fact (7). In contrast, the quadratic constraint in Lemma 2 can be applied to non-monotone, vector-valued gradient-bounded functions.

B. Computation of the smoothness margin

With the newly developed quadratic constraint in place, this subsection explains the computation for a smoothness margin of an LTI system G, whose state-space representation is given by:

$$\begin{cases} \dot{x}_G &= Ax_G + Bu \\ w &= \pi(x_G) \\ u &= e + w \end{cases} \quad (14)$$

where $x_G \in \mathbb{R}^{n_s}$ is the state (the dependence on t is omitted for simplicity). The system is assumed to be stable, i.e., A is Hurwitz. We can connect this linear system in feedback with a controller $\pi : \mathbb{R}^{n_s} \to \mathbb{R}^{n_a}$. The signal $e \in \mathbb{R}^{n_a}$ is the exploration vector introduced in reinforcement learning, and $w \in \mathbb{R}^{n_a}$ is the policy action. We are interested in certifying the set of gradient bounds $\xi \in \mathbb{R}^{n_s \times n_a}$ of π such that the interconnected system is input-output stable at all time $T \geq 0$, i.e.,

$$\int_0^T \left|y(t)\right|^2 dt \leq \gamma^2 \int_0^T \left|e(t)\right|^2 dt, \quad (15)$$

where γ is a finite upper bound for the L_2 gain. Let $A \succeq B$ or $A \succ B$ denote that $A - B$ is positive semidefinite or positive definite, respectively. To this end, define the SDP$(P, \lambda, \gamma, \xi)$ as follows:

$$\text{SDP}(P,\lambda,\gamma,\xi) : \begin{bmatrix} O(P,\lambda,\xi) & S(P) \\ S(P)^\top & -\gamma I \end{bmatrix} \prec 0, \quad (16)$$

where $P = P^\top \succeq 0$ and

$$S(P) = \begin{bmatrix} PB \\ 0 \end{bmatrix},$$
$$O(P,\lambda,\xi) = \begin{bmatrix} A^\top P + PA & PBW \\ W^\top B^\top P & 0 \end{bmatrix} + \frac{1}{\gamma}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} + M(\lambda;\xi),$$

where $M(\lambda;\xi)$ is defined in (10) and $W = \begin{bmatrix} I_m \otimes 1_{1\times n} \end{bmatrix}$ is defined in (11). We will show next that the stability of the interconnected system can be certified using linear matrix inequalities.

Theorem 3. *Let G be stable (i.e., A is Hurwitz) and $\pi \in \mathbb{R}^{n_s} \to \mathbb{R}^{n_a}$ be a bounded causal controller. Assume that:*

(i) *the interconnection of G and π is well-posed;*

(ii) *π has bounded partial derivatives on $\mathcal{B}$ (i.e., $\underline{\xi}_{ij} \leq \partial_j \pi_i(x) \leq \overline{\xi}_{ij}$, for all $x \in \mathcal{B}$, $i \in [n_a]$ and $j \in [n_s]$).*

If there exist $P \succeq 0$ and a scalar $\gamma > 0$ such that SDP(P,λ,γ,ξ) is feasible, then the feedback interconnection of G and π is stable (i.e., it satisfies (15)*).*

Proof. To proceed, we multiply $\begin{bmatrix} x_G^\top & q^\top & e^\top \end{bmatrix}^\top$ to the left and its transpose to the right of the augmented matrix in (16), and use the constraints $w = Wq$ and $y = x_G$. Then, SDP(P,λ,γ,ξ) can be written as a dissipation inequality:

$$\dot{V}(x_G) + \begin{bmatrix} x_G \\ q \end{bmatrix}^\top M(\lambda;\xi) \begin{bmatrix} x_G \\ q \end{bmatrix} < \gamma e^\top e - \frac{1}{\gamma} y^\top y,$$

where $V(x_G) = x_G^\top P x_G$ is known as the storage function, and $\dot{V}(\cdot)$ is its derivative with respect to time t. Because the second term is guaranteed to be non-negative by Lemma 2, if SDP(P,λ,γ,ξ) is feasible with a solution (P,λ,γ,ξ), we have:

$$\dot{V}(x_G) + \frac{1}{\gamma} y^\top y - \gamma e^\top e < 0, \tag{17}$$

which is satisfied at all times t. From well-posedness, the above inequality can be integrated from $t = 0$ to $t = T$, and then it follows from $P \succeq 0$ that:

$$\int_0^T |y(t)|^2 dt \leq \gamma^2 \int_0^T |e(t)|^2 dt. \tag{18}$$

Hence, the interconnected system with the RL policy π is stable. □

The above theorem requires that G be stable when there is no feedback policy π. This is automatically satisfied in many physical systems with an existing stabilizing (but not performance-optimizing) controller. In the case that the original system is not stable, one needs to first design a controller to stablize the system, which are well-studied problems in the literature (e.g., H_∞ controller synthesis [16]). Then, the result can be used to ensure stability while delegating reinforcement learning to optimize the performance of the policy under gradient bounds.

We remark that SDP(P,λ,γ,ξ) is quasiconvex, in the sense that it reduces to a standard linear matrix inequality (LMI) with a fixed γ. To solve it numerically, we start with a small γ and gradually increase it until a solution (P,λ) is found. This is repeated for multiple sets of ξ. Each iteration (i.e., LMI for a given set of γ and ξ) can be solved efficiently by interior-point methods. As an alternative to searching on γ for a given ξ, more sophisticated methods for solving the generalized eigenvalue optimization problem can be employed [23].

C. Extension to nonlinear systems with uncertainty

The previous analysis for LTI systems can be extended to a generic nonlinear system described in (3). The key idea is to model the nonlinear and potentially time-varying part $g_t(x(t))$ as an uncertain block with IQC constraints on its behavior. Specifically, consider the LTI component $\underline{G}$:

$$\begin{cases} \dot{x}_G = Ax_G + Bu + v \\ y \quad = x_G \end{cases} \tag{19}$$

where $x_G \in \mathbb{R}^{n_s}$ is the state and $y \in \mathbb{R}^{n_s}$ is the output. The linearized system is assumed to be stable, i.e., A is Hurwitz. The nonlinear part is connected in feedback:

$$\begin{cases} u \ = e + w \\ w = \pi(y) \\ v \ = g_t(y) \end{cases} \tag{20}$$

where $e \in \mathbb{R}^{n_a}$ and $w \in \mathbb{R}^{n_a}$ are defined as before, and $g_t : \mathbb{R}^{n_s} \to \mathbb{R}^{n_s}$ is the nonlinear and time-varying component. In addition to characterizing π using the Lipschitz property as in (9), we assume that $g_t : \mathbb{R}^{n_s} \to \mathbb{R}^{n_s}$ satisfies the IQC defined by (Ψ, M_g) as in Definition 14. The system Ψ has the state-space representation:

$$\begin{cases} \dot{\psi} = A_\psi \psi + B_\psi^v v + B_\psi^y y \\ z \ = C_\psi \psi + D_\psi^v v + D_\psi^y y \end{cases}, \tag{21}$$

where $\psi \in \mathbb{R}^{n_s}$ is the internal state and $z \in \mathbb{R}^{n_z}$ is the filtered output. By denoting $x = \begin{bmatrix} x_G^\top & \psi^\top \end{bmatrix}^\top \in \mathbb{R}^{2n_s}$ as the new state, one can combine (19) and (21) via reducing y and letting $w = Wq$:

$$\begin{cases} \dot{x} = \underbrace{\begin{bmatrix} A & 0 \\ B_\psi^y & A_\psi \end{bmatrix}}_{\underline{A}} x + \underbrace{\begin{bmatrix} B \\ 0 \end{bmatrix}}_{\underline{B}_e} e + \underbrace{\begin{bmatrix} BW \\ 0 \end{bmatrix}}_{\underline{B}_q} q + \underbrace{\begin{bmatrix} I \\ B_\psi^v \end{bmatrix}}_{\underline{B}_v} v \\ z = \underbrace{\begin{bmatrix} D_\psi^y & C_\psi \end{bmatrix}}_{\underline{C}} x + D_\psi^v v \end{cases}, \tag{22}$$

where $\underline{A}$, $\underline{B}_e$, $\underline{B}_q$, $\underline{B}_v$, $\underline{C}$ are matrices of proper dimensions defined above. Similar to the case of LTI systems, the objective is to find the gradient bounds on π such that the system becomes stable in the sense of (15). In the same vein, we define $\underline{\text{SDP}}(P,\lambda,\gamma,\xi)$ as:

$$\underline{\text{SDP}}(P,\lambda,\gamma,\xi) : \begin{bmatrix} O(P,\lambda,\xi) & O_v(P) & S(P) \\ O_v(P)^\top & D_\psi^{v\top} M_q D_\psi^v & 0 \\ S(P)^\top & 0 & -\gamma I \end{bmatrix} \prec 0, \tag{23}$$

where $P \succeq 0$, and

$$O(P,\lambda,\xi) = \begin{bmatrix} \underline{A}^\top P + P\underline{A} & P\underline{B}_q \\ \underline{B}_q^\top P & 0 \end{bmatrix} + \begin{bmatrix} \underline{C}^\top M_g \underline{C} & 0 \\ 0 & 0 \end{bmatrix} + M(\lambda;\xi) + \frac{1}{\gamma}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

$$O_v(P) = \begin{bmatrix} \underline{C}^\top M_q D_\psi^v + P\underline{B}_v \\ 0 \end{bmatrix},\ S(P) = \begin{bmatrix} P\underline{B}_e \\ 0 \end{bmatrix},$$

where $M(\lambda;\xi)$ is defined in (10). The next theorem provides a stability certificate for the nonlinear time-varying system (3).

Theorem 4. *Let $\underline{G}$ be stable (i.e., A in* (19) *is Hurwitz) and $\pi \in \mathbb{R}^{n_s} \to \mathbb{R}^{n_a}$ be a bounded causal controller. Assume that:*

(i) *the interconnection of $\underline{G}$, π, and g_t is well-posed;*

(ii) *π has bounded partial derivatives on $\mathcal{B}$ (i.e., $\underline{\xi}_{ij} \leq \partial_j \pi_i(x) \leq \overline{\xi}_{ij}$ for all $x \in \mathcal{B}$, $i \in [n_a]$ and $j \in [n_s]$);*

(iii) *$g_t \in \mathit{IQC}(\Psi, M_g)$, where Ψ is stable.*

If there exist $P \succeq 0$ and a scalar $\gamma > 0$ such that $\underline{SDP}(P,\lambda,\gamma,\xi)$ in (23) *is feasible, then the feedback interconnection of the nonlinear system* (3) *and π is stable (i.e., it satisfies* (15)*).*

Proof. See Appendix C. □

D. Analysis of conservatism of the stability certificate

We focus on the case where an LTI system G is interconnected with an RL policy $\pi \in \mathcal{P}(\xi)$ (i.e., a function with bounded partial gradients). This corresponds to the system (14) studied in Section III-B. To certify the stability of (14), as will be shown in the next proposition, it suffices to examine the stability of the following system:

$$\begin{cases} \dot{x}_G &= Ax_G + Bu \\ q &= \tilde{\pi}(x_G) \\ w &= Wq \\ u &= e + w \end{cases} \tag{24}$$

where $\tilde{\pi} \in \widetilde{\mathcal{P}}(\xi)$ is a function in the uncertainty set:

$$\left\{ \tilde{\pi} \;\middle|\; \underline{\xi}_{ij} x_j \leq \tilde{\pi}_{ij}(x) \leq \overline{\xi}_{ij} x_j, \forall x \in \mathbb{R}^{n_s}, i \in [n_a], j \in [n_s] \right\}. \tag{25}$$

Proposition 5. *If the system* (24) *is stable for all $\tilde{\pi} \in \widetilde{\mathcal{P}}(\xi)$, then the system* (14) *is stable for all $\pi \in \mathcal{P}(\xi)$.*

Proof. See Appendix D. □

Proposition 5 implies that one potential source of conservatism comes from the decomposition of a gradient-bounded function into a sum of sector-bounded components. Henceforth, we focus the subsequent analysis by examining (24). By considering the state-space representation of $G = \left[\begin{array}{c|cc} A & BW & B \\ \hline I & 0 & 0 \end{array}\right] = \left[\begin{array}{cc} G_{11} & G_{12} \end{array}\right]$, one can write system (24) as:

$$\begin{cases} x_G = \left[\begin{array}{cc} G_{11} & G_{12} \end{array}\right] \begin{bmatrix} q \\ e \end{bmatrix}. \\ q \;\; = \tilde{\pi}(x_G) \end{cases} \tag{26}$$

It is known that the system is input-output stable if and only if $I - G_{11}\tilde{\pi}$ is nonsingular [16]. To understand this, note that if $I - G_{11}\tilde{\pi}$ is nonsingular, then the transfer from e to x_G is given by:

$$e \mapsto x_G = H(e) = (I - G_{11}\tilde{\pi})^{-1} G_{12} e,$$

and $|x_G| = |H(e)| \leq \|(I - G_{11}\tilde{\pi})^{-1} G_{12}\| |e|$. From the previous section (in particular, Lemma 2), we know that if the function π is gradient-bounded, then the set of input/output signals belongs to $\mathcal{S}(\xi)$ given by:

$$\big\{(x,q) \mid \phi_{ij}(x,q) = (\overline{c}_{ij}^2 - c_{ij}^2)x_j^2 + 2c_{ij}q_{ij}x_j - q_{ij}^2 \geq 0, \quad \forall i \in [n_a], j \in [n_s]\big\},$$

where we use $c_{ij} = \frac{1}{2}\left(\underline{\xi}_{ij} + \overline{\xi}_{ij}\right)$, $\overline{c}_{ij} = \overline{\xi}_{ij} - c_{ij}$ for simplicity. We now show that the pair (x,q) belongs to $\mathcal{S}(\xi)$ if and only if there exists a sector-bounded function $\tilde{\pi} \in \tilde{P}(\xi)$ such that it satisfies $q = \tilde{\pi}(x)$.

Lemma 6. *Suppose that $x \in \mathbb{R}^{n_s}$ and $q \in \mathbb{R}^{n_a n_s}$, and $\overline{c}_{ij} \geq 0$ for every $i \in [n_a]$ and $j \in [n_s]$. Then, the pair (x,q) belongs to $\mathcal{S}(\xi)$ if and only if there exists an operator $\tilde{\pi} : \mathbb{R}^n \to \mathbb{R}^{n_a n_s}$, such that $q = \tilde{\pi}(x)$ and $\tilde{\pi}$ satisfies the following conditions:* ***(i)*** *$\tilde{\pi}_{ij}(x) = 0$ if $x_j = 0$, and* ***(ii)*** *$\tilde{\pi}$ is sector bounded, i.e., $(c_{ij} - \overline{c}_{ij})x_j \leq \tilde{\pi}_{ij}(x) \leq (\overline{c}_{ij} + c_{ij})x_j$ holds for all $i \in [n_a]$ and $j \in [n_s]$.*

Proof. See Appendix E. □

By slightly overloading the notations, we can extend the result of the previous lemma from static mapping to the case that $x \in L^{n_s}$ and $q \in L^{n_a * n_s}$ with the operator $\tilde{\pi} : L^{n_s} \to L^{n_a * n_s}$. Specifically, by defining $\phi_{ij}(x,q)$ in $\mathcal{S}(\xi)$ as

$$\phi_{ij}(x,q) = (\overline{c}_{ij}^2 - c_{ij}^2)\|x_j\|^2 + 2c_{ij}\langle q_{ij}, x_j\rangle - \|q_{ij}\|^2, \tag{27}$$

we can then extend the definition of $\mathcal{S}(\xi)$ to this space accordingly.

Lemma 7. *Suppose that $x \in L^{n_s}$ and $q \in L^{n_a n_s}$, and that $\overline{c}_{ij} \geq 0$ for all $i \in [n_a]$ and $j \in [n_s]$. Then, the pair (x,q) belongs to $\mathcal{S}(\xi)$ if and only if there exists an operator $\tilde{\pi} : L^{n_s} \to L^{n_a n_s}$ such that $q = \tilde{\pi}(x)$ and $\tilde{\pi}$ satisfies the following conditions:* ***(i)*** *$\tilde{\pi}_{ij}(x) = 0$ if $x_j = 0$, and* ***(ii)*** *$\tilde{\pi}$ is sector bounded, i.e., $(c_{ij} - \overline{c}_{ij})\|x_j\| \leq \big\|\tilde{\pi}_{ij}(x)\big\| \leq (\overline{c}_{ij} + c_{ij})\|x_j\|$ for all $i \in [m]$ and $j \in [n]$.*

Proof. See the Appendix F. □

The previous result indicates that the input and output pair of $\tilde{\pi}$ can be described by $\mathcal{S}(\xi)$. We show next that this set should be separated from the signal space of the dynamical system in order to ensure robust stability.

Lemma 8. *If $(G,\tilde{\pi})$ is robustly stable, then there cannot exist a nonzero $q \in L_2$ such that $x = Gq$ and $(x,q) \in \mathcal{S}(\xi)$.*

Proof. We prove this lemma by contraposition. If there exists a nonzero $q \in L_2$ such that $(x,q) \in \mathcal{S}(\xi)$, then it follows from Lemma 7 that there exists a linear operator $\tilde{\pi}$ such that $q = \tilde{\pi}(x) = \tilde{\pi}(Gq)$. This implies that the operator $(I - \tilde{\pi}G)$ is singular, and therefore, $(I - G\tilde{\pi})$ is singular, implying that the interconnected system is not robustly stable. □

The path of examining the necessity of the SDP condition (16) has become clear. Consider the set generated by the LTI system:

$$\Psi = \left\{(\phi_{ij}(x,q) : q \in L^{n_a n_s}, \|q\| = 1, x = Gq\right\}, \quad (28)$$

and the positive orthant

$$\Pi = \left\{(r_{ij}) \in \mathbb{R}^{n_a n_s} : r_{ij} \geq 0, \quad \forall i \in [n_a], j \in [n_s]\right\}. \quad (29)$$

Lemma 8 implies that the two sets Ψ and Π are separated if $(G, \tilde{\pi})$ is robustly stable. The goal is to show that there exists a separating hyperplane, whose parameters are related to the solution of (16). For simplicity, define the matrices $\Omega_{ij,x} = \text{diag}\left(\left\{\{\overline{c}_{ij}^2 - c_{ij}^2\right\}\right)$, $\Omega_{ij,q}$ and $\Omega_{ij,xq}$ with their (k,l)-th elements $[\Omega_{ij,q}]_{kl} = \begin{cases} 1 & \text{if } k = in + j \\ 0 & \text{otherwise} \end{cases}$, and $[\Omega_{ij,xq}]_{kl} = \begin{cases} c_{ij} & \text{if } k = j, l = in + j \\ 0 & \text{otherwise} \end{cases}$. To write $\phi_{ij}(x = Gq, q)$ as an inner product, define

$$T_{ij} = G^* \Omega_{ij,x} G - \Omega_{ij,q} + G^* \Omega_{ij,xq}^* + \Omega_{ij,xq} G.$$

It results from the definition (27) that

$$\begin{aligned} \phi_{ij}(x = Gq, q) &= \|Gq\|_{\Omega_{ij,x}}^2 + 2\text{Re}\left\langle \Omega_{ij,xq} Gq, q\right\rangle - \|q\|_{\Omega_{ij,q}}^2 \\ &= \left\langle q, T_{ij} q\right\rangle. \end{aligned}$$

Lemma 9. *For a given linear time-invariant operator G, the closure $\overline{\Psi}$ of Ψ defined in* (28) *is convex.*

Proof. See Appendix G. □

Now, we show that strict separation occurs when the system is robustly stable.

Lemma 10. *Suppose that $I - G\tilde{\pi}$ is nonsingular. Then, the sets Π and Ψ are strictly separated, namely $D(\Pi, \Psi) = \inf_{r \in \Pi, y \in \Psi} |r - y| > 0$.*

To prove this result, we need the following lemma.

Lemma 11. *Suppose that $D(\Pi, \Psi) = \inf_{r \in \Pi, y \in \Psi} |r - y| = 0$. Given any $\epsilon > 0$ and $t_0 \geq 0$, there exist a closed interval $[t_0, t_1]$ and two signals $x \in L^{n_s}$ and $q \in L^{n_a n_s}$ with $\|q\| = 1$ such that*

$$\phi_{ij}(x,q) \geq 0, \qquad \forall i \in [n_a], j \in [n_s] \quad (30)$$

$$\epsilon^2 > \|(I - \Gamma_{[t_0,t_1]}) Gq\| \quad (31)$$

$$\epsilon = \|x - \Gamma_{[t_0,t_1]} Gq\|_{\Omega_{ij,x}}, \quad (32)$$

where $\|q\|_\Omega = \sqrt{q^* \Omega q}$ is the scaled norm and $\Gamma_{[t_0,t_1]}$ projects the signal onto the support of $[t_0, t_1]$. With the above choice of q, x and $[t_0, t_1]$, there exists an operator $\tilde{\pi} \in \tilde{P}(\xi)$ such that $\|(I - \tilde{\pi}\Gamma_{[t_0,t_1]} G)q\| \leq C\epsilon$ for some constant $C > 0$ that depends on the sector bounds ξ.

Proof. See Appendix H. □

We are now ready to prove the strict separation result.

Proof of Lemma 10. Assume that $D(\Pi, \Psi) = \inf_{r \in \Pi, y \in \Psi} |r - y| = 0$. Consider a sequence $\epsilon_n \to 0$ as n tends to ∞. For each ϵ_n, construct signals $q^{(n)}$ with a bounded support on $[t_n, t_{n+1}]$, and $\tilde{\pi}^{(n)}$ according to Lemma 11. Define

$$\tilde{\pi} = \sum_{n=1}^{\infty} \tilde{\pi}^{(n)} \Gamma_{[t_n, t_{n+1}]}.$$

We have

$$\tilde{\pi} G q^{(n)} = \tilde{\pi}^{(n)} \Gamma_{[t_n,t_{n+1}]} G q^{(n)} + \tilde{\pi}(I - \Gamma_{[t_n,t_{n+1}]}) G q^{(n)},$$

and

$$\begin{aligned} \|(I - \pi G) q^{(n)}\| &\leq \|(I - \tilde{\pi}^{(n)} \Gamma_{[t_n,t_{n+1}]} G) q^{(n)}\| \\ &\quad + \|(I - \Gamma_{[t_n,t_{n+1}]}) G q^{(n)}\| \\ &\leq C\epsilon_n + \epsilon_n^2 \end{aligned}$$

Because $\epsilon_n \to 0$, the right-hand side approaches 0, and so does the left-hand side. Therefore, since $\|q^{(n)}\| = 1$, the mapping $I - \tilde{\pi} G$ cannot be invertible, which contradicts the robust stability assumption. This implies that Π and Ψ are strictly separable. □

To draw the connection to the SDP problem (16), observe that

$$\phi_{ij}(x,q) = \left\langle \begin{bmatrix} x \\ q \end{bmatrix}, M_\pi^{ij} \begin{bmatrix} x \\ q \end{bmatrix} \right\rangle, \quad (33)$$

where

$$[M_\pi^{ij}]_{kl} = \begin{cases} \overline{c}_{ij}^2 - c_{ij}^2 & (k,l) = (j,j) \\ c_{ij} & (k,l) = (i, i*n+j) \text{ or } (i*n+j, i) \\ -1 & (k,l) = (i*n+j, i*n+j) \end{cases},$$

and $M(\lambda; \xi) = \sum_{i \in [n_a], j \in [n_s]} \lambda_{ij} M_\pi^{ij}$ as defined in Lemma 2.

Proposition 12. *The SDP condition* (16) *is feasible if and only if there exist multipliers $\lambda_{ij} \geq 0$ and $\epsilon > 0$ such that*

$$\sum_{i \in [n_a], j \in [n_s]} \lambda_{ij} \phi_{ij}(x,q) \leq -\epsilon \|q\|^2 \quad (34)$$

for all $q \in L^{n_a n_s}$ and $x = Gq$.

Proof. Since $\phi_{ij}(x,q) = \left\langle \begin{bmatrix} x \\ q \end{bmatrix}, M_\pi^{ij} \begin{bmatrix} x \\ q \end{bmatrix} \right\rangle$, the condition (34) is equivalent to

$$\begin{bmatrix} G \\ I \end{bmatrix}^* M(\lambda; \xi) \begin{bmatrix} G \\ I \end{bmatrix} \prec 0. \quad (35)$$

By the KYP lemma, this is equivalent to the existence of $P \succeq 0$ such that:

$$\begin{bmatrix} A^\top P + PA & PBW \\ W^\top B^\top P & 0 \end{bmatrix} + M(\lambda; \xi) \prec 0. \quad (36)$$

By Schur complement, P satisfies the KYP condition if and only if it satisfies (16). Thus, the claim is shown. □

Theorem 13. *Let $\tilde{\pi} : L^{n_s} \to L^{n_a n_s}$ be a bounded causal controller such that $\tilde{\pi} \in \tilde{\mathcal{P}}(\xi)$. Assume that the interconnection of G and $\tilde{\pi}$ is well-posed. Then, the input-output stability of the feedback interconnection of system* (24) *implies that there exist $P \succeq 0$, $\gamma > 0$ and $\lambda \geq 0$ such that SDP$(P, \lambda, \gamma, \xi)$ in* (16) *is feasible.*

Proof. See Appendix I. □

E. Smoothness regularization

We propose two different smoothness penalties to be added to the objective function (1) to empirically improve learning performance on physical dynamical systems. Specifically, we use

$$L_{\text{explore}} = \sum_{t=1}^{T} \|u_{t-1} - \pi_\theta(x_t)\|^2 \tag{37}$$

as a regularization term to induce consistency during exploration. The intuition is that since the change in states between two consecutive time steps is often small, it is desirable to ensure small changes in output actions. This is closely related to another penalty term

$$L_{\text{smooth}} = \sum_{t=1}^{T} \left\| \frac{\partial}{\partial \theta} \pi_\theta(x_t) \right\|^2, \tag{38}$$

which penalizes the gradient of the policy along the trajectories. This penalty has been used in [24], which is termed "double backpropagation", and recently rediscovered in [25], [26] for image classification. Since bounded gradients lead to bounded Lipshitz constant, these penalties will induce smooth neural network functions, which is essential to ensure generalizability and, as we will show, stability. In addition, we incorporate a hard threshold (HT) approach that rescales the weight matrices at each layer by $(l^\circ / l(\pi_\theta))^{1/n_L}$ if $l(\pi_\theta) > l^\circ$, where $l(\pi_\theta)$ is the Lipschitz constant of the neural network π_θ, n_L is the number of layers of the neural network and l° is the certified Lipschitz constant. This ensures that the Lipschitz constant of the RL policy remains bounded by l°.

IV. Related work

This work is closely related to the body of works on *safe reinforcement learning*, defined as the process of learning policies that maximize performance in problems where safety is required during the learning and/or deployment [8]. A detailed literature review can be found in [8], which has categorized two main approaches by modifying: **(i)** the optimality condition with a safety factor, and **(ii)** the exploration process to incorporate external knowledge or risk metrics. Risk-aversion can be specified in the reward function, for example, by defining risk as the probability of reaching a set of unknown states in a discrete Markov decision process setting [27], [28]. Robust markov decision process (MDP) is designed to maximize rewards while safely exploring the discrete state space [29], [30]. For continuous states and actions, robust model predictive control can be employed to ensure robustness and safety constraints for the learned model with bounded errrors [31]. These methods require an accurate or estimated models for policy learning. Recently, model-free policy optimization has been successfully demonstrated in real-world tasks such as robotics, business management, smart grid and transportation [32]. Safety requirement is high in these settings. Existing approaches are based on constraint satisfaction that holds with high probability [33], [34].

The present analysis tackles the safe reinforcement learning problem from a robust control perspective, which is aimed at providing theoretical guarantees for stability [15]. Lyapunov functions are widely used to analyze and verify stability when the system and its controller are known [35], [36]. For nonlinear systems without global convergence guarantees, region of convergence is often estimated, where any state trajectory that starts within this region stays within the region for all times and converges to a target state eventually [37]. For example, recently, [38] has proposed a learning-based Lyapunov stability verification for physical systems, whose dynamics are sequentially estimated by Gaussian processes. In the same vein, [39] has employed reachability analysis to construct safe regions in the state space by solving a partial differential equation. The main challenge of these methods is to find a suitable non-conservative Lyapunov function to conduct the analysis.

The IQC framework proposed in [40] has been widely used to analyze the stability of large-scale complex systems such as aircraft control [41]. The main advantages of IQC are its computational efficiency, non-conservatism, and unified treatment of a variety of nonlinearities and uncertainties. It has also been employed to analyze the stability of small-sized neural networks in reinforcement learning [42], [43]; however, in their analysis, the exact coefficients of the neural network need to be known a priori for the static stability analysis, and a region of safe coefficients needs to be calculated at each iteration for the dynamic stability analysis. This is computationally intensive, and it quickly becomes intractable when the neural network size grows. On the contrary, because the present analysis is based on a broad characterization of control functions with bounded gradients, it does not need to access the coefficients of the neural network (or any forms of the controller). In general, robust analysis using methods like structured singular value [44] and IQC can be conservative. There are only few cases where the necessity conditions can be established, such as when the uncertain operator has a block diagonal structure of bounded singular values [16], but this set of uncertainties is much smaller than the set of performance-oriented controllers learned by RL. To this end, we are able to reduce conservatism of the results by introducing more informative quadratic constraints for those controllers, and analyze the necessity of the certificate criteria. This significantly extends the possibilities of stability-certified reinforcement learning to large and deep neural networks in nonlinear large-scale real-world systems, whose stability is otherwise impossible to be certified using existing approaches.

V. Case studies

In this section, we empirically study the stability-certified reinforcement learning in real-world problems such as flight formation [45] and power grid frequency regulation [46]. Designing an optimal controller for these systems is challenging, because they consist of interconnected subsystems that have limited information sharing, and also their underlying models are typically nonlinear and even time-varying and uncertain. Indeed, for the case of decentralized control, which aims at designing a set of local controllers whose interactions are specified by physical and informational structures, it has been long known that it amounts to an NP-hard optimization problem in general [14]. End-to-end reinforcement learning comes in handy, because it does not require model information by simply interacting with the environment while collecting rewards.

In a multi-agent setting, each agent explores and learns its own policy independently without knowing about other agents' policies [47]. For the simplicity of implementation, we consider the synchronous and cooperative scenario, where agents conduct an action at each time step and observe the reward for the whole system. Their goal is to collectively maximize the rewards (or minimize the costs) shared equally among them. The present analysis aims at offering safety certificates of existing RL algorithms when applied to real-world dynamical systems, by simply monitoring the gradients information of the neural network policy. This is orthogonal to the line of research that aims at improving the performance of the existing RL algorithms. The examples are taken from [45], [46], [48], but we deal directly with the underlying *nonlinear* physics rather than a linearized model.

A. Multi-agent flight formation

Consider the multi-agent flight formation problem [45], where each agent can only observe the relative distance from its neighbors, as illustrated in Fig. 2. The goal is to design a local controller for each aircraft such that a predefined pattern is formed as efficiently as possible.

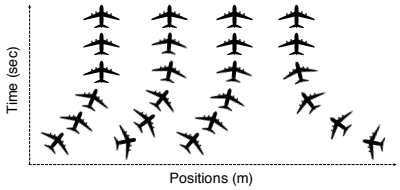


Fig. 2: Illustration of the multi-agent flight formation problem.

The physical model[4] for each aircraft is given by:

$$\ddot{z}^i(t) = u^i(t)$$
$$\ddot{\theta}^i(t) = \frac{1}{\delta}\left(\sin\theta^i(t) + u^i(t)\right),$$

where z^i, θ^i and u^i denote the horizontal position, angle and control input of aircraft i, respectively, and $\delta > 0$ characterizes the physical coupling of rolling moment and lateral acceleration. We consider 10 aircrafts in the experiments.

One particular strength of RL is that the reward function can be highly nonconvex, nonlinear, and arbitrarily designed; however, since quadratic costs are widely used in the control literature, consider the case $r(x(t), u(t)) = x(t)^\top Q x(t) + u(t)^\top R u(t)$. For the following experiments, assume that $Q = 1000 \times I_{15}$ and $R = I_4$. In addition, because the original system A has its largest eigenvalue at 0, we need a nominal static distributed controller K_d, whose primary goal is to make the largest eigenvalue of $A + BK_n$ negative. Such controller could be designed using methods such as robust control synthesis for the linearized system [15], [16].

The task for multi-agent RL is to learn the controller $u^i(t)$, which only takes inputs of the relative distances of agent i to its neighbors. For example, agent 1 can only observe $z^1(t) - z^2(t) - d$ (i.e., the 1st entry of $x(t)$); similarly, agent 2 can only observe $z^1(t) - z^2(t) - d$ and $z^2(t) - z^3(t) - d$ (i.e., the 1st and 5th entries of $x(t)$).

Stability certificate: To obtain the stability certificate, we apply the method in Section III-C. The nonzero entries of the nonlinear component $g(x(t))$ are in the form of $\sin(\theta) - \theta$, which can be treated as an uncertainty block with the slope restricted to $[-1, 0]$ for $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$; therefore, the Zames-Falb IQCs can be employed to construct (21) [20], [49]. As for the RL agents u^i, their gradient bounds can be certified according to Theorem 4. Specifically, we assume that each agent u^i is l-Lipschitz continuous, and solve (23) for a given set of γ and l. The certified gradient bounds (Lipschitz constants) are plotted in Fig. 3 using different constraints. The conservative L_2 constraint (7) is only able to certify stability for Lipschitz constants up to 0.8. By incorporating the sparsity of distributed controller, we can increase the margin to 1.2, which is satisfied throughout the learning process.

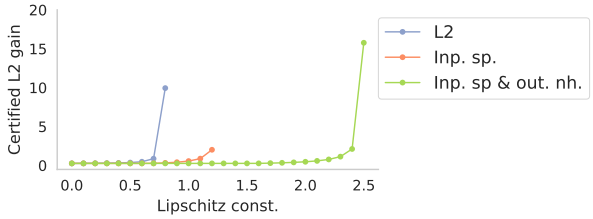


Fig. 3: Stability-certified Lipschitz constants obtained by the standard L_2 bound (L2) in (7) and the method proposed in Lemma 2, which considers input sparsity (inp. sp.) and output non-homogeneity (out. nh.).

In order to further increase the set of certifiable stable controllers, we monitor the partial gradient information for each agent and encode them as non-homogeneous gradient bounds. For instance, if $\frac{\partial \pi_i(x)}{\partial x_j}$ has been consistently positive for latest iterations, we will set $\overline{\xi}_{ij} = l$ and $\underline{\xi}_{ij} = -\epsilon l$, where $\epsilon > 0$ is a small margin, such as 0.1, to allow explorations. By performing this during learning, it would be possible to significantly enlarge the certified Lipschitz bound to up to 2.5, as shown in Fig. 3.

Policy gradient RL: To perform multi-agent reinforcement learning, we employ trust region policy optimization with natural gradients and smoothness policies. During learning, we employ the hard-thresholding step introduced in Section III-E to ensure that the gradient bounds are satisfied. The trajectories of rewards averaged over three independent experiments are shown in Fig. 4. In this example, agents with a one hidden layer neural network (each with 5 hidden units) can learn most efficiently when employed with the smoothness penalties. Without the guidance of these penalties, the linear controller and 1-layer neural network apparently cannot effectively explore the parameter space.

The learned 5-layer neural network policy is employed in an actual control task, as shown in Fig. 5. Compared to the nominal controller, the flights can be maneuvered more efficiently. In terms of the actual cost, the RL agents achieve the cost 41.0, which is about 30% lower than that of the nominal controller

[4]The cosine term in the original formulation is omitted for simplicity, though it can be incorporated in a more comprehensive treatment.

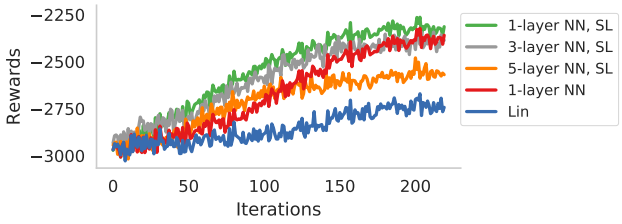
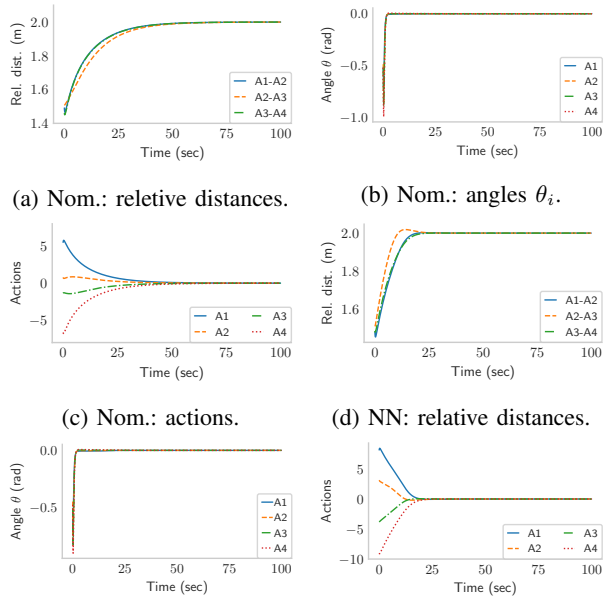


Fig. 4: Learning performance of different control structures (1-layer neural network, 5-layer neural network, and linear controller). By the inclusion of a smoothness loss (SL), the exploration becomes more effective.

(58.3). This result can be examined both in the actual state-action trajectories in Fig. 5 or the control behaviors in Fig. 6. The results indicate that RL is able to improve a given controller when the underlying system is nonlinear and unknown.

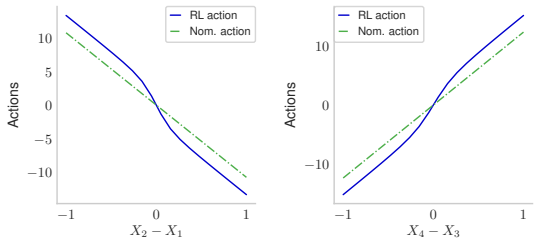


(a) Nom.: reletive distances. (b) Nom.: angles θ_i.

(c) Nom.: actions. (d) NN: relative distances.

(e) NN: angles θ_i. (f) NN: actions.

Fig. 5: State and action trajectories in a typical contral task, where the nominal controller (Nom) and the RL agents achieve costs of 58.3 and 41.0, respectively.



(a) Controller of Agent 1. (b) Controller of Agent 4.

Fig. 6: Demonstration of control outputs for the nominal action and RL agents.

B. Power system frequency regulation

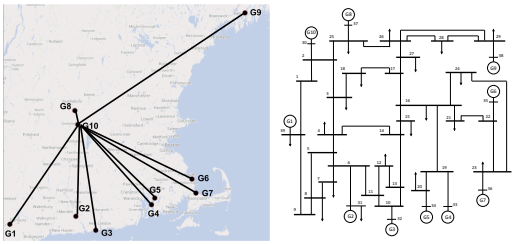
In this case study, we focus on the problem of distributed control for power system frequency regulation [46]. The IEEE 39-Bus New England Power System under analysis is shown in Fig. 7. In a distributed control setting, each generator can only share its rotor angle and frequency information with a pre-specified set of counterparts that are geographically distributed. The main goal is to optimally adjust the mechanical power input to each generator such that the phase and frequency at each bus can be restored to their nominal values after a possible perturbation. Let θ_i denote the voltage angle at a generator bus i (in rad). The physics of power systems are modeled by the per-unit swing equation:

$$m_i\ddot{\theta}_i + d_i\dot{\theta} = p_{m_i} - p_{e_i} \tag{39}$$

where p_{m_i} is the mechanical power input to the generator at bus i (in p.u.), p_{e_i} is the electrical active power injection at bus i (in p.u.), m_i is the inertia coefficient of the generator at bus i (in p.u.-sec^2/rad), and d_i is the damping coefficient of the generator at bus i (in p.u.-sec/rad). The electrical real power injection p_{e_i} depends on the voltage angle difference in a nonlinear way, as governed by the AC power flow equation:

$$p_{e_i} = \sum_{j=1}^{n} |v_i||v_j| \left(g_{ij}\cos(\theta_i - \theta_j) + b_{ij}\sin(\theta_i - \theta_j)\right) \tag{40}$$

where n is the number of buses in the system, g_{ij} and b_{ij} are the conductance and susceptance of the transmission line that connects buses i and j, v_i is the voltage phasor at bus i, and $|v_i|$ is its voltage magnitude. Because the conductance g_{ij} is typically several orders of magnitude smaller than the susceptance b_{ij}, for the simplicity of mathematical treatment, we omit the cosine term and only keep the sine term that accounts for the majority of nonlinearity. Each generator needs to make decisions on the value of the mechanical power p_{m_i} to inject in order to maintain the stability of the power system.



(a) Star-connected structure. (b) IEEE 39-bus system.

Fig. 7: Illustration of the frequency regulation problem for the New England power system. The communication among generators follows a star topology.

Let the rotor angles and the frequency states be denoted as $\theta = \begin{bmatrix}\theta_1 & \cdots & \theta_n\end{bmatrix}^\top$ and $\omega = \begin{bmatrix}\omega_1 & \cdots & \omega_n\end{bmatrix}^\top$, and the generator mechanical power injections be denoted as $p_m = \begin{bmatrix}p_{m_1} & \cdots & p_{m_n}\end{bmatrix}^\top$. Then, the state-space representation of the nonlinear system is given by:

$$\begin{bmatrix}\dot{\theta}\\ \dot{\omega}\end{bmatrix} = \underbrace{\begin{bmatrix}0 & I\\ -M^{-1}L & -M^{-1}D\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}\theta\\ \omega\end{bmatrix}}_{x} + \underbrace{\begin{bmatrix}0\\ M^{-1}\end{bmatrix}}_{B} p_m + \underbrace{\begin{bmatrix}\mathbf{0}\\ g(\theta)\end{bmatrix}}_{g(x)}$$

where $g(\theta) = \begin{bmatrix} g_1(\theta) & \cdots & g_n(\theta) \end{bmatrix}^\top$ with $g_i(\theta) = \sum_{j=1}^n \frac{b_{ij}}{m_j}\big((\theta_i - \theta_j) - \sin(\theta_i - \theta_j)\big)$, and $M = \text{diag}\left(\{m_i\}_{i=1}^n\right)$, $D = \text{diag}\left(\{d_i\}_{i=1}^n\right)$, and L is a Laplacian matrix whose entries are specified in [46, Sec. IV-B]. For linearization (also known as DC approximation), the nonlinear part $g(x)$ is assumed to be zero when the phase differences are small [46], [48]. On the contrary, we deal with this term in the stability certification to demonstrate its capability of producing non-conservative results even for nonlinear systems. Similar to the flight formation case, we assume that there exists a distributed nominal controller that stablizes the system. To conduct multi-agent RL, each controller p_{m_i} is a neural network that takes the available phases and frequencies as the input and determines the mechanical power injection at bus i. The main focus is to study the certified-gradient bounds for each agent policy in this large-scale setting.

Stability certificate: Similar to the flight formation problem, the nonlinearities in $\mathbf{g}(x)$ are in the form of $\Delta\theta_{ij} - \sin\Delta\theta_{ij}$, where $\Delta\theta_{ij} = \theta_i - \theta_j$ represents the phase difference, which has its slope restricted to $[0, 1 - \cos(\overline{\theta})]$ for every $\Delta\theta_{ij} \in [-\overline{\theta}, \overline{\theta}]$ and thus can be treated using the Zames-Falb IQC. In the smoothness margin analysis, assume that $\overline{\theta} = \frac{\pi}{3}$, which requires the phase angle difference to be within $[-\frac{\pi}{3}, \frac{\pi}{3}]$. This is a large set of uncertainties that includes both normal and abnormal operational conditions. To study the stability of the multi-agent policies, we adopt a black-box approach by simply considering the input-output constraint. By simply applying the L_2 constraint in (7), we can only certify stability for Lipschitz constants up to 0.4, as shown in Fig. 8. Because the distributed control is sparse, we can leverage it by setting the lower and upper bounds $\underline{\xi}_{ij} = \overline{\xi}_{ij} = 0$ for each agent i that does not utilize observation j, and $\overline{\xi}_{ij} = -\underline{\xi}_{ij} = l$ otherwise, where l is the Lipschitz constant to be certified. This information can be encoded in $\underline{\text{SDP}}(P, \lambda, \gamma, \xi)$ in (23), which can be solved for L up to 0.6 (doubling the certificate provided by the L_2 constraint).

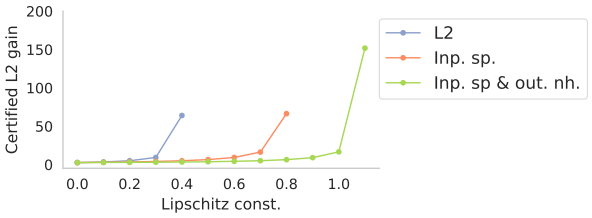
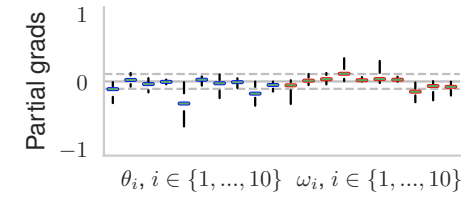
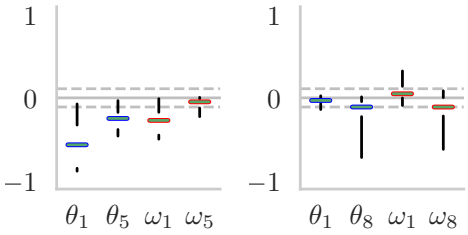


Fig. 8: Certified Lipschitz constants for the power system regulation task.

Due to the problem nature, we further observe that for each agent, the partial gradient of the policy with respect to certain observations is primarily one-sided, as shown in Fig. 9. With a band of ± 0.1, the partial gradients remain within either $[-0.1, 1]$ or $[-1, 0.1]$ throughout the learning process. This information is gleaned during the learning phase, and we can incorporate it into the partial gradient bounds (e.g., $\overline{\xi}_{ij} = -0.1l$ and $\underline{\xi}_{ij} = l$ for agent i which exhibits positive gradient with respect to observation j) to extend the certificate up to 1.1.

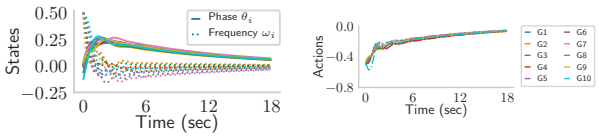


(a) G10.



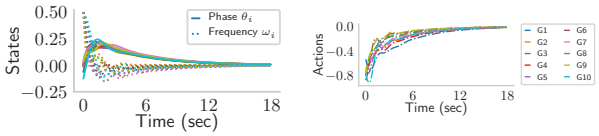
(b) G4. (c) G7.

Fig. 9: Box plots of partial gradients of individual generators (G10, G4, G7) with respect to local information. Grey dashed lines indicate ± 0.1.



(a) Nom.: θ_i and ω_i.

(b) Nom.: actions.



(c) NN: θ_i and ω_i.

(d) NN: actions.

Fig. 10: State and action trajectories of the nominal and neural network controllers for power system frequency regulation, with costs of 50.8 and 23.9, respectively.

Policy gradient RL: Similar to the flight formation task, we perform multi-agent policy gradient RL. The learned neural network controller is implemented in a typical control case, whose trajectories are shown in Fig. 10. As can be seen, the RL policies can regulate the frequencies more efficiently than the nominal controller, with a significantly lowered cost (50.8 vs. 23.9). More importantly, we compare the cases of RL with and without regulating the Lipschitz constants in Fig. 11. Without regulating the gradients, the RL is able to reach a performance slightly higher than its stability-certified counterpart. However, after about iteration 500, the performance starts to deteriorate (due to a possibly large gradient variance and high sensitivity to step size) until it completely loses the previous gains and starts to break the system. This intolerable behavior is due to the large Lipschitz gains that grow unboundedly, as shown in Fig. 12. In comparison, RL with regulated gradient bounds is able to make a substantial improvement, and also exhibits a

more stable behavior.

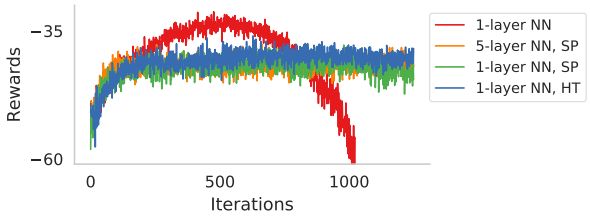
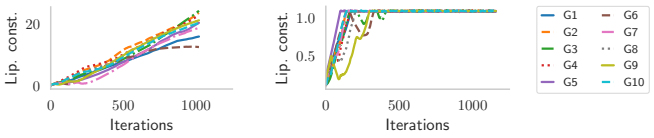


Fig. 11: Long-term performance of RL for agents with regulated gradients by soft penalty (SP) and hard thresholding (HT). The RL agents without regulating the gradients exhibit "dangerous" behaviors in the long run.



(a) No gradient regulation. (b) Gradient regulation.

Fig. 12: Trajectories of Lipschitz constants with and without regulation.

VI. Conclusions

In this paper, we focused on the challenging task of ensuring the stability of reinforcement learning in real-world dynamical systems. By solving the proposed SDP feasibility problem, we offered a preventative certificate of stability for a broad class of neural network controllers with bounded gradients. Furthermore, we analyzed the (non)conservatism of the certificate, which was demonstrated in the empirical investigation of decentralized nonlinear control tasks, including multi-agent flight formation and power grid frequency regulation. Results indicated that the set of stability-certified controllers was significantly larger than what the existing approaches could offer, and that the RL agents would substantially improve the performance of nominal controllers while staying within the safe set. Most importantly, regulation of gradient bounds was able to improve on-policy learning stability and avoid "catastropic" effects caused by the unregulated high gains. The present study represents a key step towards safe deployment of reinforcement learning in mission-critical real-world systems.

References

[1] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[2] S. M. Kakade, "A natural policy gradient," in *Advances in neural information processing systems*, 2002, pp. 1531–1538.

[3] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. of the International Conference on Machine Learning*, 2015, pp. 1889–1897.

[4] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[6] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[7] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. of the International Conference on Machine Learning*, 2016, pp. 1928–1937.

[8] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[10] I. Stoica, D. Song, R. A. Popa, D. Patterson, M. W. Mahoney, R. Katz, A. D. Joseph, M. Jordan, J. M. Hellerstein, J. E. Gonzalez *et al.*, "A Berkeley view of systems challenges for AI," *arXiv preprint arXiv:1712.05855*, 2017.

[11] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[12] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *Proc. of AAAI*, 2018.

[13] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 854–863.

[14] L. Bakule, "Decentralized control: An overview," *Annual reviews in control*, vol. 32, no. 1, pp. 87–98, 2008.

[15] K. Zhou, J. C. Doyle, K. Glover *et al.*, *Robust and optimal control*. Prentice hall New Jersey, 1996, vol. 40.

[16] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. Springer Science & Business Media, 2013, vol. 36.

[17] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[18] A. Zemouche and M. Boutayeb, "On LMI conditions to design observers for lipschitz nonlinear systems," *Automatica*, vol. 49, no. 2, pp. 585–591, 2013.

[19] F. H. Clarke, *Optimization and nonsmooth analysis*. SIAM, 1990, vol. 5.

[20] G. Zames and P. Falb, "Stability conditions for systems with monotone and slope-restricted nonlinearities," *SIAM Journal on Control*, vol. 6, no. 1, pp. 89–108, 1968.

[21] M. G. Safonov and V. V. Kulkarni, "Zames-Falb multipliers for MIMO nonlinearities," in *Proc. of the American Control Conference*, vol. 6, 2000, pp. 4144–4148.

[22] W. P. Heath and A. G. Wills, "Zames-Falb multipliers for quadratic programming," in *Proc. of the IEEE Conference on Decision and Control*, 2005, pp. 963–968.

[23] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994, vol. 15.

[24] H. Drucker and Y. Le Cun, "Improving generalization performance using double backpropagation," *IEEE Transactions on Neural Networks*, vol. 3, no. 6, pp. 991–997, 1992.

[25] A. G. Ororbia II, D. Kifer, and C. L. Giles, "Unifying adversarial training algorithms with data gradient regularization," *Neural computation*, vol. 29, no. 4, pp. 867–887, 2017.

[26] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[27] S. P. Coraluppi and S. I. Marcus, "Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes," *Automatica*, vol. 35, no. 2, pp. 301–309, 1999.

[28] P. Geibel and F. Wysotzki, "Risk-sensitive reinforcement learning applied to control under constraints," *Journal of Artificial Intelligence Research*, vol. 24, pp. 81–108, 2005.

[29] T. M. Moldovan and P. Abbeel, "Safe exploration in Markov decision processes," in *Proc. of the International Conference on Machine Learning*, 2012, pp. 1451–1458.

[30] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust Markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[31] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, pp. 1216–1226, 2013.

[32] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[33] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with Gaussian processes," in *Proc. of the International Conference on Machine Learning*, 2015, pp. 997–1005.

[34] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. of the International Conference on Machine Learning*, 2017, pp. 22–31.

[35] T. J. Perkins and A. G. Barto, "Lyapunov design for safe reinforcement learning," *Journal of Machine Learning Research*, vol. 3, no. Dec, pp. 803–832, 2002.

[36] R. Bobiti and M. Lazar, "A sampling approach to finding lyapunov functions for nonlinear discrete-time systems," in *Proc. of the IEEE European Control Conference*, 2016, pp. 561–566.

[37] H. K. Khalil, "Noninear systems," *Prentice-Hall, New Jersey*, vol. 2, no. 5, pp. 5–1, 1996.

[38] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in Neural Information Processing Systems*, 2017, pp. 908–919.

[39] A. K. Akametalu, J. F. Fisac, J. H. Gillula, S. Kaynama, M. N. Zeilinger, and C. J. Tomlin, "Reachability-based safe learning with Gaussian processes," in *Proc. of the IEEE Conference on Decision and Control*, 2014, pp. 1424–1431.

[40] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Transactions on Automatic Control*, vol. 42, no. 6, pp. 819–830, 1997.

[41] J. M. Fry, M. Farhood, and P. Seiler, "IQC-based robustness analysis of discrete-time linear time-varying systems," *International Journal of Robust and Nonlinear Control*, vol. 27, no. 16, pp. 3135–3157, 2017.

[42] R. M. Kretchmara, P. M. Young, C. W. Anderson, D. C. Hittle, M. L. Anderson, and C. Delnero, "Robust reinforcement learning control," in *Proc. of the IEEE American Control Conference*, vol. 2, 2001, pp. 902–907.

[43] C. W. Anderson, P. M. Young, M. R. Buehner, J. N. Knight, K. A. Bush, and D. C. Hittle, "Robust reinforcement learning control using integral quadratic constraints for recurrent neural networks," *IEEE Transactions on Neural Networks*, vol. 18, no. 4, pp. 993–1002, 2007.

[44] A. Packard and J. Doyle, "The complex structured singular value," *Automatica*, vol. 29, no. 1, pp. 71–109, 1993.

[45] J. Hauser, S. Sastry, and G. Meyer, "Nonlinear control design for slightly non-minimum phase systems: Application to v/stol aircraft," *Automatica*, vol. 28, no. 4, pp. 665–679, 1992.

[46] G. Fazelnia, R. Madani, A. Kalbat, and J. Lavaei, "Convex relaxation for optimal distributed control problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 206–221, 2017.

[47] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," in *Innovations in multi-agent systems and applications-1*. Springer, 2010, pp. 183–221.

[48] S. Fattahi, G. Fazelnia, J. Lavaei, and M. Arcak, "Transformation of optimal centralized controllers into near-globally optimal static distributed controllers," *IEEE Transactions on Automatic Control*, pp. 1–1, 2018.

[49] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

[50] P. Seiler, "Stability analysis with dissipation inequalities and integral quadratic constraints," *IEEE Transactions on Automatic Control*, vol. 60, no. 6, pp. 1704–1709, 2015.

APPENDIX

A. Overview of IQC framework

The IQC theory is celebrated for systematic and efficient stability analysis of a large class of uncertain, dynamic, and interconnected systems [40]. It unifies and extends classical passitivity-based multiplier theory, and has close connections to dissipativity theory in the time domain [50].

To state the IQC framework, some terminologies are necessary. We define the space $L_2^n[0,\infty) = \{x : \int_{t=0}^{\infty} |x(t)|_2^2 dt < \infty\}$ for signals supported on $t \geq 0$, where n denotes the spatial dimension of $x(t)$, and the extended space $L_{2e}^n[0,\infty) = \{x : \int_{t=0}^{T} |x(t)|_2^2 dt < \infty, \forall T \geq 0\}$ (we will use L_2 and L_{2e} if it is not necessary to specify the dimension and signal support), where we use x to denote the signal in general and $x(t)$ to denote its value at time t. For a vector or matrix, we use superscript $*$ to denote its conjugate transpose. An operator is causal if the current output does not depend on future inputs. It is bounded if it has a finite L_2 gain. Let $\Phi : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator on a Hilbert space. Then, its Hilbert adjoint is the operator $\Phi^* : \mathcal{H} \to \mathcal{H}$ such that $\langle \Phi x, y\rangle = \langle x, \Phi^* y\rangle$ for all $x, y \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product. It is *self-adjoint* if $\Phi = \Phi^*$.

Consider the system (see also Fig. 1)

$$y = G(u) \tag{41}$$

$$u = \Delta(y) + e, \tag{42}$$

where G is the transfer function of a causal and bounded LTI system (i.e., it maps input $u \in L^{n_a}$ to output $y \in L^{n_o}$ through the internal state dynamics $\dot{x} = Ax(t) + Bu(t)$), $e \in L^{n_a}$ is the disturbance, and $\Delta : L^{n_o} \to L^{n_a}$ is a bounded and causal function that is used to represent uncertainties in the system. IQC provides a framework to treat uncertainties such as nonlinear dynamics, model approximation and identification errors, time-varying parameters and disturbance noise, by using their input-output characterizations.

Definition 14 (Integral quadratic constraints)**.** *Consider the signals $w \in L_2$ and $y \in L_2$ associated with Fourier transforms $\hat{w}$ and $\hat{y}$, and $w = \Delta(y)$, where Δ is a bounded and causal operator. We present both the frequency- and time-domain IQC definitions:*

(a) *(Frequency domain) Let Π be a bounded and self-adjoint operator. Then, Δ is said to satisfy the IQC defined by Π (i.e., $\Delta \in$ IQC(Π)) if:*

$$\sigma_\Pi(\hat{y}, \hat{w}) = \int_{-\infty}^{\infty} \begin{bmatrix} \hat{y}(j\omega) \\ \hat{w}(j\omega) \end{bmatrix}^* \Pi(j\omega) \begin{bmatrix} \hat{y}(j\omega) \\ \hat{w}(j\omega) \end{bmatrix} d\omega \geq 0. \tag{43}$$

(b) *(Time domain) Let (Ψ, M) be any factorization of $\Pi = \Psi^* M \Psi$ such that Ψ is stable and $M = M^\top$. Then, Δ is said to satisfy the* hard IQC *defined by (Ψ, M) (i.e., $\Delta \in$ IQC(Ψ, M)) if:*

$$\int_0^T z(t)^\top M z(t) dt \geq 0, \qquad \forall\, T \geq 0, \tag{44}$$

where $z = \Psi \begin{bmatrix} y \\ w \end{bmatrix}$ is the filtered output given by the stable operator Ψ. If instead of requiring nonnegativity at each time T, the nonnegativity is considered only when $T \to \infty$, then the corresponding condition is called soft IQC.

As established in [50], the time- and frequency-domain IQC definitions are equivalent if there exists $\Pi = \Psi^* M \Psi$ as a spectral factorization of Π with $M = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$ such that Ψ and Ψ^{-1} are stable.

Example 15 (Sector IQC)**.** *A single-input single-output uncertainty $\Delta : \mathbb{R} \to \mathbb{R}$ is called "sector bounded" between $[\alpha, \beta]$ if $\alpha y(t) \leq \Delta(y(t)) \leq \beta y(t)$, for all $y \in \mathbb{R}$ and $t \geq 0$. It thus satisfies the sector IQC(Ψ, M) with $\Psi = I$ and $M = \begin{bmatrix} -2\alpha\beta & \alpha+\beta \\ \alpha+\beta & -2 \end{bmatrix}$. It also satisfies IQC$(\Pi)$ with $\Pi = M$ defined above.*

Example 16 (L_2 gain bound)**.** *A MIMO uncertainty $\Delta : \mathbb{R}^n \to \mathbb{R}^m$ has the L_2 gain γ if $\int_0^\infty \|w(t)\|^2 dt \le \gamma^2 \int_0^\infty \|y(t)\|^2 dt$, where $w(t) = \Delta(y(t))$. Thus, it satisfies IQC(Ψ, M) with $\Psi = I_{n+m}$ and $M = \begin{bmatrix} \lambda\gamma^2 I_n & 0 \\ 0 & -\lambda I_m \end{bmatrix}$, where $\lambda > 0$. It also satisfies IQC(Π) with $\Pi = M$ defined above. This can be used to characterize nonlinear operators with fast time-varying parameters.*

Before stating a stability result, we define the system (41)–(42) (see Fig. 1) to be well-posed if for any $e \in L_{2e}$, there exists a solution $u \in L_{2e}$, which depends causally on e. A main IQC result for stability is stated below:

Theorem 17 ([40])**.** *Consider the interconnected system* (41)–(42)*. Assume that:* ***(i)*** *the interconnected system $(G, \tau\Delta)$ is well posed for all $\tau \in [0, 1]$;* ***(ii)*** *$\tau\Delta \in$ IQC(Π) for $\tau \in [0, 1]$; and* ***(iii)*** *there exists $\epsilon > 0$ such that*

$$\begin{bmatrix} \hat{G}(j\omega) \\ I(j\omega) \end{bmatrix}^* \Pi(j\omega) \begin{bmatrix} \hat{G}(j\omega) \\ I(j\omega) \end{bmatrix} \le -\epsilon I, \qquad \forall\, \omega \in [0, \infty). \quad (45)$$

Then, the system (41)–(42) *is input-output stable (i.e., finite L_2 gain).*

The above theorem requires three technical conditions. The well-posedness condition is a generic property for any acceptable model of a physical system. The second condition is implied if $\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^* & \Pi_{22} \end{bmatrix}$ has the properties $\Pi_{11} \succeq 0$ and $\Pi_{22} \preceq 0$. The third condition is central, and it requires checking the feasibility at every frequency, which represents a main obstacle. As discussed in Section sec:compute, this condition can be equivalently represented as a linear matrix inequality (LMI) using the Kalman-Yakubovich-Popov (KYP) lemma. In general, the more IQCs exist for the uncertainty, the better characterization can be obtained. If $\Delta \in$ IQC(Π_k), $k \in [n_K]$, where n_K is the number of IQCs satisfied by Δ, then it is easy to show that $\Delta \in$ IQC$(\sum_{k=1}^{n_K} \tau_k \Pi_k)$, where $\tau_k \ge 0$; thus, the stability test (46) becomes a convex program, i.e., to find $\tau_k \ge 0$ such that:

$$\begin{bmatrix} \hat{G}(j\omega) \\ I(j\omega) \end{bmatrix}^* \left(\sum_{k=1}^{n_K} \tau_k \Pi_k(j\omega) \right) \begin{bmatrix} \hat{G}(j\omega) \\ I(j\omega) \end{bmatrix} \le -\epsilon I,\ \forall\, \omega \in [0, \infty). \quad (46)$$

The counterpart for the frequency-domain stability condition in the time-domain can be stated using a standard dissipation argument [50].

B. Proof of Lemma 2

Proof. For a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ that is differentiable with bounded partial derivatives on $\mathcal{B}$ (i.e., $\underline{\xi}_{ij} \le \partial_j f_i(x) \le \overline{\xi}_{ij}$ for all $x \in \mathcal{B}$), there exist functions $\delta_{ij} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ bounded by $\underline{\xi}_{ij} \le \delta_{ij}(x, y) \le \overline{\xi}_{ij}$ for all $i \in [m]$ and $j \in [n]$ such that

$$f(x) - f(y) = \begin{bmatrix} \sum_{j=1}^n \delta_{1j}(x, y)(x_j - y_j) \\ \vdots \\ \sum_{j=1}^n \delta_{mj}(x, y)(x_j - y_j) \end{bmatrix}. \quad (47)$$

By defining $q_{ij} = \delta_{ij}(x, y)(x_j - y_j)$, since $\left(\delta_{ij}(x, y) - c_{ij}\right)^2 \le \overline{c}_{ij}^2$, it follows that

$$\begin{bmatrix} x_j - y_j \\ q_{ij} \end{bmatrix}^\top \begin{bmatrix} \overline{c}_{ij}^2 - c_{ij}^2 & c_{ij} \\ c_{ij} & -1 \end{bmatrix} [\star] \ge 0. \quad (48)$$

The result follows by introducing nonnegative multipliers $\lambda_{ij} \ge 0$, and the fact that $f_i(x) - f_i(y) = \sum_{j=1}^m q_{ij}$. □

C. Proof of Theorem 4

Proof. The proof is in the same vein as that of Theorem 3. The main technical difference is the consideration of the filtered state ψ and the output z to impose IQC constraints on the nonlinearities $g_t(y)$ in the dynamical system [40]. The dissipation inequality follows by multiplying both sides of the matrix in (23) by $\begin{bmatrix} x^\top & q^\top & v^\top & e^\top \end{bmatrix}^\top$ and its transpose:

$$\dot{V}(x) + z^\top M_g z + \begin{bmatrix} x_G \\ q \end{bmatrix}^\top M_\pi \begin{bmatrix} x_G \\ q \end{bmatrix} < \gamma e^\top e - \frac{1}{\gamma} y^\top y,$$

where x and z are defined in (22), and $V(x) = x^\top P x$ is the storage function with $\dot{V}(\cdot)$ as its time derivative. The second term on the left side is non-negative because $g_t \in$ IQC(Ψ, M_g), and the third term is non-negative due to the smoothness quadratic constraint in Lemma 2. Thus, if there exists a feasible solution $P \succeq 0$ to $\underline{\text{SDP}}(P, \lambda, \gamma, \xi)$, integrating the inequality from $t = 0$ to $t = T$ yields that:

$$\int_0^T |y(t)|^2 dt \le \gamma^2 \int_0^T |e(t)|^2 dt. \quad (49)$$

Hence, the nonlinear system interconnected with the RL policy π is certifiably stable in the sense of a finite L_2 gain. □

D. Proof of Proposition 5

Proof. It suffices to show that for any $\pi \in \mathcal{P}(\xi)$, there exists a policy $\tilde{\pi} \in \widetilde{\mathcal{P}}(\xi)$ such that $\pi = W\tilde{\pi}$. Let $y_j^0 = \begin{bmatrix} 0 & \cdots & 0 & y_{j+1} & \cdots & y_{n_s} \end{bmatrix} \in \mathbb{R}^{n_s}$ for every $j \in \{0, 1, ..., n_s\}$, and $y_0^0 = y$, $y_{n_s}^0 = 0$. Then, one can write:

$$\pi_i(y) = \sum_{j=1}^{n_s} \pi_i(y_{j-1}^0) - \pi_i(y_j^0) = \sum_{j=1}^{n_s} \tilde{\pi}_{ij}(y),$$

where $\tilde{\pi}_{ij}(y)$ satisfies

$$\frac{\tilde{\pi}_{ij}(y)}{y_j} = \frac{\pi_i(y_{j-1}^0) - \pi_i(y_j^0)}{|y_{j-1}^0 - y_j^0|} \in [\underline{\xi}_{ij}, \overline{\xi}_{ij}]$$

if $y_j \ne 0$ and $\tilde{\pi}_{ij}(y) = 0$ if $y_j = 0$. The bound is due to the mean-value theorem and the bounds on the partial derivatives of π_i. Since the above argument is valid for all $i \in [n_a]$, it means that $\tilde{\pi} \in \widetilde{\mathcal{P}}(\xi)$, and $\pi = W\tilde{\pi}$. □

E. Proof of Lemma 6

Proof. To show the sufficiency direction, conditions (i) and (ii) yield that

$$\left(\overline{c}_{ij} x_j\right)^2 \ge \left(\left(\frac{\pi_{ij}(x)}{x_j} - c_{ij} \right) x_j \right)^2 = \left(q_{ij} - c_{ij} x_j\right)^2.$$

By rearranging the above inequality, it can be concluded that $(x, q) \in \mathcal{S}(\xi)$.

For the necessary direction, note that the condition $\phi_{ij}(x, q) \geq 0$ is equivalent to $\left|q_{ij} - c_{ij}x_j\right| \leq \left|\overline{c}_{ij}x_j\right|$. Thus, we have $\pi_{ij}(x) = 0$ if $x_j = 0$. Since $\overline{c}_{ij} \geq 0$, one can obtain $\left|\frac{q_{ij}}{x_j} - c_{ij}\right| \leq \overline{c}_{ij}$, which is equivalent to the sector bounds. □

F. Proof of Lemma 7

Proof. For the sufficiency condition, since $\tilde{\pi}_{ij}$ is sector bounded, and $\tilde{\pi}_{ij}(x) = 0$ if $x_j = 0$, without loss of generality, assume that $c_{ij} \leq 0$. One can write

$$\begin{aligned}
\left\|\overline{c}_{ij}x_j\right\|^2 &\geq \left\|\left|\frac{\left\|\tilde{\pi}_{ij}(x)\right\|}{\|x_j\|} - c_{ij}\right| x_j\right\|^2 \\
&\geq \left\|\left(\frac{\left\|\tilde{\pi}_{ij}(x)\right\|}{\|x_j\|} - c_{ij}\right) x_j\right\|^2 \\
&= \|c_{ij}x_j\|^2 + \|\tilde{\pi}_{ij}(x)\|^2 - 2\left\langle c_{ij}x_j, \frac{\left\|\tilde{\pi}_{ij}(x)\right\|}{\|x_j\|}x_j\right\rangle \\
&= \|c_{ij}x_j\|^2 + \|\tilde{\pi}_{ij}(x)\|^2 - 2c_{ij}\|x_j\|\left\|\tilde{\pi}_{ij}(x)\right\| \\
&\geq \|c_{ij}x_j\|^2 + \|\tilde{\pi}_{ij}(x)\|^2 - 2c_{ij}\left\langle\tilde{\pi}_{ij}(x), x_j\right\rangle \\
&= \left\|q_{ij} - c_{ij}x_j\right\|^2.
\end{aligned}$$

By rearranging the above inequality, it can be concluded that $(x, q) \in \mathcal{S}(\xi)$.

For the necessary direction, we can construct $\tilde{\pi}(y) = q\frac{\langle y,x\rangle}{|x|^2}$ for all $y \in L^{n_s}$. This leads to $\tilde{\pi}(x) = q$, and the condition $\phi_{ij}(x, q) \geq 0$ is equivalent to $\left\|q_{ij} - c_{ij}x_j\right\| \leq \overline{c}_{ij}\left\|x_j\right\|$. Thus, we have $\tilde{\pi}_{ij}(x) = 0$ if $x_j = 0$. Without loss of generality, assume that $c_{ij} \leq 0$. Therefore, $\|q_{ij}\| \leq \overline{c}_{ij}\|x_j\| + c_{ij}\|x_j\|$ and $\|q_{ij}\| \geq -\overline{c}_{ij}\|x_j\| + c_{ij}\|x_j\|$, which are equivalent to the sector bound condition. □

G. Proof of Lemma 9

Proof. Because G is time-invariant, by denoting D_τ as the delay operator at scale τ, we obtain $D_\tau^* T_{ij} D_\tau = T_{ij}$. Let $y = \phi(q)$ and $\tilde{y} = \phi(\tilde{q})$ be the elements of Ψ, with $\|q\| = \|\tilde{q}\| = 1$. By considering $q_\tau = \sqrt{\alpha}q + \sqrt{1-\alpha}D_\tau\tilde{q}$, one can write

$$\begin{aligned}
\phi_{ij}(q_\tau) &= \alpha\left\langle T_{ij}q, q\right\rangle + (1-\alpha)\left\langle T_{ij}D_\tau\tilde{q}, D_\tau\tilde{q}\right\rangle \\
&\qquad + 2\alpha\sqrt{1-\alpha}\mathrm{Re}\left\langle T_{ij}q, D_\tau\tilde{q}\right\rangle \\
&= \alpha\phi_{ij}(q) + (1-\alpha)\phi_{ij}(\tilde{q}) + 2\alpha\sqrt{1-\alpha}\mathrm{Re}\left\langle T_{ij}q, D_\tau\tilde{q}\right\rangle.
\end{aligned}$$

By letting $\tau \to \infty$, we obtain $\mathrm{Re}\left\langle T_{ij}q, D_\tau\tilde{q}\right\rangle \to 0$, where $\mathrm{Re}(x)$ denotes the real part of a complex vector x. Thus,

$$\lim_{\tau\to\infty}\phi_{ij}(q_\tau) = \alpha\phi_{ij}(q) + (1-\alpha)\phi_{ij}(\tilde{q})$$

and $\lim_{\tau\to\infty}\|q_\tau\|^2 = \alpha\|q\|^2 + (1-\alpha)\|\tilde{q}\|^2 = 1$. Therefore,

$$\lim_{\tau\to\infty}\phi\left(\frac{q_\tau}{\|q_\tau\|}\right) = \alpha y + (1-\alpha)\tilde{y} \in \overline{\Psi}.$$

□

H. Proof of Lemma 11

Proof. For a given $\epsilon > 0$, by hypothesis, there exists $q \in L^{n_a n_s}$ with $\|q\| = 1$ satisfying $\phi_{ij}(x, q) > -\epsilon^2$ for all $i \in [n_a]$ and $j \in [n_s]$, i.e.,

$$\epsilon^2 + \|Gq\|^2_{\Omega_{ij,x}} + 2\mathrm{Re}\left\langle\Omega_{ij,xq}Gq, q\right\rangle > \|q\|^2_{\Omega_{ij,q}},$$

where $\Omega_{ij,x}$ and $\Omega_{ij,xq}$ are defined previously. Clearly, if q is truncated to a sufficiently long interval, and q is rescaled to have a unit norm, the above inequality will still hold. Since $Gq \in L^{n_s}$, by possibly enlarging the truncation interval to $[t_0, t_1]$, we obtain (31), and

$$\epsilon^2 + \|\Gamma_{[t_0,t_1]}Gq\|^2_{\Omega_{ij,x}} + 2\mathrm{Re}\left\langle\Omega_{ij,xq}\Gamma_{[t_0,t_1]}Gq, q\right\rangle > \|q\|^2_{\Omega_{ij,q}},$$

Next, we choose $\eta \in L^{n_s}$ such that $\|\eta\|^2_{\Omega_{ij,x}} = \epsilon^2$, and that η is orthogonal to $\Gamma_{[t_0,t_1]}Gq$ and $\Omega^*_{ij,xq}q$ for all $i \in [n_a]$ and $j \in [n_s]$. Then, by considering $x = \Gamma_{[t_0,t_1]}Gq + \eta$, we obtain

$$\|x\|^2_{\Omega_{ij,x}} = \|\Gamma_{[t_0,t_1]}Gq + \eta\|^2_{\Omega_{ij,x}} = \epsilon^2 + \|\Gamma_{[t_0,t_1]}Gq\|^2_{\Omega_{ij,x}},$$

which leads to $\phi_{ij}(x, q) \geq 0$ and (32). Now, we can invoke Lemma 7 to construct $\tilde{\pi} \in \mathcal{P}(\xi)$ based on (30) such that $\tilde{\pi}$ becomes sector bounded and $q = \tilde{\pi}x$. Then,

$$(I - \tilde{\pi}\Gamma_{[t_0,t_1]}G)q = \tilde{\pi}(x - \Gamma_{[t_0,t_1]}Gq).$$

Let $\|\tilde{\pi}\| \leq C$ (which depends on the sector bounds). Then,

$$\|(I - \tilde{\pi}\Gamma_{[t_0,t_1]}G)q\| \leq C\epsilon$$

. □

I. Proof of Theorem 13

Proof. Since the system is input-output stable, the sets Π and $\overline{\Psi}$ are strictly separable due to Lemma 10. Since both Π and $\overline{\Psi}$ are convex (Lemma 9), there exist a strictly separating hyperplane parametrized by $\lambda \in \mathbb{R}^{mn}$ and scalars α, β, such that

$$\langle\lambda, \phi\rangle \leq \alpha < \beta \leq \langle\lambda, y\rangle$$

for all $\phi \in \overline{\Psi}$ and $y \in \Pi$. Since $\langle\lambda, y\rangle$ is bounded from below, we must have $\lambda \geq 0$, and without loss of generality, we can set $\beta = 0$ and $\alpha < 0$. This condition is equivalent to (34), and by Proposition 12, this implies that the SDP condition is feasible. □

Ming Jin is a postdoctoral researcher in the Department of Industrial Engineering and Operations Research at University of California, Berkeley. He received his doctoral degree from EECS department at University of California, Berkeley in 2017. His research interests are optimization, learning and control with applications to sustainable infrastructures. He was the recipient of the Siebel scholarship, 2018 Best Paper Award of Building and Environment, 2015 Best Paper Award at the International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2016 Best Paper Award at the International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Electronic and Computer Engineering Department Scholarship, School of Engineering Scholarship, and University Scholarship at the Hong Kong University of Science and Technology.

Javad Lavaei is an Associate Professor in the Department of Industrial Engineering and Operations Research at UC Berkeley. He obtained the Ph.D. degree in Control & Dynamical Systems from California Institute of Technology, and was a postdoctoral scholar at Electrical Engineering and Precourt Institute for Energy of Stanford University for one year. He has won several awards, including Presidential Early Career Award for Scientists and Engineers given by the White House, DARPA Young Faculty Award, Office of Naval Research Young Investigator Award, Air Force Office of Scientific Research Young Investigator Award, NSF CAREER Award, DARPA Director's Fellowship, Office of Naval Research's Director of Research Early Career Grant, Google Faculty Award, Governor General's Gold Medal given by the Government of Canada, and Northeastern Association of Graduate Schools Master's Thesis Award. He is a recipient of the 2015 INFORMS Optimization Society Prize for Young Researchers, the 2016 Donald P. Eckman Award given by the American Automatic Control Council, the 2016 INFORMS ENRE Energy Best Publication Award, and the 2017 SIAM Control and Systems Theory Prize. Javad Lavaei is an associate editor of the IEEE Transactions on Automatic Control, IEEE Transactions on Smart Grid and of the IEEE Control Systems Letters, and serves on the conference editorial boards of the IEEE Control Systems Society and European Control Association.