IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Stability-Certified Reinforcement Learning: A Control-Theoretic Perspective

MING JIN[1], JAVAD LAVAEI[2]
[1]Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: jinming@vt.edu)
[2]Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail: lavaei@berkeley.edu)

Corresponding author: Ming Jin (e-mail: jinming@vt.edu).

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."

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 G is the environment and π_θ is the policy. The goal of RL is to maximize the expected return:

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

where $r(x, u)$ is the reward at state $x \in \mathbb{R}^{n_s}$ and action $u \in \mathbb{R}^{n_a}$, $\rho \in (0, 1]$ is the future discount factor, and the expectation is taken over the policy π_θ as well as the initial state distribution and dynamics 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 reinforcement learning [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 robust control, Lyapunov functions are widely used to analyze and verify stability under uncertainty [11], [12]. For nonlinear systems without global stability guarantees, region of attraction and reachability analysis have been employed for local convergence analysis [13]–[15]. The main challenge of these methods is that the robustness guarantee can be conservative due to coarse constraints on nonlinearity such as those based on Lipschitz constants [16], [17], leading to a limited search space for safe policies. To mitigate this issue, the integral quadratic constraint (IQC) framework proposed in [18] can be employed, which has been widely used to analyze the stability of large-scale complex systems, such as aircraft control [19]. Nevertheless, existing techniques can be computational intensive for deep neural networks, and establishing necessary conditions for robustness has been limited to only a few cases (e.g., block-diagonal structured uncertainty operators with bounded singular values [20]).

To address this issue, we introduce a more informative quadratic constraint and analyze the necessity of the certificate criterion as an extension of the preliminary conference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

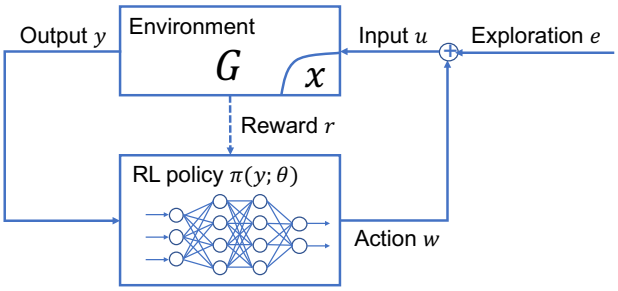


FIGURE 1: The goal of RL is to maximize expected rewards through interaction and exploration. In this study, we address the challenge of maintaining the stability of the interconnected system during the learning and control process.

version [21]. This opens up new possibilities to safely apply deep RL to nonlinear large-scale real-world systems, whose stability is otherwise impossible to be certified using existing approaches. As an overview, we propose a "safety set" of policies $\mathcal{P}(\xi)$ based on 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 $y \in \mathbb{R}^{n_s}$ is the output vector, and n_a, n_s are the dimensions of the input and output, respectively, $\partial_j \pi_i$ is the partial derivative of π_i for the j-th input. Our framework designs a set of 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. Importantly, this work bridges the robust control with reinforcement learning to provide provably guarantees of stability during the learning and control process. Key contributions of the present study are as follows:

- Development of a general framework to certify the stability of reinforcement learning for nonlinear systems, with new strategies to incorporate the stability requirement into RL;
- A new characterization of the safety set of policies based on point-wise IQC, as well as the analysis of its non-conservativeness property;
- Numerical evaluation on decentralized control problems for flight formation and power grid frequency regulation.

The rest of the paper is organized as follows. We formulate the problem in Section II. Main results for the stability analysis are presented in Section III, where we also analyze the conservatism of the certificate. Section IV presents numerical studies on two nonlinear decentralized control tasks. Conclusions are drawn in Section V.

Notations

Let $A \succeq B$ or $A \succ B$ denote that $A - B$ is positive semidefinite or positive definite, respectively. Also, $\mathbb{R}$ and $\mathbb{R}_+$ represent the sets of real and nonnegative real numbers, respectively. The notation $[n]$ shows the set $\{1, ..., n\}$. We use x to denote a vector, and boldface $\boldsymbol{x}$ to denote a signal with value $x(t)$ at each time t. Also, we use $x_{0:T}$ to denote the signal over the time interval $[0, T]$. We use $\{x_{ij}\}_{i\in[m],j\in[n]}$ to denote a collection of entries x_{ij} in a vector form, or simply $\{x_{ij}\}$ if the ranges of i and j are clear from the context. By convention, we arrange the entries in $\{x_{ij}\}_{i\in[m],j\in[n]}$ such that x_{ij} is at location $k = (i-1)n + j$. Similarly, we use $\{\boldsymbol{x}_{ij}\}$ to denote a collection of signals. The i-th entry of x is written as x_i or $[x]_i$, and the i-th dimension of a signal is $\boldsymbol{x}_i$ or $[\boldsymbol{x}]_i$. For norms, we denote $|x| = \sqrt{x_1^2 + \cdots + x_n^2}$ as the 2-norm of a vector x, and $\|\boldsymbol{x}\| = \sqrt{\int_0^\infty |x(t)|^2 dt}$ as the 2-norm of a signal. When $\|\cdot\|$ is applied to an operator, it denotes an induced norm from $\mathcal{L}$ to $\mathcal{L}$, where $\mathcal{L}$ is the vector space of signals with bounded 2-norm, i.e., $\|\boldsymbol{x}\| = \sqrt{\int_0^\infty |x(t)|^2 dt} < \infty$ for $\boldsymbol{x} \in \mathcal{L}$. In the case we need to specify n as the spatial dimension of a signal, we will use $\mathcal{L}^n$. Also, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \int_0^\infty x(t)^\top y(t) dt$ denotes the inner product between two signals in $\mathcal{L}$; therefore, $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. The notation $\|\boldsymbol{x}\|_\Omega = \sqrt{\langle \boldsymbol{x}, \Omega \boldsymbol{x} \rangle}$ is used for an adjoint operator Ω. Given a matrix $M \in \mathbb{R}^{m\times n}$, $\mathrm{diag}(M)$ denotes a diagonal matrix whose diagonal elements are given by the entries $M_{11}, \ldots, M_{1n}, M_{21}, \ldots, M_{2n}, \ldots, M_{m1}, \ldots, M_{mn}$. We use $\otimes$ to denote the Kronecker product.

II. PROBLEM FORMULATION

Consider a 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, $f_t(\cdot, \cdot)$ can be a time-varying and nonlinear function. In this work, we focus on an important class of systems described by:

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

where $f_t(\cdot, \cdot)$ comprises a linear time-invariant (LTI) component modeled by a matrix $A \in \mathbb{R}^{n_s \times n_s}$ that is Hurwitz (i.e., every eigenvalue of A has strictly negative real part) and a control matrix $B \in \mathbb{R}^{n_s \times n_a}$, as well as a slowly time-varying component $g_t(\cdot)$ that captures the nonlinearity and uncertainty of the original system.[2] The output $y(t) = Cx(t) \in \mathbb{R}^{n_s}$ is a linear function of the states, where $C \in \mathbb{R}^{n_s \times n_s}$ may have a sparsity pattern in the context of decentralized controls [22]. Nevertheless, since we can account for the sparsity pattern in the design of the control policy, we henceforth assume that $y(t) = x(t)$ for simplicity. The control input based on RL is

$$u(t) = \pi_t(y(t); \theta_t) + e(t), \tag{5}$$

where $\pi_t(y(t); \theta_t)$ is usually 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 input perturbation that is assumed to have a bounded energy over time ($\|\boldsymbol{e}\|_2 = \sqrt{\int |e(t)|_2^2 dt} \le$

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

[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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

∞). Let $\{\pi_t | t \in [0,T]\}$ be the trajectory of policies deployed in the system over the time interval $[0,T]$. We propose the following notion of dynamic stability adapted from the classical definition of the L_2 gain [20], [23].[3]

Definition 1 (Dynamic input-output stability)**.** *Given a dynamical system G, the L_2 gain of the system G with the input $u(t) = \pi_t(y(t);\theta_t) + e(t)$ is defined as the worst-case ratio between the total output energy and the total input perturbation energy:*

$$\gamma(G, \{\pi_t | t \in [0,T]\}) = \sup_{e \in \mathcal{L}} \frac{\|y_{0:T}\|_2}{\|e_{0:T}\|_2}, \tag{6}$$

where L_2 is the set of all square-summable signals and $\|y_{0:T}\|_2 = \sqrt{\int_0^T |y(t)|_2^2 dt}$ is the total energy within the time window $[0,T]$. If $\lim_{T\to\infty} \gamma(G, \{\pi_t | t \in [0,T]\})$ is bounded by a finite number Γ, then the closed-loop system is said to be dynamically input-output stable (with finite gain Γ).

The stability-certified RL problem is thus to find an optimal policy π_θ that maximizes the expected reward $\eta(\pi_\theta)$ defined in (1) while ensuring that the system is dynamically input-output stable, i.e., $\gamma(G, \{\pi_t | t \in [0,T]\}) < \Gamma$ for a given finite gain Γ. Our approach is to delineate a safety set of controllers $\mathcal{P}(\xi)$ based on partial gradients, such that as long as $\pi_t \in \mathcal{P}(\xi)$ for all $t \in [0,T]$, we can guarantee that $\gamma(G, \{\pi_t | t \in [0,T]\})$ is bounded by Γ. After identifying the safety set, one can impose it via a constraint during RL, and then the problem can be solved by any arbitrary RL algorithm (e.g., policy gradient [1], [2], Q-learning [4], [24], actor-critic [6], [7]). Furthermore, the certification can be regarded as an $\mathcal{S}$-procedure (c.f., [25]), and we analyze its non-conservatism by showing that this condition is necessary for the robustness of a surrogate system that is closely related to the original system.

III. MAIN RESULTS

A. QUADRATIC CONSTRAINTS ON GRADIENT-BOUNDED FUNCTIONS

First, we introduce the time-domain definition of IQC [26]:

Definition 2 (IQC in the time-domain)**.** *Consider the signals $\boldsymbol{w} \in \mathcal{L}$ and $\boldsymbol{y} \in \mathcal{L}$ with $\boldsymbol{w} = \Delta(\boldsymbol{y})$, where Δ is a bounded and causal operator. Let Ψ be a stable LTI system and $M = M^\top$ be a symmetric matrix. Then, Δ is said to satisfy the hard IQC defined by (Ψ, M), denoted by $\Delta \in \mathrm{IQC}(\Psi, M)$, if:*

$$\int_0^T z(t)^\top M z(t) dt \geq 0, \qquad \forall\, T \geq 0, \tag{7}$$

where $\boldsymbol{z} = \Psi \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{w} \end{bmatrix}$ is the filtered output given by Ψ. Furthermore, if

$$z(t)^\top M z(t) \geq 0, \qquad \forall t \geq 0 \tag{8}$$

then Δ is said to satisfy the point-wise IQC.

Recall that a function $h(x)$ is *Lipschitz continuous* with constant $\xi > 0$ if and only if it satisfies the following quadratic constraint for all $x_\alpha, x_\beta \in \mathbb{R}^n$:

$$\begin{bmatrix} x_\alpha - x_\beta \\ h(x_\alpha) - h(x_\beta) \end{bmatrix}^\top \begin{bmatrix} \xi^2 I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} x_\alpha - x_\beta \\ h(x_\alpha) - h(x_\beta) \end{bmatrix} \geq 0, \tag{9}$$

which corresponds to the point-wise IQC (8) with $\Psi = I$ and $M = \begin{bmatrix} \xi^2 I & 0 \\ 0 & -I \end{bmatrix}$. 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. For example, consider the function $h : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$h(x) = \left[\tanh(0.5x_1) - a x_1, \sin(x_2)\right]^\top, \tag{10}$$

where $|a| \leq 0.1$ is a deterministic but unknown parameter with a bounded magnitude. Clearly, the function has the Lipschitz constant 1; however, the above characterization ignores the non-homogeneity of $h_1(x)$ (i.e., the first output of $h(x)$) and $h_2(x)$, as well as the sparsity of the dependence on x. Indeed, $h_1(x)$ only depends on x_1 with its slope restricted to $[-0.1, 0.6]$ for all possible $|a| \leq 0.1$, and $h_2(x)$ only depends on x_2 with its slope restricted to $[-1, 1]$. In the context of control synthesis, the non-homogeneity of outputs often arises from physical constraints, and the sparsity of inputs can be due to distributed local information. To explicitly address these requirements, we propose the following quadratic constraint.

Lemma 1 (Function with bounded partial gradients)**.** *Consider an arbitrary vector-valued function $h : \mathbb{R}^n \to \mathbb{R}^m$ that is differentiable with bounded partial derivatives (i.e., $\underline{\xi}_{ij} \leq \partial_j h_i(x) \leq \overline{\xi}_{ij}$ for all $x \in \mathbb{R}^n$). Define the vectors $c \in \mathbb{R}^{nm}$ and $\bar{c} \in \mathbb{R}^{nm}$ with the entries $c_{ij} = \frac{1}{2}\left(\underline{\xi}_{ij} + \overline{\xi}_{ij}\right)$ and $\bar{c}_{ij} = \overline{\xi}_{ij} - c_{ij}$ for $i \in [m]$ and $j \in [n]$. Let $W = \left[I_m \otimes 1_{1\times n}\right]$. Then, for all $\lambda \in \mathbb{R}_+^{n\times m}$ and $x_\alpha, x_\beta \in \mathbb{R}^n$, there exists a vector-valued function $q : \mathbb{R}^{2n} \to \mathbb{R}^{nm}$ (with entries $q_{ij}(\cdot,\cdot)$ for $i \in [m]$ and $j \in [n]$) with the property that*

$$h(x_\alpha) - h(x_\beta) = W q(x_\alpha, x_\beta), \tag{11}$$

such that the following quadratic constraint is satisfied:

$$\begin{bmatrix} x_\alpha - x_\beta \\ q(x_\alpha, x_\beta) \end{bmatrix}^\top M(\lambda;\xi) \begin{bmatrix} x_\alpha - x_\beta \\ q(x_\alpha, x_\beta) \end{bmatrix} \geq 0 \tag{12}$$

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

$$\begin{bmatrix} \mathrm{diag}\left(\left\{\sum_{i=1}^m \lambda_{ij}(\bar{c}_{ij}^2 - c_{ij}^2)\right\}_{j=1}^n\right) & U(\lambda,c)^\top \\ U(\lambda,c) & \mathrm{diag}\,(-\lambda) \end{bmatrix}, \tag{13}$$

and $U(\lambda,c) \in \mathbb{R}^{n\times mn}$ is given by

$$\begin{bmatrix} \mathrm{diag}\left(\{-\lambda_{1j}c_{1j}\}_{j=1}^n\right) & \cdots & \mathrm{diag}\left(\{-\lambda_{mj}c_{mj}\}_{j=1}^n\right) \end{bmatrix}.$$

[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. Since we assume that $y(t) = x(t)$, the input-output stability is equivalent to internal stability.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

Proof. See Appendix V-A. □

The above bound is a direct consequence of standard tools in real analysis [27]. To understand this result, it can be observed that (12) is equivalent to:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_{ij}\Big((\bar{c}_{ij}^2-c_{ij}^2)([x_\alpha]_j-[x_\beta]_j)^2 + 2c_{ij}q_{ij}([x_\alpha]_j-[x_\beta]_j)-q_{ij}^2\Big)\geq 0 \tag{14}$$

for $\lambda_{ij}\geq 0$ and $h_i(x_\alpha)-h_i(x_\beta)=\sum_{j=1}^{n}q_{ij}$, where we omit the dependence of q_{ij} on x_α and x_β for notational simplicity. Since (14) holds for all $\lambda_{ij}\geq 0$, it is equivalent to the condition that $(\bar{c}_{ij}^2-c_{ij}^2)([x_\alpha]_j-[x_\beta]_j)^2+2c_{ij}q_{ij}([x_\alpha]_j-[x_\beta]_j)-q_{ij}^2\geq 0$ for all $i\in[m]$ and $j\in[n]$, which is a direct result of the bounds imposed on the partial derivatives of h_i. If we apply it to the example function (10), we can specify that $\underline{\xi}_{11}=-0.1,\bar{\xi}_{11}=0.6,\underline{\xi}_{22}=-1,\bar{\xi}_{22}=1$, and all the other bounds $(\underline{\xi}_{12},\bar{\xi}_{12},\underline{\xi}_{21},\bar{\xi}_{21})$ are zero. This clearly yields a more informative constraint than merely relying on the Lipschitz constraint (9). In fact, for a differentiable l-Lipschitz function, we have $\bar{\xi}_{ij}=-\underline{\xi}_{ij}=l$, and by limiting the choice of $\lambda_{ij}=\begin{cases}\lambda & \text{if } i=1\\ 0 & \text{if } i\neq 1\end{cases}$, (14) is reduced to (9). However, as illustrated in this example and the experiments in Section IV, since the condition in Lemma 1 can incorporate richer information about the problem structure, it often gives rise to less restrictive stability bounds compared to (9).

Remark 1: The constraint introduced above can be considered as a point-wise IQC; however, it is unconventional in the sense that it involves an intermediate unknown function $q(\cdot,\cdot)$ that is related to the output $h(\cdot)$. For stability analysis, let $x_\beta=x^*\in\mathbb{R}^n$ be the equilibrium point, and without loss of generality, assume that $x^*=0$ and $h(x^*)=0$. 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 (12) can be written as

$$\sum_{ij}\lambda_{ij}\phi_{ij}(x,q)\geq 0,\qquad \forall\lambda_{ij}\geq 0, \tag{15}$$

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

To simplify the mathematical treatment, we have focused on differentiable functions in Lemma 1; 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 [28, Chap. 2]. In brief, by re-assigning the bounds on partial derivatives to uniform bounds on the set of generalized partial derivatives, the constraint (12) can be directly applied.

In relation to the existing IQCs, this constraint has wider applications for the characterization of gradient-bounded functions. The Zames-Falb IQC introduced in [29] has been widely used for single-input single-output (SISO) functions $h:\mathbb{R}\to\mathbb{R}$, but it requires the function to be monotone with the slope restricted to $[\alpha,\beta]$ with $\alpha\geq 0$, i.e., $0\leq\alpha\leq\frac{h(x_\alpha)-h(x_\beta)}{x_\alpha-x_\beta}\leq\beta$ whenever $x_\alpha\neq x_\beta$. The multi-input multi-output (MIMO) extension holds true only if the nonlinear function $h:\mathbb{R}^n\to\mathbb{R}^n$ is restricted to be the gradient of a convex real-valued function [30], [31]. As for the sector IQC, the scalar version can not be used (because it requires $h_i(x)=0$ whenever there exists $j\in[n]$ such that $x_j=0$, which is extremely restrictive), and the vector version is in fact (9). In contrast, the quadratic constraint in Lemma 1 can be applied to non-monotone, vector-valued gradient-bounded functions.

B. STABILITY CERTIFICATION

In this part, we assume that the nonlinear component $g_t(x)$ is zero and leave the generalization to the next subsection. It is desirable to study the stability certification of RL for an LTI system G, whose state-space representation is given by:

$$\begin{cases}\dot{x}_G &= Ax_G+Bu\\ w &= \pi_t(x_G)\\ u &= e+w\end{cases} \tag{16}$$

where $x_G\in\mathbb{R}^{n_s}$ is the state (the dependence on t is omitted for simplicity), and the policy π_t is changing due to RL. As before, A is Hurwitz. The goal is to certify the safety set $\mathcal{P}(\xi)$ such that as long as π_t remains in $\mathcal{P}(\xi)$, the RL-controlled system is stable with some constant γ and c_γ:

$$\int_0^T\big|y(t)\big|^2dt\leq\gamma^2\int_0^T\big|e(t)\big|^2dt+c_\gamma. \tag{17}$$

To this end, define the SDP(P,λ,γ,ξ) as follows:

$$\text{SDP}(P,\lambda,\gamma,\xi):\begin{bmatrix}O(P,\lambda,\xi) & S(P)\\ S(P)^\top & -\gamma I\end{bmatrix}\prec 0, \tag{18}$$

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 $M(\lambda;\xi)$ and $W=\begin{bmatrix}I_m\otimes 1_{1\times n}\end{bmatrix}$ are defined in Lemma 1. We will next show that the stability can be certified using the aforementioned linear matrix inequalities (LMIs).

Theorem 1. *Let A be Hurwitz and $\pi_t\in\mathbb{R}^{n_s}\to\mathbb{R}^{n_a}$ be a time-varying causal controller. Assume that π_t has bounded partial derivatives (i.e., $\underline{\xi}_{ij}\leq\partial_j[\pi_t]_i(x)\leq\bar{\xi}_{ij}$, for all $t\in[0,T]$, $x\in\mathbb{R}^{n_s}$, $i\in[n_a]$ 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 RL-controlled system* (16) *is stable with the gain γ.*

Proof. To proceed, we choose a q according to Lemma 1 and 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 (18), and use the constraints

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

$w = Wq$ and $y = x_G$. Then, SDP$(P, \lambda, \gamma, \xi)$ 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. Denote the initial state of the system by x_0. Because the second term is guaranteed to be non-negative at all times t by Lemma 1, if SDP$(P, \lambda, \gamma, \xi)$ is feasible with a solution $(P, \lambda, \gamma, \xi)$, we have:

$$\dot{V}(x_G) + \frac{1}{\gamma} y^\top y - \gamma e^\top e < 0, \tag{19}$$

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

$$\int_0^T |y(t)|^2 dt - x_0^\top P x_0 \leq \gamma^2 \int_0^T |e(t)|^2 dt, \tag{20}$$

which completes the proof. □

We remark that SDP$(P, \lambda, \gamma, \xi)$ is quasiconvex, in the sense that it reduces to a standard 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 [32].

C. EXTENSION TO NONLINEAR SYSTEMS WITH UNCERTAINTY

In the context of RL, we often need to deal with systems with nonlinear dynamics and/or unmodeled dynamics. We model the nonlinear, uncertain and potentially time-varying part of the system with $g_t(x(t))$ in (3) and regard it as an uncertain block with IQC constraints. Specifically, consider the LTI component $\underline{G}$:

$$\begin{cases} \dot{x}_G = Ax_G + Bu + v \\ y \quad = x_G \end{cases} \tag{21}$$

where A is assumed to be Hurwitz. The nonlinear part is moved to the feedback:

$$\begin{cases} u \ = e + w \\ w = \pi_t(y) \\ v \ = g_t(y) \end{cases} \tag{22}$$

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. 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 2 (see [18], [33] for some examples). The system Ψ has the state-space representation:

$$\begin{cases} \dot{\psi} = A_\psi \psi + B_\psi^v v + B_\psi^y y \\ z \ = C_\psi \psi + D_\psi^v v + D_\psi^y y \end{cases}, \tag{23}$$

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 (21) and (23) via reducing y and letting $w = Wq$:

$$\begin{cases} \dot{x} = \underbrace{\begin{bmatrix} A & 0 \\ B_\psi^y & A_\psi \end{bmatrix}}_{\underline{A}} x + \underbrace{\begin{bmatrix} B \\ 0 \end{bmatrix}}_{\underline{B}_e} e + \underbrace{\begin{bmatrix} BW \\ 0 \end{bmatrix}}_{\underline{B}_q} q + \underbrace{\begin{bmatrix} I \\ B_\psi^v \end{bmatrix}}_{\underline{B}_v} v \\ z = \underbrace{\begin{bmatrix} D_\psi^y & C_\psi \end{bmatrix}}_{\underline{C}} x + D_\psi^v v \end{cases}, \tag{24}$$

where $\underline{A}$, $\underline{B}_e$, $\underline{B}_q$, $\underline{B}_v$, $\underline{C}$ are matrices of proper dimensions defined above. We define $\underline{\text{SDP}}(P, \lambda, \gamma, \xi)$ as:

$$\underline{\text{SDP}}(P, \lambda, \gamma, \xi) : \begin{bmatrix} O(P,\lambda,\xi) & O_v(P) & S(P) \\ O_v(P)^\top & D_\psi^{v\top} M_q D_\psi^v & 0 \\ S(P)^\top & 0 & -\gamma I \end{bmatrix} \prec 0, \tag{25}$$

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 (13). The next theorem provides a stability certificate for the nonlinear time-varying system (3).

Theorem 2. *Let A in (21) be Hurwitz, and $\pi_t \in \mathbb{R}^{n_s} \to \mathbb{R}^{n_a}$ be a time-varying causal controller. Assume that:*

(i) *π_t has bounded partial derivatives (i.e., $\underline{\xi}_{ij} \leq \partial_j [\pi_t]_i(x) \leq \overline{\xi}_{ij}$ for all $t \in [0,T]$, $x \in \mathbb{R}^{n_s}$, $i \in [n_a]$ and $j \in [n_s]$);*

(ii) *$g_t \in \text{IQC}(\Psi, M_g)$ for all $t \in [0,T]$, where Ψ is a stable LTI system.*

If there exist $P \succeq 0$ and a scalar $\gamma > 0$ such that $\underline{\text{SDP}}(P, \lambda, \gamma, \xi)$ given in (25) is feasible, then the RL-controlled system (3) is stable with the gain γ.

Proof. See Appendix V-B. □

Theorem 2 requires A to be stable, but this is not a conservative assumption. Indeed, any arbitrary system can be decomposed (in infinitely many ways) into a stable component Ax and a second part $g_t(x)$. In other words, $g_t(x)$ gives the flexibility to perform this decomposition. Alternatively, one can find a linear nominal model for the system (using linearization of the known part of the dynamics), stabilize it via an internal controller to arrive at a stable matrix A, and then design a main RL controller to deal with the remaining dynamics $g_t(x)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

The present study models the unknown system as a known nominal model plus an unknown (un-modelled) dynamics with weak assumptions on the prior knowledge about the system. This is standard in control theory and indeed it is theoretically impossible to design a high-performance stabilizing controller without prior knowledge about the system. If such knowledge is not available, the only option is to perform system identification to gain such knowledge, which requires extensive data collection and sufficient system excitation. Note that having the nonlinearity satisfying IQC is different from having the nonlinearity being known. For example, if we make an assumption that an unknown parameter of the system has norm less than 1, then there are still infinitely many possibilities for that unknown parameter. IQC serves the same purpose. In the two case studies provided in the paper (see Sec. IV), it is NP-hard to design an optimal distributed controller even for a known model of the system. In those examples, it is easy to define the nominal part and the nonlinear part, and then check the IQC. This way we can effectively handle the NP-hardness of the problem. In other words, one can ignore the nonlinearity to break down the NP-hardness of the controller design and then use IQC to make up for the discarded nonlinear part through an RL controller.

The stability analysis fundamentally depends on the uncertainty of the system—as the system becomes more uncertain (e.g., due to time variation or finite sample estimation), the stability guarantee becomes more conservative. The proposed method is a general framework that can address a variety of uncertainties, both in the system model and in the reinforcement learning policy. As we impose more and tighter constraints, the results become less conservative. Therefore, as opposed to being restrictive assumptions, the IQC constraints imposed on the nonlinear, uncertain, and potentially time-varying components of the system incorporate useful information to make the stability analysis less conservative.

D. ANALYSIS OF CONSERVATISM OF THE STABILITY CERTIFICATE

We focus on system (16) where an LTI system is interconnected with a fixed RL policy. Without loss of generality, we also assume that $\overline{\xi}_{ij} = -\underline{\xi}_{ij}$ for all $i \in [n_a]$ and $j \in [n_s]$ to streamline the presentation. To certify stability of the original system, as will be shown in the next proposition, it suffices to examine the stability of the following system:

$$\begin{cases} \boldsymbol{x}_G = \begin{bmatrix} G_{11} & G_{12} \end{bmatrix} \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{e} \end{bmatrix}, \\ \boldsymbol{q} \quad = \tilde{\pi}(\boldsymbol{x}_G) \end{cases} \tag{26}$$

where $G = \left[\begin{array}{c|cc} A & BW & B \\ \hline I & 0 & 0 \end{array}\right] = \begin{bmatrix} G_{11} & G_{12} \end{bmatrix}$, and $\tilde{\pi} \in \widetilde{\mathcal{P}}(\xi)$ is a *functional* in the uncertainty set

$$\widetilde{\mathcal{P}}(\xi) = \Big\{ \tilde{\pi} \mid \|\tilde{\pi}_{ij}(\boldsymbol{x})\| \le \overline{\xi}_{ij} \|\boldsymbol{x}_j\|, \quad \forall \boldsymbol{x} \in L^{n_s}, i \in [n_a], j \in [n_s] \Big\}. \tag{27}$$

Recall that the notation $\boldsymbol{x}$ implies a signal, and it represents $x(t)$ for t from 0 to infinity. This means that although the argument of $\pi(\cdot)$ is a vector, the argument of $\tilde{\pi}(\cdot)$ is a signal. Therefore, $\tilde{\pi}(\cdot)$ is a functional extension of the function π. By convention, we interpret the output of a function applied to a signal, e.g. $\pi(\boldsymbol{x})$, as point-wise operation at each time instance, i.e. $\{\pi(x(t)) | t \in [0, \infty)\}$.

Proposition 1. *If the system* (26) *is stable for all* $\tilde{\pi} \in \widetilde{\mathcal{P}}(\xi)$*, then the system* (16) *is stable for all* $\pi \in \mathcal{P}(\xi)$*.*

Proof. See Appendix V-C. □

Now, let $\mathcal{S}(\xi) = \big\{(\boldsymbol{x}, \boldsymbol{q}) \mid \widetilde{\phi}_{ij}(\boldsymbol{x}, \boldsymbol{q}) \ge 0, \forall i \in [n_a], j \in [n_s]\big\}$, where

$$\widetilde{\phi}_{ij}(\boldsymbol{x}, \boldsymbol{q}) = \overline{\xi}_{ij}^2 \|\boldsymbol{x}_j\|^2 - \|\boldsymbol{q}_{ij}\|^2. \tag{28}$$

We now show that the pair $(\boldsymbol{x}, \boldsymbol{q})$ belongs to $\mathcal{S}(\xi)$ if and only if there exists a sector-bounded function $\tilde{\pi} \in \widetilde{P}(\xi)$ such that it satisfies $\boldsymbol{q} = \tilde{\pi}(\boldsymbol{x})$.

Lemma 2. *Suppose that* $\boldsymbol{x} \in \mathcal{L}^{n_s}$ *and* $\boldsymbol{q} \in \mathcal{L}^{n_a n_s}$*. Then, the pair* $(\boldsymbol{x}, \boldsymbol{q})$ *belongs to* $\mathcal{S}(\xi)$ *if and only if there exists an operator* $\tilde{\pi} : \mathcal{L}^{n_s} \to \mathcal{L}^{n_a n_s}$ *such that* $\boldsymbol{q} = \tilde{\pi}(\boldsymbol{x})$ *and* $\tilde{\pi}$ *satisfies the following conditions: (i)* $\tilde{\pi}_{ij}(\boldsymbol{x}) = \boldsymbol{0}$ *if* $\boldsymbol{x}_j = \boldsymbol{0}$*, and (ii)* $\big\|\tilde{\pi}_{ij}(\boldsymbol{x})\big\| \le \overline{\xi}_{ij} \|\boldsymbol{x}_j\|$ *for all* $i \in [n_a]$ *and* $j \in [n_s]$*.*

Proof. See the Appendix V-D. □

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

Lemma 3. *If* $(G, \tilde{\pi})$ *is robustly stable, then there cannot exist a nonzero* $\boldsymbol{q} \in L$ *such that* $\boldsymbol{x} = G\boldsymbol{q}$ *and* $(\boldsymbol{x}, \boldsymbol{q}) \in \mathcal{S}(\xi)$*.*

Proof. We prove this lemma by contraposition. If there exists a nonzero $\boldsymbol{q} \in \mathcal{L}$ such that $(\boldsymbol{x}, \boldsymbol{q}) \in \mathcal{S}(\xi)$, then it follows from Lemma 2 that there exists a linear operator $\tilde{\pi}$ such that $\boldsymbol{q} = \tilde{\pi}(\boldsymbol{x}) = \tilde{\pi}(G\boldsymbol{q})$. 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. □

Consider the set generated by the LTI system:

$$\Lambda = \Big\{ \{\widetilde{\phi}_{ij}(\boldsymbol{x}, \boldsymbol{q})\} \in \mathbb{R}^{n_a n_s} | \boldsymbol{q} \in \mathcal{L}^{n_a n_s}, \|\boldsymbol{q}\| = 1, \boldsymbol{x} = G\boldsymbol{q} \Big\}, \tag{29}$$

and the positive orthant

$$\Pi = \big\{ \{r_{ij}\} \in \mathbb{R}^{n_a n_s} | r_{ij} \ge 0, \quad \forall i \in [n_a], j \in [n_s] \big\}. \tag{30}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

Lemma 3 implies that the two sets Λ and Π are separated if $(G, \tilde{\pi})$ is robustly stable.

To prove the necessity of the stability certificate, we first aim to show in Lemma 4 that there exists a strict separating hyperplane when the system is robustly stable. We then draw the connection in Proposition 2 between the existence of the strict separating hyperplane and the feasibility of the SDP condition (18). These two results together imply that the stability certificate derived from (18) is necessary for any robustly stable system controlled by RL (Theorem 3).

Define $\Omega_{ij,x} = \text{diag}\left(\left\{\overline{\xi}_{ij}^2\right\}\right)$ and $\Omega_{ij,q}$ as a matrix with its (k,l)-th element equals to

$$[\Omega_{ij,q}]_{kl} = \begin{cases} 1 & \text{if } k = in_s + j \\ 0 & \text{otherwise} \end{cases}.$$

Thus, we can write $\widetilde{\phi}_{ij}(\boldsymbol{x} = G\boldsymbol{q}, \boldsymbol{q})$ as an inner product:

$$\widetilde{\phi}(\boldsymbol{x} = G\boldsymbol{q}, \boldsymbol{q}) = \|G\boldsymbol{q}\|^2_{\Omega_{ij,x}} - \|\boldsymbol{q}\|^2_{\Omega_{ij,q}} = \langle \boldsymbol{q}, T_{ij}\boldsymbol{q} \rangle,$$

where $T_{ij} = G^*\Omega_{ij,x}G - \Omega_{ij,q}$. Now, we show that strict separation occurs when the system is robustly stable.

Lemma 4. *Suppose that $I - G\tilde{\pi}$ is nonsingular. Then, the sets Π and Λ are strictly separated, namely $D(\Pi, \Lambda) = \inf_{r\in\Pi, y\in\Lambda} |r - y| > 0$.*

Proof. Assume that $D(\Pi, \Psi) = \inf_{r\in\Pi, y\in\Psi} |r - y| = 0$. Consider a sequence $\epsilon_k \to 0$ as k tends to ∞. For each ϵ_k, we construct the signals $\boldsymbol{q}^{(k)}$ with a bounded support on the time interval $[t_k, t_{k+1}]$, such that $\|\boldsymbol{q}^{(k)}\| = 1$ and that there exists a signal $\boldsymbol{x}^{(k)} \in \mathcal{L}^{n_s}$ satisfying:

$$\widetilde{\phi}_{ij}(\boldsymbol{x}^{(k)}, \boldsymbol{q}^{(k)}) \geq 0, \qquad \forall i \in [n_a], j \in [n_s] \tag{31a}$$

$$\epsilon_k^2 > \|(I - \Gamma_{[t_k,t_{k+1}]})G\boldsymbol{q}^{(k)}\| \tag{31b}$$

$$\epsilon_k = \|\boldsymbol{x}^{(k)} - \Gamma_{[t_k,t_{k+1}]}G\boldsymbol{q}^{(k)}\|_{\Omega_{ij,x}}, \tag{31c}$$

where $\Gamma_{[t_k,t_{k+1}]}$ projects the signal onto the support of $[t_k, t_{k+1}]$. We also construct a functional $\tilde{\pi}^{(k)} \in \tilde{P}(\xi)$ such that $\boldsymbol{q}^{(k)} = \tilde{\pi}^{(k)}\boldsymbol{x}^{(k)}$ and $\|(I - \tilde{\pi}^{(k)}\Gamma_{[t_k,t_{k+1}]}G)\boldsymbol{q}^{(k)}\| \leq C\epsilon_k$ for some constant $C > 0$ that depends on the sector bounds ξ (the existence of such $\boldsymbol{q}^{(k)}$ and $\tilde{\pi}^{(k)}$ is shown in Lemma 6 in Appendix V-F). Now, define

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

Then, $\|\tilde{\pi}\| \leq C_\pi$, where $C_\pi > 0$ is a finite number that depends on ξ. We have

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

and

$$\begin{aligned}\|(I - \tilde{\pi}G)\boldsymbol{q}^{(k)}\| &\leq \|(I - \tilde{\pi}^{(k)}\Gamma_{[t_k,t_{k+1}]}G)\boldsymbol{q}^{(k)}\| \\ &\quad + C_\pi\|(I - \Gamma_{[t_k,t_{k+1}]})G\boldsymbol{q}^{(k)}\| \\ &\leq C\epsilon_k + C_\pi\epsilon_k^2\end{aligned}$$

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

Next, we draw the connection to the SDP problem (18). Observe that

$$\widetilde{\phi}_{ij}(\boldsymbol{x}, \boldsymbol{q}) = \left\langle \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{q} \end{bmatrix}, M_\pi^{ij} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{q} \end{bmatrix} \right\rangle, \tag{32}$$

where

$$[M_\pi^{ij}]_{kl} = \begin{cases} \overline{\xi}_{ij}^2 & (k,l) = (j,j) \\ 0 & (k,l) = (i, in_s+j) \text{ or } (in_s+j, i) \\ -1 & (k,l) = (in_s+j, in_s+j) \end{cases},$$

and $M(\lambda; \xi) = \sum_{i\in[n_a], j\in[n_s]} \lambda_{ij} M_\pi^{ij}$ as defined in Lemma 1.

Proposition 2. *The SDP condition* (18) *is feasible if and only if there exist multipliers $\lambda_{ij} \geq 0$ and $\epsilon > 0$ such that*

$$\sum_{i\in[n_a], j\in[n_s]} \lambda_{ij}\widetilde{\phi}_{ij}(\boldsymbol{x}, \boldsymbol{q}) \leq -\epsilon\|\boldsymbol{q}\|^2 \tag{33}$$

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

Proof. By (32), the condition (33) is equivalent to

$$\begin{bmatrix} G \\ I \end{bmatrix}^* M(\lambda; \xi) \begin{bmatrix} G \\ I \end{bmatrix} \prec 0. \tag{34}$$

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

$$\begin{bmatrix} A^\top P + PA & PBW \\ W^\top B^\top P & 0 \end{bmatrix} + M(\lambda; \xi) \prec 0. \tag{35}$$

By Schur complement, P satisfies the KYP condition if and only if it satisfies (18), which proves the statement. □

Finally, we are able to show the necessity of the stability certificate below.

Theorem 3. *Let $\tilde{\pi} : \mathcal{L}^{n_s} \to \mathcal{L}^{n_a n_s}$ be a bounded causal controller such that $\tilde{\pi} \in \tilde{\mathcal{P}}(\xi)$. Then, the input-output stability of the feedback interconnection of system* (26) *implies that there exist $P \succeq 0$, $\gamma > 0$ and $\lambda \geq 0$ such that SDP$(P, \lambda, \gamma, \xi)$ in* (18) *is feasible.*

Proof. Since the system is input-output stable, the sets Π and $\overline{\Psi}$ are strictly separable due to Lemma 4. Since both Π and $\overline{\Psi}$ are convex (see Lemma 5 in Appendix V-E), there exist a strictly separating hyperplane parametrized by $\lambda \in \mathbb{R}^{mn}$ and scalars α, β, such that

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

for all $\phi \in \overline{\Psi}$ and $\nu \in \Pi$. Since $\langle \lambda, \nu \rangle$ 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 (33), and by Proposition 2, this implies that the SDP condition is feasible. □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

E. REINFORCEMENT LEARNING WITH STABILITY REGULARIZATION

The key takeaway message from the previous analysis is that in order to ensure stability during the learning and control process, it is critical to regulate the magnitudes of the partial gradients. In this section, we use trust region policy optimization (TRPO) [3] as an example to illustrate our approach to addressing this requirement. Note that our methods can be applied to other types of RL algorithms as well.

1) Stability penalty

In this method, we consider adding a penalty term to the RL objective function:

$$L_{\text{pen}}(\theta;\xi) = \sum\nolimits_{t=1}^{T} \mathbb{1}\left(\partial_j[\pi_\theta]_i(x_t) \in [\underline{\xi}_{ij}, \bar{\xi}_{ij}]\right), \quad (36)$$

where $\mathbb{1}(\cdot)$ is the indicator function that aims to keep the partial gradients along the trajectories inside the bounds by softly penalizing it. This term is closely related to the term used in the literature

$$\sum\nolimits_{t=1}^{T} \left|\partial_j[\pi_\theta]_i(x_t)\right|^2. \quad (37)$$

Both terms encourage the behavior that small changes in the input should not change the output drastically, and they are different from typical regularization terms designed for the weights θ (e.g., weight decay in neural network). Interestingly, the penalty (37) was termed "double backpropagation" in [34], and recently rediscovered in [35], [36] to improve the generalization performance of image classification tasks. The present study provides theoretical and empirical justification of this penalty along with (36) to improve stability of the RL-controlled dynamical system.

For policy gradient with TRPO, by manipulating the expected return $\eta(\pi)$ using the identity proposed in [37], the following "surrogate objective" $\eta_{\text{pol}}(\theta)$ can be designed:

$$\eta_{\text{pol}}(\theta) = \mathbb{E}\left[\frac{\pi_\theta(u|x)}{\pi_{\text{old}}(u|x)}\Lambda^{\pi_{\text{old}}}(x,u)\right], \quad (38)$$

where the expectation is taken over the old policy $\pi_{\text{old}} = \pi_{\theta_{\text{old}}}$, the ratio inside the expectation is also known as the importance weight, and $\Lambda^{\pi_{\text{old}}}(x,u)$ is the advantage function given by:

$$\Lambda^{\pi_{\text{old}}}(x,u) = \mathbb{E}\left[r(x,u) + \rho V^{\pi_{\text{old}}}(x') - V^{\pi_{\text{old}}}(x)\right], \quad (39)$$

where the expectation is with respect to the dynamics $x' \sim \mathcal{T}(x,u)$ (in our case, the dynamics is a nonlinear dynamical system, and x' is the next state given the current state x and input u), and it measures the improvement of taking action u at state x over the old policy in terms of the value function $V^{\pi_{\text{old}}}$. Combined with (36), the modified objective is given by:

$$\eta_{\text{new}}(\theta) = \eta_{\text{pol}}(\theta) - \omega L_{\text{pen}}(\theta), \quad (40)$$

where $\omega \geq 0$ is the regularization coefficient whose value is selected such that the scale of the corresponding term is about $[0.01, 0.05]$ of the surrogate loss value $\eta_{\text{pol}}(\theta)$. In practice, the modified objective $\eta_{\text{new}}(\theta)$ can be estimated using trajectories sampled from π_{old}.

2) Hard thresholding

After each gradient step, we obtain an upper bound on the Lipschitz constant $\bar{l}(\pi_\theta)$ of the updated neural network π_θ (e.g., using the simple approach introduced in [38]). If the upper bound is greater than the certified bound l°, then we execute a hard thresholding (HT) step that rescales the weight matrices at each layer by $(l^\circ/l(\pi_\theta))^{1/n_L}$ if $l(\pi_\theta) > l^\circ$, where n_L is the number of layers of the neural network. The goal is to ensure that the Lipschitz constant of the RL policy remains bounded by l°.

Remark 2: We note that the proposed approaches are applicable to existing RL algorithms (e.g., policy gradient, Q-learning, and actor-critic), since we either modify the objective function (1) or the resulting policy directly. For neural networks, it remains an open problem how to calculate the Lipschitz constant or partial gradients exactly [39]. In some cases, the bounds given by the simple approach proposed in [38] can be conservative. Alternatively, one can estimate the bounds on partial gradients using existing trajectories, which is reasonable as long as the trajectory does not deviate significantly from the history.

IV. CASE STUDIES

In this section, we empirically study the stability-certified RL in two important problems: flight formation [40] and power grid frequency regulation [41]. 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 distributed 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 [22]. 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 [42]. 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 stability certificates when applying RL to dynamical systems, which is orthogonal to the line of research that aims at improving the performance of the existing RL algorithms. For illustration, we adopt an approach based on policy gradient that combines TRPO [3] with natural gradient [2] and the two regularization techniques proposed in Section III-D. For both the multi-agent formation and power grid frequency regulation, we employ the method proposed in [41] to design the nominal controller, which is distributed and stabilizing for the nominal LTI system without uncertainty and nonlinearity, since those two parts together with reward maximization are handled through RL.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

A. MULTI-AGENT FLIGHT FORMATION

Consider the multi-agent flight formation problem [40], 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.

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, we designed a nominal static distributed controller K_n to make the largest eigenvalue of $A + BK_n$ negative [41].

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 $x(t)$).

1) 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 (23) [29], [43]. As for the RL agents u^i, their gradient bounds can be certified according to Theorem 2. Specifically, we assume that each agent u^i is l-Lipschitz continuous, and solve (25) for a given set of γ and l. The certified gradient bounds (Lipschitz constants) are plotted in Fig. 3 using different constraints. The L_2 constraint (9) can only certify stability for Lipschitz constants up to 0.8. By setting the bounds $\overline{\xi}_{ij}$ and $\underline{\xi}_{ij}$ to 0 for agent i that does not access input j due to input sparsity, we can increase the certified value to 1.2.

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 $\partial_j \pi_i(x)$ 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, e.g. 0.1, to allow explorations. As a result, we can significantly enlarge the certified Lipschitz bound to up to 2.5, as shown in Fig. 3.

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

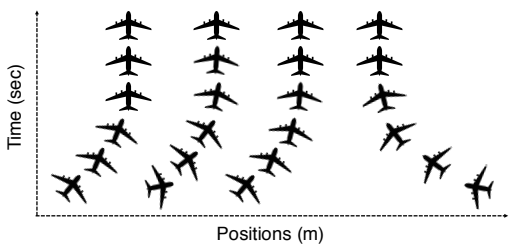


FIGURE 2: Illustration of the multi-agent flight formation problem.

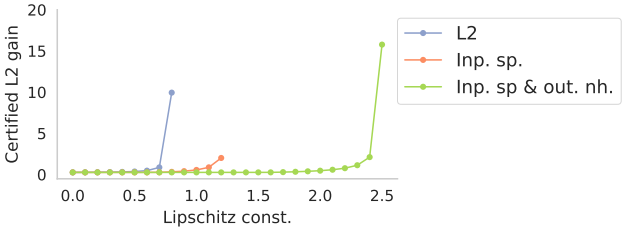


FIGURE 3: Stability-certified Lipschitz constants obtained by the standard L_2 bound (L2) in (9) and the method proposed in Lemma 1, which considers input sparsity (inp. sp.) and output non-homogeneity (out. nh.).

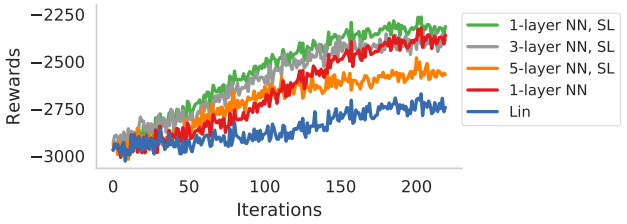


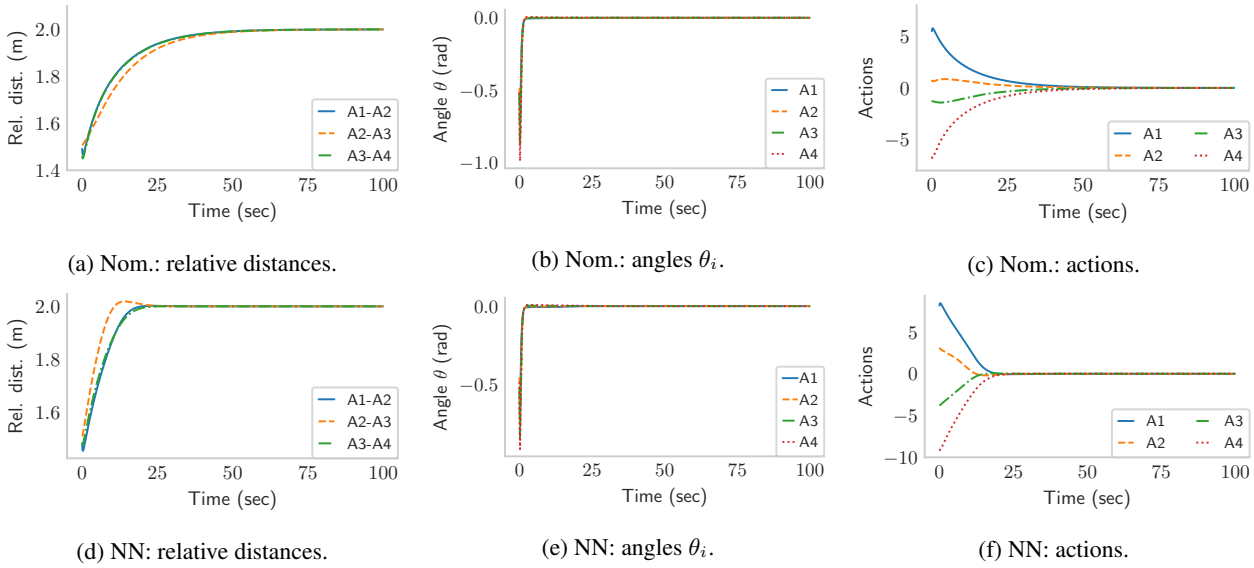
FIGURE 4: Learning performance of different control structures (1-layer neural network, 5-layer neural network, and linear controller). The outputs of the last layer of neural networks are used as the system inputs. By the inclusion of the stability regularization (SL), the exploration becomes more effective.

2) RL performance

The trajectories of rewards averaged over 20 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 with stability regularization, which significantly outperforms the linear controller.

The learned 5-layer neural network policy is employed in an actual control task, as shown in Fig. 5. Compared to the nominal controller (i.e., no RL control), 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 (58.3). This result can be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

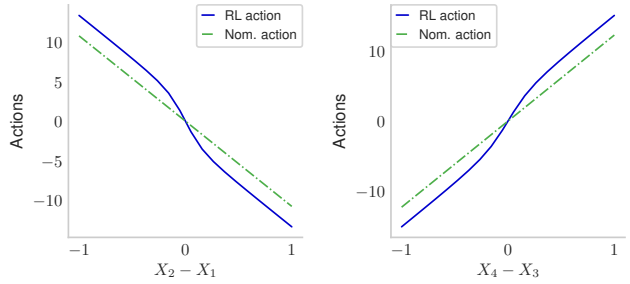


(a) Nom.: relative distances. (b) Nom.: angles θ_i. (c) Nom.: actions.

(d) NN: relative distances. (e) NN: angles θ_i. (f) NN: actions.

FIGURE 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. In the legend, we use "A1-A2" to denote the relative distance between agents A1 and A2.

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

FIGURE 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 [41]. 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 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:

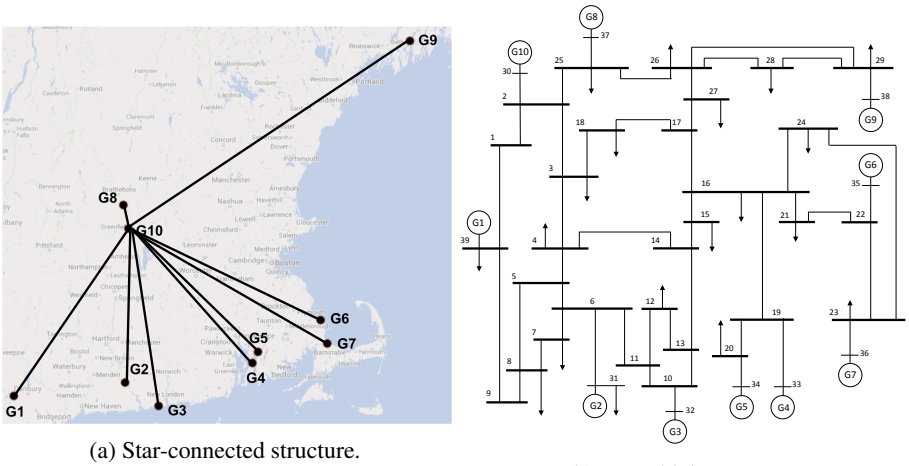
$$\begin{bmatrix}\dot\theta \\ \dot\omega\end{bmatrix} = \underbrace{\begin{bmatrix}0 & I \\ -M^{-1}L & -M^{-1}D\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}\theta \\ \omega\end{bmatrix}}_{x} + \underbrace{\begin{bmatrix}0 \\ M^{-1}\end{bmatrix}}_{B} p_m + \underbrace{\begin{bmatrix}\mathbf{0} \\ g(\theta)\end{bmatrix}}_{g(x)}$$

where $g(\theta) = \begin{bmatrix}g_1(\theta) & \cdots & g_n(\theta)\end{bmatrix}^\top$ with $g_i(\theta) = \sum_{j=1}^n \frac{b_{ij}}{m_j}\left((\theta_i-\theta_j) - \sin(\theta_i-\theta_j)\right)$, 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 [41, 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 [41], [44]. 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 stabilizes the system. To conduct multi-agent RL, each controller p_{m_i} is a neural network learned online.

1) Stability certificate

The nonlinearities in $g_i(\theta)$ 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 $\Delta\theta_{ij} \in [-\overline{\theta}, \overline{\theta}]$ (here, assume $\overline{\theta} = \frac{\pi}{3}$ to include both normal and abnormal operational conditions); thus, we can apply the Zames-Falb IQC. By simply applying the $\mathcal{L}_2$ constraint in (9), we can only certify stability for Lipschitz constants up to 0.4, as shown in Fig. 8. With input sparsity, we can certify l up to 0.6. With output non-homogeneity, which is visualized in Fig. 9, we can further extend the certificate up to 1.1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access



(a) Star-connected structure.

(b) IEEE 39-bus system.

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

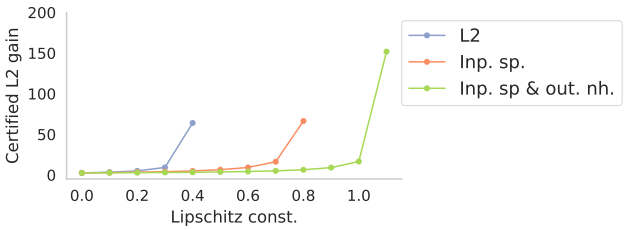


FIGURE 8: Certified Lipschitz constants for power system frequency regulation.

2) RL performance

First, we visualize the behavior of the learned neural network controller in a typical control case (Fig. 10). The nominal controller is designed by the method in [41] for the LTI nominal system and does not maximize the reward. As can be seen, the RL policies can regulate the frequencies more efficiently than the nominal controller (i.e. no RL), 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 behave erratically. 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 preserve stability of the interconnected system at the same time.

C. DISCUSSIONS

While the case studies focus on the application of the framework to systems with nonlinearities, the method can be readily deployed to handle uncertainties (which may arise due to finite sample estimation in system identification), since the unmodeled parts can be principally characterized using IQC constraints [18]. In practice, to certify the stability of any system with the feedback control of reinforcement learning agents, one needs to summarize the information into the linear parts of the model and impose as many valid constraints as possible on the unmodeled parts to reduce the conservativeness.

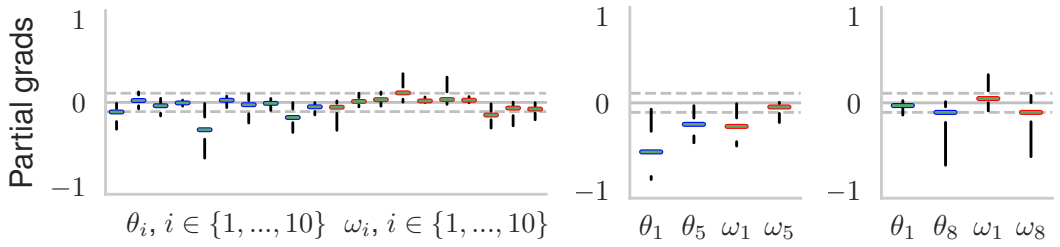
One limitation of the present study is the requirement of the system nominal system matrix A to be Hurwitz. Even though it is not difficult for many applications to design a controller that is stabilizing as showcased in the experiments, this requirement may exclude some possibilities for reinforcement learning agents to control highly unstable systems. In particular, if K_n is required to be distributed, finding a stabilizing K_n even for a known LTI system is NP-hard in the worst case. The second limitation is that the analysis is restricted to global asymptotic stability – it is interesting to extend the analysis to local stability.

V. CONCLUSIONS

In this paper, we focused on the challenging task of ensuring the stability of RL in real-world dynamical systems. The present study makes the following contributions:

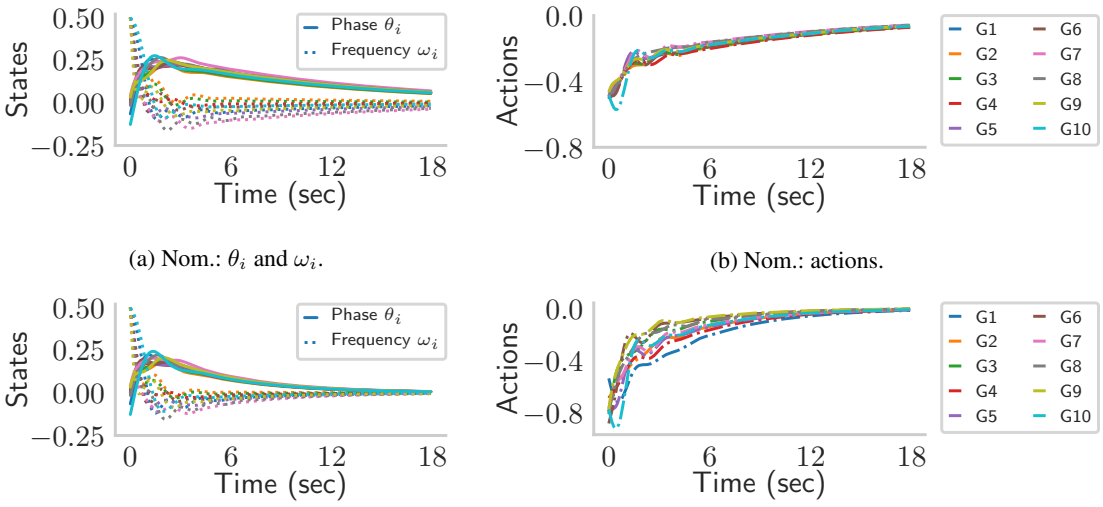
- Development of a new quadratic constraint based on partial gradients, which can be integrated into a SDP-based problem to certify stability of RL-controlled systems;
- Analysis of the (non)conservatism of the certification condition, which is shown to be almost necessary and sufficient under some assumptions;
- Evaluation of the framework for decentralized nonlinear control tasks, which demonstrates that the proposed approach can certify policies with Lipschitz constants that are about 3 times larger than those certifiable by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access



(a) G10. (b) G4. (c) G7.

FIGURE 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.

FIGURE 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.

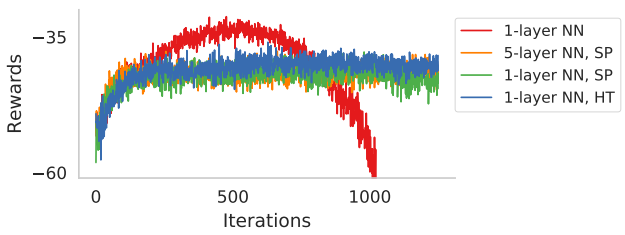


FIGURE 11: Long-term performance of RL for agents with regulated gradients by soft penalty (SP) and hard thresholding (HT). The RL agents without regulating the gradients exhibit "dangerous" behaviors in the long run.

existing methods for the flight formation and power system frequency regulation problems.

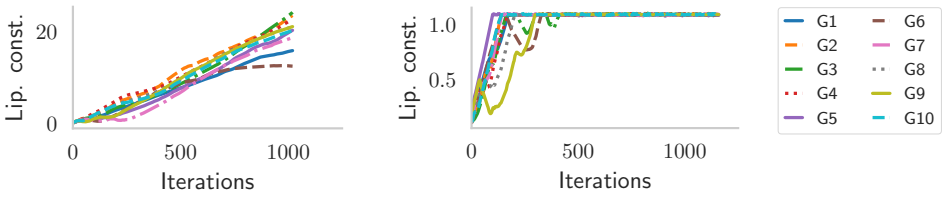
The proposed approach can systematically address nonlinearity in the neural network policy and uncertainty/time-variation in the underlying system. A key benefit is the ability to certify a much larger set of controllers for exploration by reinforcement learning. Moreover, unlike some existing techniques that require the controller to be static for the stability analysis, our method allows it to change over time, which is crucial for RL since the policies are continuously updated with new data. Most importantly, the regulation of gradient bounds was shown to improve on-policy learning performance and avoid "catastrophic" effects caused by the unregulated high gains. The present study represents a key step towards safe deployment of RL 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.



(a) No gradient regulation.

(b) Gradient regulation.

FIGURE 12: Trajectories of Lipschitz constants with and without stability regulation.

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, 2018.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] M. Jin and J. Lavaei, "Control-theoretic analysis of smoothness for stability-certified reinforcement learning," in *Proc. of the IEEE Conference on Decision and Control*, 2018, pp. 6840–6847.

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

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

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[25] S. V. Gusev and A. L. Likhtarnikov, "Kalman-popov-yakubovich lemma and the s-procedure: A historical essay," *Automation and Remote Control*, vol. 67, no. 11, pp. 1768–1810, 2006.

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

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

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

[29] 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.

[30] 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.

[31] 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.

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

[33] A. Helmersson, "An iqc-based stability criterion for systems with slowly varying parameters," *IFAC Proceedings Volumes*, vol. 32, no. 2, pp. 3183–3188, 1999.

[34] 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.

[35] 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.

[36] 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.

[37] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *Proc. of the International Conference on Machine Learning*, 2002, pp. 267–274.

[38] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. of the International Conference on Learning Representations*, 2014.

[39] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas, "Efficient and accurate estimation of Lipschitz constants for deep neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 423–11 434.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

APPENDIX

A. PROOF OF LEMMA 1

Proof. For a vector-valued function $h : \mathbb{R}^n \to \mathbb{R}^m$ that is differentiable with bounded partial derivatives (i.e., $\underline{\xi}_{ij} \leq \partial_j h_i(x) \leq \overline{\xi}_{ij}$), there exist functions $\delta_{ij} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ bounded by $\underline{\xi}_{ij} \leq \delta_{ij}(x_\alpha, x_\beta) \leq \overline{\xi}_{ij}$ for $i \in [m]$ and $j \in [n]$ such that

$$h(x_\alpha) - h(x_\beta) = \begin{bmatrix} \sum_{j=1}^n \delta_{1j}(x_\alpha, x_\beta)([x_\alpha]_j - [x_\beta]_j) \\ \vdots \\ \sum_{j=1}^n \delta_{mj}(x_\alpha, x_\beta)([x_\alpha]_j - [x_\beta]_j) \end{bmatrix}. \tag{41}$$

By defining $q_{ij} = \delta_{ij}(x_\alpha, x_\beta)([x_\alpha]_j - [x_\beta]_j)$, since $\big(\delta_{ij}(x_\alpha, x_\beta) - c_{ij}\big)^2 \leq \overline{c}_{ij}^2$, it follows that

$$\begin{bmatrix} [x_\alpha]_j - [x_\beta]_j \\ q_{ij} \end{bmatrix}^\top \begin{bmatrix} \overline{c}_{ij}^2 - c_{ij}^2 & c_{ij} \\ c_{ij} & -1 \end{bmatrix} \begin{bmatrix} [x_\alpha]_j - [x_\beta]_j \\ q_{ij} \end{bmatrix} \geq 0. \tag{42}$$

The result follows by introducing nonnegative multipliers $\lambda_{ij} \geq 0$, and the fact that $h_i(x_\alpha) - h_i(x_\beta) = \sum_{j=1}^n q_{ij}$. □

B. PROOF OF THEOREM 2

Proof. The proof is in the same vein as that of Theorem 1. 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 [18]. The dissipation inequality follows by multiplying both sides of the matrix in (25) 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 (24), 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 $g_t \in \text{IQC}(\Psi, M_g)$, and the third term is non-negative due to the smoothness quadratic constraint in Lemma 1. 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:

$$\int_0^T |y(t)|^2 dt \leq \gamma^2 \int_0^T |e(t)|^2 dt. \tag{43}$$

Hence, the nonlinear system interconnected with the RL policy π is certifiably stable in the sense of a finite $\mathcal{L}_2$ gain. □

C. PROOF OF PROPOSITION 1

Proof. First, we define the set

$$\overline{\mathcal{P}}(\xi) = \Big\{ \bar{\pi} \mid |\bar{\pi}_{ij}(x)| \leq |\overline{\xi}_{ij} x_j|, \quad \forall x \in \mathbb{R}^{n_s}, i \in [n_a], j \in [n_s] \Big\}. \tag{44}$$

Since the constraint on $\bar{\pi}$ in $\overline{\mathcal{P}}(\xi)$ is point-wise at each time instance, it is straightforward to verify that $\overline{\mathcal{P}}(\xi) \subseteq \widetilde{\mathcal{P}}(\xi)$. It suffices to show that for any $\pi \in \mathcal{P}(\xi)$, there exists a policy $\bar{\pi} \in \overline{\mathcal{P}}(\xi)$ such that $\pi = W\bar{\pi}$. 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:

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

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

$$\left| \frac{\bar{\pi}_{ij}(y)}{y_j} \right| = \left| \frac{\pi_i(y_{j-1}^0) - \pi_i(y_j^0)}{|y_{j-1}^0 - y_j^0|} \right| \leq \overline{\xi}_{ij}$$

if $y_j \neq 0$ and $\bar{\pi}_{ij}(y) = 0$ 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 $\bar{\pi} \in \overline{\mathcal{P}}(\xi) \subseteq \widetilde{\mathcal{P}}(\xi)$, and $\pi = W\bar{\pi}$. □

D. PROOF OF LEMMA 2

Proof. For the sufficiency condition, since $\tilde{\pi}_{ij}$ is sector bounded, and $\tilde{\pi}_{ij}(\boldsymbol{x}) = 0$ if $\boldsymbol{x}_j = 0$, we have

$$\left\| \overline{\xi}_{ij} \boldsymbol{x}_j \right\|^2 \geq \left\| \frac{\left\| \tilde{\pi}_{ij}(\boldsymbol{x}) \right\|}{\|\boldsymbol{x}_j\|} \boldsymbol{x}_j \right\|^2 = \|\boldsymbol{q}_{ij}\|^2.$$

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

For the necessary direction, we can construct $\tilde{\pi}(\boldsymbol{y}) = \boldsymbol{q} \frac{\langle \boldsymbol{y}, \boldsymbol{x} \rangle}{\|\boldsymbol{x}\|^2}$ for all $\boldsymbol{y} \in \mathcal{L}^{n_s}$. This leads to $\tilde{\pi}(\boldsymbol{x}) = \boldsymbol{q}$, and the condition $\widetilde{\phi}_{ij}(\boldsymbol{x}, \boldsymbol{q}) \geq 0$ is equivalent to $\left\| \boldsymbol{q}_{ij} \right\| \leq \overline{\xi}_{ij} \left\| \boldsymbol{x}_j \right\|$. Thus, we have $\tilde{\pi}_{ij}(\boldsymbol{x}) = 0$ if $\boldsymbol{x}_j = 0$ and the sector bound condition is satisfied. □

E. STATEMENT AND PROOF OF LEMMA 5

Lemma 5. *For a given linear time-invariant operator G, the closure $\overline{\Lambda}$ of Λ defined in* (29) *is convex.*

Proof. Because G is time-invariant, by denoting D_τ as the delay operator at scale τ, we obtain $D_\tau^* T_{ij} D_\tau = T_{ij}$. For simplicity, let $\widetilde{\phi}(\boldsymbol{q})$ denote $\widetilde{\phi}(G\boldsymbol{q}, \boldsymbol{q})$, since we restrict the first argument of $\widetilde{\phi}$ to only depend on $\boldsymbol{q}$. Let $\boldsymbol{y} = \widetilde{\phi}(\boldsymbol{q})$ and $\tilde{\boldsymbol{y}} = \widetilde{\phi}(\tilde{\boldsymbol{q}})$ be the elements of Λ, with $\|\boldsymbol{q}\| = \|\tilde{\boldsymbol{q}}\| = 1$. By considering $\boldsymbol{q}_\tau = \sqrt{\alpha}\boldsymbol{q} + \sqrt{1-\alpha} D_\tau \tilde{\boldsymbol{q}}$, one can write

$$\begin{aligned} \widetilde{\phi}_{ij}(\boldsymbol{q}_\tau) &= \alpha \left\langle T_{ij}\boldsymbol{q}, \boldsymbol{q} \right\rangle + (1-\alpha) \left\langle T_{ij} D_\tau \tilde{\boldsymbol{q}}, D_\tau \tilde{\boldsymbol{q}} \right\rangle \\ &\quad + 2\alpha\sqrt{1-\alpha} \mathrm{Re} \left\langle T_{ij}\boldsymbol{q}, D_\tau \tilde{\boldsymbol{q}} \right\rangle \\ &= \alpha \widetilde{\phi}_{ij}(\boldsymbol{q}) + (1-\alpha) \widetilde{\phi}_{ij}(\tilde{\boldsymbol{q}}) + 2\alpha\sqrt{1-\alpha} \mathrm{Re} \left\langle T_{ij}\boldsymbol{q}, D_\tau \tilde{\boldsymbol{q}} \right\rangle. \end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045114, IEEE Access

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

$$\lim_{\tau\to\infty} \widetilde{\phi}_{ij}(\boldsymbol{q}_\tau) = \alpha\widetilde{\phi}_{ij}(\boldsymbol{q}) + (1-\alpha)\widetilde{\phi}_{ij}(\tilde{\boldsymbol{q}})$$

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

$$\lim_{\tau\to\infty} \widetilde{\phi}\left(\frac{\boldsymbol{q}_\tau}{\|\boldsymbol{q}_\tau\|}\right) = \alpha\boldsymbol{y} + (1-\alpha)\tilde{\boldsymbol{y}} \in \overline{\Lambda}.$$

□

F. STATEMENT AND PROOF OF LEMMA 6

Lemma 6. *Suppose that $D(\Pi,\Lambda) = \inf_{r\in\Pi, y\in\Lambda}|r-y| = 0$. Given any $\epsilon > 0$ and $t_0 \geq 0$, there exist a closed interval $[t_0, t_1]$ and two signals $\boldsymbol{x} \in \mathcal{L}^{n_s}$ and $\boldsymbol{q} \in \mathcal{L}^{n_a n_s}$ with $\|\boldsymbol{q}\| = 1$ such that*

$$\widetilde{\phi}_{ij}(\boldsymbol{x},\boldsymbol{q}) \geq 0, \qquad \forall i \in [n_a], j \in [n_s] \tag{45}$$

$$\epsilon^2 > \|(I - \Gamma_{[t_0,t_1]})G\boldsymbol{q}\| \tag{46}$$

$$\epsilon = \|x - \Gamma_{[t_0,t_1]}G\boldsymbol{q}\|_{\Omega_{ij,x}}, \tag{47}$$

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

Proof. For a given $\epsilon > 0$, since $D(\Pi,\Lambda) = 0$, there exists $\boldsymbol{q} \in L^{n_a n_s}$ with $\|\boldsymbol{q}\| = 1$ satisfying $\widetilde{\phi}_{ij}(\boldsymbol{x},\boldsymbol{q}) > -\epsilon^2$ for all $i \in [n_a]$ and $j \in [n_s]$, i.e.,

$$\epsilon^2 + \|G\boldsymbol{q}\|^2_{\Omega_{ij,x}} > \|\boldsymbol{q}\|^2_{\Omega_{ij,q}},$$

where $\Omega_{ij,x}$ is defined previously. Clearly, if $\boldsymbol{q}$ is truncated to a sufficiently long interval, and $\boldsymbol{q}$ is rescaled to have a unit norm, the above inequality will still hold. Since $G\boldsymbol{q} \in \mathcal{L}^{n_s}$, by possibly enlarging the truncation interval to $[t_0,t_1]$, we obtain (46), and

$$\epsilon^2 + \|\Gamma_{[t_0,t_1]}G\boldsymbol{q}\|^2_{\Omega_{ij,x}} > \|\boldsymbol{q}\|^2_{\Omega_{ij,q}},$$

Next, we choose $\boldsymbol{\eta} \in \mathcal{L}^{n_s}$ such that $\|\boldsymbol{\eta}\|^2_{\Omega_{ij,x}} = \epsilon^2$, and that $\boldsymbol{\eta}$ is orthogonal to $\Gamma_{[t_0,t_1]}G\boldsymbol{q}$. Then, by considering $\boldsymbol{x} = \Gamma_{[t_0,t_1]}G\boldsymbol{q} + \boldsymbol{\eta}$, we obtain

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

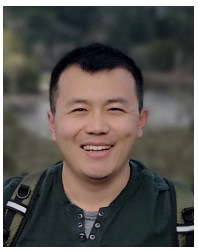
which leads to $\widetilde{\phi}_{ij}(\boldsymbol{x},\boldsymbol{q}) \geq 0$ and (47). Now, we can invoke Lemma 2 to construct $\tilde{\pi} \in \mathcal{P}(\xi)$ based on (45) such that $\tilde{\pi}$ becomes sector bounded and $\boldsymbol{q} = \tilde{\pi}\boldsymbol{x}$. Then,

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

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

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

. □



MING JIN is an Assistant Professor in the Department of Electrical and Computer Engineering at Virginia Tech. He received his doctoral degree from EECS department at University of California, Berkeley in 2017 and was a postdoctoral researcher in the Department of Industrial Engineering and Operations Research at University of California, Berkeley. 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.