REVIEW OF METRICS TO MEASURE THE STABILITY, ROBUSTNESS AND RESILIENCE OF REINFORCEMENT LEARNING

Laura L. Pullum

Mathematics and Computer Science Division, Oak Ridge National Laboratory, Bethel Valley Road, Oak Ridge, USA

ABSTRACT

Reinforcement learning (RL) has received significant interest in recent years, primarily because of the success of deep RL in solving many challenging tasks, such as playing chess, Go, and online computer games. However, with the increasing focus on RL, applications outside gaming and simulated environments require an understanding of the robustness, stability, and resilience of RL methods. To this end, we conducted a comprehensive literature review to characterize the available literature on these three behaviors as they pertain to RL. We classified the quantitative and theoretical approaches used to indicate or measure robustness, stability, and resilience behaviors. In addition, we identified the actions or events to which the quantitative approaches attempted to be stable, robust, or resilient. Finally, we provide a decision tree that is useful for selecting metrics to quantify behavior. We believe that this is the first comprehensive review of stability, robustness, and resilience, specifically geared toward RL.

KEYWORDS

Reinforcement Learning, Resilience, Robustness, Stability

1. INTRODUCTION

Recent literature on the robustness of machine-learning models has focused almost entirely on the robustness of deep neural networks for imaging applications. However, at the time of this study, there were no published surveys on the robustness of reinforcement learning (RL). We pursued this review because of the increasing use of RL, particularly in control systems. Along with robustness, stability and resilience are included. Stability was included because the term has been used interchangeably with robustness, and resilience was included because the term has been used as a state beyond robustness.

RL involves agents that act in an environment and experience a reward for their actions. The agent learns the policy that maximizes the cumulative reward. Formally, consider an agent operating at time $t \in \{1, \dots, T\}$. At time t, the agent is in environment state s_t and produces an action $a_t \in A$. The agent then observes a new state s_{t+1} and receives reward $r_t \in R$. A set of possible actions A can be discrete or continuous. The goal of reinforcement learning is to find a policy $\pi(a_t|s_t)$ for choosing an action in state s_t to maximize the utility function or (expected return). [252]

David C. Wyld et al. (Eds): MLSC, ITCSS, ACSTY, SOFE, NATP, BDAB - 2023
pp. 59-78, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130205

$$J(\pi) = \boldsymbol{E}_{s_0,a_0,\ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \qquad (1)$$

Where $0 \leq \gamma \leq 1$ is a discount factor, $a_t \sim \pi(a_t|s_t)$ is drawn from the policy, and $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$ is generated by environmental dynamics. The state value function

$$V^{\pi}(s_t) = \boldsymbol{E}_{a_t,s_{t+1},\ldots}\left[\sum_{i=0}^{\infty} \gamma^i r(s_{t+1}, a_{t+1})\right] \qquad (2)$$

is the expected return by policy π from state s_t. The state action function

$$Q^{\pi}(s_t, a_t) = \boldsymbol{E}_{s_{t+1},a_{t+1},\ldots}\left[\sum_{i=0}^{\infty} \gamma^i r(s_{t+1}, a_{t+1})\right] \qquad (3)$$

is the expected return by policy π after taking action a_t at state s_t. [252].

The objective of this study is to present a systematic review of RL literature to identify metrics for measuring the stability, robustness, and resilience of RL. We limit RL to general reinforcement learning and not to specialized RL, such as inverse RL. We reviewed studies that attempted to measure or otherwise characterize stability and robustness. and resilience of RL, seeking metrics for these behaviors.

We searched computer science and technical literature databases for eligible papers, combining RL, behavior terms, and terms related to measuring, metrics, and quantification. The result comprised 16,015 items, and after removal of duplications and extraneous material, a collection of 546 items was established. Through the process of elimination described in full in this paper, we reduced the set to 248 papers. We systematically reviewed 248 papers and presented the results in this analysis.
We classified the papers by behavior (i.e., stability (*n*=76), robustness (*n*=169), and resilience (*n*=3)), and identified the primary domains of application as robotics, network systems, power system control, and vehicle/traffic control and navigation. We identified approaches to determine or measure each behavior individually and across behaviors. The approaches were categorized as quantitative or theoretical, and the quantitative approaches were further classified as being applied internally (e.g., in training) or externally (e.g., performance measures on outputs) to the model. The metrics, approaches, and objectives were identified for each paper reviewed. The objective indicates the metric or approach intended to be stable, robust, or resilient. We close by indicating the need to define stability, robustness, and resilience behaviors for RL and identify quantitative and theoretical approaches to achieve measurement and determination of these behaviors.
There is a rich set of domains (i.e., 53 identified in this survey) in which the measurement of RL stability, robustness, and resilience has been conducted. The domains ranged from robotics and network systems to sheep herding and fish behavior. The most frequently mentioned domains include robotics, general control, and network systems, with numerous studies not specifying a domain. Many studies used Gym [254] and other environments for demonstration purposes. Though the search focused on the quantitative measurement of stability, robustness, and resilience, theoretical approaches were identified as well. The quantitative approaches were categorized as internal or external depending on where the evaluation was conducted in the model. Internal measures quantified the performance of the training and external measures quantified the ultimate performance of the model.

The goal of this systematic review is to identify metrics for measuring the stability, robustness, and resilience of RL. To initiate the search for this review, we identified keywords and phrases related to reinforcement learning, the *behaviors* of interest (stability, robustness, and resilience), and measurement. The *key phra*se is reinforcement learning. The *measurement* keywords are metric, measure, index, score, quantifier and indicator.

We believe that this is the first comprehensive review of stability, robustness, and resilience specifically geared toward RL. The remainder of this paper is organized as follows. Section 2 describes the methods used in the systematic review. Section 3 presents the results of the review. Section 4 discusses the results of the review and introduces a decision tree for metric selection based on the review.

2. METHODS

Keywords salient to RL, system behavior, and measurement were identified for the research topic. The typical search was of the form:

<Key Phrase> + <Behavior> + <Measurement>

with <Key Phrase>, <Behavior> and <Measurement> defined above. A specific example is

“reinforcement learning” AND robust* AND (“metric” OR “measure” OR “index” OR “score” OR “quantifier” OR “indicator”)

Multiple searches were conducted using bibliographic databases covering broad areas of computer science, physical and biological sciences, and engineering. The information sources used in this study are the open-access arXiv covering 1991-present and the subscription services Scopus (1823-present) and Web of Science (1900-present). No restrictions were placed on the publication date or language. Journal articles, books, books in a series, book sections or chapters, edited books, theses and dissertations, conference papers, and technical reports containing keywords and phrases were included in the search. The publication date of the returned search results is bound by the dates of coverage of each database and the date on which the search was performed; however, all searches were completed by October 31, 2020. The range of dates for the documents ultimately included in the review was from 2002 to 2020.

The queried databases yielded 16,015 citations. Irrelevant citations were also retrieved. We excluded extraneous studies, resulting in a collection of 699 publications. Furthermore, the removal of duplicate papers resulted in 580 publications. Citations for “full conference proceedings were removed if the relevant paper(s) within the associated conference were otherwise collected, resulting in 546 publications. Further refinement excluded publications that were not on RL, which were not on the searched behavior, or those that had no metrics or theoretical content, resulting in 248 documents. We systematically reviewed 248 papers, and the results are presented in this analysis.

The 248 papers that made it through the screening process were grouped by search behavior: stability, robustness, and resilience. We also identified papers on one behavior that mentioned one or both other behaviors. Some studies that mentioned other behaviors did so interchangeably. For instance, stability and robustness have been used interchangeably in several studies, which can lead to some confusion in the definitions of these behaviors. The primary domains of application were identified and categorized as robotics, network systems, general control systems, Gym [254], and other environments. We also identified publications that mentioned the RL policy.

The primary focus of this study was to identify approaches to determine or measure each behavior. Of course, most publications reviewed focused on quantitative approaches because of the search terms used. Those that use a theoretical approach provide additional insight into the behavior-determination problem. The quantitative approaches were further classified as being applied internally (e.g., in training) or externally (e.g., performance measures on outputs) to the model.

Metrics, approaches, and objectives were identified for each study (see Figure 1). The objective indicates the metric or approach intended to be stable and robust, or resilient.

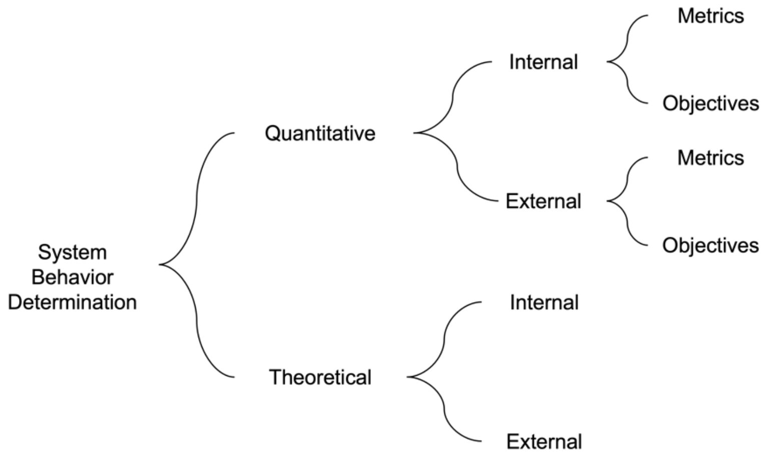

Figure 1. Categorization and resulting metrics, approaches, and objectives

There is little agreement in the literature on the definitions of stability, robustness, and resilience. In fact, there are few distinct definitions of these behaviors. In this review, we used the following definitions:

Stability is a property of the learning algorithm (i.e., a small change in the training set results in a similar model) and refers to the ranking of the variance of a model [253]. For example, if we use the variance of the loss function over all datasets as a performance measure, we test a set of models. The smallest loss indicated a more stable model. Given this definition, stability analysis is an application of sensitivity analysis to machine learning.

Robustness, when used with respect to computer software, refers to an operating system or other program that performs well not only under ordinary conditions but also under unusual conditions that stress its designers' assumptions (http://www.linfo.org/robust.html). Robustness is a property of the model and is measured by, for example, loss over all datasets (as opposed to the variance of the loss).

Throughout the literature, *resilience* has been used interchangeably with robustness; however, it is used most often with production machine learning systems to indicate robustness to different datasets and different data added to the dataset.

3. RESULTS AND ANALYSIS

Publications were categorized by behavior as follows: stability (*n*=76) [4-80], robustness (*n*=169) [81-169], and resilience (*n*=3) [1-3]. Studies on one behavior often mention other behaviors, especially stability and robustness. Resilience was mentioned in five stability papers and 11 robustness papers. Robustness was mentioned in 50 stability papers and in one resilience paper. Stability was mentioned in 104 Robustness papers and in all (3) Resilience papers.

Given the recent explosion of literature on the robustness of neural networks to adversarial attacks, one might expect it to be a cornerstone of the robustness papers reviewed herein. The term "adversarial" was mentioned in a quarter (*n*=61, *N*=248) of the papers reviewed. That is, 1 resilience paper, 56 robustness papers, and 4 stability papers mention "adversarial". Some papers on one behavior used one of the other behaviors interchangeably, notably stability and robustness, specifically [91, 93, 105, 145, 146, 179, 194, 225, and 237] and generally in several other articles.

3.1. Application Domains

The publication application domains are provided in the supplementary information and summarized in Figure 2. The primary domains were robotics, with 16.4% (*n*=44) of the total citations (*N*=268), followed by network systems and general control (*n*=7.8%, *n*=21), with 9.3% (*n*=25) using Gym or other environments as their experimental domain. Just as many (*n*=25, 9.3%) papers did not specify a domain. These top 5 (*n*=53) domains comprised over 50% (52.9%, *n*=136) of citations. Most (52.8%, *n*=28) domains (*n*=53) had a single citation.

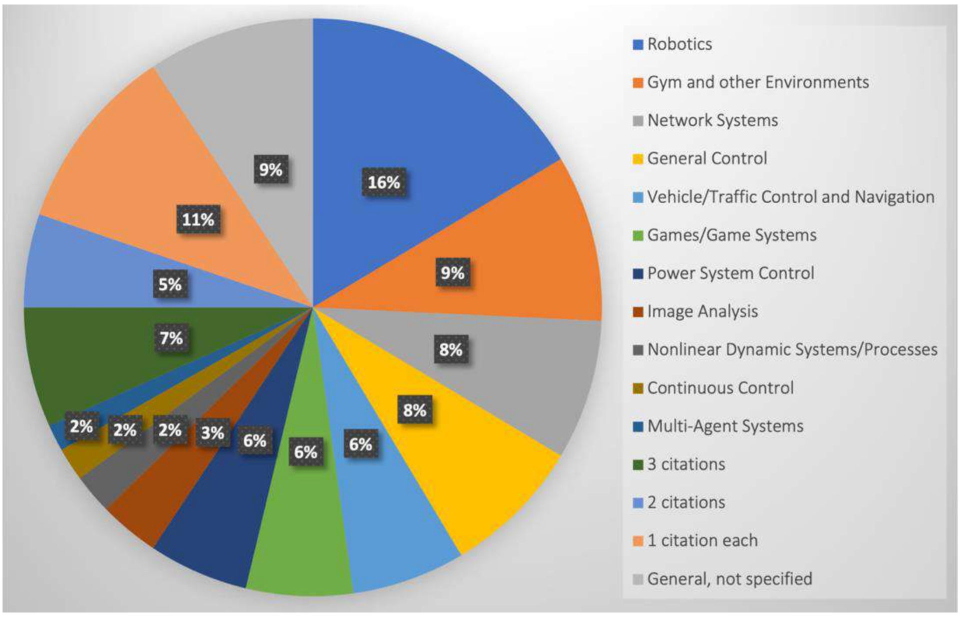

Figure 2. Application Domain Categories

3.2. Reinforcement Learning Policies

Tweny-one (21) RL policies were mentioned in the articles. Most documents did not identify the policies used. Of the 21 types of policies mentioned, the top 4 – Actor-Critic (*n*=18), Q-learning (*n*=16), Proximal Policy Optimization (PPO) (*n*=8), and Adaptive Critic Design (*n*=5) comprised 72.3% of the total citations that included policy (*N*=65).

3.3. Approach to Determining or Measuring Behavior

The publications' approaches to determining or measuring each behavior are categorized as either quantitative or theoretical. Most of the publications focused on quantitative approaches (*n*=205, 82.0%), which is understandable given that the search focused on quantifying behaviors. For publications on stability behavior, there was an almost even split between the quantitative (*n*=42) and theoretical (*n*=43) approaches. However, publications on robustness behavior have primarily focused on quantitative approaches (*n*=160) vice theoretical (*n*=35). All (3) publications on resilience applied quantitative approaches.

3.3.1. Types of Quantitative Approaches

Next, we further categorized the quantitative approaches according to whether they were focused internal or external to the model. Internal quantitative approaches measure aspects within the model, such as its training and associated measures, including the value of rewards over time or the number of episodes until convergence. External quantitative approaches measure performance-related aspects of a model, such as variations in accuracy or throughput. Most (n=142, 63.1%) quantitative approaches were categorized as performance-related or external measures. Of these, most (n=103) were for robustness, followed by those for stability (n=36). The 3 papers on resilience focused on performance-related quantitative measures. Robustness also led to internal approaches (n=69) with stability (n=14). This is primarily due to the large number of robustness papers (n=170) and paucity of resilience papers (n=3). Of the robustness papers, 40.0% (n=69) contained internal quantitative measures, and 60.6% contained external quantitative measures. The stability values were 18.2% and 46.8%, respectively.

3.3.2. Types of Internal Quantitative Approaches

Looking at the types of internal quantitative approaches, we see a narrow set of aspects considered in the papers. These metrics are specifically designed to measure stability rather than the variance of the output. They measured the variation in training performance. The vast majority (n=75, 88.2%) of the internal quantitative approaches calculated the reward- or score-based metrics. Other types of internal quantitative approaches include two each of policy entropy, variations in control strategy approximation weights, and convergence rate, and one each of policy weight, calculation of the Lyapunov stability criteria, and calculation of the Wasserstein function lower bound. In RL context, convergence refers to the stability of the learning process (and the underlying model) over time [11].

3.3.3. Types of External Quantitative Approaches

External or performance-based quantitative approaches for measuring behaviors primarily (n=39) used deviations or variations in performance-related metrics other than precision, accuracy, or recall (Figure 3). The next highest category (n=28) of quantitative metrics used error, failure, and success rates. Statistics on the performance of the tracking or estimation error follow, with n=23 papers. Papers in the network domain used network-related metrics (n=15) to measure behavior. Statistics on precision, accuracy, and recall (n=12) were also used. Five papers used variance in loss or regret estimation, three papers used game-related performance measures to quantify behavior, and two papers each used bounds on or the size of the stability region and terminal wealth and inventory. Eighteen (18) additional different types of external quantitative metric categories were represented by a single paper each.

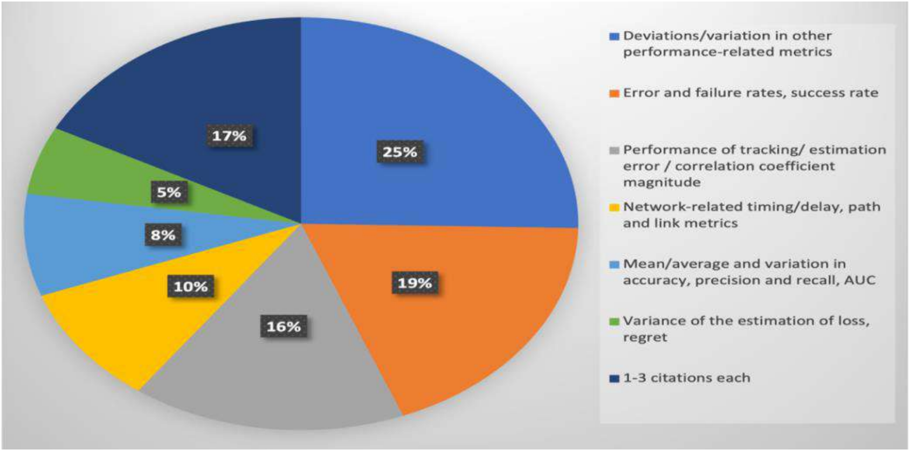

Figure 3. External Quantitative Metrics

3.3.4. Quantitative Approach Objectives

An additional aspect reviewed was to determine to what actions or events were the quantitative approaches attempting to be stable, robust, or resilient. We call this the <*behavior*> *objective*. The <*behavior*> objective category (see Figure 4), with the highest number of citations, was geared toward handling changes in the operational environment, dynamic environment, or network (*n*=41). Papers that did not specifically state their objectives comprised the next most populous category (*n*=35). The objective of handling uncertainties and disturbances in the environment also contained *n*=35 papers. The remaining objectives included input variation/perturbations (*n*=20), differences between training and test or operational environments (*n*=19), differences or uncertainties in model parameters (*n*=16), adversarial attack (*n*=14), different domains, environments, or settings (*n*=8), errors or failures in the operational environment (*n*=5), differences in training datasets or initializations (*n*=5), high variability (*n*=2), and one paper each in systematic pressure, spamming, incomplete data, and unknown control coefficients.

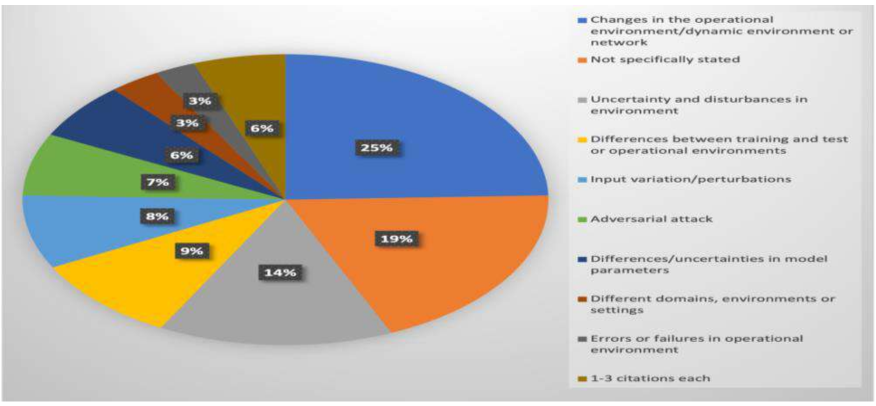

Figure 4. Quantitative <*behavior*> Objectives

3.3.5. Types of Theoretical Approaches

Most of the theoretical approaches in the papers reviewed were based on the Lyapunov theory (*n*=50, 61.0%) (Figure 5). The next highest types of theoretical approaches used are convergence to Nash equilibrium (*n*=10) and value-based guarantees, such as error and output deviation bounds (*n*=8). Of the remainder, three papers used the Wasserstein distance to explore stability, three studies proved that the methods were doubly robust, two papers proved that the methods exhibited Lipschitz continuity, and stochastic stability theory to prove stability, stability guarantees, policy-based guarantees, regret bounds, minimization of the Jacobian on input, and per-episode Bellman-error regret guarantees/bounds were used by a single paper each to establish the stability of the RL methods discussed.

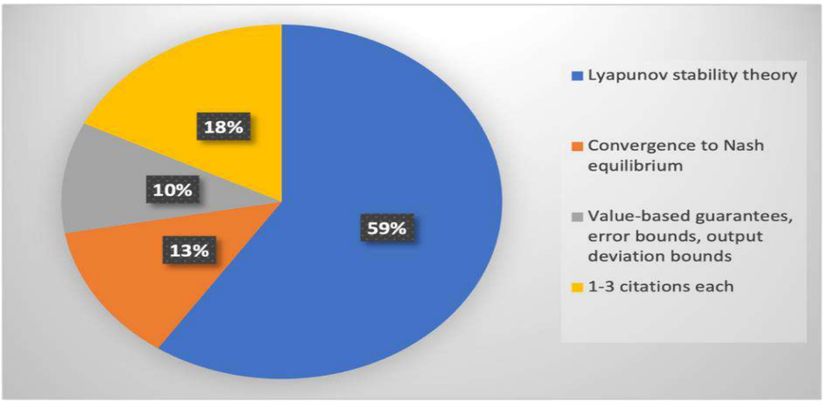

Figure 5. Theoretical approaches

3.3.6. Theoretical Approach Objectives

We also reviewed the <*behavior*> objective for theoretical papers (Figure 6). Most papers (*n*=42, 54.5%) on theoretical approaches did not state their objectives. Of the few that did, changes or dynamics in the operational environment were the most frequent objective (*n*=10), followed by differences or uncertainties in model parameters (*n*=7), adversarial attack (*n*=6), error or failure (*n*=5), differences between training and test or operational environments (*n*=2), input variation (*n*=2), and one each for domain shifts, different function approximation architectures, and differences in quantization levels.

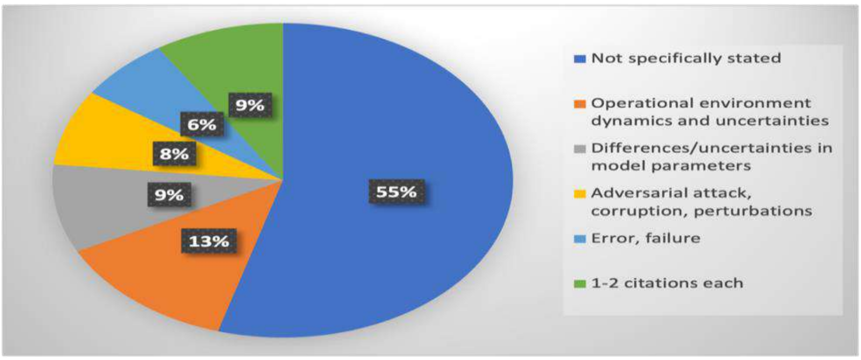

Figure 6. Theoretical <*behavior*> Objectives

4. Discussion

Our study was conducted to characterize the published methods of measuring or determining the stability, robustness, or resilience of RL. Of an initial collection of 16,015 items, 248 papers met the inclusion criteria and were systematically reviewed. Approaches to measuring or determining behavior are classified as either quantitative or theoretical. Quantitative approaches were further classified as internal or external depending on whether they evaluated the training, test, or operational phases. For both categories of quantitative approaches, we categorized the metrics used, with internal approaches primarily using the reward or score (and statistics on the same) and external approaches primarily using variations in performance-related metrics (although not precision, accuracy, or recall). The theoretical approaches were dominated by Lyapunov stability theory. We further characterized the objectives of stability, robustness, and resilience. Quantitative approaches to measuring behavior focused on the ability to handle differences in the operational environment, whereas most theoretical approaches to determining behavior did not specifically state an objective. However, the objective of the theoretical approaches can be implied using Lyapunov stability theory, that is, to prove the stability of the system. Lyapunov was used, regardless of whether the article was on stability or robustness.

To determine the metric to use, we developed a decision tree based on the information obtained in this literature review. It is a collapsible tree, so that branches are not exposed unless selected, and open branches can be closed or collapsed. There are several levels in the decision tree, starting with i) behavior (stability, robustness, or resilience); ii) the domain; iii) a list of quantitative and theoretical objectives; iv) the next level divides the metrics into external, internal, and theoretical metrics; and v) the last level, that is, the leaves, is the set of metrics for that branch of the decision tree. For example, suppose we want to find a suitable metric to measure the robustness of a control system expected to face changes in the operational environment. From the metric decision tree shown in Figure 7, we can see that the first selection is for a robustness metric. This selection displays the domains in which the robustness metrics are described. Selecting the General Control domain reveals 9 objectives, including the objective “Dynamic Environment.” An external metric found in the literature for this case is “blood glucose response” which is not applicable for this control system. The more appropriate metrics and approaches are the size of the stability region, value-based guarantees, error bounds, and Lyapunov stability theory and calculation. Any or all of these can be used to measure the robustness of a general control system in a dynamic operational environment.

Supplementary information for this review is provided at https://arxiv.org/pdf/2203.12048.pdf, including a) PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [251] diagrams for Stability, Robustness and Resilience, respectively; b) the data reduction methodology for Stability, Robustness and Resilience, respectively; and the PRISMA checklist. In addition, the site provides detailed tables of the results described in Section 3.

ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
DOE	Department of Energy
ORNL	Oak Ridge National Laboratory
PPO	Proximal Policy Optimization
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RL	Reinforcement Learning
US	United States

ACKNOWLEDGMENTS

The author would like to acknowledge Rama Vasudevan, PhD of the Oak Ridge National Laboratory (ORNL) for intellectual discussions and collaborative research on reinforcement learning. The author would also like to thank Nathan Martindale (ORNL) for assistance in improving the functionality and usability of the decision tree.

This work was funded initially by the AI Initiative at the Oak Ridge National Laboratory and subsequently funded by the US Department of Energy, National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation Research and Development (NA-22). This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy (DOE). The publisher, by accepting the article for publication, acknowledges that the US government retains a non-exclusive, paid up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for U.S. Government purposes. The DOE will provide public access to these results in accordance with the DOE Public Access Plan *(http://energy.gov/downloads/doe-public-access-plan)*.

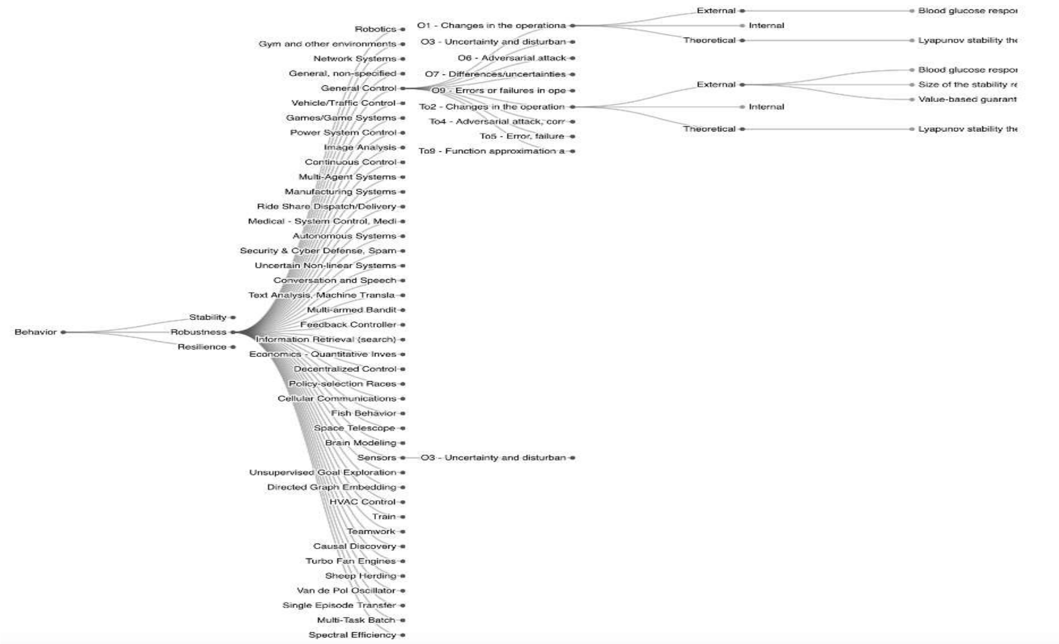

Figure 7. Metric Selection Decision Tree Section

REFERENCES

[1] V. Behzadan and A. Munir, "Adversarial exploitation of emergent dynamics in smart cities," Proc 2018 IEEE Intl Smart Cities Conf, doi 10.1109/ISC2.2018. 8656789.

[2] S. Enjalber and F. Vanderhaegen, "A hybrid reinforced learning system to estimate resilience indicators," Eng Appl of AI, vol. 64, pp. 295-301, 2017.

[3] M. Bunyakitanon, et al, "End-to-end performance-based autonomous vnf placement with adopted reinforcement learning," IEEE Trans on Cognitive Comms and Networks, vol. 6, no. 2, pp. 534-547, 2020, doi 10.1109/TCCN.2020.2988486.

[4] Z. Dong, X. Huang, Y. Dong, and Z. Zhang, "Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system," J. Appl Energy, vol. 259, 2020, doi 10.1016/j.apenergy.2019.114193.

[5] G. Wen, et al, "Optimized adaptive nonlinear tracking control using actor-critic reinforcement learning strategy," IEEE Trans on Industrial Informatics, vol. 15, no. 9, pp. 4969-4977, 2019.

[6] B. Muneeswari and M.S.K. Manikandan. "Energy efficient clustering and secure routing using reinforcement learning for three-dimensional mobile ad hoc networks," IET Commun, vol. 13, no. 12, pp. 1828-1839, 2019.

[7] B. A. G. de Oliveira, C. A. P. da S. Martins, F. Magalhaes, L. Fabricio, and W. Goes, "Difference based metrics for deep reinforcement learning algorithms," IEEE Access, vol. 7, pp. 159141-159149, 2019.

[8] K. Zhang, et al, “Policy search in infinite-horizon discounted reinforcement learning: advances through connections to non-convex optimization,” in Proc: 53rd CISS, Baltimore, MD, USA, 2019.

[9] Z. Du, W. Wang, Z. Yan, W. Dong, and W. Wang, “Variable admittance control based on fuzzy reinforcement learning for minimally invasive surgery manipulator,” Sensors, vol. 17, no. 4, 2017.

[10] H. Jiang, et al, “Optimal tracking control for completely unknown nonlinear discrete-time Markov jump systems using data-based reinforcement learning method,” Neurocomputing, vol. 194, pp. 176-182, 2016.

[11] S. Abdallah, “Why global performance is a poor metric for verifying convergence of multi-agent learning,” arXiv:0904. 2320v1 [cs.MA] 15 April 2009.

[12] N. Talele and K. Byl, “Mesh-based tools to analyze deep reinforcement learning policies for underactuated biped locomotion,” arXiv:1903.12311v2 [cs.RO] 1 November 2019.

[13] Y.-L. Tuan, J. Zhang, Y. Li, and H.-y. Lee, “Proximal policy optimization and its dynamic version for sequence generation,” arXiv:1808.07982v1 [cs.CL] 24 August 2018.

[14] A. Serhani, et al, “AQ-Routing: mobility-, stability-aware adaptive routing protocol for data routing in MANET–IoT systems,” Cluster Comp, vol. 23, pp. 13-27, 2020, doi 10.1007/s10586-019-02937-x.

[15] Z. Dong, et al, “Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system,” Applied Energy, vol. 259, 2020, doi 10.1016/j.apenergy. 2019.114193.

[16] H. Zhang, K. Zhang, Y. Cai, and J. Han, “Adaptive fuzzy fault-tolerant tracking control for partially unknown systems with actuator faults via integral reinforcement learning method,” IEEE Trans on Fuzzy Systems, vol. 27, no. 10, 2019, doi 10.1109/TFUZZ.2019.2893211.

[17] D. Cohen, S. M. Jordan, and W. B. Croft, “Learning a better negative sampling policy with deep neural networks for search,” in ICTIR ’19, Santa Clara, CA, USA, 2019, doi 10.1145/3341981.3344220.

[18] C. Mu, et al, “Q-learning solution for optimal consensus control of discrete-time multiagent systems using reinforcement learning,” J Frank Inst, vol. 356, pp. 6946–6967, 2019, 10.1016/j.jfranklin.2019.06.0070016-0032.

[19] X. Tang, et al, “A deep value-network based approach for multi-driver order dispatching,” in KDD 19, Anchorage, AK, USA, 2019, doi 10.1145/3292500.3330724.

[20] M. Abouheaf, and W. Gueaieb, “Model-free adaptive control approach using integral reinforcement learning,” in Proc. IEEE Intl Symp on Robotic and Sensors Environ, 2019.

[21] D. Seo, H. Kim, and D. Kim, “Push recovery control for humanoid robot using reinforcement learning,” Third IEEE IRC, 2019, doi 10.1109/IRC.2019.00102.

[22] Y. Lv, X. Ren, and J. Na, “Online Nash-optimization tracking control of multi-motor driven load system with simplified RL scheme,” ISA Trans, 2019, doi 10.1016/j.isatra.2019.08.025.

[23] L. Tang, Y.-J. Liu, and C. L. P. Chen, “Adaptive critic design for pure-feedback discrete-time MIMO systems preceded by unknown backlashlike hysteresis,” IEEE Trans on Neural Networks and Learning Syst., vol. 29, no. 11, 2018, doi 10.1109/TNNLS.2018.2805689.

[24] P. Mertikopoulos, and W. H. Sandholm, “Riemannian game dynamics,” J of Econ Theory, vol. 177, pp. 315-364, 2018, doi 10.1016/j.jet.2018.06.002.

[25] D. Liu, and G.-H. Yang, “Model-free adaptive control design for nonlinear discrete-time processes with reinforcement learning techniques,” Intl J of Systems Science, vol. 49, no. 11, pp. 2298-2308, 2018, doi 10.1080/00207721.2018.1498557.

[26] A. Bentaleb, et al, “ORL-SDN: Online reinforcement learning for SDN-enabled HTTP adaptive streaming,” ACM Trans. Multimedia Comput. Commun. Appl, vol. 14, no. 3, 2018, Art. no. 71, doi 10.1145/3219752.

[27] Y. Hu and B. Si, “A reinforcement learning neural network for robotic manipulator control,” Neural Computation, vol. 30, no. 7, pp. 1983-2004, 2018, doi 10.1162/neco_a_01079.

[28] Y. Mei, et al, “Chaotic time series prediction based on brain emotional learning model and self-adaptive genetic algorithm,” Acta Physica Sinica, vol. 67, no. 8,2018, doi 10.7498/aps. 67.20172104.

[29] Z.-W. Hong, S.-Y. Su, T.-Y. Shann, Y.-H. Chang, and C.-Y. Lee, “A deep policy inference Q-network for multi-agent systems,” in Proc. AAMAS 2018, Stockholm, Sweden, 2018.

[30] Y. Xiong, H. Chen, M. Zhao, and B. An, “HogRider: Champion agent of Microsoft Malmo collaborative AI challenge,” in AAAI-18, pp. 4767-4774, 2018.

[31] W. Wu and L. Gao, “Posture self-stabilizer of a biped robot based on training platform and reinforcement learning,” Robotics and Autonomous Systems, vol. 98, pp. 42-55, 2017, doi 10.1016/j.robot. 2017.09.001.

[32] M. Boushaba, A. Hafid, and M. Gendreau, “Node stability-based routing in wireless mesh networks,” J of Network and Computer Appl, vol. 93, pp. 1-12, 2017, doi 10.1016/j.jnca.2017.02.010.
[33] G. C. Chasparis, “Stochastic stability analysis of perturbed learning Automata with constant step-size in strategic-form games,” in Proc. ACC, Seattle, WA, USA, 2017, pp. 4607-4612.
[34] N. W. Prins, J. C. Sanchez and A. Prasad, “Feedback for reinforcement learning based brain-machine interfaces using confidence metrics,” J of Neural Eng, 2017, doi 10.1088/1741-2552/aa6317.
[35] R. Song, F. L. Lewis, and Q. Wei, “Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games,” IEEE Trans on Neural Networks and Learning Systems, vol. 28, no. 3, 2017, doi 10.1109/TNNLS.2016.2582849.
[36] R. Yousefian, et al, “Hybrid transient energy function-based real-time optimal wide-area damping controller,” IEEE Trans on Industry Appls, vol. 53, no. 2, 2017, doi 10.1109/TIA.2016.2624264.
[37] F. Tatari, M.-B. Naghibi-Sistani, and K. G. Vamvoudakis, “Distributed learning algorithm for non-linear differential graphical games,” Trans of the Inst of Measurement and Control, pp. 1-10, 2015.
[38] C. Lu, J. Huang, and J. Gong, “Reinforcement learning for ramp control: an analysis of learning parameters,” Promet – Traffic & Transportation, vol. 28, no. 4, pp. 371-381, 2016.
[39] K. G. Vamvoudakis, “Optimal trajectory output tracking control with a Q-learning algorithm,” in Proc of the American Control Conf, pp. 5752-5757, 2016, doi 10.1109/ACC.2016. 7526571.
[40] P. H. M. Rêgo, et al, “Convergence of the standard RLS method and UDUT factorisation of covariance matrix for solving the algebraic Riccati equation of the DLQR via heuristic approximate dynamic programming,” Intl J of Systems Science, 2013, doi 10.1080/00207721.2013.844283.
[41] D. Liu and Q. Wei, “Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems,” IEEE Trans on Neural Networks and Learning Systems, vol. 25, no. 3, 2014.
[42] A. Alharbi, A. Al-Dhalaan, and M. Al-Rodhaan, “Q-routing in cognitive packet network routing protocol for MANETs,” in Proc NCTA-2014, pp. 234-243, 2014, doi 10.5220/0005082902340243.
[43] R. Yousefian and S. Kamalasadan, “An approach for real-time tuning of cost functions in optimal system-centric wide area controller based on adaptive critic design,” in IEEE PESGM, 2014, doi 10.1109/PESGM.2014.6939224.
[44] B. Dong and Y. Li, “Decentralized reinforcement learning robust optimal tracking control for time varying constrained reconfigurable modular robot based on ACI and Q-function,” Mathematical Problems in Eng, 2013, Art. no. 387817, doi 10.1155/2013/387817.
[45] C. Teixeira, et al, “Biped locomotion - improvement and adaptation,” in Proc. ICARSC, Espinho, Portugal, 2014.
[46] N. T. Luy, et al, “Reinforcement learning-based intelligent tracking control for wheeled mobile robot,” Trans of the Inst of Measurement and Control, vol. 36, no. 7, pp. 868–877, 2014, doi 10.1177/0142331213509828.
[47] L. vS. Hager, et al, “Series-parallel approach to on-line observer based neural control of a helicopter system,” in Proc 19th World Congress the Intl Fedn of Autom Cntrl, Cape Town, South Africa, 2014.
[48] Q. Wei and D. Liu, “Data-driven neuro-optimal temperature control of water-gas shift reaction using stable iterative adaptive dynamic programming,” IEEE Trans on Industrial Electr, vol. 61, no. 11, 2014.
[49] D. Zhao, B. Wang, and D. Liu, “A supervised Actor-Critic approach for adaptive cruise control,” Soft Comput, vol. 17, pp. 2089–2099, 2013, doi 10.1007/s00500-013-1110-y.
[50] M. Kashki, et al, “Power system dynamic stability enhancement using optimum design of PSS and static phase shifter based stabilizer,” Arab J Sci Eng, vol. 38, pp. 637–650, 2013, doi 10.1007/s13369-012-0325-z.
[51] C. Li, R. Lowe, and T. Ziemke, “Humanoids learning to walk: a natural CPG-actor-critic architecture,” Frontiers in Neurorobotics, vol. 7, 2013, Art. no. 5, doi 10.3389/fnbot.2013.00005.
[52] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, “Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality,” Automatica, vol. 48, no. 8, pp. 1598-1611, 2012, doi 10.1016/j.automatica.2012.05.074.
[53] P. Moradi, et al, “Automatic skill acquisition in reinforcement learning using graph centrality measures,” Intelligent Data Analysis, vol. 16, no. 1, pp. 113-135, 2012, doi 10.3233/IDA-2011-0513.
[54] S. Bhasin, et al, “Asymptotic tracking by a reinforcement learning-based adaptive critic controller,” J Control Theory Appl, vol. 9, no. 3, pp. 400-409, 2011, doi 10.1007/s11768-011-0170-8.
[55] R. Hafner and M. Riedmiller, “Reinforcement learning in feedback control: Challenges and benchmarks from technical process control,” Machine Learning, vol. 84, pp. 137-169, 2011, doi 10.1007/s10994-011-5235-x.

[56] N. T. Luy, “Reinforcement learning-based tracking control for wheeled mobile robot,” in IEEE Intl Conf on Systems, Man, and Cybernetics, Seoul, Korea, 2012.
[57] P. Shih, B. C. Kaul, S. Jagannathan, and J. A. Drallmeier, “Reinforcement-learning-based output-feedback control of nonstrict nonlinear discrete-time systems with application to engine emission control,” IEEE Trans on Systems, Man, and Cybernetics—Part B: Cybernetics, vol. 39, no. 5, 2009.
[58] M. J. L. Boada, et al, “Active roll control using reinforcement learning for a single unit heavy vehicle,” Intl J of Heavy Vehicle Systems, vol. 16, no. 4, pp. 412-430, 2009, doi 10.1504/IJHVS.2009.027413.
[59] L. Guo, Y. Zhang, and J.-L. Hu, “Adaptive HVDC supplementary (lamping controller based on reinforcement learning,” Electric Power Automation Equip, vol. 27, no. 10, pp. 87-91, 2007.
[60] C.-K. Lin, “A reinforcement learning adaptive fuzzy controller for robots,” Fuzzy Sets and Systems, vol. 137, no. 3, pp. 339-352, 2003, doi 10.1016/ S0165-0114(02)00299-3.
[61] S. Jagannathan, “Adaptive critic neural network-based controller for nonlinear systems,” in Proc 2002 IEEE Intl Symp on Intelligent Control, Vancouver, Canada, 2002.
[62] B. H. Kaygisiz, A. M. Erkmen, and I. Erkmen, “Smoothing stability roughness of fractal boundaries using reinforcement learning,” in IFAC Procs Vols, vol. 15, no. 1, pp. 481-485, 2002.
[63] J. N. Li, et al, “Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes,” IEEE Trans on Cybernetics, vol. 50, no. 9, pp. 4132-4145, 2020, doi 10.1109/TCYB.2019.2950262.
[64] K. Zhang, H. G. Zhang, Y. L. Cai, and R. Su, “Parallel optimal tracking control schemes for mode-dependent control of coupled Markov jump systems via integral RL method,” IEEE Trans on Automation Science and Eng, vol. 17, no. 3, pp. 1332-1342, 2020, doi 10.1109/TASE.2019.2948431.
[65] Q. Zhang, et al, “Route planning and power management for phevs with reinforcement learning,” IEEE Trans on Vehicular Tech, vol. 69, no. 5, pp. 4751-4762, 2020, doi 10.1109/TVT.2020.2979623.
[66] A. Serhani, et al, “AQ-Routing: mobility-, stability-aware adaptive routing protocol for data routing in MANET-IoT systems,” Cluster Comp, v 23, no. 1, pp. 13-27, 2020, doi 10.1007/s10586-019-02937-x.
[67] Z. Dong, et al, “Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system,” Appl Energy, vol. 259, 2020, doi 10.1016/j.apenergy.2019.114193.
[68] Q. Wang, “Integral reinforcement learning control for a class of high-order multivariable nonlinear dynamics with unknown control coefficients,” IEEE Access, vol. 8, pp. 86223-86229, 2020, doi 10.1109/ACCESS.2020.2993265.
[69] J. Zhang, Z. Peng, J. Hu, Y. Zhao, R. Luo, B. K. Ghosh, “Internal reinforcement adaptive dynamic programming for optimal containment control of unknown continuous-time multi-agent systems,” Neurocomputing, vol. 413, pp. 85-95, 2020, doi 10.1016/j.neucom.2020. 06.106.
[70] A. Mitriakov, et al, “Staircase traversal via reinforcement learning for active reconfiguration of assistive robots,” in Proc IEEE Intl Conf on Fuzzy Systems, 2020, doi 10.1109/FUZZ48607.2020. 9177581.
[71] Y. Lv, et al, “Online Nash-optimization tracking control of multi-motor driven load system with simplified RL scheme,” ISA Trans, vol. 98, pp. 251-262, 2020, doi 10.1016/j.isatra.2019. 08.025.
[72] C. E. Thornton, et al, “Deep reinforcement learning control for radar detection and tracking in congested spectral environments,” IEEE Trans on Cognitive Commun and Networking, 2020, doi 10.1109/TCCN.2020.3019605.
[73] J. D. Prasanna, et al, “Reinforcement learning based virtual backbone construction in manet using connected dominating sets,” J of Critical Reviews, vol. 7, no. 9, pp. 146-152, 2020, doi 10.31838/jcr.07.09.28.
[74] J. Pongfai, X. Su, H. Zhang, and W. Assawinchaichote, “PID controller autotuning design by a deterministic Q-SLP algorithm,” IEEE Access, vol. 8, pp. 50010-50021, 2020, doi 10.1109/ACCESS.2020.2979810.
[75] S. Hoppe and M. Toussaint, “Q graph-bounded Q-learning: stabilizing model-free O-policy deep reinforcement learning,” arXiv:2007.07582v1 [cs.LG] 15 Jul 2020.
[76] P. Osinenko, L. Beckenbach, T. Göhrt, and S. Streif, “A reinforcement learning method with closed-loop stability guarantee,” arXiv:2006.14034v1 [math.OC] 24 Jun 2020.
[77] M. Han, L. Zhang, J. Wang, and W. Pan, “Actor-critic reinforcement learning for control with stability guarantee,” arXiv:2004.14288v3 [cs.RO] 15 Jul 2020.

[78] S. A. Khader, H. Yin, P. Falco and D. Kragic, "Stability-guaranteed reinforcement learning for contact-rich manipulation," arXiv:2004.10886v2 [cs.RO] 27 Sep 2020.
[79] M. Han, Y. Tian, L. Zhang, J. Wang, and W. Pan, "H infinity model-free reinforcement learning with robust stability guarantee," arXiv:1911.02875v3 [cs.LG], 2019.
[80] C. Tessler, N. Merlis, and S. Mannor, "Stabilizing deep reinforcement learning with conservative updates," arXiv:1910. 01062v2 [cs.LG], 2019.
[81] N. Abuzainab, et al, "QoS and jamming-aware wireless networking using deep reinforcement learning," arXiv:1910.05766v1 [cs.NI] 13 October 2019.
[82] M. Ahn, "ROBEL: Robotics benchmarks for learning with low-cost robots," arXiv:1909.11639v3 [cs.RO] 16 Dec 2019.
[83] V. Dhiman, S. Banerjee, B. Griffin, J. M. Siskind, and J. J. Corso, "A critical investigation of deep reinforcement learning for navigation," arXiv:1802.02274v2 [cs.RO], 2018.
[84] N. Naderializadeh, et al, "When multiple agents learn to schedule: a distributed radio resource management framework," arXiv:1906.08792v1 [cs.LG] 20 Jun 2019.
[85] K. Nguyen, H. Daumé III, and J. Boyd-Graber, "Reinforcement learning for bandit neural machine translation with simulated human feedback," arXiv:1707.07402v4 [cs.CL] 11 Nov 2017.
[86] N. Talele and K. Byl, "Mesh-based tools to analyze deep reinforcement learning policies for underactuated biped locomotion," arXiv:1903.12311v2 [cs.RO] 1 Nov 2019.
[87] M. Turchetta, A. Krause, and S. Trimpe, "Robust model-free reinforcement learning with multi-objective Bayesian optimization," arXiv:1910.13399v1 [cs.RO] 29 Oct 2019.
[88] Y. Yuan and K. Kitani, "Ego-pose estimation and forecasting as real-time PD control," arXiv:1906.03173v2 [cs.CV] 4 Aug 2019.
[89] B. Muneeswari and M. S. K. Manikandan, "Energy efficient clustering and secure routing using reinforcement learning for three-dimensional mobile ad hoc networks," IET Commun, vol. 13, no. 12, pp. 1828-1839, 2019, doi 10.1049/iet-com.2018.6150.
[90] B. Zhao, et al, "Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration," IEEE Trans on Sys Man Cybernetics-Syst, vol. 48, no. 10, 2018.
[91] Y. Zhang, et al, "Optimal design of residual-driven dynamic compensator using iterative algorithms with guaranteed convergence," IEEE Trans on Systems, Man, and Cybernetics: Systems, vol. 46, no. 4, 2016, doi 10.1109/TSMC.2015.2450203.
[92] M. Tokic, "Adaptive epsilon-greedy exploration in reinforcement learning based on value differences," in Lecture Notes in Artificial Intelligence, 33rd Annual German Conf on AI, Karlsruhe, Germany, 2010.
[93] Y. Xiong, L. Guo, Y. Huang, and L. Chen, "Intelligent thermal control strategy based on reinforcement learning for space telescope," J of Thermophysics and Heat Transfer, vol. 34, no. 1, pp. 37-44, 2020.
[94] R. F. Isa-Jara, G. J. Meschino, and V. L. Ballarin, "A comparative study of reinforcement learning algorithms applied to medical image registration," in IFMBE Procs, pp. 281-289, 2020.
[95] F. Guo, et al, "A reinforcement learning decision model for online process parameters optimization from offline data in injection molding," Applied Soft Computing J, vol. 85, 2019, doi 10.1016/j.asoc.2019.105828.
[96] S. Li, et al, "Design and implementation of aerial communication using directional antennas: Learning control in unknown communication environments," IET Control Theory and Apps, vol. 13, no. 17, pp. 2906-2916, 2019, doi 10.1049/iet-cta.2018.6252.
[97] X. Tang, et al, "A deep value-network based approach for multi-driver order dispatching," in Proc of the ACM SIGKDD Intl Conf on Knowl Discovery and Data Mining, pp. 1780-1790, 2019, doi 10.1145/3292500.3330724.
[98] A. Chowdhury, et al, "DA-DRLS: Drift adaptive deep reinforcement learning based scheduling for IoT resource management," J of Network and Computer Appls, vol. 138, pp. 51-65, 2019, doi 10.1016/j.jnca.2019.04.010.
[99] X. Wang, et al, "UAV first view landmark localization with active reinforcement learning," Pattern Recognition Letters, vol. 125, pp. 549-555, 2019, doi 10.1016/j. patrec.2019.03.011.
[100] B. Lütjens, et al, "Safe reinforcement learning with model uncertainty estimates," in Procs IEEE Intl Conf on Robotics and Automation, pp. 8662-8668, 2019, doi 10.1109/ICRA.2019.8793611.

[101] A. Balakrishnan and J. V. Deshmukh, "Structured reward functions using STL," in Proc of the 2019 22nd ACM Intl Conf on Hybrid Systems: Comput and Control, pp. 270-271, 2019, doi 10.1145/3302504.3313355.

[102] C. Tang, W. Zhu, and X. Yu, "Deep hierarchical strategy model for multi-source driven quantitative investment," IEEE Access, vol. 7, pp. 79331-79336, 2019, doi 10.1109/ACCESS.2019.2923267.

[103] Q. Cheng, X. Wang, Y. Niu, and L. Shen, "Reusing source task knowledge via transfer approximator in reinforcement transfer learning," Symmetry, vol. 11, no. 1, 2019, doi 10.3390/sym11010025.

[104] Y.-S. Jeon, H. Lee, and N. Lee, "Robust MLSD for wideband SIMO systems with one-bit ADCs: reinforcement-learning approach," in Proc ICC Workshops, pp. 1-6, 2018, doi 10.1109/ICCW.2018.8403665, 2018.

[105] X, Yang and H. He, "Self-learning robust optimal control for continuous-time nonlinear systems with mismatched disturbances," Neural Networks, vol. 99, pp. 19-30, 2018, doi 10.1016/j.neunet.2017.11.022.

[106] H. Jiang, H. Zhang, Y. Cui, and G. Xiao, "Robust control scheme for a class of uncertain nonlinear systems with completely unknown dynamics using data-driven reinforcement learning method," Neurocomputing, vol. 273, pp. 68-77, 2018, doi 10.1016/j.neucom.2017. 07.058.

[107] H. Shayeghi and A. Younesi, "An online Q-learning based multi-agent LFC for a multi-area multi-source power system including distributed energy resources," Iranian J of Electrical and Electronic Engineering, vol. 13, no. 4, pp. 385-398, 2017, doi 10.22068/IJEEE. 13.4.385.

[108] D. Zhao, Y. Ma, Z. Jiang, and Z. Shi, "Multiresolution airport detection via hierarchical reinforcement learning saliency model," IEEE J of Selected Topics in Applied Earth Observ and Remote Sensing, vol. 10, no. 6, pp. 2855-2866, 2017, doi 10.1109/JSTARS.2017.2669335.

[109] A. W. Tow, S. Shirazi, J. Leitner., N. Sünderhauf, M. Milford, and B. Upcroft, "A robustness analysis of deep Q networks," in Australasian Conf on Robotics and Automation, pp. 116-125, 2016.

[110] E. Hatami, and H. Salarieh, "Adaptive critic-based neuro-fuzzy controller for dynamic position of ships," Scientia Iranica, vol. 22, no. 1, pp. 272-280, 2015.

[111] J. Xiang and Z. Chen, "Adaptive traffic signal control of bottleneck subzone based on grey qualitative reinforcement learning algorithm," in Proc ICPRAM, vol. 2, pp. 295-301, 2015.

[112] R. Padmanabhan, et al, "Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning," Biomed Signal Process and Control, vol. 22, pp. 54-64, 2015, doi 10.1016/j.bspc.2015.05.013.

[113] R. Bruno, et al, "Robust adaptive modulation and coding (AMC) selection in LTE systems using reinforcement learning," in IEEE Vehicular Technology Conf, 2014, doi 10.1109/VTCFall.2014.6966162.

[114] N. Jamali, P. Kormushev, S.R. Ahmadzadeh, and D. G. Caldwell, "Covariance analysis as a measure of policy robustness," in OCEANS, Taipei, Taiwan, 2014, doi 10.1109/OCEANS-TAIPEI.2014.6964339.

[115] S. Tati, S. Silvestri, T. He, and T. L. Porta, "Robust network tomography in the presence of failures," in Proc Intl Conf on Distributed Computing Systems, pp. 481-492, 2014, doi 10.1109/ICDCS.2014.56.

[116] N. T. Luy, N. T. Thanh, and H. M. Tri, "Reinforcement learning-based robust adaptive tracking control for multi-wheeled mobile robots synchronization with optimality," in Proc 2013 IEEE Workshop on Robotic Intelligence in Info Structured Space, pp. 74-81, 2013, doi 10.1109/RiiSS.2013.6607932.

[117] M. Kashki, M. A. Abido, and Y. L. Abdel-Magid, "Power system dynamic stability enhancement using optimum design of PSS and static phase shifter based stabilizer," Arabian J for Science and Engineering, vol. 38, no. 3, pp. 637-650, 2013, doi 10.1007/s13369-012-0325-z.

[118] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer, "Exploration in model-based reinforcement learning by empirically estimating learning progress," Advances in Neural Info Process Systems, vol. 1, pp. 206-214, 2012.

[119] M. S. Llorente and S. E. Guerrero, "Increasing retrieval quality in conversational recommenders," IEEE Trans on Knowl and Data Eng, vol. 24, no. 10, pp. 1876-1888, 2012, doi 10.1109/TKDE. 2011.116.

[120] F. Maes, et al, "Learning to play K-armed bandit problems," in 4th Intl Conf on Agents and AI, pp. 74-81, 2012.

[121] S. Bhasin, et al, "Asymptotic tracking by a reinforcement learning-based adaptive critic controller," J of Control Theory and Apps, vol. 9, no. 3, pp. 400-409, 2011, doi 10.1007/s11768-011-0170-8.

[122] A. Tjahjadi, et al, "Robustness analysis of genetic network programming with reinforcement learning," in Proc Jt 5th Intl Conf on Soft Comp and Intell Sys and 11th Intl Symp on Advanced Intelligent Systems, pp. 594-601, 2010.
[123] S. A. Kulkarni and G. R. Rao, "Vehicular ad hoc network mobility models applied for reinforcement learning routing algorithm," Comms in Computer and Info Science, pp. 230-240, 2010, doi 10.1007/978-3-642-14825-5_20.
[124] C. Molina, et al, "Maximum entropy-based reinforcement learning using a confidence measure in speech recognition for telephone speech," IEEE Trans on Audio, Speech and Language Processing, vol. 18, no. 5, pp. 1041-1052, 2010, doi 10.1109/TASL.2009.2032618.
[125] N. T. Luy, et al, "Robust reinforcement learning-based tracking control for wheeled mobile robot," in 2nd Intl Conf on Computer and Automation Eng, pp. 171-176, 2010, doi 10.1109/ICCAE.2010. 5451973.
[126] V. Heidrich-Meisner and C. Igel, "Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search," in Proc of the 26th Intl Conf on Machine Learning, pp. 401-408, 2009.
[127] H. Satoh, "A nonlinear approach to robust routing based on reinforcement learning with state space compression and adaptive basis construction," IEICE Trans on Fundamentals of Electronics, Comms and Comp Sci, vol. 7, pp. 1733-1740, 2008, doi 10.1093/ietfec/e91-a.7.1733.
[128] K. Conn, and R. A. Peters, "Reinforcement learning with a supervisor for a mobile robot in a real-world environment," in Proc of the 2007 IEEE Intl Symp on Computl Intell in Robotics and Automation, pp. 73-78, 2007, doi 10.1109/CIRA.2007.382878.
[129] X.-S. Wang, et al, "A proposal of adaptive PID controller based on reinforcement learning," J of China Univ of Mining and Tech, vol. 17, no. 1, pp. 40-44, 2007, doi 10.1016/S1006-1266(07)60009-1.
[130] JB. Leem, and H. Y. Kim, "Action-specialized expert ensemble trading system with extended discrete action space using deep reinforcement learning," PLOS One, vol. 15, no. 7, 2020, doi 10.1371/journal.pone.0236178.
[131] Y. Xiong, et al, "Intelligent thermal control strategy based on reinforcement learning for space telescope," J of Thermophysics and Heat Transfer, vol. 34, no. 1, pp. 37-44, 2020, doi 10.2514/1.T5774.
[132] A. Balakrishnan and J. V. Deshmukh, "Structured reward functions using STL," Proc HSCC '19, pp. 270-271, 2019. doi 10.1145/3302504.3313355.
[133] G. Chen, et al, "Distributed non-communicating multi-robot collision avoidance via map-based deep reinforcement learning," Sensors, vol. 20, no. 17, 2020, doi 10.3390/s20174836.
[134] C. Sun, X. Li, and C. Belta, "Automata guided semi-decentralized multi-agent reinforcement learning," in Proc of the American Control Conf, pp. 3900-3905, 2020, doi 10.23919/ACC45564. 2020.9147704.
[135] X. Wang and X. Ye, "Optimal robust control of nonlinear uncertain system via off-policy integral reinforcement learning," in Proc of the Chinese Control Conf, pp. 1928-1933, 2020, doi 10.23919/CCC50068.2020.9189626.
[136] Z. Yan, J. Ge, Y. Wu, L. Li, and T. Li, "Automatic virtual network embedding: A deep reinforcement learning approach with graph convolutional networks," IEEE J on Selected Areas in Commun, vol. 38, no. 6, pp. 1040-1057, 2020. doi 10.1109/JSAC.2020. 2986662.
[137] K. Alhazmi and S. M. Sarathy, "Continuous control of complex chemical reaction network with reinforcement learning," in Proc ECC, pp. 1066-1068, 2020.
[138] A. Ghasemkhani, et al, "DeepGrid: robust deep reinforcement learning-based contingency management," in Proc IGST, 2020, doi 10.1109/ISGT45199. 2020.9087633.
[139] A. Pitti, M. Quoy, C. Lavandier, and S. Boucenna, "Gated spiking neural network using Iterative Free-Energy Optimization and rank-order coding for structure learning in memory sequences (INFERNO GATE)," Neural Networks, vol. 121, pp. 242-258, 2020, doi 10.1016/j.neunet.2019.09.023.
[140] M. Vecerik, et al, "S3K: self-supervised semantic keypoints for robotic manipulation via multi-view consistency," arXiv:2009.14711, 2020.
[141] Y. Jiao, et al, "Learning to swim in potential flow," arXiv:2009.14280, 2020.
[142] P. Almási, R. Moni, and B. Gyires-Tóth, "Robust reinforcement learning-based autonomous driving agent for simulation and real world," arXiv:2009.11212, 2020.

[143] W. Ding, B. Chen, B. Li, K. J. Eun, and D. Zhao, “Multimodal safety-critical scenarios generation for decision-making algorithms evaluation,” arXiv:2009.08311, 2020.
[144] G. Schamberg, M. Badgeley, and E. N. Brown, “Controlling level of unconsciousness by titrating propofol with deep reinforcement learning,” arXiv:2008.12333, 2020.
[145] B. Pang and Z.-P. Jiang, “Robust reinforcement learning: a case study in linear quadratic regulation,” arXiv:2008.11592, 2020.
[146] T. Kobayashi and W. E. L. Ilboudo, “t-Soft update of target network for deep reinforcement learning,” arXiv:2008.10861, 2020.
[147] A. Zavoli and L. Federici, “Reinforcement learning for low-thrust trajectory design of interplanetary missions,” arXiv:2008.08501, 2020.
[148] O. Limoyo, et al, “Heteroscedastic uncertainty for robust generative latent dynamics,” arXiv:2008.08157, 2020.
[149] W. Zhao, J. P. Queralta, L. Qingqing, and T. Westerlund, “Towards closing the sim-to-real gap in collaborative multi-robot deep reinforcement learning,” arXiv:2008.07875, 2020.
[150] X. Qu, Y.-S. Ong, A. Gupta, and Z. Sun, “Defending adversarial attacks without adversarial attacks in deep reinforcement learning,” arXiv:2008.06199, 2020.
[151] P. Swazinna, S. Udluft, and T. Runkler, “Overcoming model bias for robust offline deep reinforcement learning,” arXiv:2008.05533, 2020.
[152] I. Ahmed, H. Khorasgani, and G. Biswas, “Comparison of model predictive and reinforcement learning methods for fault tolerant control,” arXiv:2008.04403, 2020.
[153] G. Kovač, A. Laversanne-Finot, and P.-Y. Oudeyer, “GRIMGEP: learning progress for robust goal sampling in visual deep reinforcement learning,” arXiv:2008.04388, 2020.
[154] J. L. Zhu, et al, “Adversarial directed graph embedding,” arXiv:2008.03667, 2020.
[155] X. Ma, S. Chen, D. Hsu, and W. S. Lee, “Contrastive variational model-based reinforcement learning for complex observations,” arXiv:2008.02430, 2020.
[156] T. Oikarinen, et al, “Robust deep reinforcement learning through adversarial loss,” arXiv:2008.01976, 2020.
[157] E. Vinitsky, et al, “Robust reinforcement learning using adversarial populations,” arXiv:2008.01825, 2020.
[158] H. Park, etal, “Understanding the stability of deep control policies for biped locomotion,” arXiv:2007.15242, 2020.
[159] K. Steverson, J. Mullin, and M. Ahiskali, “Adversarial robustness for machine learning cyber defenses using log data,” arXiv:2007.14983, 2020.
[160] X. Chen, et al, “Same-day delivery with fairness,” arXiv:2007.09541, 2020.
[161] X. Chen, Y. Duan, Z. Chen, H. Xu, Z. Chen, X. Liang, T. Zhang, and Z. Li, “CATCH: context-based meta reinforcement learning for transferrable architecture search,” arXiv:2007.09380, 2020.
[162] L. Zhang, H. Xiong, O. Ma, and Z. Wang, “Multi-robot cooperative object transportation using decentralized deep reinforcement learning,” arXiv:2007.09243, 2020.
[163] K. L. Tan, Y. Esfandiari, X. Y. Lee, Aakanksha, and S. Sarkar, “Robustifying reinforcement learning agents via action space adversarial training,” arXiv:2007.07176, 2020.
[164] A. Stooke, et al, “Responsive safety in RL by PID Lagrangian methods,” arXiv:2007.03964, 2020.
[165] K. Abe and Y. Kaneko, “Off-policy exploitability-evaluation and equilibrium-learning in two-player zero-sum Markov games,” arXiv:2007.02141, 2020.
[166] X. Wang, et al, “Falsification-based robust adversarial reinforcement learning,” arXiv:2007.00691, 2020.
[167] H. Lee, M. Girnyk, and J. Jeong, “Deep reinforcement learning approach to MIMO precoding problem: optimality and robustness,” arXiv:2006.16646, 2020.
[168] D. Xu, M. Agarwal, E. Gupta, F. Fekri, and R. Sivakumar, “Accelerating reinforcement learning agent with eeg-based implicit human feedback,” arXiv:2006.16498, 2020.
[169] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, and X. Guan, “Multi-agent deep reinforcement learning for hvac control in commercial buildings,” arXiv:2006.14156, 2020.
[170] A. Gleave, et al, “Quantifying differences in reward functions,” arXiv:2006.13900, 2020.
[171] R. Raileanu, M. Goldstein, D. Yarats, I. Kostrikov, and R. Fergus, “Automatic data augmentation for generalization in deep reinforcement learning,” arXiv:2006.12862, 2020.
[172] H. Liu and W. Wu, “Online multi-agent reinforcement learning for decentralized inverter-based volt-var control,” arXiv:2006.12841, 2020.
[173] Y. Zou and X. Lu, “Gradient-EM Bayesian meta-learning,” arXiv:2006.11764, 2020.

[174]K. Panaganti and D. Kalathil, "Model-free robust reinforcement learning with linear function approximation," arXiv:2006.11608, 2020.
[175]A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," arXiv:2006.10742, 2020.
[176]A. Rahman, et al, "Open ad hoc teamwork using graph-based policy learning," arXiv:2006.10412, 2020.
[177]H. Jeong, et al, "Learning to track dynamic targets in partially known environments," arXiv:2006.10190, 2020.
[178]K.-P. Ning and S.-J. Huang, "Reinforcement learning with supervision from noisy demonstrations," arXiv:2006.07808, 2020.
[179]Y. Dou, et al, "Robust spammer detection by nash reinforcement learning," arXiv:2006.06069, 2020.
[180]X. Huang, F. Zhu, L. Holloway, and A. Haidar, "Causal discovery from incomplete data using an encoder and reinforcement learning," arXiv:2006.05554, 2020.
[181]Y. Chow, et al, "Variational model-based policy optimization," arXiv:2006.05443, 2020.
[182]T. Jafferjee, E. Imani, E. Talvitie, M. White, and M. Bowling, "Hallucinating value: a pitfall of dyna-style planning with imperfect environment models," arXiv:2006.04363, 2020.
[183]Y. Tian, et al, "Real-time model calibration with deep reinforcement learning," arXiv:2006.04001, 2020.
[184]N. Kallus and M. Uehara, "Efficient evaluation of natural stochastic policies in offline reinforcement learning," arXiv:2006.03886, 2020.
[185]L. Hou, et al, "Robust reinforcement learning with Wasserstein constraint," arXiv:2006.00945, 2020.
[186]J. Zhi and J.-M. Lien, "Learning to herd agents amongst obstacles: training robust shepherding behaviors using deep reinforcement learning," arXiv:2005.09476, 2020.
[187]Y. Chandak, et al, "Optimizing for the future in non-stationary MDPs," arXiv:2005.08158, 2020.
[188]Y. Ding, et al, "Mutual information maximization for robust plannable representations," arXiv:2005.08114, 2020.
[189]S. Totaro, I. Boukas, A. Jonsson, and B. Cornélusse, "Lifelong control of off-grid microgrid with model based reinforcement learning," arXiv:2005.08006, 2020.
[190]Z. Xie, et al, "ALLSTEPS: curriculum-driven learning of stepping stone skills," arXiv:2005.04323, 2020.
[191]R. Singh, Q. Zhang, and Y. Chen, "Improving robustness via risk averse distributional reinforcement learning," arXiv:2005.00585, 2020.
[192]I. Kostrikov, D. Yarats, and R. Fergus, "Image augmentation is all you need: regularizing deep reinforcement learning from pixels," arXiv:2004.13649, 2020.
[193]J. Z. Chen, "Reinforcement learning generalization with surprise minimization," arXiv:2004.12399, 2020.
[194]P. D. Ngo and F. Godtliebsen, "Data-driven robust control using reinforcement learning," arXiv:2004.07690, 2020.
[195]M. Everett, et al, "Certified adversarial robustness for deep reinforcement learning," arXiv:2004.06496, 2020.
[196]M. Koren and M. J. Kochenderfer, "Adaptive stress testing without domain heuristics using go-explore," arXiv:2004.04292, 2020.
[197]B. Anahtarci, et al, "Q-Learning in regularized mean-field games," arXiv:2003.12151, 2020.
[198]B. Lindenberg, et al, "Distributional reinforcement learning with ensembles," arXiv:2003.10903, 2020.
[199]Q. Shen, et al, "Deep reinforcement learning with robust and smooth policy," arXiv:2003.09534, 2020.
[200]H. Zhang, H. Chen, C. Xiao, B. Li, M. Liu, D. Boning, and C.-J. Hsieh, "Robust deep reinforcement learning against adversarial perturbations on state observations," arXiv:2003.08938, 2020.
[201]X. Guo, et al, "A general framework for learning mean-field games," arXiv:2003.06069, 2020.
[202]A. Touati, A. A. Taiga, and M. G. Bellemare, "Zooming for efficient model-free reinforcement learning in metric spaces," arXiv:2003.04069, 2020.
[203]S. Gao, P. Dong, Z. Pan, and G Y. Li, "Reinforcement learning based cooperative coded caching under dynamic popularities in ultra-dense networks," arXiv:2003.03758, 2020.
[204]J. Lin, K. Dzeparoska, S. Q. Zhang, A. Leon-Garcia, and N Papernot, "On the robustness of cooperative multi-agent reinforcement learning," arXiv:2003.03722, 2020.

[205] E. Derman, and S. Mannor, “Distributional robustness and regularization in reinforcement learning,” arXiv:2003.02894, 2020.
[206] T. Spooner, R. Savani, “Robust market making via adversarial reinforcement learning,” arXiv:2003.01820, 2020.
[207] M. Chancán and M. Milford, “MVP: unified motion and visual self-supervised learning for large-scale robotic navigation,” arXiv:2003.00667, 2020.
[208] W. E. L. Ilboudo, et al, “TAdam: a robust stochastic gradient optimizer,” arXiv:2003.00179, 2020.
[209] A. Tschantz, et al, “Reinforcement learning through active inference,” arXiv:2002.12636, 2020.
[210] S. Kuutti, et al, “Training adversarial agents to exploit weaknesses in deep cntrl policies,” arXiv:2002.12078, 2020.
[211] N. D. Nguyen, T. T. Nguyen, and S. Nahavandi, “A visual communication map for multi-agent deep reinforcement learning,” arXiv:2002.11882, 2020.
[212] C.-H. H. Yang, J. Qi, P.-Y. Chen, Y. Ouyang, I-T. D. Hung, C.-H. Lee, and X. Ma, “Enhanced adversarial strategically-timed attacks against deep reinforcement learning,” arXiv:2002.09027, 2020.
[213] T. Sun, et al “Adaptive temporal difference learning with linear function approximation,” arXiv:2002.08537, 2020.
[214] N. Naderializadeh, J. Sydir, M. Simsek, and H. Nikopour, “Resource management in wireless networks via multi-agent deep reinforcement learning,” arXiv:2002.06215, 2020.
[215] P. Kamalaruban, Y.-T. Huang, Y.-P. Hsieh, P. Rolland, C. Shi, and V. Cevher, “Robust reinforcement learning via adversarial training with langevin dynamics,” arXiv:2002.06063, 2020.
[216] N. Kallu, and M. Uehara, “Statistically efficient off-policy policy gradients,” arXiv:2002.04014, 2020.
[217] G. Lee, B. Hou, S. Choudhury, and S. S. Srinivasa, “Bayesian residual policy optimization: scalable Bayesian reinforcement learning with clairvoyant experts,” arXiv:2002.03042, 2020.
[218] V. Pacelli and A. Majumdar, “Learning task-driven control policies via information bottlenecks,” arXiv:2002.01428, 2020.
[219] J. Yao, et al, “Policy gradient based quantum approx optimization algorithm,” arXiv:2002.01068, 2020.
[220] D. Nishio, et al, “Discriminator soft actor critic without extrinsic rewards,” arXiv:2001.06808, 2020.
[221] T. Dai, K. Arulkumaran, T. Gerbert, S. Tukra, F. Behbahani, and A. A. Bharath, “Analysing deep reinforcement learning agents trained with domain randomisation,” arXiv:1912.08324, 2019.
[222] X. Zhang, J. Liu, X. Xu, S. Yu, and H. Chen, “Learning-based predictive control for nonlinear systems with unknown dynamics subject to safety constraints,” arXiv:1911.09827, 2019.
[223] T. Lykouris, et al, “Corruption robust exploration in episodic reinforcement learning,” arXiv:1911.08689, 2019.
[224] S. Salter, et al, “Attention-privileged reinforcement learning,” arXiv:1911.08363, 2019.
[225] M. Han, et al, “H∞ model-free reinforcement learning with robust stability guarantee,” arXiv:1911.02875, 2019.
[226] B. Lütjens, et al, “Certified adversarial robustness for deep reinforcement learning,” arXiv:1910.12908, 2019.
[227] M. Uehara, et al, “Minimax Weight and Q-Function Learning for Off-Policy Evaluation,” arXiv:1910.12809, 2019.
[228] S. Li, and O. Bastani, “Robust model predictive shielding for safe reinforcement learning with stochastic dynamics,” arXiv:1910.10885, 2019.
[229] R. B. Slaoui, et al, “Robust visual domain randomization for reinforcement learning,” arXiv:1910.10537, 2019.
[230] K. Zhang, B. Hu, and T. Başar, “Policy optimization for H2 linear control with H∞ robustness guarantee: implicit regularization and global convergence,” arXiv:1910.09496, 2019.
[231] Z. Liu, et al, “Regularization matters in policy optimization,” arXiv:1910.09191, 2019.
[232] J. Yang, et al, “Single episode policy transfer in reinforcement learning,” arXiv:1910.07719, 2019.
[233] S. Chen, et al, “Zap q-learning with nonlinear function approximation,” arXiv:1910.05405, 2019.
[234] E. Schwartz, G. Tennenholtz, C. Tessler, and S. Mannor, “Language is power: representing states using natural language in reinforcement learning,” arXiv:1910.02789, 2019.
[235] D. Yarats, A. Zhang, I. Kostrikov, B. Amos, J. Pineau, and R. Fergus, “Improving sample efficiency in model-free reinforcement learning from images,” arXiv:1910.01741, 2019.
[236] G. Kalweit, M. Huegle, and J. Boedecker, “Composite q-learning: multi-scale q-function decomposition and separable optimization,” arXiv:1909.13518, 2019.

[237] M. Ryu, et al, "CAQL: continuous action q-learning," arXiv:1909.12397, 2019.

[238] J. Li, et al, "Multi-task batch reinforcement learning with metric learning," arXiv:1909.11373, 2019.

[239] M. Shen, and J. P. How, "Robust opponent modeling via adversarial ensemble reinforcement learning in asymmetric imperfect-information games," arXiv:1909.08735, 2019.

[240] N. Kallus and M. Uehara. "Double reinforcement learning for efficient off-policy evaluation in Markov decision processes," arXiv:1908.08526, 2019.

[241] J. Roy, P. Barde, F. G. Harvey, D. Nowrouzezahrai, and C. Pal, "Promoting coordination through policy regularization in multi-agent deep reinforcement learning," arXiv:1908.02269, 2019.

[242] Y. Urakami, A. Hodgkinson, C. Carlin, R. Leu, L. Rigazio, and P. Abbeel, "DoorGym: A scalable door opening environment and baseline agent," arXiv:1908.01887, 2019.

[243] Q. Wang, K. Feng, X. Li, and S. Jin, "PrecoderNet: hybrid beamforming for millimeter wave systems with deep reinforcement learning," arXiv:1907.13266, 2019.

[244] M. Bogdanovic, et al, "Learning variable impedance control for contact sensitive tasks," arXiv:1907.07500, 2019.

[245] D. J. Mankowitz, et al, "Robust reinforcement learning for continuous control with model misspecification," arXiv:1906.07516, 2019.

[246] A. C. Li, et al, "Sub-policy adaptation for hierarchical reinforcement learning," arXiv:1906.05862, 2019.

[247] M. Assran, J. Romoff, N. Ballas, J. Pineau, and M. Rabbat, "Gossip-based actor-learner architectures for deep reinforcement learning," arXiv:1906.04585, 2019.

[248] B. Gravell, P. M. Esfahani, and T. Summers, "Learning robust control for LQR systems with multiplicative noise via policy gradient," arXiv:1905.13547, 2019.

[249] A. Francis, A. Faust, H.-T. L. Chiang, J. Hsu, J. C. Kew, M. Fiser, and T.-W. E. Lee, "Long-range indoor navigation with PRM-RL," arXiv:1902.09458, 2019.

[250] J. Wang, Y. Liu, and B. Li. "Reinforcement learning with perturbed rewards," arXiv:1810.01032, 2018.

[251] D. Moher, et al, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," PLoS Med, vol. 6, no. 7, e1000097, 2009, DOI 10.1371/journal.pmed1000097.

[252] Y. Liu, Y., et al, "Stein Variational Policy Gradient," arXiv:1704.02399v1 [cs.LG], April 2017.

AUTHORS

LAURA L. PU LLUM (IEEE M'86–SM'03) received the B.S. degree in mathematics from the University of Alabama in Huntsville (UAH) in 1983, the M.S. degree in operations research from UAH in 1987, the MBA from the Southeastern Institute of Technology (SIT), Huntsville, AL in 1990, the D.Sc. in systems engineering and operations research from SIT in 1992, and the M.S. in geology from the University of Tennessee (Knoxville) in 2015.Since 1982, she worked in industry (large and small businesses), non-profit research institutes, and academia. She was most recently a Senior Research Scientist at Oak Ridge National Laboratory, Oak Ridge, TN, USA. She is the author of 2 books (Software Fault Tolerance Techniques and Implementation, Artech House, 2001 and Guidance for the Verification and Validation of Neural Networks, Wiley, 2007), numerous book chapters and hundreds of articles/papers. Throughout her career, she has conducted research and development to enhance and ensure the dependability of intelligent and complex systems, including those incorporating machine learning, artificial intelligence and autonomy. Her current research is in the robustness and confidence of classifiers of non-traditional images and signals..

Dr. Pullum has received several best paper awards and certificates of appreciation from the software and artificial intelligence standards development organizations on which she has served.