Reinforcement Learning for Safety-Critical Applications

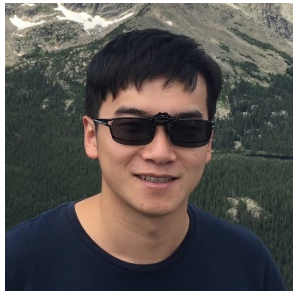
Enrique Mallada

Johns Hopkins University

T. Zheng

A. Castellano

H. Min



P. You

J. Bazerque

ISC Seminar, JHU Applied Physics Laboratory

Jan 11, 2024

Networks, Dynamics, and Learning Laboratory

Electrical and Computer Engineering

Johns Hopkins University

JOHNS HOPKINS MATHEMATICAL INSTITUTE for DATA SCIENCE

JOHNS HOPKINS INSTITUTE for ASSURED AUTONOMY

JOHNS HOPKINS RALPH O'CONNOR SUSTAINABLE ENERGY INSTITUTE

LABORATORY FOR Computational Sensing + Robotics

NetD-Lab Members

PhD Students

Agustin Castellano
PhD, 2^{nd} year

Roy Siegelman
PhD, AMS, 3^{trd} year

Yue Shen
PhD, 4^{th} year

Rajni Bansal* PhD
2023, Postdoc UCSD

Tianqi Zheng
PhD 2023, Amazon

Hancheng Min PhD
2023, Postdoc UPenn

Jay Guthrie
PhD 2022, Apple

Charalampos Avraam*
PhD 2021, Postdoc NYU

Yan Jiang, PhD
2021, Postdoc UW

MSE and Undergraduates

Mengnan Zhao,
MSE 2018
PhD, Rice University

Zachary Nelson,
MSE 2018
Lockheed Martin

Elijah Pivo, BS ’18,
RA 2019-2020
PhD, MIT IDSS

Jesse Rines, BS ’18
US Navy
Nuclear Energy

Aurik Sarker, BS’18
JHU APL

Eliza Cohn BS’20
Optum->Columbia (Ph.D.)

Postdocs and Research Scientists

Dhajanajay (DJ) Annand
RSc 2022-Present

Pengcheng You*
Postdoc 2019-2021
Asst. Prof. Pekin University

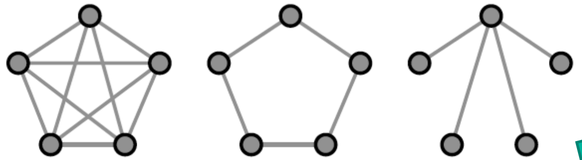
M. Devrim Kaba
RSc 2018-2020
eBay

Mohammad Hajiesmaili
Postdoc 2017-2018
Asst. Prof. UMASS CS

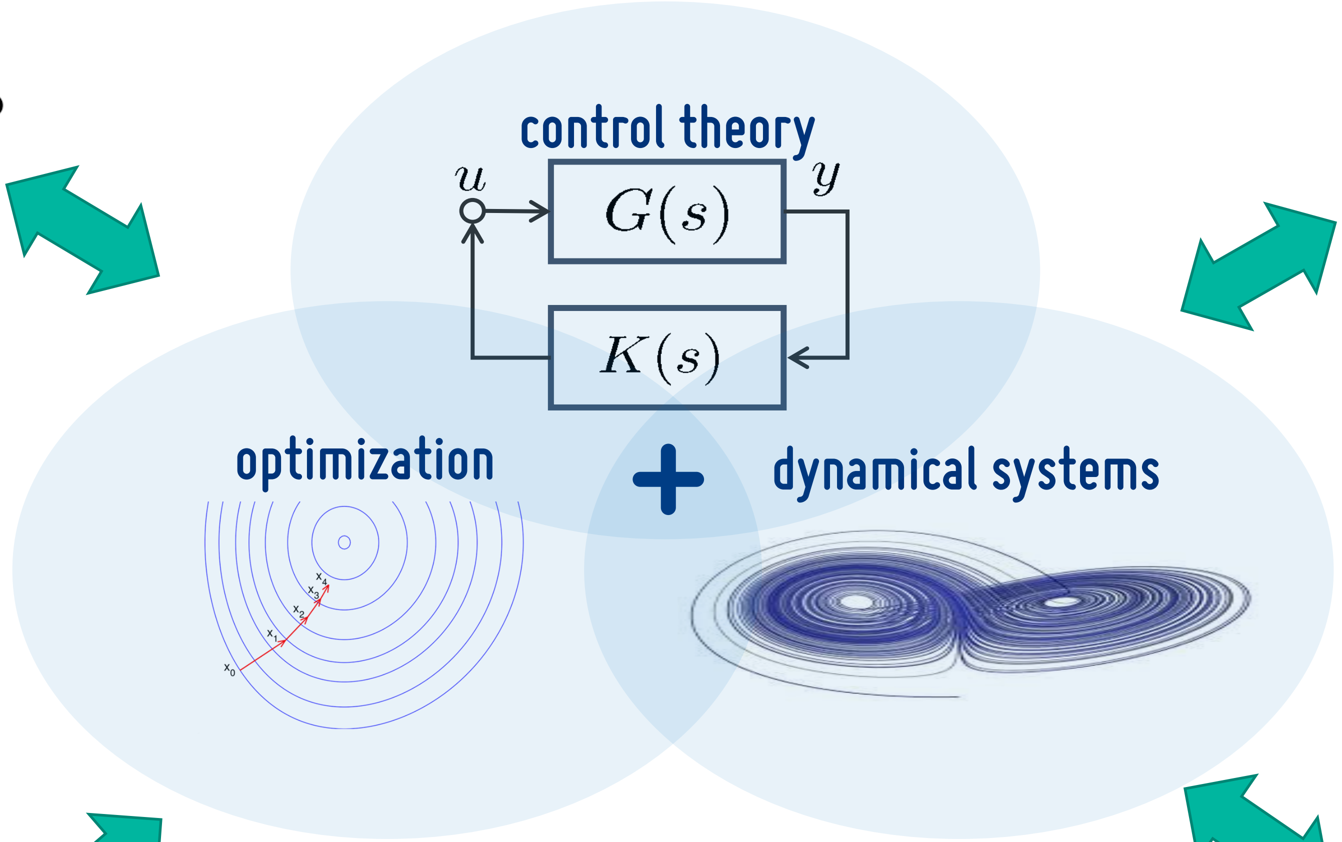
* co-advised

NetD-Lab: Research Areas

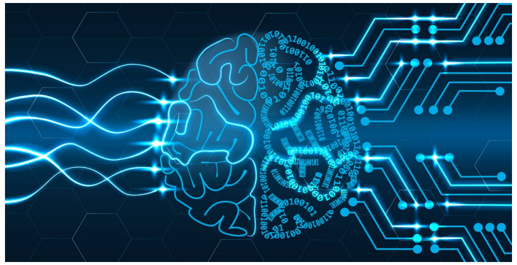
graph theory



economics



machine learning

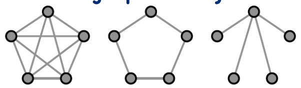


statistics

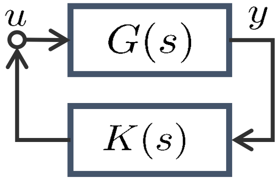
NetD-Lab: Research Methodology

Core Research – Networked Dynamical Systems

graph theory

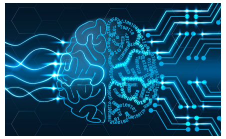


control theory

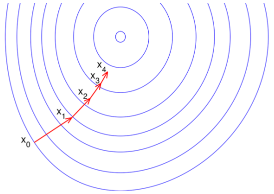


economics

machine learning

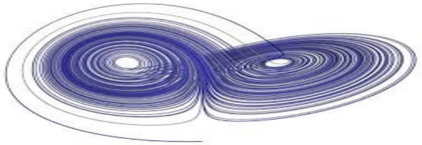


optimization



+

dynamical systems



statistics

Insights

+

Solutions

Constraints

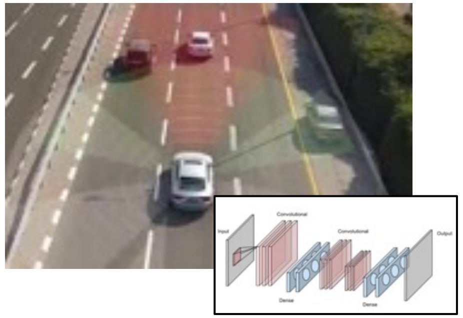
+

Limitations

Applications – Engineering Systems

information networks

power networks



safety-critical systems

Recent Projects

Learning, Dynamics, and Control

- Learning safety via recurrence
- Reinforcement Learning with almost sure constraints
- Online C-RL via regularized saddle flow dynamics
- The role of overparameterization in learning dynamics of deep networks

Power System Operations and Control

- Frequency Shaping Control
- Spectral Clustering and model reduction in network dynamical systems
- A Market Mechanism for energy storage
- Market power in two-stage Market
- Market Dynamics: Stability, Efficiency, and Incentive Alignment

Recent Projects

Learning, Dynamics, and Control

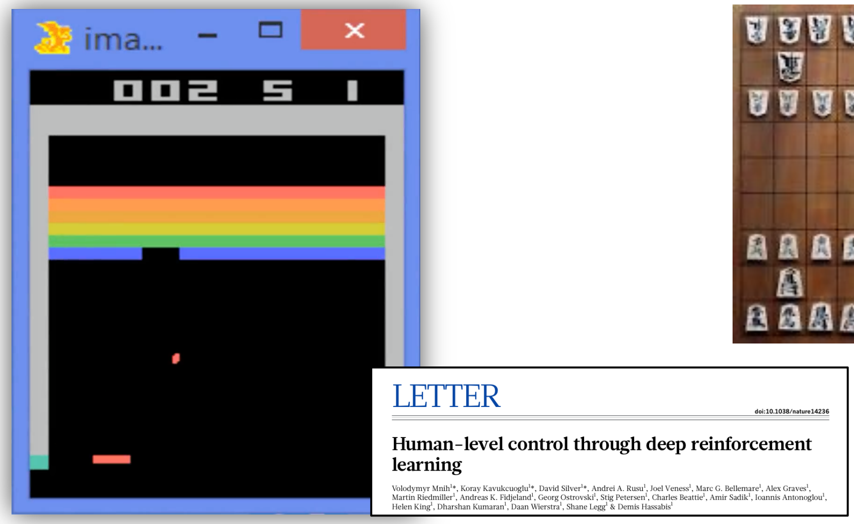
- Learning safety via recurrence
- **Reinforcement Learning with almost sure constraints**
- **Online C-RL via regularized saddle flow dynamics**
- The role of overparameterization in learning dynamics of deep networks

Power System Operations and Control

- Frequency Shaping Control
- Spectral Clustering and model reduction in network dynamical systems
- A Market Mechanism for energy storage
- Market power in two-stage Market
- Market Dynamics: Stability, Efficiency, and Incentive Alignment

A World of Success Stories

2017 Google DeepMind's DQN



LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodymyr Mnih[1*], Koray Kavukcuoglu[1*], David Silver[1*], Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

2017 AlphaZero – Chess, Shogi, Go

2019 AlphaStar – Starcraft II

Article

Grandmaster level in StarCraft II using multi-agent reinforcement learning

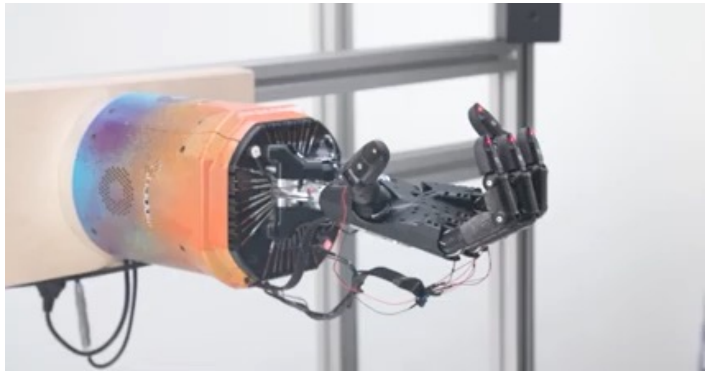
https://doi.org/10.1038/s41586-019-1724-z

Received: 30 August 2019

Accepted: 10 October 2019

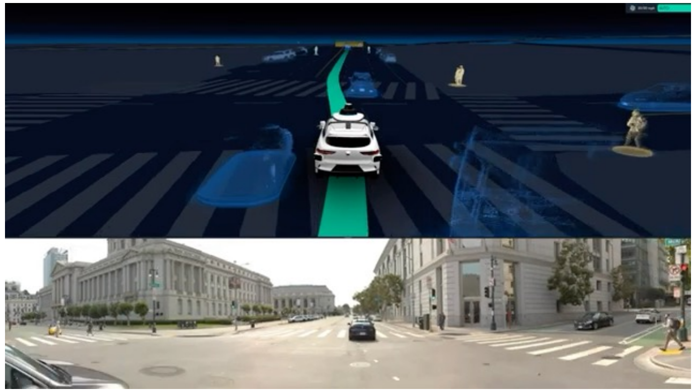
Published online: 30 October 2019

OpenAI – Rubik's Cube



Boston Dynamics

Waymo



Reality Kicks In

Angry Residents, Abrupt Stops: Waymo Vehicles Are Still Causing Problems in Arizona

RAY STERN | MARCH 31, 2021 | 8:26AM

GARY MARCUS BUSINESS 08.14.2019 09:00 AM

DeepMind's Losses and the Future of Artificial Intelligence

Alphabet's DeepMind unit, conqueror of Go and other games, is losing lots of money. Continued deficits could imperil investments in AI.

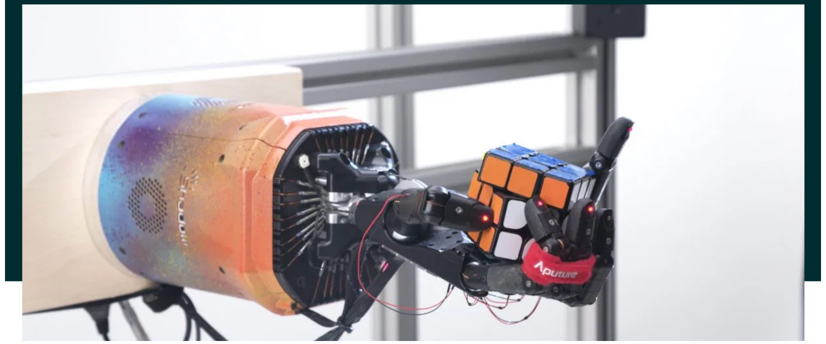
AARIAN MARSHALL BUSINESS 12.07.2020 04:06 PM

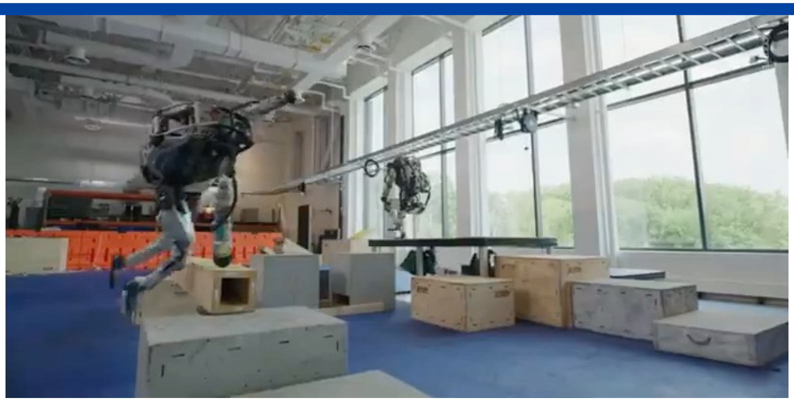
Uber Gives Up on the Self-Driving Dream

Can we adapt reinforcement learning algorithms to address physical systems challenges?

OpenAI dis

Kyle Wiggers @Kyle_L_Wiggers July 16, 2021 11:24 AM





d

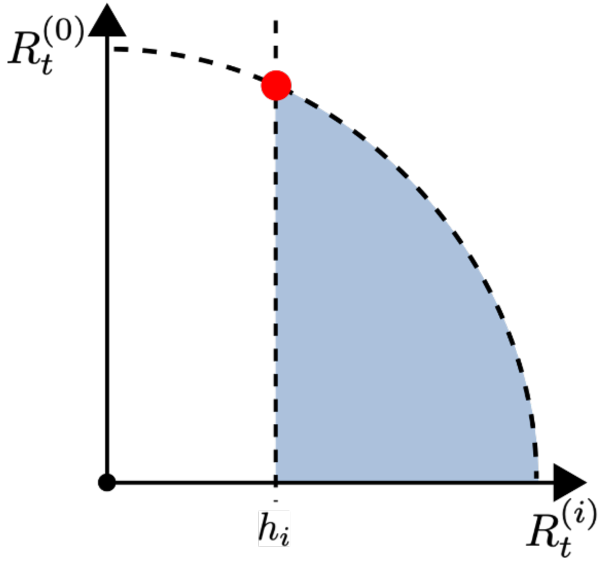
woman did not recognize that pedestrians jaywalk

The automated car lacked "the capability to classify an object as a pedestrian unless that object was near a crosswalk," an NTSB report said.

Challenges of RL for Physical Systems

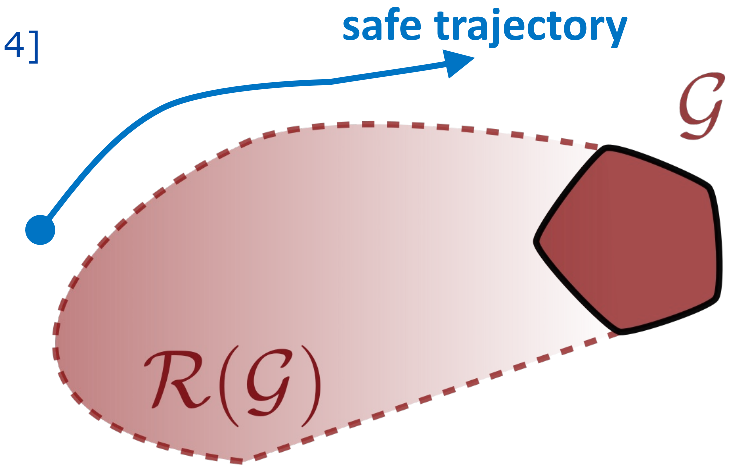
- Physical systems must meet multiple objectives
  - Need to trade off between the different goals
  - Constrained RL allows to explore the Pareto Front [1,2]

$$\max_{\pi}\ (1-\gamma)\mathbb{E}_{\pi,S_0\sim q}\left[\sum_{t=0}^{+\infty}\gamma^t R_{t+1}^{(0)}\right]$$
$$\text{s.t.}\quad (1-\gamma)\mathbb{E}_{\pi,S_0\sim q}\left[\sum_{t=0}^{+\infty}\gamma^t R_{t+1}^{(i)}\right]\geq h_i,\ \forall i\in[n]$$



- Failures have a qualitatively different impact
  - Expectation constraints cannot meet safety requirements
  - Hard (almost sure) constraints can guarantee safety [3,4]

$$\max_{\pi}\ \mathbb{E}_{\pi,S_0\sim q}\left[\sum_{t=0}^{+\infty}\gamma^t R_{t+1}\right]$$
$$\text{s.t.}\quad \mathbb{P}_{\pi,S_0\sim q}\left[S_t\notin\mathcal{G}\right]=1,\ \forall t\geq 0$$



[1] Zheng, You, and M, Constrained reinforcement learning via dissipative saddle flow dynamics, Asilomar 2022
[2] You, and M, Saddle flow dynamics: Observable certificates and separable regularization, ACC 2021
[3] Castellano, Min, Bazerque, M, Reinforcement Learning with Almost Sure Constraints, L4DC 2022
[4] Castellano, Min, Bazerque, M, Learning to Act Safely with Limited Exposure and Almost Sure Certainty, IEEE TAC, 2023
[5] Castellano, Min, Bazerque, M, Correct-by-design Safety Critics Using Non-contractive Bellman Operators, submitted

[Submitted on 30 Sep 2020]

Saddle Flow Dynamics: Observable Certificates and Separable Regularization

Pengcheng You, Enrique Mallada

arXiv > math > arXiv:2009.14714

[Submitted on 3 Dec 2022]

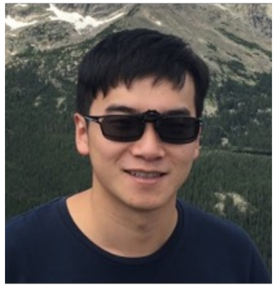
Constrained Reinforcement Learning via Dissipative Saddle Flow Dynamics

Tianqi Zheng, Pengcheng You, Enrique Mallada

arXiv > cs > arXiv:2212.01505

Tianqi Zheng

amazon



Pengcheng You

北京大學 PEKING UNIVERSITY

Outline

- Intro to Constrained RL
- Dissipative Saddle Flows for Bilinear Saddles
- Solving Constrained RL via D-SGDA

Constrained Reinforcement Learning

Goal: Given initial state $S_0 \sim q$, find policy $\pi^* \in \Pi_\theta$ that solves:

$$\max_{\pi \in \Pi_\theta} \quad V_q^{(0)}(\pi) \quad \text{s.t.} \quad V_q^{(i)}(\pi) \geq h_i, \quad \forall i \in [n]$$

where $V_q^{(i)}(\pi) \coloneqq (1-\gamma)\mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}^{(i)}\right]$.

General Approach: Lagrange relaxation

$$\max_{\pi \in \Pi_\theta} \min_{\mu \geq 0} L(\pi, \mu) := V_q^{(0)}(\pi) + \sum_{i=1}^{n} \mu_i (V_q^{(i)}(\pi) - h_i)$$

Non-convex yet has zero duality gap! [1],[2]

[1] S Paternain, L Chamon, M Calvo-Fullana, and A Ribeiro. Constrained reinforcement learning has zero duality gap. NeurIPS 2019
[2] E. Altman. Constrained Markov decision processes. Vol. 7. CRC press 1999

Constrained Reinforcement Learning

Goal: Given initial state $S_0 \sim q$, find policy $\pi^* \in \Pi_\theta$ that solves:

$$\max_{\pi \in \Pi_\theta} \quad V_q^{(0)}(\pi) \quad \text{s.t.} \quad V_q^{(i)}(\pi) \geq h_i, \quad \forall i \in [n]$$

where $V_q^{(i)}(\pi) \coloneqq (1-\gamma)\mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}^{(i)}\right]$.

General Approach: Lagrange relaxation

$$\min_{\mu \geq 0} \max_{\pi \in \Pi_\theta} L(\pi, \mu) := V_q^{(0)}(\pi) + \sum_{i=1}^{n} \mu_i (V_q^{(i)}(\pi) - h_i)$$

Non-convex yet has zero duality gap! [1],[2]

[1] S Paternain, L Chamon, M Calvo-Fullana, and A Ribeiro. Constrained reinforcement learning has zero duality gap. NeurIPS 2019
[2] E. Altman. Constrained Markov decision processes. Vol. 7. CRC press 1999

Prior Work: Algorithms for Constrained RL [1]-[8]

Use primal and/or dual methods of the form:

$$\pi_{k+1} = \begin{cases} \pi_k + \eta \nabla_\pi \tilde{L}(\pi_k, \mu_k; \zeta_k) \\ \arg\max_\pi \tilde{L}(\pi, \mu_k; \zeta_k) \end{cases} \qquad \mu_{k+1} = \begin{cases} \mu_k - \eta \nabla_\mu \tilde{L}(\pi_k, \mu_k; \zeta_k) \\ \arg\min_{\mu \geq 0} \tilde{L}(\pi_k, \mu; \zeta_k) \end{cases}$$

where $\tilde{L}(\pi, \mu; \zeta) \coloneqq L(\pi, \mu; \zeta) + \Omega(\pi, \mu; \zeta)$ is a regularized Lagrangian

- Parametrization of Π_θ: Soft-max [1,4], occupancy measures [2,3], greedy.
- Horizon: Infinite γ-discounting [1-4], finite H [5-7], or average [8]

- Regret:

value

$$\mathbb{E}\left[\sum_{k=0}^{T-1} V_q^{(0)}(\pi^*) - V_q^{(0)}(\pi_k)\right] = \mathcal{O}(T^{\frac{1}{2}})$$

constraint satisfaction

$$\mathbb{E}\left[\sum_{k=1}^{T-1} c_i - V_q^{(i)}(\pi_k)\right] = \mathcal{O}(T^p), \; p \in [0, 3/4)$$

- Policy: Iterates π_k lack convergence guarantees: Instead $\hat{\pi}_T = \sum_{t=0}^{T-1} \alpha_k \pi_k \to \pi^*$ [2,3]

[1] D Ding, K Zhang, T Basar, and M Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. NeurIPS 2020
[2] Y Chen, J Dong, Z Wang, A Primal-Dual Approach to Constrained Markov Decision Processes, arXiv:2101.10895, 2021
[3] Q Bai, A S Bedi, M Agarwal, A Koppel, V Aggarwal. Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Primal-Dual Approach, AAAI 2022
[4] T Xu, Y Liang, and G Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. ICML 2021
[5] D Ding, X Wei, Z Yang, Z Wang, and M Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. AISTATS 2021
[6] H Wei, X Liu, and L Ying. A provably-efficient model-free algorithm for constrained markov decision processes. arXiv:2106.01577 2021.
[7] T Liu, R Zhou, D Kalathil, P Kumar, and C Tian. "Learning policies with zero or bounded constraint violation for constrained MDPs." NeurIPS 2021
[8] M Calvo-Fullana, S Paternain, L Chamon, and A Ribeiro. State augmented C-RL: Overcoming the limitations of learning with rewards. arXiv:2102.11941 2021

Prior Work: Algorithms for Constrained RL [1]-[8]

Use primal and/or dual methods of the form:

$$\pi_{k+1} = \begin{cases} \pi_k + \eta \nabla_\pi \tilde{L}(\pi_k, \mu_k; \zeta_k) \\ \arg\max_\pi \tilde{L}(\pi, \mu_k; \zeta_k) \end{cases} \qquad \mu_{k+1} = \begin{cases} \mu_k - \eta \nabla_\mu \tilde{L}(\pi_k, \mu_k; \zeta_k) \\ \arg\min_{\mu \geq 0} \tilde{L}(\pi_k, \mu; \zeta_k) \end{cases}$$

where $\tilde{L}(\pi, \mu; \zeta) \coloneqq L(\pi, \mu; \zeta) + \Omega(\pi, \mu; \zeta)$ is a regularized Lagrangian

- Parametrization of Π_θ: Soft-max [1,4], occupancy measures [2,3], greedy.
- Horizon: Infinite γ-discounting [1-4], finite H [5-7], or average [8]

- Regret:

value

$$\mathbb{E}\left[\sum_{k=0}^{T-1} V_q^{(0)}(\pi^*) - V_q^{(0)}(\pi_k)\right] = \mathcal{O}(T^{\frac{1}{2}})$$

constraint satisfaction

$$\mathbb{E}\left[\sum_{k=1}^{T-1} c_i - V_q^{(i)}(\pi_k)\right] = \mathcal{O}(T^p),\ p \in [0, 3/4)$$

- Policy: Iterates π_k lack convergence guarantees: Instead $\hat{\pi}_T = \sum_{t=0}^{T-1} \alpha_k \pi_k \to \pi^*$ [2,3]

Question: Can we achieve convergence of the policy iterates $\pi_k \to \pi^*$ $a.s.$, or is learning from rewards a fundamental limitation?

Towards convergent π_k iterates – Good news

Good news: Non-convexity of $L(\pi, \mu)$ is not so bad…

- There exists a convex parametrization Π_θ that makes it convex-concave

$$\max_{\pi} \ (1-\gamma)\mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}^{(0)}\right]$$
$$\text{s.t.} \quad (1-\gamma)\mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}^{(i)}\right] \geq h_i, \ \forall i \in [n]$$

$$\Updownarrow$$

- **LP Formulation:[1]**

$$\max_{\lambda \geq 0} \ \sum_a \lambda_a^T r_a^{(0)}$$
$$\text{s.t.} \quad \sum_a \lambda_a^T r_a^{(i)} \geq h_i, \ \forall i \in [n] \qquad (\mu_i)$$
$$\sum_a (I - \gamma P_a^T)\lambda_a = (1-\gamma) q \qquad (v)$$

$$\boxed{\pi(a|s) = \frac{\lambda_{s,a}}{\sum_{a'} \lambda_{s,a'}}}$$

- where $\lambda_{s,a} = (1-\gamma)\sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\pi, S_0 \sim q}(S_t = s, A_t = a)$ is the occupancy measure

[1] E. Altman. Constrained Markov decision processes. Vol. 7. CRC press 1999

Towards convergent π_k iterates – Bad news

Bad news: Non-strictness of $L(\lambda, \mu, v)$

- **LP Formulation:**
- Outline

$$\begin{aligned} \max_{\lambda \geq 0} \quad & \textstyle\sum_a \lambda_a^T r_a^{(0)} \\ \text{s.t.} \quad & \textstyle\sum_a \lambda_a^T r_a^{(i)} \geq h_i, \ \forall i \in [n] \qquad & (\mu_i) \\ & \textstyle\sum_a (I - \gamma P_a^T) \lambda_a = (1-\gamma) q \qquad & (v) \end{aligned} \Bigg\} \text{ dual vars}$$

- where $\lambda_{s,a} = (1-\gamma) \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\pi, S_0 \sim q}(S_t = s, A_t = a)$ is the occupancy measure

- **Bilinear Lagrangian:**
  - Lacks strict convexity/concavity necessary for convergence of primal-dual algorithms

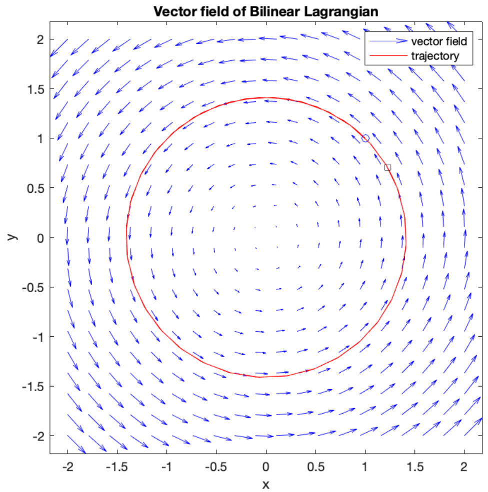
$$\min_{\mu \geq 0, v} \max_{\lambda \geq 0} L(\lambda, \mu, v) = \lambda^T M \begin{bmatrix} \mu \\ v \end{bmatrix}$$

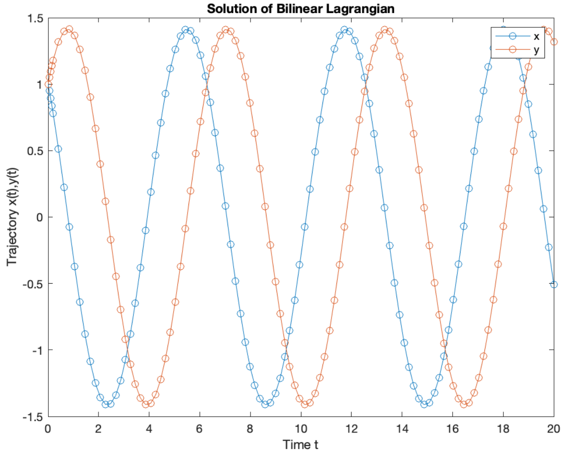
Outline

- Intro to Constrained RL
- Dissipative GDA Flows for Convex-concave L
- Solving Constrained RL via D-SGDA

Warm-up: Scalar Case

- We start by looking at a Naïve GDA Flow on a scalar bilinear Lagrangian
  - Min-max Problem:
    $$\min_x \max_y L(x,y) := xy \quad x, y \in \mathbb{R}$$
  - Saddle-point at $(x^*, y^*) = (0,0)$
  - Naïve Gradient Descent-Ascent (GDA) Flow
    $$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x L(x,y) \\ +\nabla_x L(x,y) \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$
  - Energy Dissipation:
    $$V(x,y) = \tfrac{1}{2}x^2 + \tfrac{1}{2}y^2, \quad \dot{V}(x,y) = x(-y) \ + yx \equiv 0$$





Remark: Behavior generalizes for general non-strict convex-concave Lagrangians [1].

[1] T Holding, and I Lestas. Stability and instability in saddle point dynamics—Part I." IEEE TAC 2020
[2] A Cherukuri, B Gharesifard, and J Cortes. Saddle-point dynamics: conditions for asymptotic stability of saddle points." SIAM JC&O 2017
[3] A Cherukuri, E Mallada, S Low, and J Cortés. The role of convexity in saddle-point dynamics: Lyapunov function and robustness." IEEE TAC 2017

Naïve GDA Flow Scalar Case

	Naïve GDA Flow
Lagrangian	$L(x, y) = xy$
Dynamics	
Energy Function	
Energy Dissipation	
Asympt. Behavior	

Dissipative GDA Flow Algorithm

- Given general convex-concave $L(x, y)$, we consider

$$\hat{L}(x, \hat{x}, y, \hat{y}) = L(x, y) + \frac{\rho}{2}\|x - \hat{x}\|^2 - \frac{\rho}{2}\|y - \hat{y}\|^2$$

- Remarks:
  - If (x^*, y^*) is a saddle point of L, then (x^*, x^*, y^*, y^*) is a saddle point of $\hat{L}$.
  - $\hat{L}$ is neither strictly convex, nor strictly concave (don't worry)

- **Dissipative GDA Flow:**
  - Just apply Naïve GDA on $\hat{L}(x, \hat{x}, y, \hat{y})$!

$$\dot{x} = -\nabla_x L(x, y) - \rho(x - \hat{x}) \qquad \dot{y} = +\nabla_y L(x, y) - \rho(y - \hat{y})$$

$$\dot{\hat{x}} = -\rho(\hat{x} - x) \qquad \dot{\hat{y}} = -\rho(\hat{y} - y)$$

Dissipative GDA Flow Algorithm

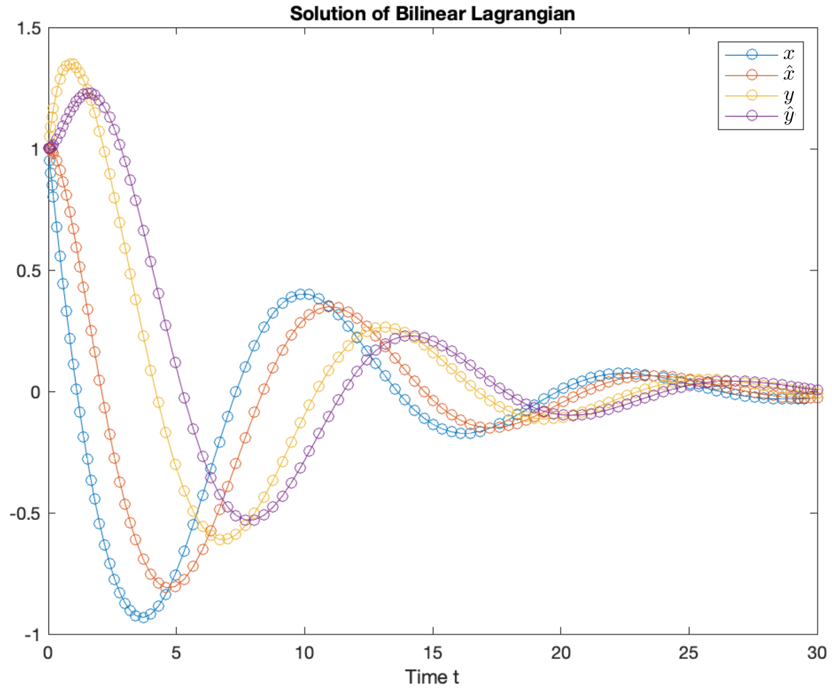
- **Dissipative GDA Flow:**
  - Just apply Naïve GDA on $\hat{L}(x,\hat{x},y,\hat{y}) = L(x,y) + \frac{\rho}{2}\|x-\hat{x}\|^2 - \frac{\rho}{2}\|y-\hat{y}\|^2$!

$$\dot{x} = -\nabla_x L(x,y) - \rho(x-\hat{x}) \qquad \dot{y} = +\nabla_y L(x,y) - \rho(y-\hat{y})$$

$$\dot{\hat{x}} = -\rho(\hat{x}-x) \qquad \dot{\hat{y}} = -\rho(\hat{y}-y)$$

- Scalar case:
  - $\hat{L}(x,\hat{x},y,\hat{y}) = xy + \frac{\rho}{2}(x-\hat{x})^2 + \frac{\rho}{2}(y-\hat{y})^2$

$$\begin{bmatrix}\dot{x}\\ \dot{\hat{x}}\\ \dot{y}\\ \dot{\hat{y}}\end{bmatrix} = \begin{bmatrix}-\rho & \rho & -1 & 0\\ \rho & -\rho & 0 & 0\\ 1 & 0 & -\rho & \rho\\ 0 & 0 & \rho & -\rho\end{bmatrix}\begin{bmatrix}x\\ \hat{x}\\ y\\ \hat{y}\end{bmatrix}$$



Dissipative GDA Flow Scalar Case

	Naïve GDA Flow	Dissipative GDA Flow
Lagrangian	$L(x, y) = xy$	
Dynamics	$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x L(x, y) \\ +\nabla_y L(x, y) \end{bmatrix}$	
Energy Function	$V(x, y) = \frac{1}{2}(x^2 + y^2)$	
Energy Dissipation	$\dot{V} \equiv 0$	
Asympt. Behavior	$V(t) \equiv c$	

General Analysis of Dissipative GDA Flows

Theorem [You, M ACC 21]

Consider the minimax problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} L(x, y)$$

where $L(x, y)$ is convex-concave, and the sets $\mathcal{X}$ and $\mathcal{Y}$ are convex polyhedral.

Then, for any initial feasible point $(x_0, \hat{x}_0, y_0, \hat{y}_0)$ the Dissipative GDA Flow

$$\dot{x} = \Pi_{\mathcal{X},x}\left[-\nabla_x L(x,y) - \rho(x - \hat{x})\right] \qquad \dot{y} = \Pi_{\mathcal{Y},y}\left[+\nabla_y L(x,y) - \rho(y - \hat{y})\right]$$

$$\dot{\hat{x}} = -\rho(\hat{x} - x) \qquad \dot{\hat{y}} = -\rho(\hat{y} - y)$$

converges to some saddle point.

- Remarks:
  - Convergence is guaranteed point-wise, to *some saddle point*
  - Proof uses LaSalle on the same dissipation property $\dot{V} \leq -\rho \left|\left|\dot{\hat{x}}\right|\right|^2 - \rho \left|\left|\dot{\hat{y}}\right|\right|^2$
  - For unconstrained bilinear problems *convergence is exponential*

[You, M ACC 21] P You, Pengcheng, and E Mallada. Saddle flow dynamics: Observable certificates and separable regularization, ACC 2021

Outline

- Intro to Constrained RL
- Dissipative GDA Flows for Convex-concave L
- Solving Constrained RL via D-SGDA

Dissipative GDA for Constrained MDPs

- **LP Formulation of C-RL**

$$\max_{\lambda \geq 0} \sum_a \lambda_a^T r_a^{(0)}$$

$$\text{s.t.} \quad \sum_a \lambda_a^T r_a^{(i)} \geq h_i, \ \forall i \in [n]$$

$$\sum_a (I - \gamma P_a^T) \lambda_a = (1-\gamma) q$$

$$\Longleftrightarrow \quad \min_{\mu \geq 0, v} \max_{\lambda \geq 0} L(\lambda, \mu, v) = \lambda^T M \begin{bmatrix} \mu \\ v \end{bmatrix}$$

$$\Downarrow$$

$$\min_{\mu \geq 0, v, \hat{\mu}, \hat{v}} \max_{\lambda \geq 0, \hat{\lambda}} L(\lambda, \mu, v) + \frac{\rho}{2} \left(\|\mu - \hat{\mu}\|^2 + \|v - \hat{v}\|^2 - \|\lambda - \hat{\lambda}\|^2 \right)$$

$$\Downarrow$$

D-GDA Flow

$$\dot{v} = \sum_a (I - \underline{\gamma P_a^T}) \lambda_a - (1-\gamma) q - \rho (v - \hat{v}) \qquad \dot{\hat{v}} = -\rho(\hat{v} - v)$$

$$\dot{\mu}_i = \Pi_{\mathbb{R}_+}\left[\mu;\ h_i - \sum_a \underline{\lambda_a^T r_a^{(i)}} - \rho(\mu_i - \hat{\mu}_i)\right] \qquad \dot{\hat{\mu}}_i = -\rho(\hat{\mu}_i - \mu_i)$$

$$\dot{\lambda}_a = \Pi_{\Delta}\left[\lambda_a;\ \underline{r_a^{(0)}} - (I - \underline{\gamma P_a}) v + \sum_{i \in [n]} \mu_i \underline{r_a^{(i)}} - \rho(\lambda_a - \hat{\lambda}_a)\right] \qquad \dot{\hat{\lambda}}_a = -\rho(\hat{\lambda}_a - \lambda_a)$$

unknowns

Dissipative Stochastic GDA for Constrained RL

- Oracle: At each time t sample $S_0 \sim q$, $(S_t, A_t) \sim \xi$, $S_{t+1} \sim \mathbb{P}(\cdot\,|S_t, A_t)$:
- **DS-GDA Update:**

$$v^{t+1} = v^t + \alpha^t \left[\mathbb{1}_{\{\xi(S_t,A_t)>0\}} \frac{\lambda^t_{S_t,A_t}}{\xi(S_t,A_t)} (\mathbf{e}_{S_t} - \gamma \mathbf{e}_{S_{t+1}}) - (1-\gamma)\mathbf{e}_{S_0} - \rho(v^t - \hat{v}^t) \right], \qquad \hat{v}^{t+1} = \hat{v}^t - \alpha^t \rho(\hat{v}^t - v^t)$$

$$\mu_i^{t+1} = \left[\mu_i^t + \alpha^t (h_i - \mathbb{1}_{\{\xi(S_t,A_t)>0\}} \frac{\lambda^t_{S_t,A_t} R^{(i)}_{t+1}}{\xi(S_t,A_t)} - \rho(\mu_i^t - \hat{\mu}_i^t) \right], \qquad \hat{\mu}_i^{t+1} = \mu_i^t - \alpha^t \rho(\hat{\mu}_i^t - \mu_i^t)$$

$$\lambda_a^{t+1} = \left[\lambda_a^t + \alpha^t \left(\mathbb{1}_{\{\xi(S_t,A_t)>0\,\&\,A_t=a\}} \frac{\sum_{i=1}^n \mu_i^t R^{(i)}_{t+1} + \gamma v^t_{S_{t+1}} - v^t_{S_t}}{\xi(S_t,A_t)} \mathbf{e}_{S_t} - \rho(\lambda_a^t - \hat{\lambda}_a^t) \right) \right]^+, \quad \hat{\lambda}_a^{t+1} = \lambda_a^t - \alpha^t \rho(\hat{\lambda}_a^t - \lambda_a^t)$$

Theorem [Zheng, You, M '22]

Under mild assumptions, as $t \to \infty$ the sequence (λ^t, μ^t, v^t) generated by S-GDA converges to the optimal solution to the C-RL LP Problem.

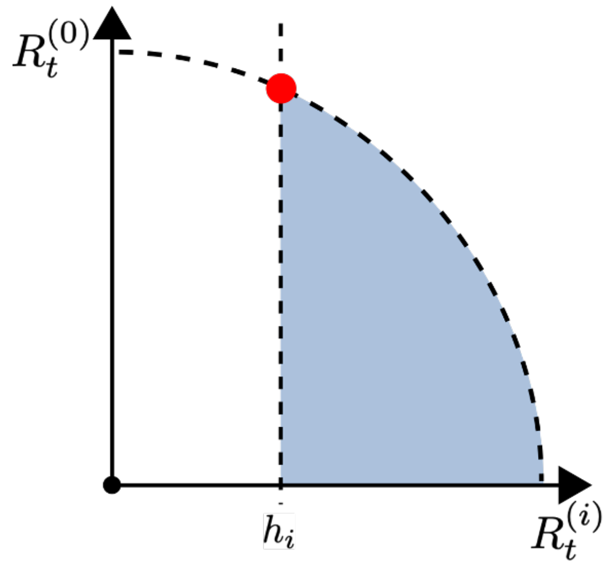
In particular, the iterates $\pi_t(a|s) = \frac{\lambda^t_{s,a}}{\sum_{a'} \lambda^t_{s,a'}} \to \pi^*\ a.s.$

[Zheng, You, M '22] T Zheng P You, and E Mallada. Constrained reinforcement learning via dissipative saddle flow dynamics Asilomar 2022

Challenges of RL for Physical Systems

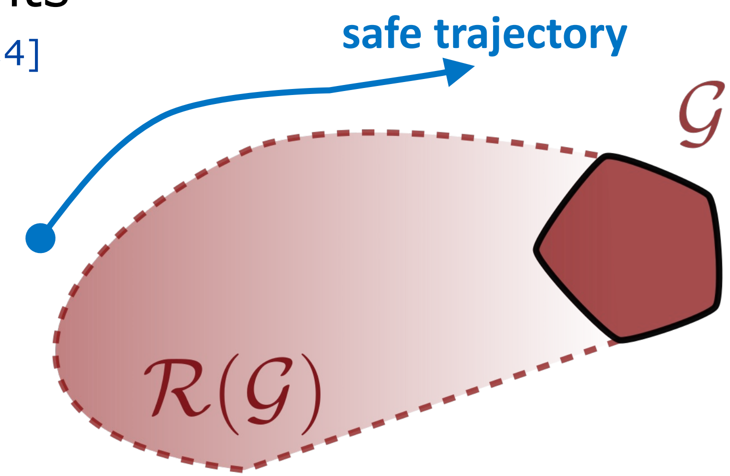
- Physical systems must meet multiple objectives
  - Need to trade off between the different goals
  - Constrained RL allows to explore the Pareto Front [1,2]

$$\max_{\pi} \ (1-\gamma)\mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}^{(0)}\right]$$
$$\text{s.t.} \quad (1-\gamma)\mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}^{(i)}\right] \geq h_i, \ \forall i \in [n]$$



- Failures have a qualitatively different impact
  - Expectation constraints cannot meet safety requirements
  - Hard (almost sure) constraints can guarantee safety [3,4]

$$\max_{\pi} \ \mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$
$$\text{s.t.} \quad \mathbb{P}_{\pi, S_0 \sim q}\left[S_t \notin \mathcal{G}\right] = 1, \ \forall t \geq 0$$



[1] Zheng, You, and M, Constrained reinforcement learning via dissipative saddle flow dynamics, Asilomar 2022
[2] You, and M, Saddle flow dynamics: Observable certificates and separable regularization, ACC 2021
[3] Castellano, Min, Bazerque, M, Reinforcement Learning with Almost Sure Constraints, L4DC 2022
[4] Castellano, Min, Bazerque, M, Learning to Act Safely with Limited Exposure and Almost Sure Certainty, IEEE TAC, 2023
[5] Castellano, Min, Bazerque, M, Correct-by-design Safety Critics Using Non-contractive Bellman Operators, submitted

[Submitted on 9 Dec 2021 (v1), last revised 7 Apr 2022 (this version, v2)]

Reinforcement Learning with Almost Sure Constraints

Agustin Castellano, Hancheng Min, Juan Bazerque, Enrique Mallada

arXiv > cs > arXiv:2112.05198

[Submitted on 18 May 2021 (v1), last revised 25 May 2021 (this version, v2)]

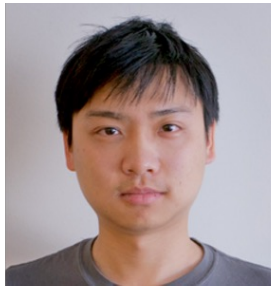
Learning to Act Safely with Limited Exposure and Almost Sure Certainty

Agustin Castellano, Hancheng Min, Juan Bazerque, Enrique Mallada

arXiv > eess > arXiv:2105.08748

Agustin Castellano

Johns Hopkins University



Hancheng Min

Penn University of Pennsylvania

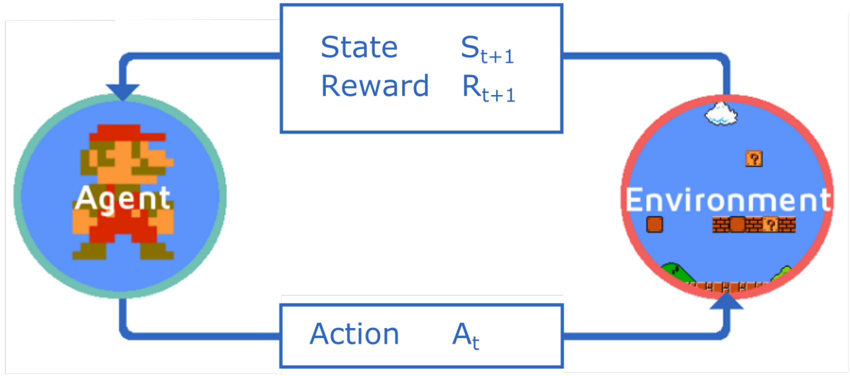
Juan Bazerque

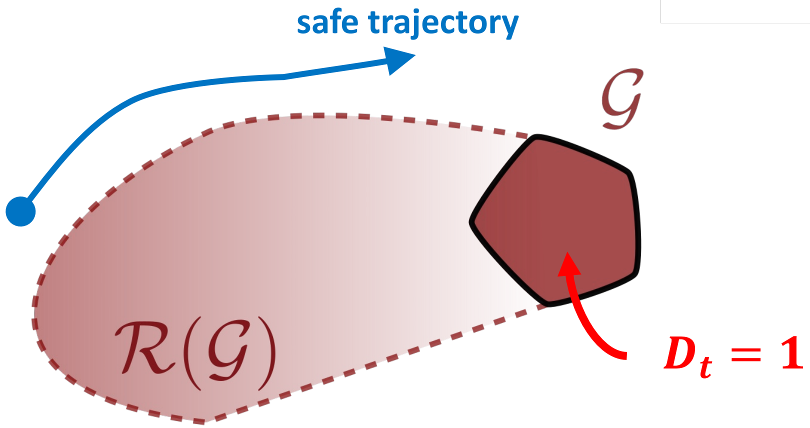
University of Pittsburgh

Reinforcement Learning for Safety-Critical Systems

$$\max_{\pi} \; \mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

$$\text{s.t.} \quad \mathbb{P}_{\pi, S_0 \sim q}\left[S_t \notin \mathcal{G}\right] = 1, \; \forall t \geq 0$$





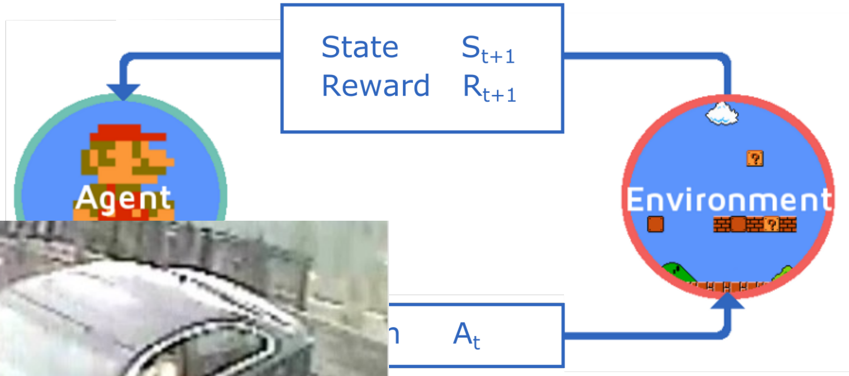
Challenges of SC-RL:

- Avoiding unsafe regions requires anticipation
  - A car at 100 mph at 10 feet from a wall still hasn't hit the wall!

Reinforcement Learning for Safety-Critical Systems

$$\max_{\pi} \ \mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

$$\text{s.t.} \quad \mathbb{P}_{\pi, S_0 \sim q}\left[\quad \right]$$



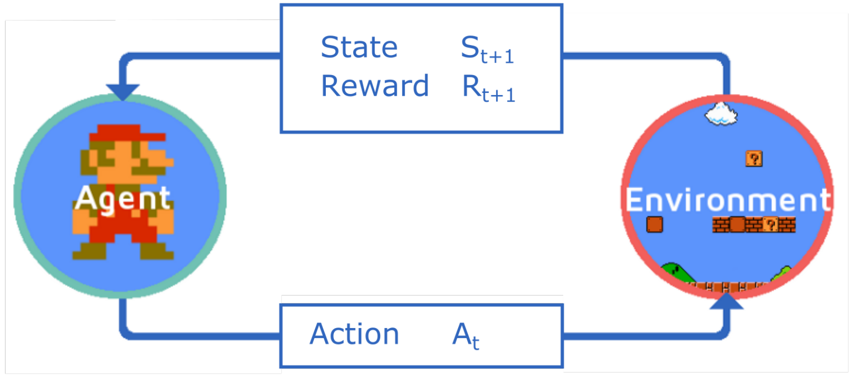
Challenges of S

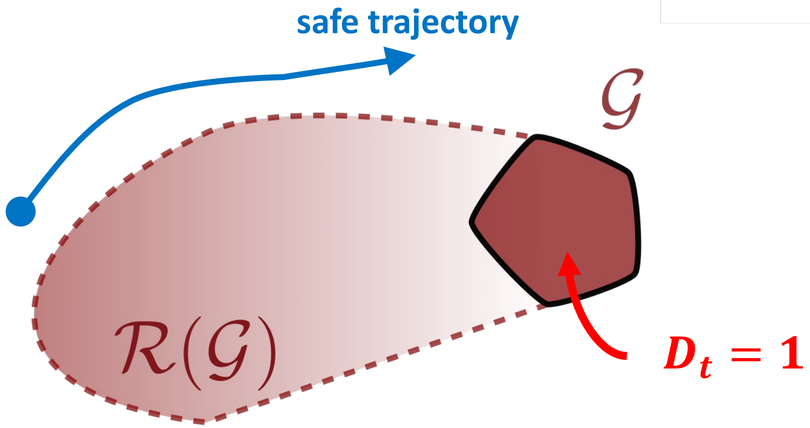
- Avoiding unsafe r
  - A car at 100 mph at 10 feet from a wall still hasn't hit the wall!

Reinforcement Learning for Safety-Critical Systems

$$\max_{\pi} \; \mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

$$\text{s.t.} \quad \mathbb{P}_{\pi, S_0 \sim q}\left[S_t \notin \mathcal{G}\right] = 1, \; \forall t \geq 0$$

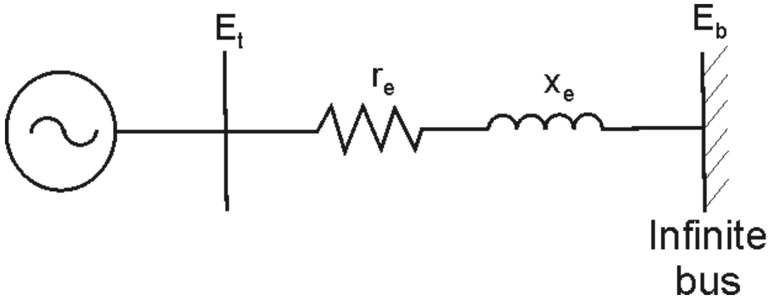




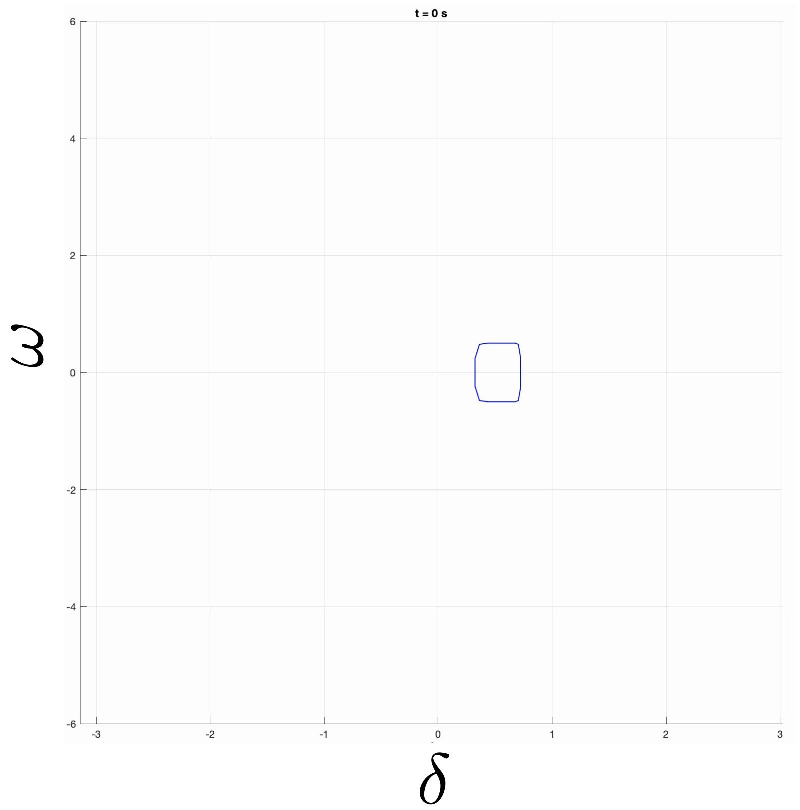
Challenges of SC-RL:

- Avoiding unsafe regions requires anticipation
  - A car at 100 mph at 10 feet from a wall still hasn't hit the wall!
  - Model-based ➔ Reachability Theory
- Model-free:
  - Constraints not given a priori: Need to learn from experience!
  - Constraint violations are inevitable ➔ Maybe not all constraints can be learned online

Example: Transient Stability in Power Systems

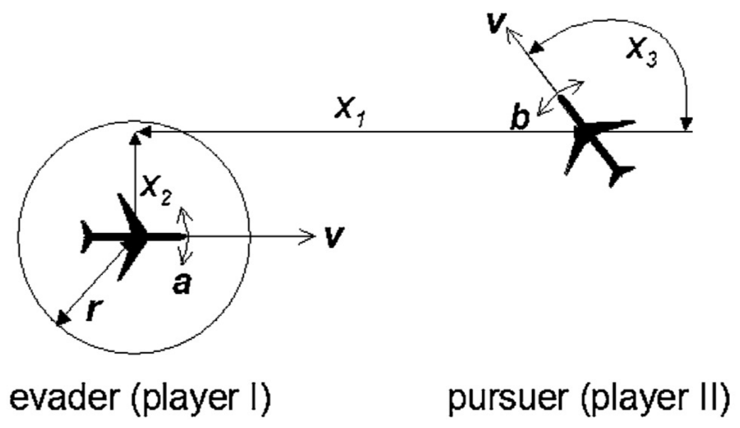


$$\begin{cases} \dot{\delta} = \omega \\ \dot{\omega} = \frac{1}{M}\left(u - D\omega - P_e \sin\delta\right) \\ u \in \left[u_{\min}, u_{\max}\right] \end{cases}$$

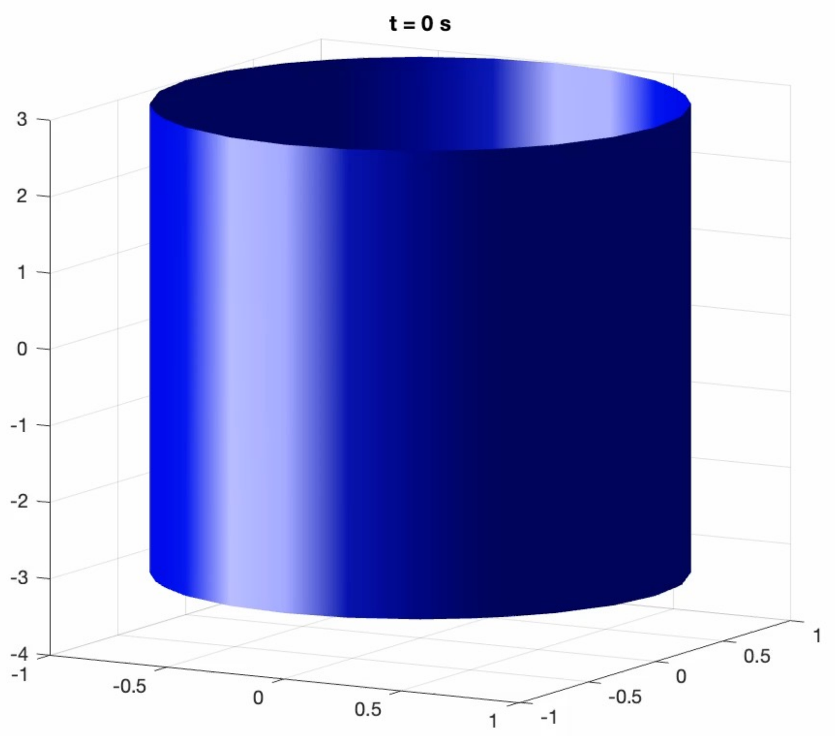


- **Q:** Which states can reach a neighborhood of the stable equilibrium?

Example: Air Collision Avoidance



$$\begin{cases} \dot{x}_1 = -v + v\cos x_3 + a x_2 \\ \dot{x}_2 = v \sin x_3 - a x_2 \\ \dot{x}_3 = b - a \\ b, a \in [-1, 1] \end{cases}$$



- **Q:** From which states can the evader avoid collision?

Related Work

Reachability Theory[1-2]

- **Model-based:** Via Hamilton Jacobi Issacs Equations (cont. time), or iterative set updates (discrete time).
- **Constraints:** Provides hard/almost sure guarantees
- **Output:** Finds the *maximum control invariant set (M-CIS) outside* $\mathcal{G}$

Control Barrier Functions (CBF)[3-4]

- **Model-based:** Requires knowledge of dynamics and *finding such CBF!*
- **Constraints:** Provides hard/almost sure guarantees
- **Output:** Possibly conservative CIS

Safety Critics (SC)[5-7]

- **Model-free:** Q-Learning-like algorithms, computes function such that $Q_{safe}(s, a) \geq \eta_{thresh} \Rightarrow$ "$safety$"
- **Constraints:** Provides soft/approximate guarantees, depending on discounting factor $\gamma \in (0,1)$
- **Output:** *Converges to maximum CIS as* $\gamma \to 1$

[1] I Mitchell, A Bayen, and C Tomlin. "A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games." IEEE TAC, 2005
[2] D Bertsekas. "Infinite time reachability of state-space regions by using feedback control." IEEE TAC, 1972
[3] A Ames, X Xu, J Grizzle, and P Tabuada, "Control barrier function based quadratic programs for safety critical systems," IEEE TAC, 2017.
[4] A Ames, S Coogan, M Egerstedt, G Notomista, K Sreenath, and P Tabuada. "Control barrier functions: Theory and applications" ECC, 2019
[5] J Fisac, N Lugovoy, V Rubies-Royo, S Ghosh, and C Tomlin, "Bridging Hamilton-Jacobi safety analysis and reinforcement learning," ICRA, 2019.
[6] K Srinivasan, B Eysenbach, S Ha, J Tan, and C Finn. "Learning to be safe: Deep RL with a safety critic." arXiv preprint arXiv:2010.14603 (2020).
[7] B Thananjeyan, A Balakrishna, S Nair, M Luo, K Srinivasan, M Hwang, J E Gonzalez, J Ibarz, C Finn, and K Goldberg. Recovery RL: Safe reinforcement learning with learned recovery zones. IEEE Robotics and Automation Letters, 2021

Related Work

Reachability Theory[1-2]

- **Model-based:** Via Hamilton Jacobi Issacs Equations (cont. time), or iterative set updates (discrete time).
- **Constraints:** Provides hard/almost sure guarantees
- **Output:** Finds the *maximum control invariant set (M-CIS) outside* $\mathcal{G}$

Control Barrier Functions (CBF)[3-4]

- **Model-based:** Requires knowledge of dynamics and *finding such CBF!*
- **Constraints:** Provides hard/almost sure guarantees
- **Output:** Possibly conservative CIS

Safety Critics (SC)[5-7]

- **Model-free:** Q-Learning-like algorithms, computes function such that $Q_{safe}(s,a) \geq \eta_{thresh} \Rightarrow$ "$safety$"
- **Constraints:** Provides soft/approximate guarantees, depending on discounting factor $\gamma \in (0,1)$
- **Output:** *Converges to maximum CIS as* $\gamma \to 1$

Method	Model-free	Constraint Type	Set Size
Reachability Theory[1-2]	No	Hard	Maximal
Control Barrier Functions[3-4]	No	Hard	Subset
Safety Critics[5-7]	Yes	Soft/Approx.	Maximal

Our Work

Reachability Theory[1-2]

- **Model-based:** Via Hamilton Jacobi Issacs Equations (cont. time), or iterative set updates (discrete time).
- **Constraints:** Provides hard/almost sure guarantees
- **Output:** Finds the *maximum control invariant set (M-CIS) outside* $\mathcal{G}$

Control Barrier Functions (CBF)[3-4]

- **Model-based:** Requires knowledge of dynamics and *finding such CBF!*
- **Constraints:** Provides hard/almost sure guarantees
- **Output:** Possibly conservative CIS

Safety Critics (SC)[5-7]

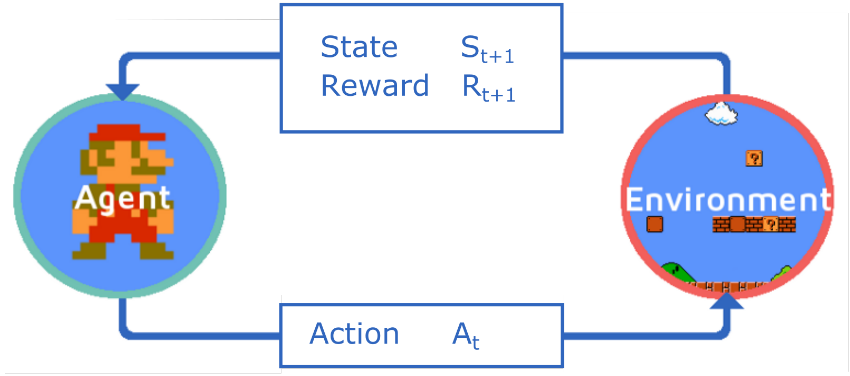
- **Model-free:** Q-Learning-like algorithms, computes function such that $Q_{safe}(s,a) \geq \eta_{thresh} \Rightarrow$ "$safety$"
- **Constraints:** Provides soft/approximate guarantees, depending on discounting factor $\gamma \in (0,1)$
- **Output:** *Converges to maximum CIS as* $\gamma \to 1$

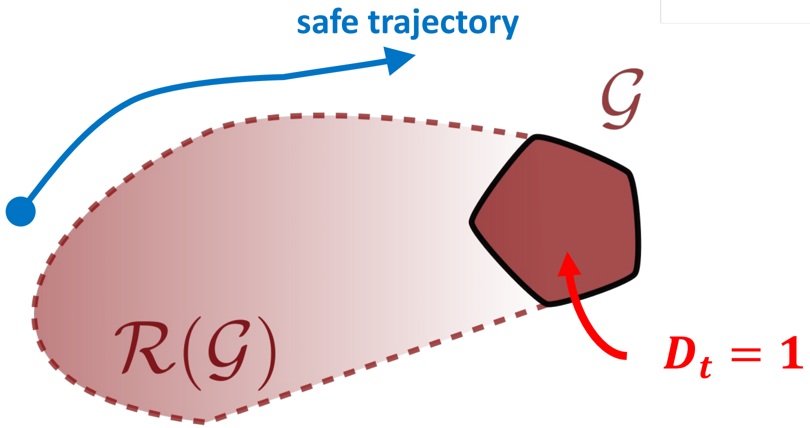
Method	Model-free	Constraint Type	Set Size
Reachability Theory[1-2]	No	Hard	Maximal
Control Barrier Functions[3-4]	No	Hard	Subset
Safety Critics[5-7]	Yes	Soft/Approx.	Maximal
Ours	Yes	Hard	Maximal and Subsets

Reinforcement Learning for Safety-Critical Systems

$$\max_{\pi} \; \mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

$$\text{s.t.} \quad \mathbb{P}_{\pi, S_0 \sim q}\left[S_t \notin \mathcal{G}\right] = 1, \; \forall t \geq 0$$





Methodology:

- Enhance RL with **logical** feedback naturally arising from constraint violations

$$S_t \in \mathcal{G} \Leftrightarrow D_t = 1$$

- Decouple **feasibility** from optimality: **Separation Principle**
- Develop algorithms for learning fixed points of **non-contractive operators**

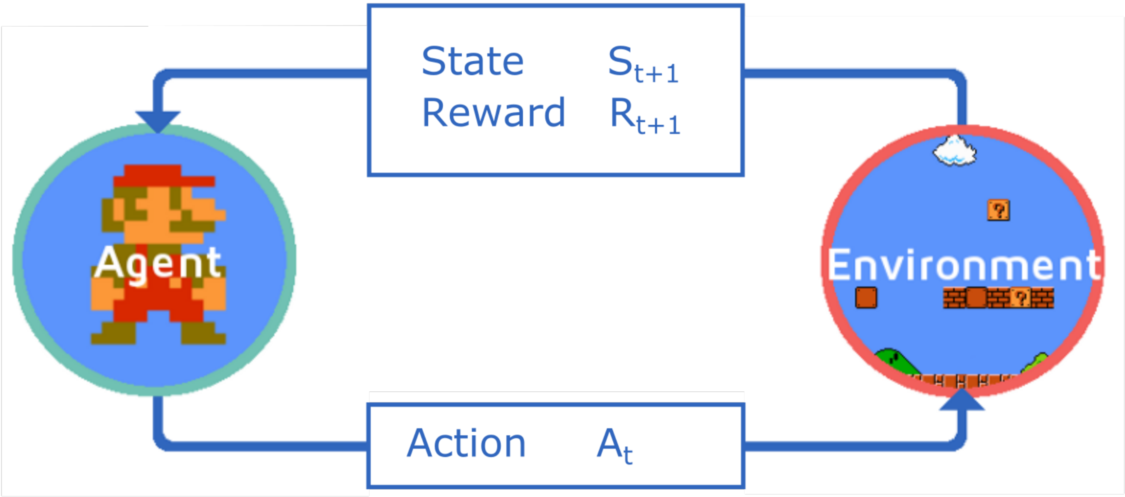
Outline

- Separation Principle for Joint Safety & Optimality
- Learning Safety with Limited Failures
- One-sided Bellman Equations for Continuous States

Recap: RL with Almost Sure Constraints

$$\max_{\pi} \ \mathbb{E}_{\pi, S_0 \sim q}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

$$\text{s.t.} \quad \mathbb{P}_{\pi, S_0 \sim q}\left[S_t \notin \mathcal{G}\right] = 1, \ \forall t \geq 0 \iff D_{t+1} = 0 \text{ almost surely } \forall t$$



- Damage indicator $D_t \in \{0,1\}$ turns on $(D_t = 1)$ when constraints are violated

Formulation via hard barrier indicator

Safe RL problem:

$$V^*(s) := \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \mid S_0 = s\right]$$
$$\text{s.t.: } D_{t+1} = 0 \text{ almost surely } \forall t$$

$\sim$

Equivalent unconstrained formulation:

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} + \underbrace{\log[1 - D_{t+1}]}_{\begin{array}{ll} 0 & \textit{if } D_{t+1} = 0 \\ -\infty & \textit{if } D_{t+1} = 1 \end{array}} \mid S_0 = s\right]$$

Questions/Comments:

- Is this just a standard RL problem with $\tilde{R}_{t+1} = R_{t+1} + \log(1 - D_{t+1})$?
- Standard MDP assumptions for Value Iteration, Bellman's Eq., Optimality Principle, etc., do not hold!
- Not to mention convergence of stochastic approximations.

Key idea: Separate the problem of safety from optimality

Hard Barrier Action-Value Functions

Consider the Q-function for a given policy π,

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\left(\gamma^t R_{t+1} + \log(1 - D_{t+1})\right) \;\middle|\; S_0 = s, A_0 = a\right]$$

and define the hard-barrier function

$$B^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\log(1 - D_{t+1}) \;\middle|\; S_0 = s, A_0 = a\right]$$

Notes on $\boldsymbol{B^\pi(s,a)}$:

- $\boldsymbol{B^\pi(s,a) \in \{0, -\infty\}}$
- Summarizes safety information
  - $\boldsymbol{B^\pi(s,a) = 0}$ iff π is safe after choosing $A_t = a$ when $S_t = s$
- It is independent of the reward process

Separation Principle

Theorem (Separation principle)

Assume rewards R_{t+1} are bounded almost surely for all t. Then for every policy π:

$$Q^\pi(s,a) = Q^\pi(s,a) + B^\pi(s,a)$$

In particular, for optimal π_*

$$Q^*(s,a) = Q^*(s,a) + B^*(s,a)$$

Approach: Learn feasibility (encoded in B^*) independently from optimality.

Optimal Hard Barrier Action-Value Function

Theorem (Safety Bellman Equation for B^*)

Let $B^*(s,a) := \max_\pi B^\pi(s,a)$, then the following holds:

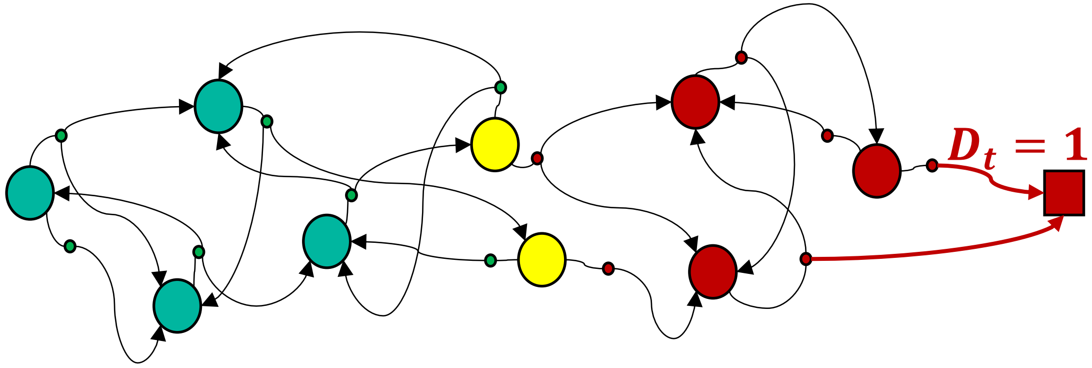
$$B^*(s,a) = \mathbb{E}\left[-\log(1 - D_{t+1}) + \max_{a'} B^*(S_{t+1}, a') \;\middle|\; S_0 = s, A_0 = a \right]$$

Understanding $B^*(s,a)$:

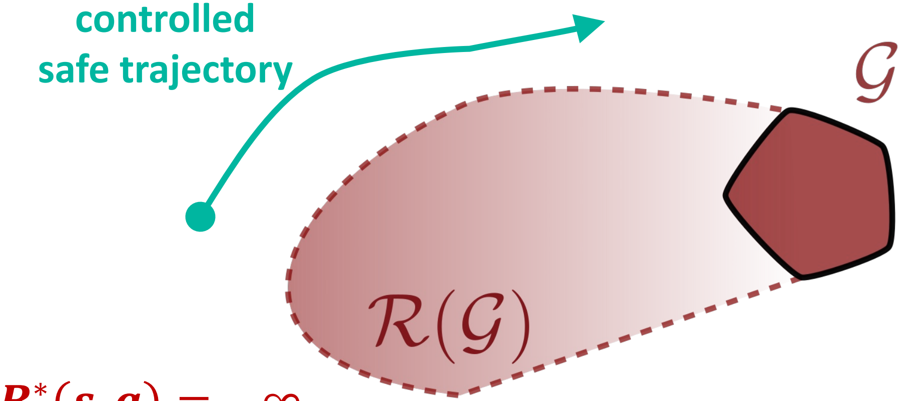
$B^*(s,a) \in \{0, -\infty\}$ summarizes safety information of the entire MDP

- $B^*(s,a) = 0$ if $\exists$ safe π after choosing $A_t = a$ when $S_t = s$ **Control Invariant**
- $B^*(s,a) = -\infty$ if no safe policy exists after choosing $A_t = a$ when $S_t = s$ **Unsafe**

Discrete States



Continuous States



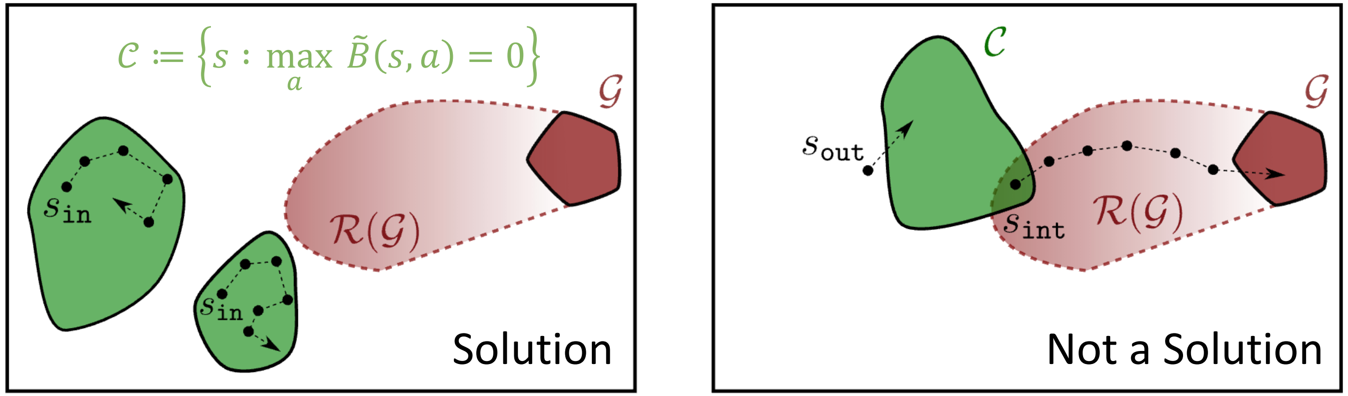
- $V^*(s) = \max_a B^*(s,a) = 0$
- $V^*(s) = \max_a B^*(s,a) = -\infty$

Properties of Safety Bellman Equation

Understanding the Solutions to the Safety Bellman Equation (SBE):

$$\tilde{B}(s,a) = \mathbb{E}\left[-\log(1 - D_{t+1}) + \max_a \tilde{B}(S_{t+1}, a) \;\middle|\; S_0 = s, A_0 = a\right]$$

- SBE can have **multiple solutions,** including $\tilde{\boldsymbol{B}}(\boldsymbol{s},\boldsymbol{a}) = -\infty$, for all pairs (s,a)
- If the function $\tilde{\boldsymbol{B}}$ is a solution to the SBE, then:
  - The set $\mathcal{C} \coloneqq \left\{s : \max_a \tilde{B}(s,a) = 0\right\}$ is a *control invariant safe set*
  - $\mathcal{C}$ is *maximal*: If $S_0 \notin \mathcal{C}$, then S_t never reaches $\mathcal{C}$ for all policies π



Outline

- Separation Principle for Joint Safety & Optimality
- Learning Safety with Limited Failures
- One-sided Bellman Equations for Continuous States

Learning the barrier in finite MDPs…

Algorithm 3: `barrier_update`

B-function (initialized as all-zeroes);
Input: (s, a, s', d)
Output: Barrier-function $B(s, a)$
$B(s, a) \leftarrow B(s, a) + \log(1 - d) + \max_{a'} B(s', a')$

Pros:

- Wraps around learning algorithms (Q-learning, SARSA)
- Use the B to trim the exploration set and avoid repeating unsafe actions

…with a generative model:

- Sample a transition (s, a, s', d) according to the MDP. Update barrier function.

Algorithm 5: Barrier Learner Algorithm

Data: Constrained Markov Decision Process $\mathcal{M}$
Result: Optimal action-value function B^*
$\boxed{\text{Initialize } B^{(0)}(s, a) = 0, \forall (s, a) \in \mathcal{S} \times \mathcal{A}}$ — Initially, all (s, a)-pairs are "safe"
for $t = 0, 1, \cdots$ **do**
  $\boxed{\text{Draw } (s_t, a_t) \sim \text{Unif}(\{(s, a) : B^{(t)}(s, a) \neq -\infty\})}$ — Draw (s, a)-pair uniformly among those considered to be "safe" at time t
  Sample transition (s_t, a_t, s'_t, d_t) according to
   $P(S_1 = s'_t, D_1 = d_t | S_0 = s_t, A_0 = a_t)$
  $\boxed{B^{(t+1)} \leftarrow \texttt{barrier\_update}(B^{(t)}, s_t, a_t, s'_t, d_t)}$ — Update barrier function
end

Convergence in Expected Finite Time

Theorem (Safety Guarantee): Let $T = \min_t \{B^{(t)} = B^*\}$, then

$$\mathbb{E}T \leq (L+1)\frac{|S||A|}{\mu}\left(\sum_{k=1}^{|S||A|}\frac{1}{k}\right)$$

- After $T = \min_t \{B^{(t)} = B^*\}$, all "unsafe" (s,a)-pairs are detected
- μ: Lower bound on the non-zero transition probability

$$\mu = min\{p(s', d|s, a): p(s', d|s, a) \neq 0\}$$

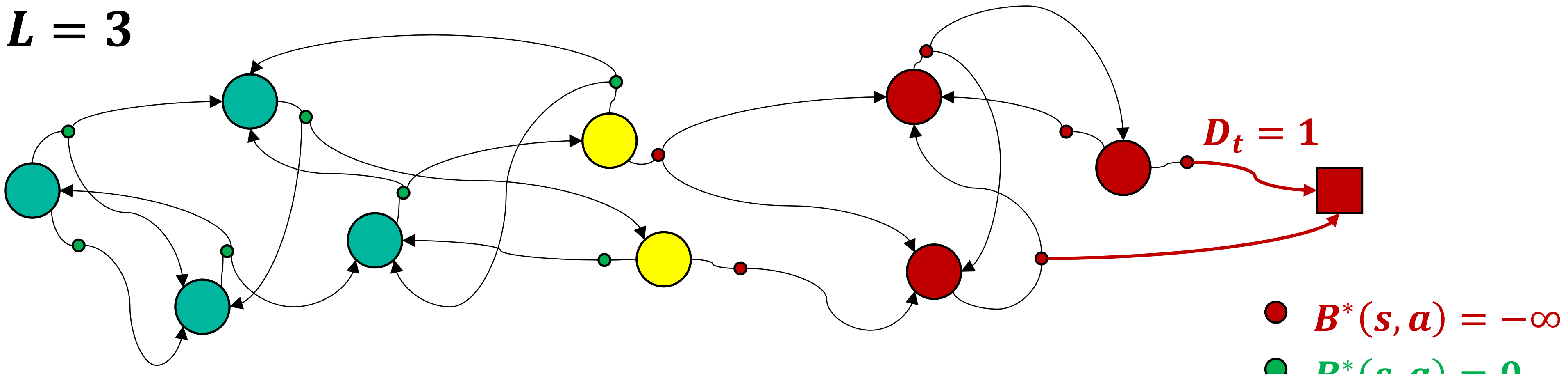
- L: **Lag of the MDP**

$$L = \max_{\substack{(s,a) \\ B^*(s,a)=-\infty}} \left\{ \begin{array}{l} \underline{\text{Minimum}} \text{ number of transitions} \\ \text{needed to observe damage,} \\ \text{starting from unsafe } (s,a) \end{array} \right\}$$

Lag of the MDP: L

$$L = \max_{\substack{(s,a) \\ B^*(s,a) = -\infty}} \left\{ \begin{array}{l} \underline{\text{Minimum}} \text{ number of transitions needed to} \\ \text{observe damage, starting from unsafe } (s,a) \end{array} \right\}$$

$L = 3$



Sample Complexity of Safety

Theorem (Sample Complexity): With at least $1-\delta$ probability, the algorithm learns optimal barrier function B^* after

$$(L+1)\frac{|S||A|}{\mu}\left(\sum_{k=1}^{|S||A|}\frac{1}{k}\right)\log\frac{1}{\delta}$$

iterations

- Concentration of sum of exponential random variables
- **Much more sample-efficient** than "learning an ϵ-optimal policy with $1-\delta$ probability" (Li et al. 2020)

$$N=\frac{|S||A|}{(1-\gamma)^4\varepsilon^2}\log^2\left(\frac{|S||A|}{(1-\gamma)\varepsilon\delta}\right)$$

Sample Complexity of Safety

Theorem (Sample Complexity): With at least $1-\delta$ probability, the algorithm learns optimal barrier function B^* after

$$(L+1)\frac{|S||A|}{\mu}\left(\sum_{k=1}^{|S||A|}\frac{1}{k}\right)\log\frac{1}{\delta}$$

iterations

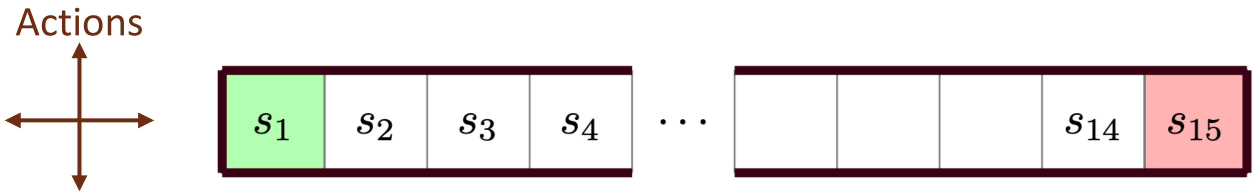
- Concentration of sum of exponential random variables

- If the Barrier Function is learnt first, then learning an ϵ-optimal policy takes

$$N' = \frac{|S_{safe}||A_{safe}|}{(1-\gamma)^4\varepsilon^2}\log^2\left(\frac{|S_{safe}||A_{safe}|}{(1-\gamma)\varepsilon\delta}\right)$$

samples (**Trimming the MDP by learning the barrier**)

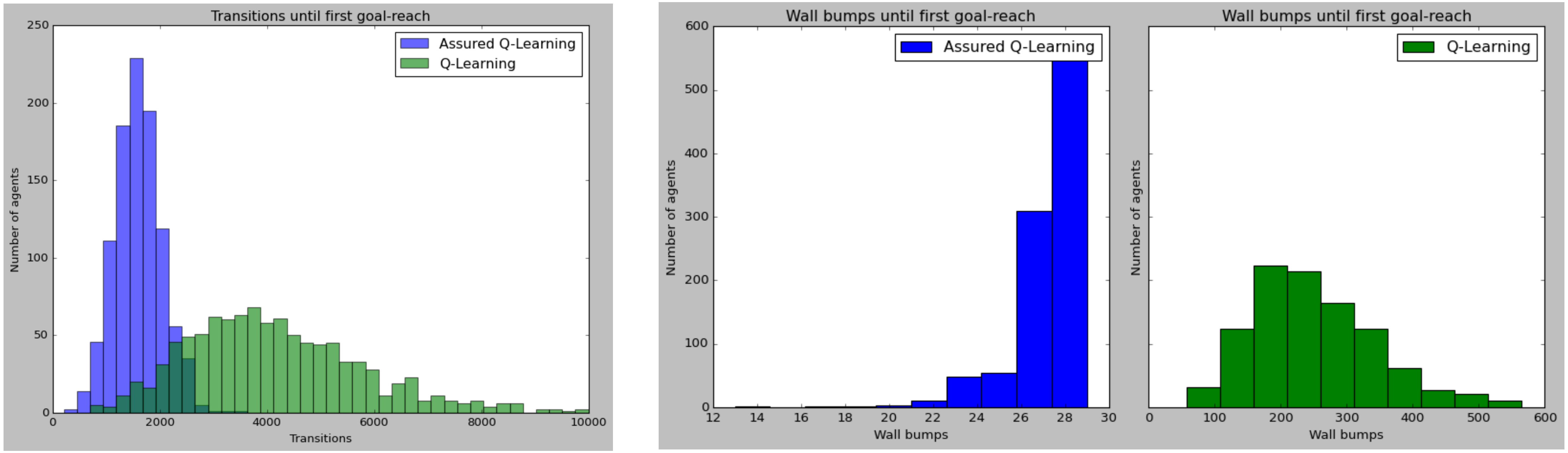
Numerical Experiments



Goal: Reach the end of the aisle ($R_{t+1} = 10$)

Touching the wall gives $D_{t+1} = 1$, resets the episode.

Results



Why does Assured Q-learning perform much better?

If $D_{t+1} = 1 \Longrightarrow B_\pi(s,a) = -\infty \Longrightarrow$ Never take action a at s again!

Takeaways:

- Adding constraints to the problem can accelerate learning
- Barrier function avoids actions that lead to further wall bumps

Outline

- Separation Principle for Joint Safety & Optimality
- Learning Safety with Limited Failures
- One-sided Bellman Equations for Continuous States

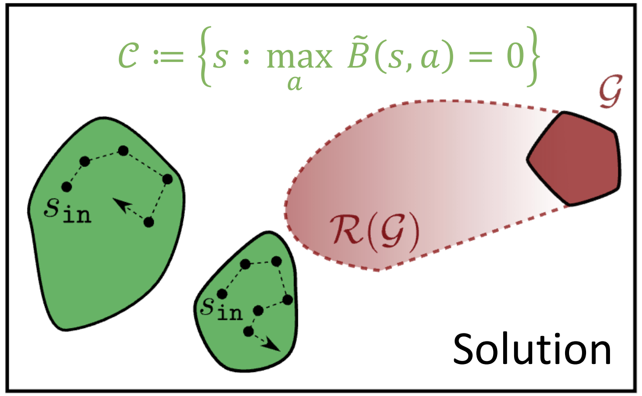
Recall: Properties of Safety Bellman Equation

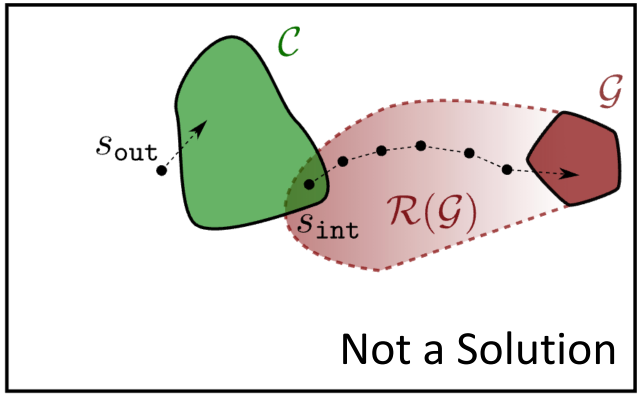
Understanding the Solutions to the Safety Bellman Equation (SBE):

$$\tilde{B}(s,a) = \mathbb{E}\left[-\log(1 - D_{t+1}) + \max_a \tilde{B}(S_{t+1}, a) \;\middle|\; S_0 = s, A_0 = a\right]$$

Understanding the Solutions to the Safety Bellman Equation (SBE):

- SBE can have **multiple solutions,** including $\boldsymbol{\tilde{B}(s,a)} = -\infty$, for all pairs (s,a)
- If the function $\boldsymbol{\tilde{B}}$ is a solution to the SBE, then:
  - The set $\mathcal{C} \coloneqq \left\{s : \max_a \tilde{B}(s,a) = 0\right\}$ is a *control invariant safe set*
  - ~~$\mathcal{C}$ is *maximal*: If $S_0 \notin \mathcal{C}$, then S_t never reaches $\mathcal{C}$ for all policies π~~





Problem: Maximal solutions can be very close to unsafe region $\mathcal{R}(\mathcal{G})$

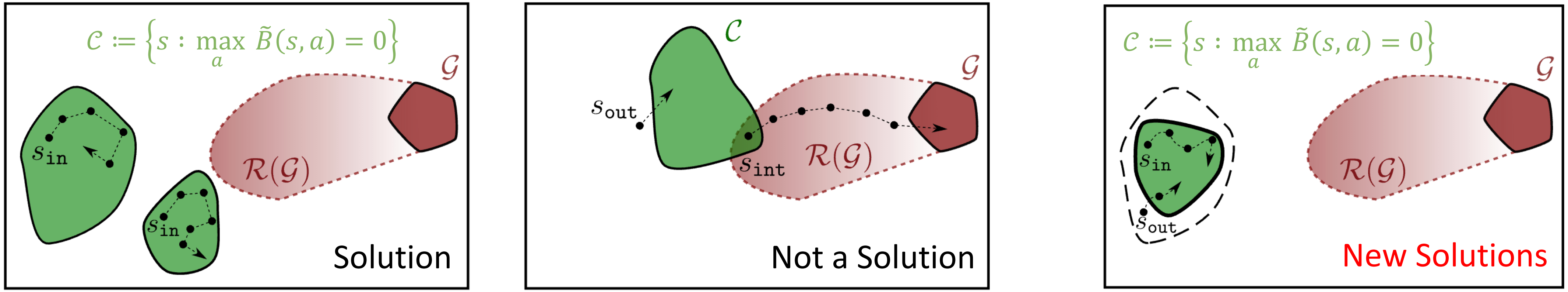
One-Sided Safety Bellman Equation

Theorem (One-Sided Safety Bellman Equation)

Let $\tilde{\boldsymbol{B}}(\boldsymbol{s}, \boldsymbol{a})$ be a solution of the following set of inequalities:

$$\tilde{B}(s,a) \leq \mathbb{E}\left[-\log(1-D_{t+1}) + \max_{a'} \tilde{B}\,(S_{t+1}, a')|S_0 = s, A_0 = a\right]$$

The set $\mathcal{C} \coloneqq \left\{s : \max_{a}\ \tilde{B}(s,a) = 0\right\}$ is a *control invariant safe set,* not necessarily maximal



Learning CIS Using Deep Neural Nets

Algorithm Summary

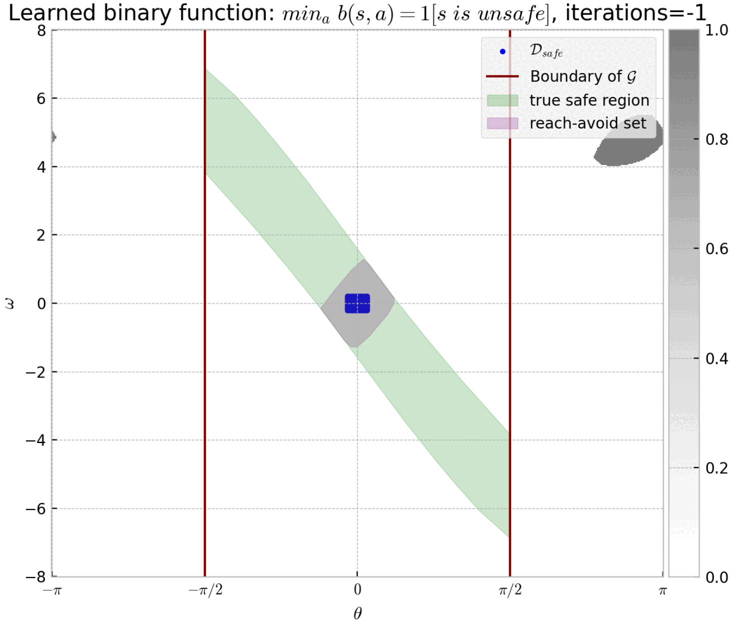
- Uses *axiomatic data* $(\mathrm{s}, \mathrm{a}, \mathrm{d}, \mathrm{s}') \in \mathcal{D}_{safe}$ known to be safe
- *Initialize* $\hat{b}^{\theta}(s, a) = 0$, where $\hat{b}(s, a) = 1 - e^{B(s,a)}$ (all presumed safe)
- *At each iteration*, take N episodes starting from $\mathcal{D}_{safe}$
  - Behavioral policy: *uniform safe policy*

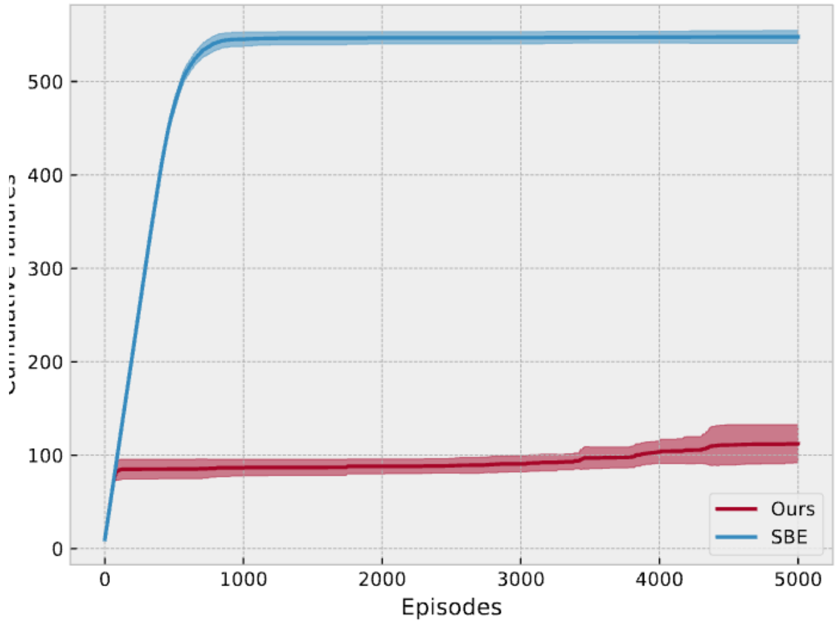
$$\pi^{\theta}(a|s) = \begin{cases} 0 & \text{if } \hat{b}^{\theta}(s,a) = 1 \\ 1/\sum_{a' \in \mathcal{A}} \mathbb{1}\{\hat{b}^{\theta}(s,a') = 0\} & \text{if } \hat{b}^{\theta}(s,a) = 0 \end{cases}$$

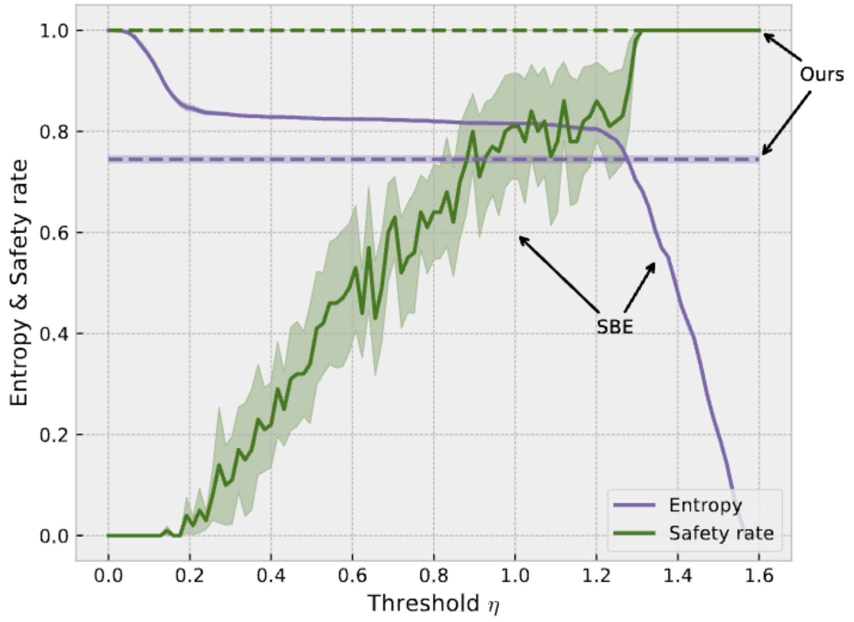
  - Train NN using SGD *until fully fitting the data*
  - Start a new iteration, and repeat

Numerical Illustration

Control Engineer Favorite's: Inverted Pendulum







SBE = Fisac's '19 Safety Critic

Summary and future work

- **Methodologies to Adapt Reinforcement Learning to Safety-Critical Systems**

- **C-RL via Dissipative Saddle Flows**
  - Investigate methods to learn saddle-points in deterministic and stochastic settings
  - Proposed **a general methodology** to ensure convergence to saddle points of general convex-concave functions
  - Application to Constrained RL problems
  - **Takeaways:**
    - **Dissipative GDA** guarantees convergence on a **wide family of minimax problems**
    - When combined with stochastic approximations (D-SGDA) renders **convergent policy iterates** $\pi_k \to \pi^*$ **a.s.**

- **RL with Almost Sure Constraints**
  - Treat constraints separately or in parallel (Barrier Learner)
  - *Finite State-Spaces:* Can **characterize** all feasible policies ($D_t \equiv 0$) with **finite mistakes**
  - *Continuous State-Spaces:* Requires learning using Bellman equations with non-unique solutions
  - **Takeaways:**
    - **Learning feasible policies** is simpler **than learning** the optimal ones
    - Adding **constraints** makes **optimal policies, easier to find**
    - **One-sided Safe Bellman** can be used to find CISs that are not maximal

Thanks!

Related Publications:

[1] Castellano, Min, Bazerque, M, *Reinforcement Learning with Almost Sure Constraints,* **L4DC, 2022**

[2] Castellano, Min, Bazerque, M, *Learning to Act Safely with Limited Exposure and Almost Sure Certainty,* **IEEE TAC, 2023**

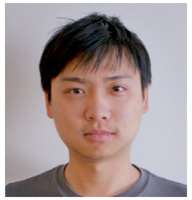
[3] Castellano, Min, Bazerque M, Correct-by-design Safety Critics Using Non-contractive Bellman Operators, **submitted**

Tianqi Zheng

amazon

Agustin Castellano

JOHNS HOPKINS UNIVERSITY



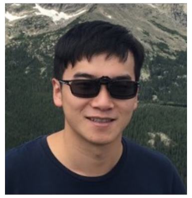
Hancheng Min

Penn UNIVERSITY of PENNSYLVANIA

Enrique Mallada

mallada@jhu.edu

http://mallada.ece.jhu.edu



Pengcheng You

北京大学 PEKING UNIVERSITY

Juan Bazerque

University of Pittsburgh