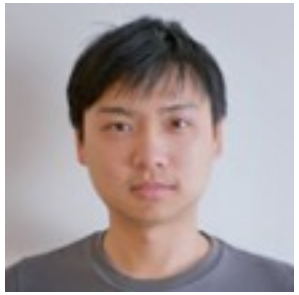
# Reinforcement Learning for Safety-Critical Applications

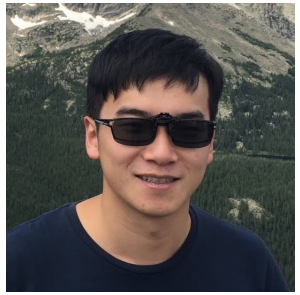
**Enrique Mallada**

JOHNS HOPKINS UNIVERSITY


T. Zheng


A. Castellano


H. Min

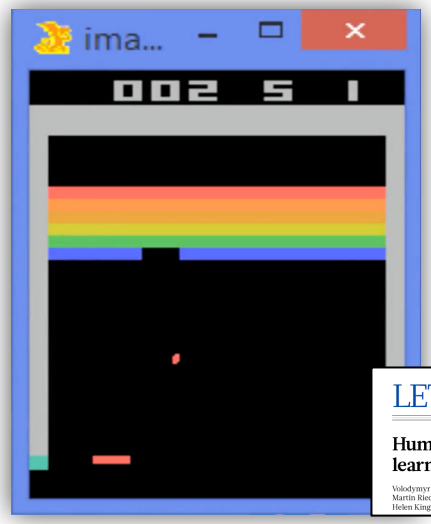

P. You


J. Bazerque

**Jun 12, 2024**

# A World of Success Stories

2017 Google DeepMind's DQN




LETTER

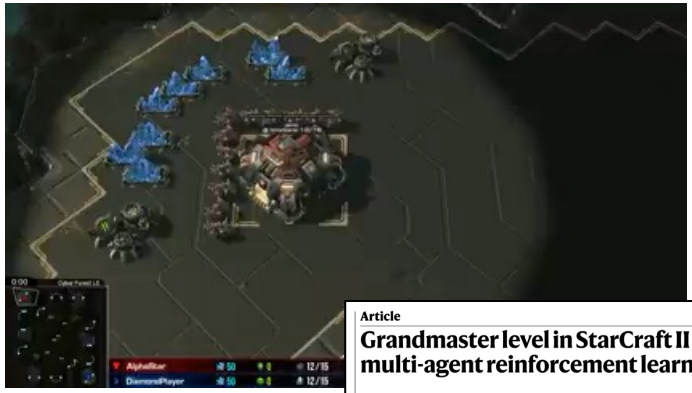
doi:10.1038/nature14236

**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



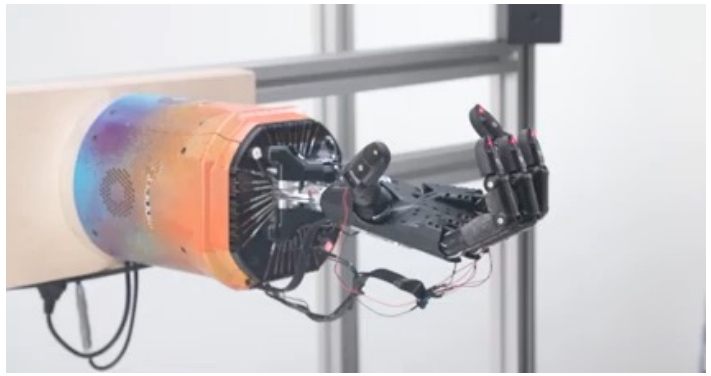
Article

**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

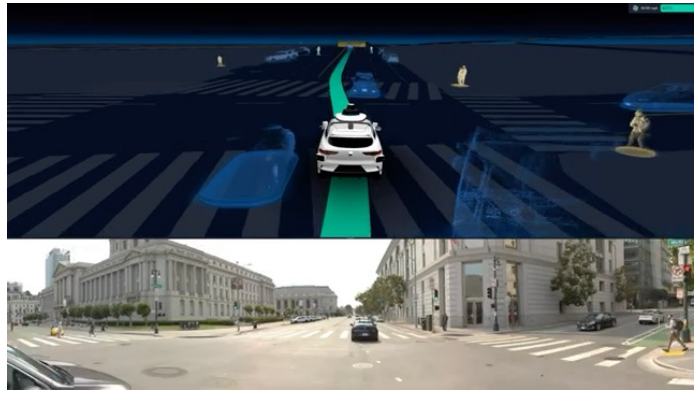
Oriol Vinyals[1,2*], Igor Babuschkin[1,2], Wojciech M. Czarnecki[1,2], Michaël Mathieu[1,2], Andrew Dudzik[1,2], Junyoung Chung[1,2], David H. Choi[1,2], Richard Powell[1,2], Timo Ewalds[1,2], Petko Georgiev[1,2], Junhyuk Oh[1,2], Dan Horgan[1,2], Manuel Kroiss[1,2], Ivo Danihelka[1,2], Aja Huang[1,2], Laurent Sifre[1,2], Trevor Cai[1,2], John P. Agapiou[1,2], Max Jaderberg[1], Alexander S. Vezhnevets[1], Rémi Leblond[1], Tobias Pohlen[1], Valentin Dalibard[1], David Budden[1], Yury Sulsky[1], James Molloy[1], Tom L. Paine[1], Caglar Gulcehre[1], Ziyu Wang[1], Tobias Pfaff[1], Yuhuai Wu[1], Roman Ring[1], Dani Yogatama[1], Dario Wünsch[1], Katrina McKinney[1], Oliver Smith[1], Tom Schaul[1], Timothy Lillicrap[1], Koray Kavukcuoglu[1], Demis Hassabis[1], Chris Apps[1,2] & David Silver[1,2*]

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.

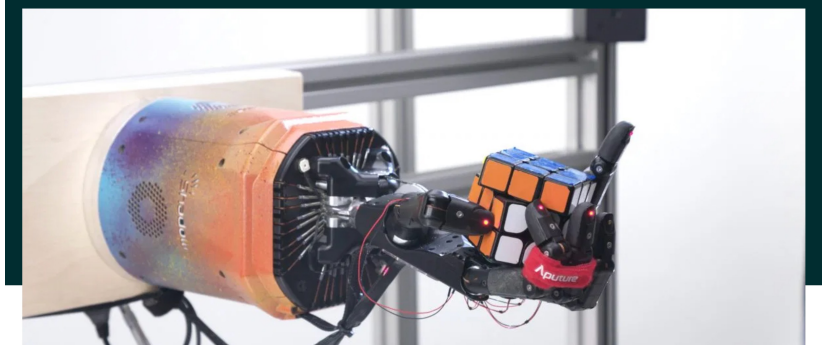
AARIAN MARSHALL BUSINESS 12.07.2020 04:06 PM

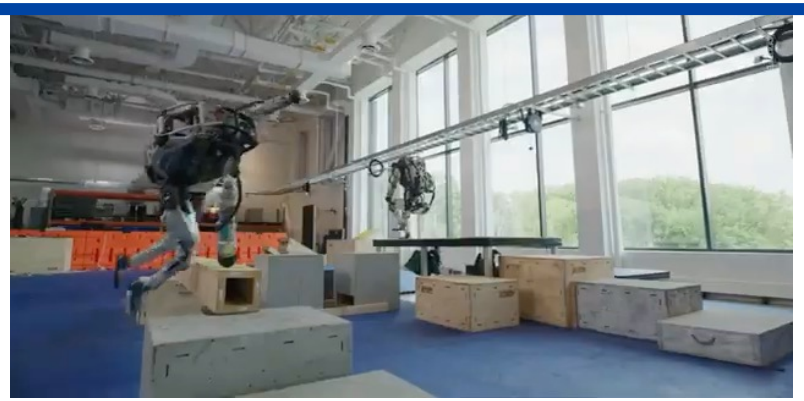
## Uber Gives Up on the Self-Driving Dream

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

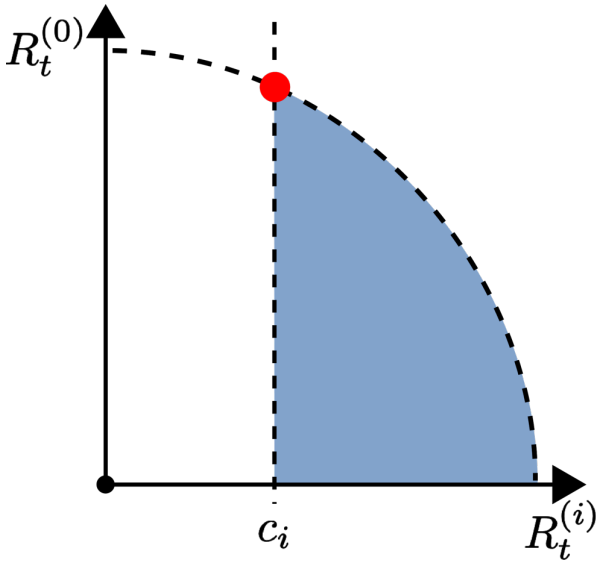
woman did not recognize that pedestrians jaywalk

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

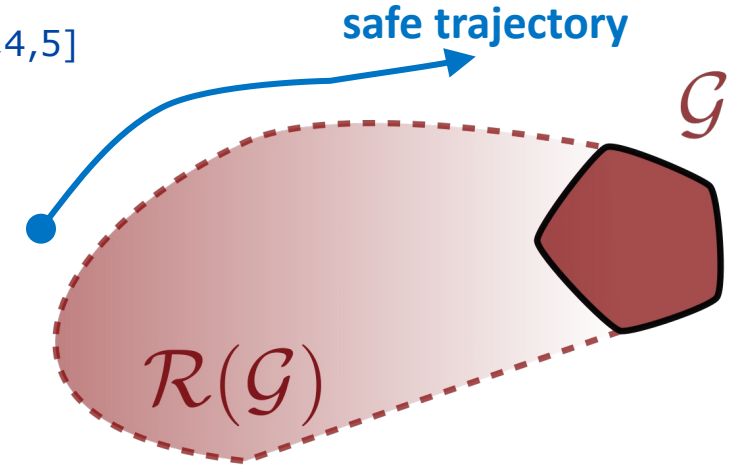
- Physical systems must meet multiple objectives
  - Need to trade off between the different goals
  - Constrained RL allows to explore the Pareto Front [1,2]

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




- Failures have a qualitatively different impact
  - Expectation constraints cannot meet safety requirements
  - Hard (almost sure) constraints can guarantee safety [3,4,5]

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





[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)]*

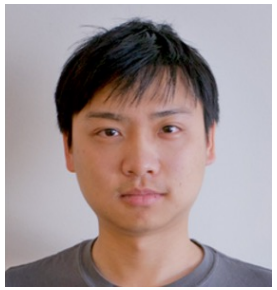
## Learning to Act Safely with Limited Exposure and Almost Sure Certainty

Agustin Castellano, Hancheng Min, Juan Bazerque, Enrique Mallada

arXiv > eess > arXiv:2105.08748

**Agustin Castellano**

JOHNS HOPKINS UNIVERSITY



**Hancheng Min**

Penn
University of Pennsylvania

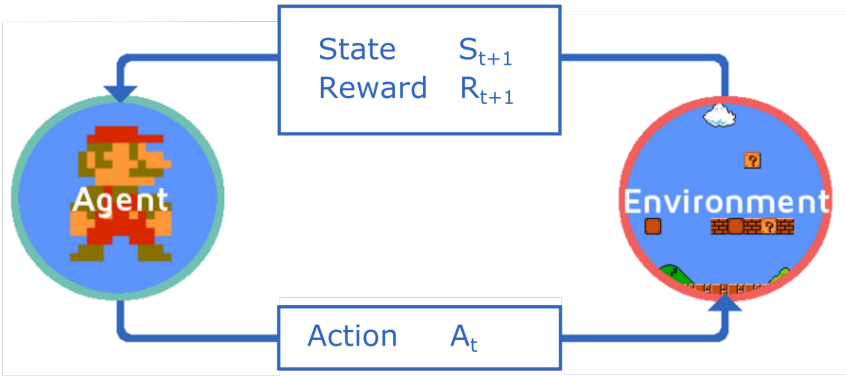
**Juan Bazerque**

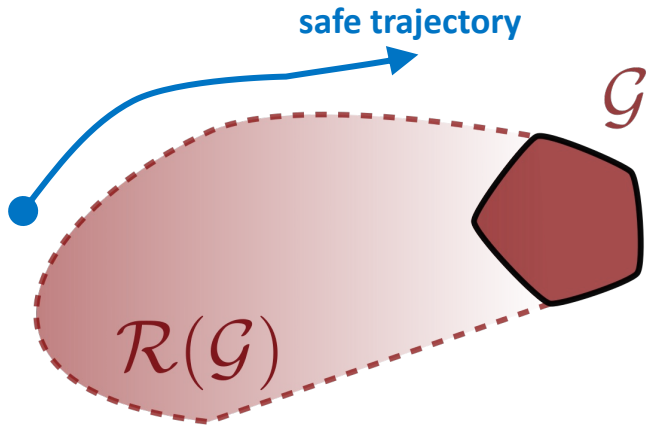
University of Pittsburgh

# Reinforcement Learning for Safety-Critical Systems

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

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



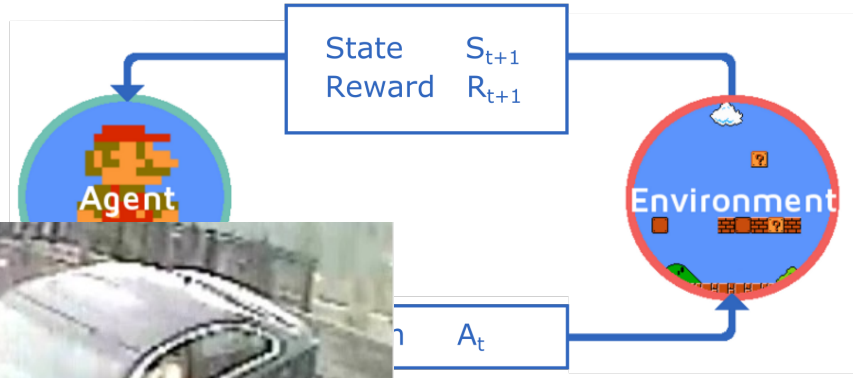

**Challenges of SC-RL:**

- 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

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

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





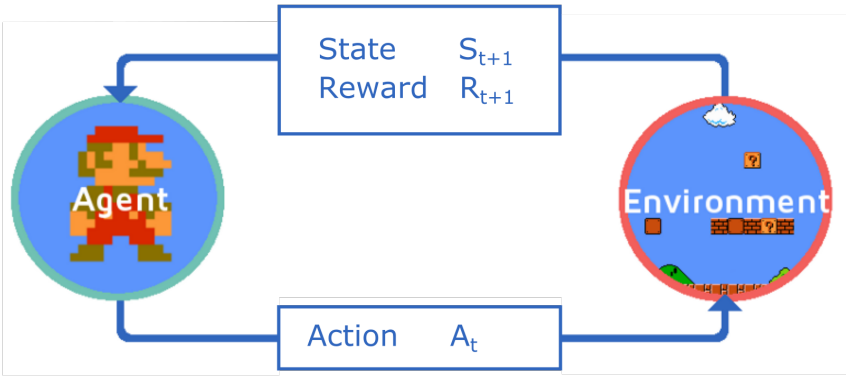
**Challenges of S**

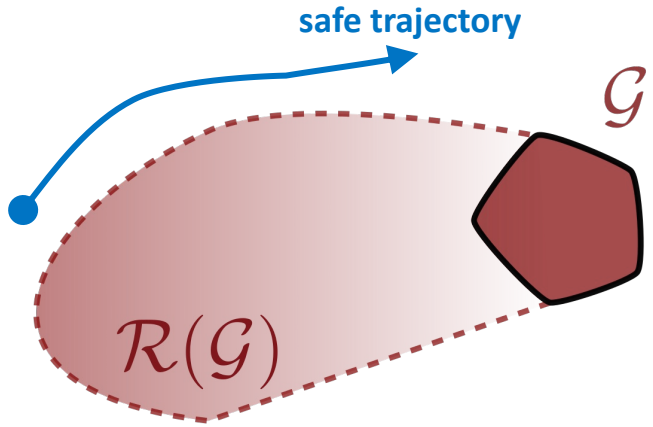
- Avoiding unsafe r
  - A car at 100 mph at 10 feet from a wall still hasn't hit the wall!

# 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

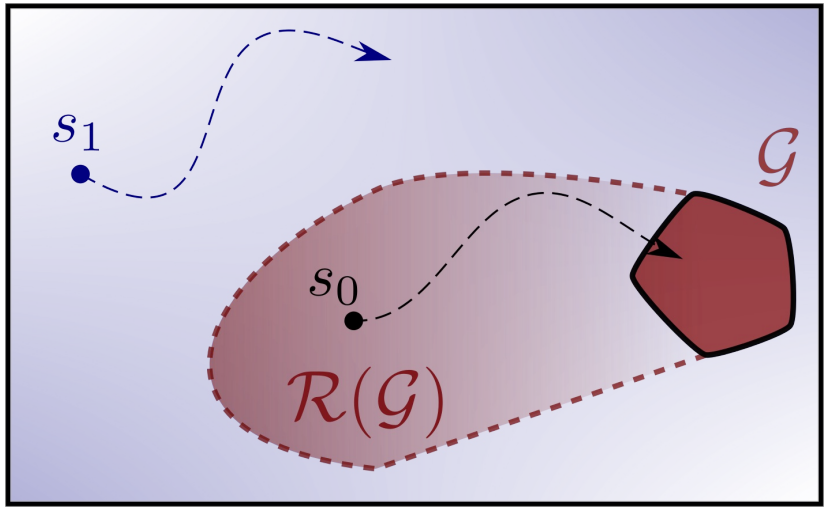
# Reachability Theory

Consider a controlled system

$$\dot{s} = f(s, a, d)$$

$a(\cdot)$ : control/actions

$d(\cdot)$ : disturbance/adversary



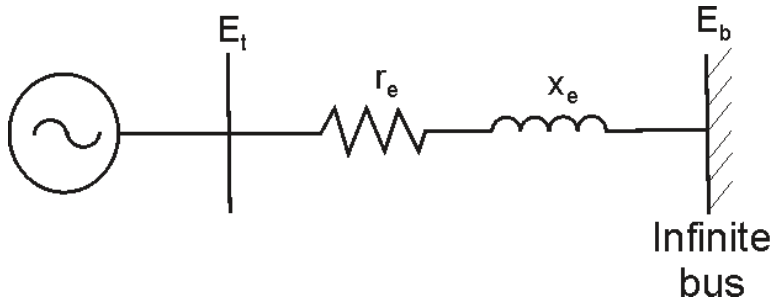

Three flavor of reachability w.r.t a target set $\mathcal{G}$:

1. **Reach Problems** $\mathcal{G}$: set of *goal* states
   → which states can reach $\mathcal{G}$?
   → which states can reach $\mathcal{G}$ and stay forever (c.f. invariance)?
   - E.g.: $\mathcal{G}$ is a neighborhood of a system's desired operating point.

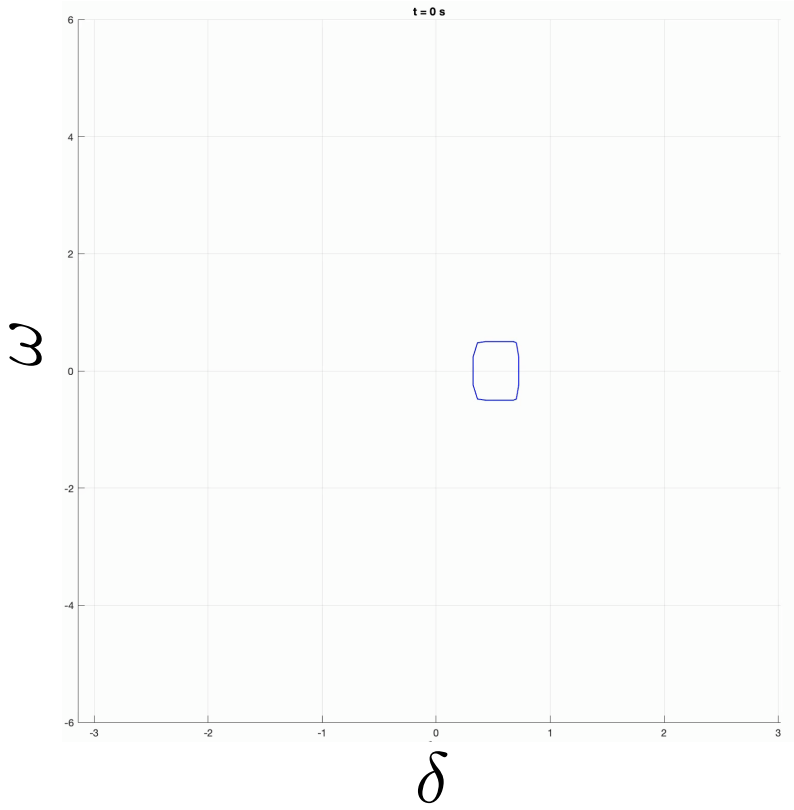
2. **Avoid Problems** $\mathcal{G}$: set of *unsafe* states
   → which states inevitably visit $\mathcal{G}$?
   - E.g.: $\mathcal{G}$ is a set of buses' voltages outside [.95, 1.05] p.u., lines thermal limits.

1. **Reach-avoid problems**: combination of previous

# Example: Transient Stability in Power Systems

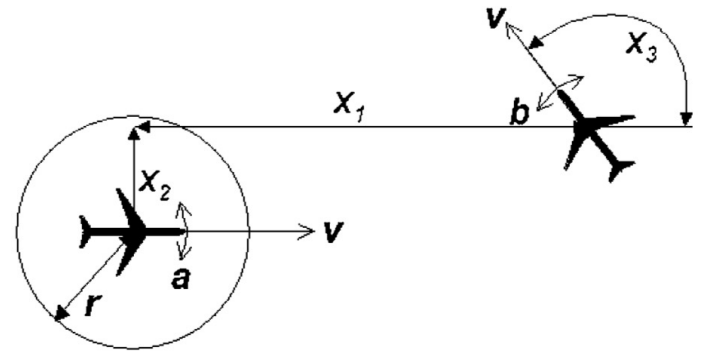



$$\begin{cases} \dot{\delta} = \omega \\ \dot{\omega} = \frac{1}{M}\left(u - D\omega - P_e \sin\delta\right) \\ u \in [u_{\min}, u_{\max}] \end{cases}$$



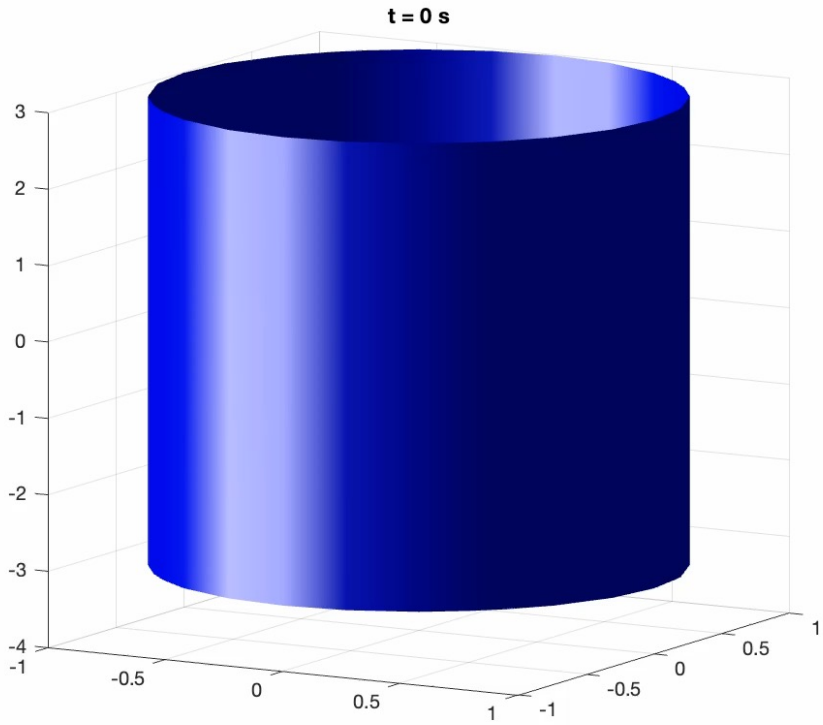

- **Q:** Which states can reach a neighborhood of the stable equilibrium?

# Example: Air Collision Avoidance




evader (player I) pursuer (player II)

$$\begin{cases} \dot{x}_1 = -v + v\cos x_3 + ax_2 \\ \dot{x}_2 = v\sin x_3 - ax_2 \\ \dot{x}_3 = b - a \\ b, a \in [-1, 1] \end{cases}$$

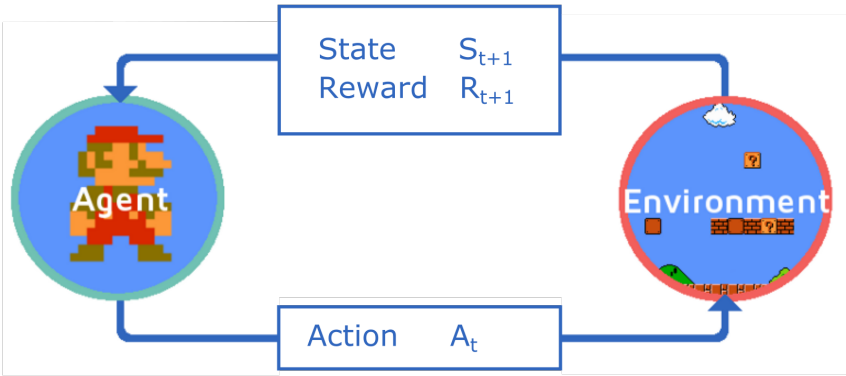


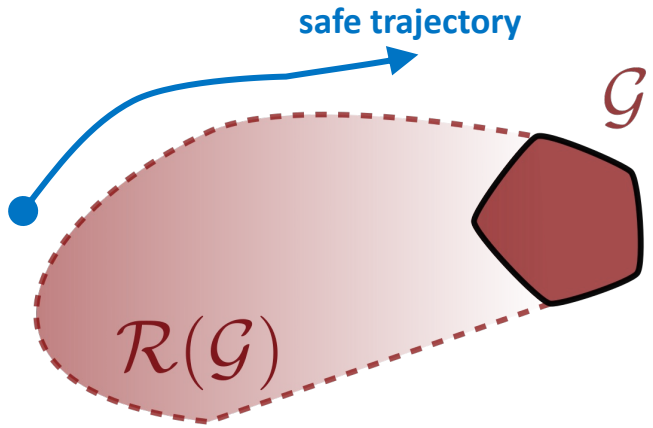

- **Q:** From which states can the evader avoid collision?

# 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

# 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]**

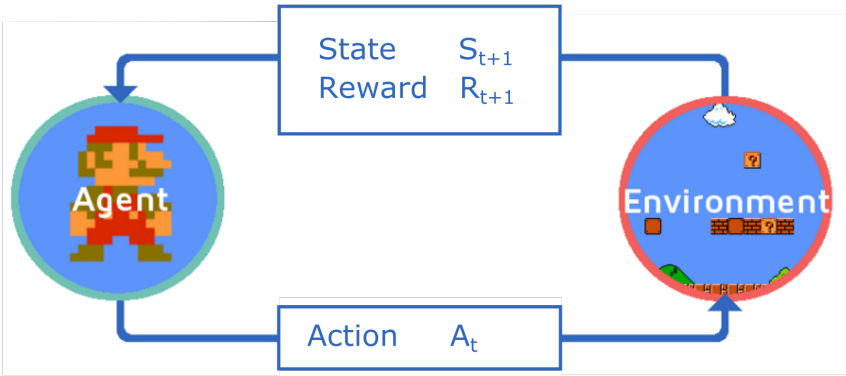
- **Model-free:** Q-Learning-like algorithms, computes function such that $Q_{safe}(s,a) \geq \eta_{thresh} \Rightarrow$ "$safety$"
- **Constraints:** Provides soft/approximate guarantees, depending on discounting factor $\gamma \in (0,1)$
- **Output:** *Converges to maximum CIS as* $\gamma \to 1$

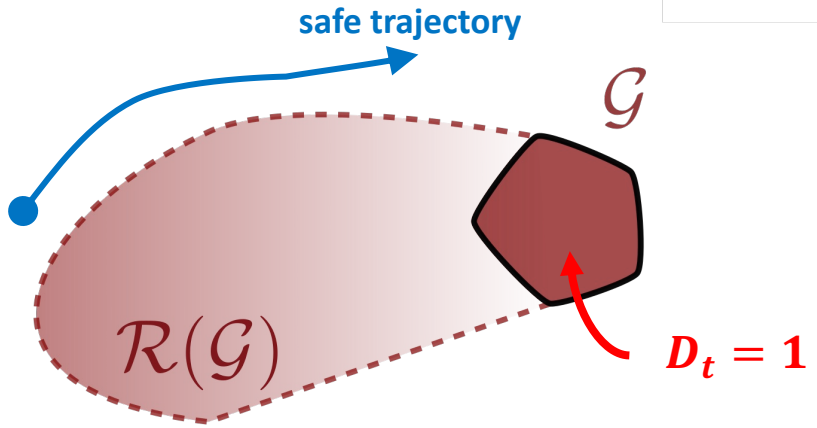
| Method | Model-free | Constraint Type | Set Size |
|---|---|---|---|
| Reachability Theory[1-2] | No | Hard | Maximal |
| Control Barrier Functions[3-4] | No | Hard | Subset |
| Safety Critics[5-7] | Yes | Soft/Approx. | Maximal |
| Ours | Yes | Hard | Maximal and Subsets |

# 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**

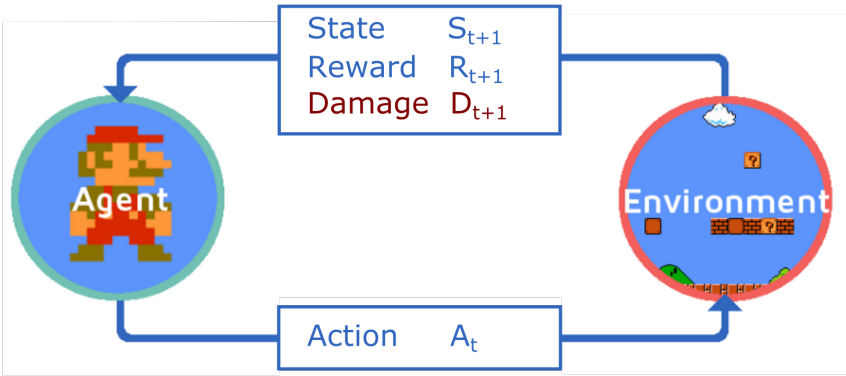
# Outline

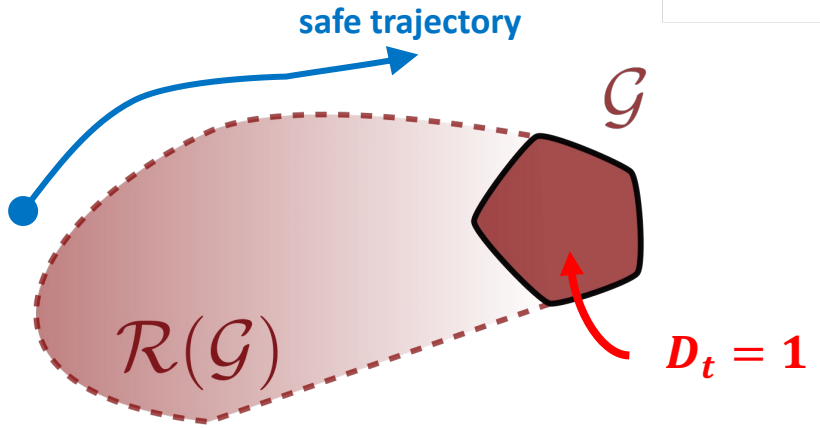
- Problem Setup and Motivation
- Separation Principle for Joint Safety & Optimality
- One-sided Bellman Equations for Continuous States

# Recap: RL 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

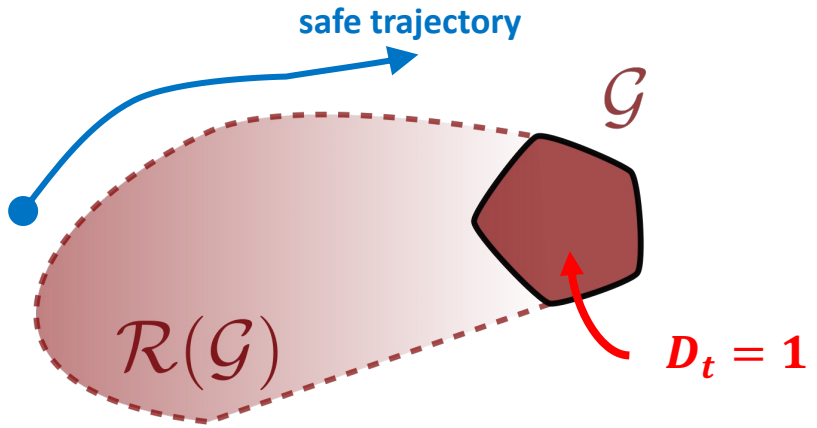
$$S_t \in \mathcal{G} \Leftrightarrow D_t = 1$$

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

# Recap: RL for Safety-Critical Systems

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

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




**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**

# Formulation via hard barrier indicator

Safe RL problem:

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

$$\text{s.t.} \quad D_{t+1} = 0 \ \text{ almost surely } \forall t$$

$\sim$

Equivalent unconstrained formulation:

$$\max_{\pi} \quad \mathbb{E}_{\pi, S_i \sim q}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} + \underbrace{\log(1 - D_{t+1})}_{\begin{matrix} 0 & if\ D_{t+1} = 0 \\ -\infty & if\ D_{t+1} = 1 \end{matrix}}\right]$$

**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 $\pi$,

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

and define the hard-barrier function

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

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

- $B^{\pi}(s,a) \in \{0, -\infty\}$
- Summarizes safety information
  - $B^{\pi}(s,a) = 0$ iff $\pi$ 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 $\pi$:

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

In particular, for optimal $\pi_*$

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

**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:

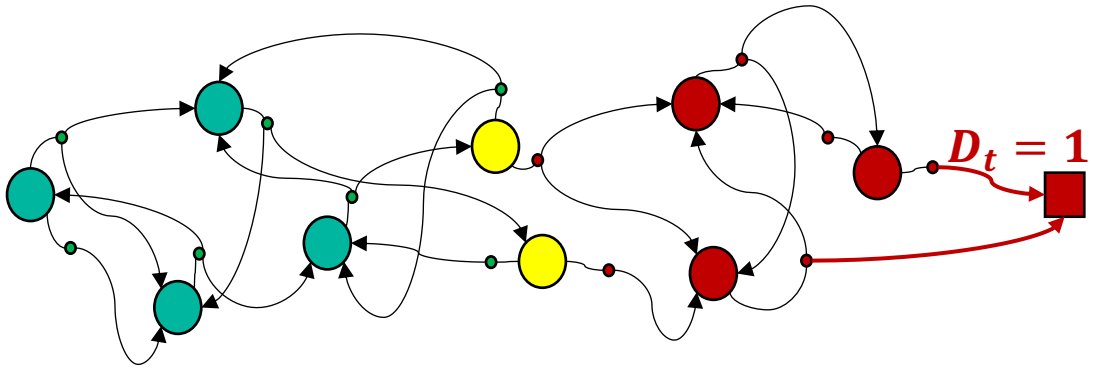
$$B^*(s,a) = \mathbb{E}\left[ -\log(1 - D_{t+1}) + \max_{a'} B^*(S_{t+1}, a') \mid S_0 = s, A_0 = a \right]$$

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

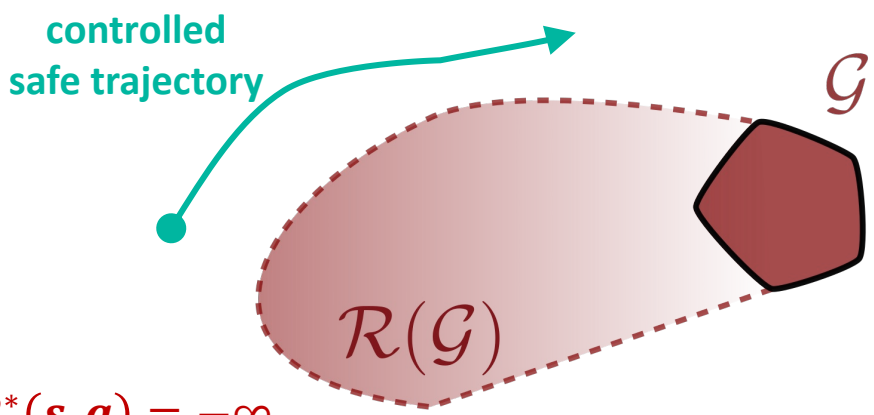
$B^*(s,a) \in \{0, -\infty\}$ summarizes safety information of the entire MDP

- $B^*(s,a) = 0$ if ∃ safe $\pi$ 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**



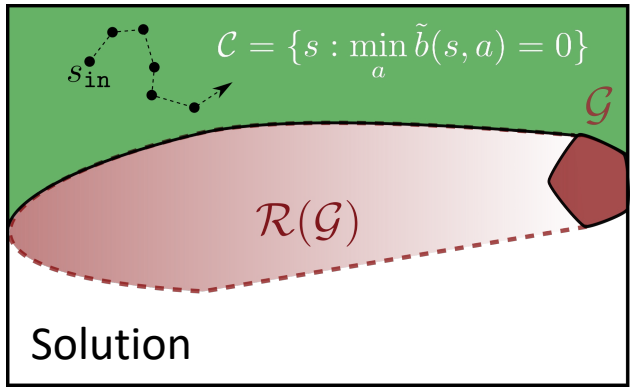

- $V^*(s) = \max_a B^*(s,a) = 0$
- $V^*(s) = \max_a B^*(s,a) = -\infty$

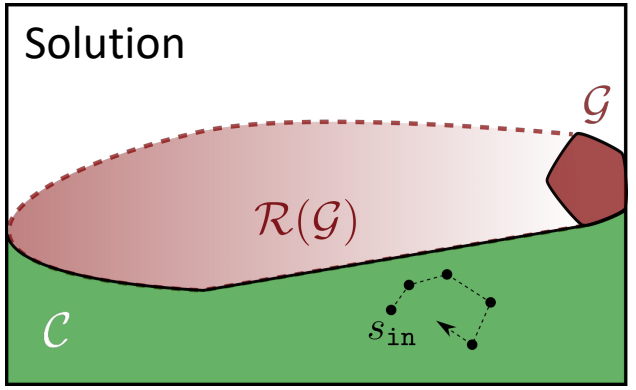
# Properties of Safety Bellman Equation

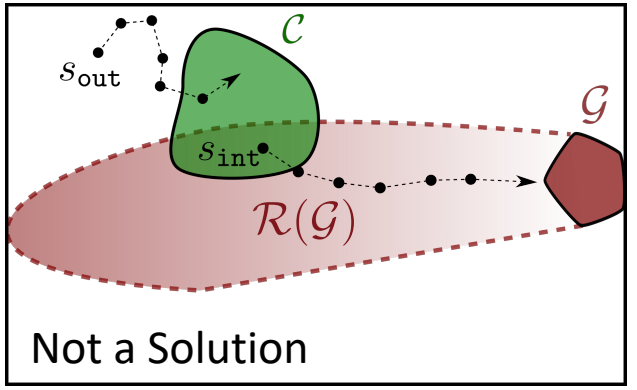
**Understanding the Solutions to the Safety Bellman Equation (SBE):**

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

- SBE can have **multiple solutions,** including $\widetilde{\boldsymbol{B}}(\boldsymbol{s}, \boldsymbol{a}) = -\infty$, for all pairs $(s, a)$
- If the function $\widetilde{\boldsymbol{B}}$ 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 $\pi$



# Outline

- Problem Setup and Motivation
- Separation Principle for Joint Safety & Optimality
- One-sided Bellman Equations for Continuous States

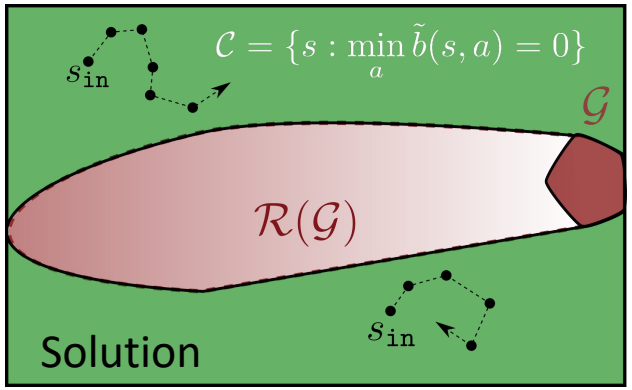
# Recall: Properties of Safety Bellman Equation

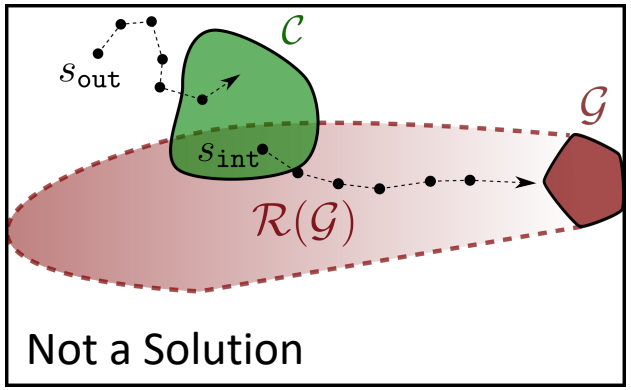
**Understanding the Solutions to the Safety Bellman Equation (SBE):**

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

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

- SBE can have **multiple solutions,** including $\widetilde{\boldsymbol{B}}(\boldsymbol{s}, \boldsymbol{a}) = -\infty$, for all pairs $(s, a)$
- If the function $\widetilde{\boldsymbol{B}}$ 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 $\pi$~~




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

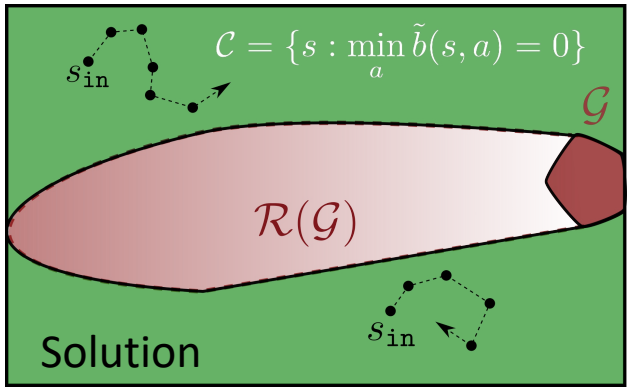
# One-Sided Safety Bellman Equation

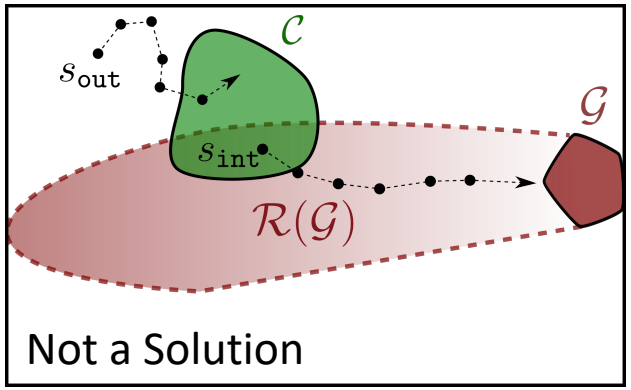
**Theorem** (One-Sided Safety Bellman Equation)

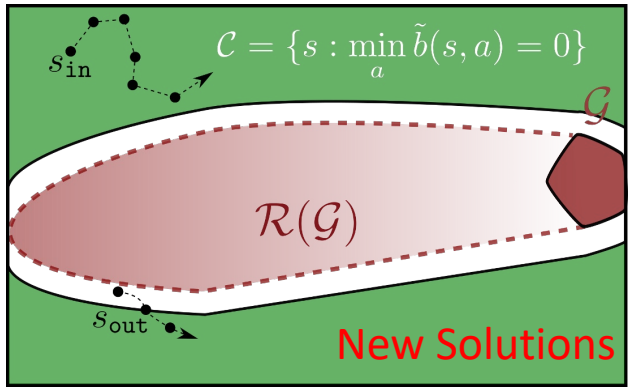
Let $\widetilde{\boldsymbol{B}}(\boldsymbol{s}, \boldsymbol{a})$ be a solution of the following set of inequalities:

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

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



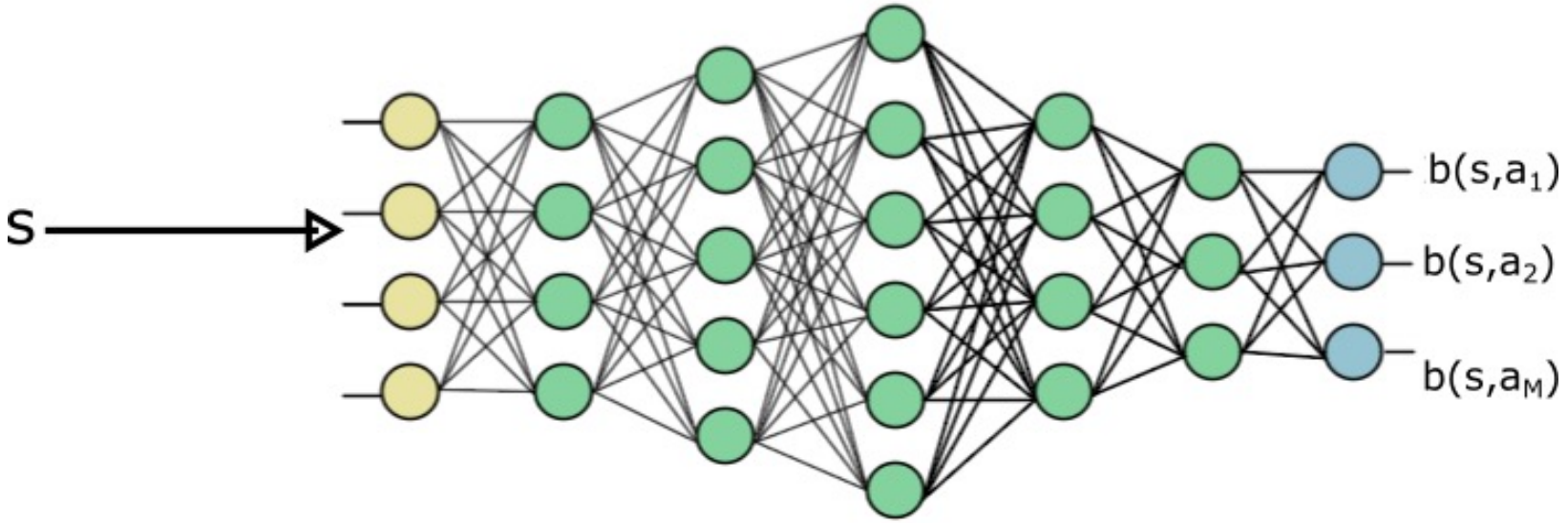



# Learning Solutions to Bellman Inequalities

## Architecture

- akin to Q-Learning



# Learning Solutions to Bellman Inequalities

## Algorithm Summary

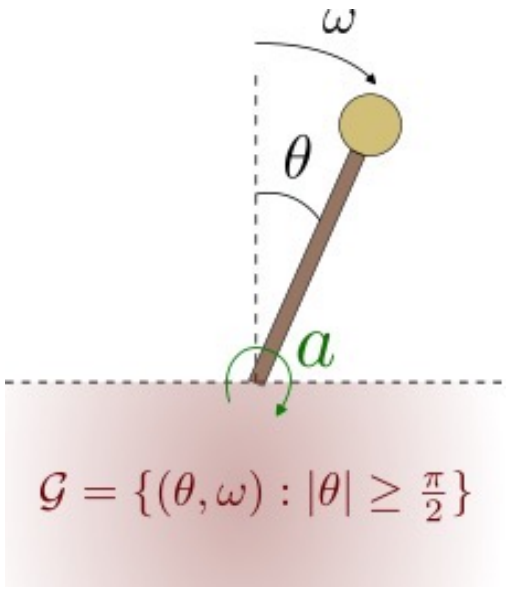
- *Require:*
  - *Axiomatic data* $(\mathrm{s}, \mathrm{a}, \mathrm{d}, \mathrm{s}') \in \mathcal{D}_{safe}$ (dataset of safe transitions)
- *Initialize:*
  - $\hat{b}^{\theta}(s,a) = 0$, where $\hat{b}(s,a) = 1 - e^{B(s,a)}$ (all presumed safe)
- *At each iteration:*
  - Take N episodes starting from $\mathcal{D}_{safe}$
  - Behavioral policy: *uniform safe policy*

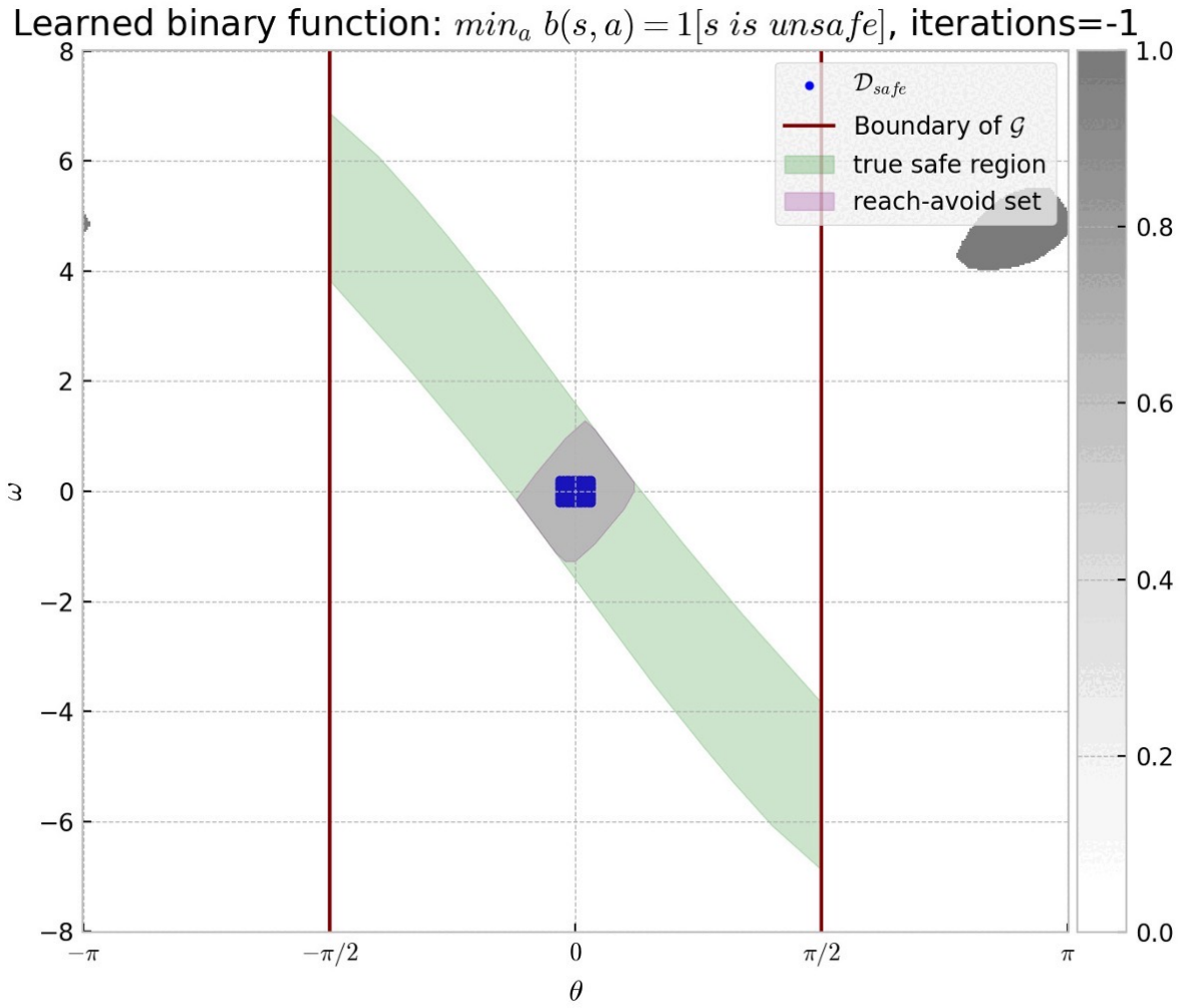
$$\pi^{\theta}(a|s) = \begin{cases} 0 & \text{if } \hat{b}^{\theta}(s,a) = 1 \\ 1/\sum_{a' \in \mathcal{A}} \mathbb{1}\{\hat{b}^{\theta}(s,a') = 0\} & \text{if } \hat{b}^{\theta}(s,a) = 0 \end{cases}$$

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

# Numerical Illustration

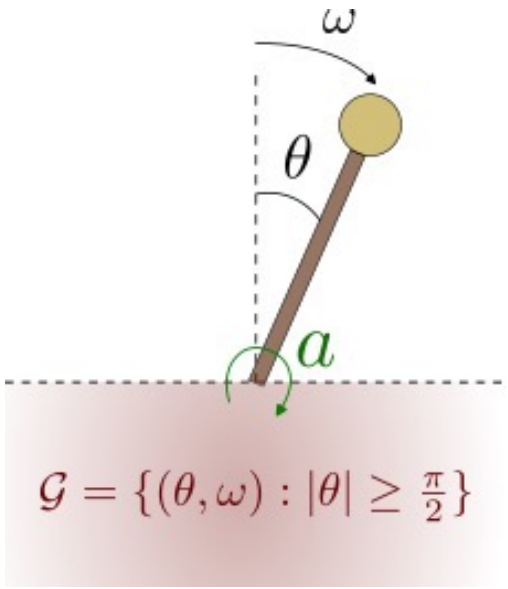
*Control Engineer Favorite's:* Inverted Pendulum

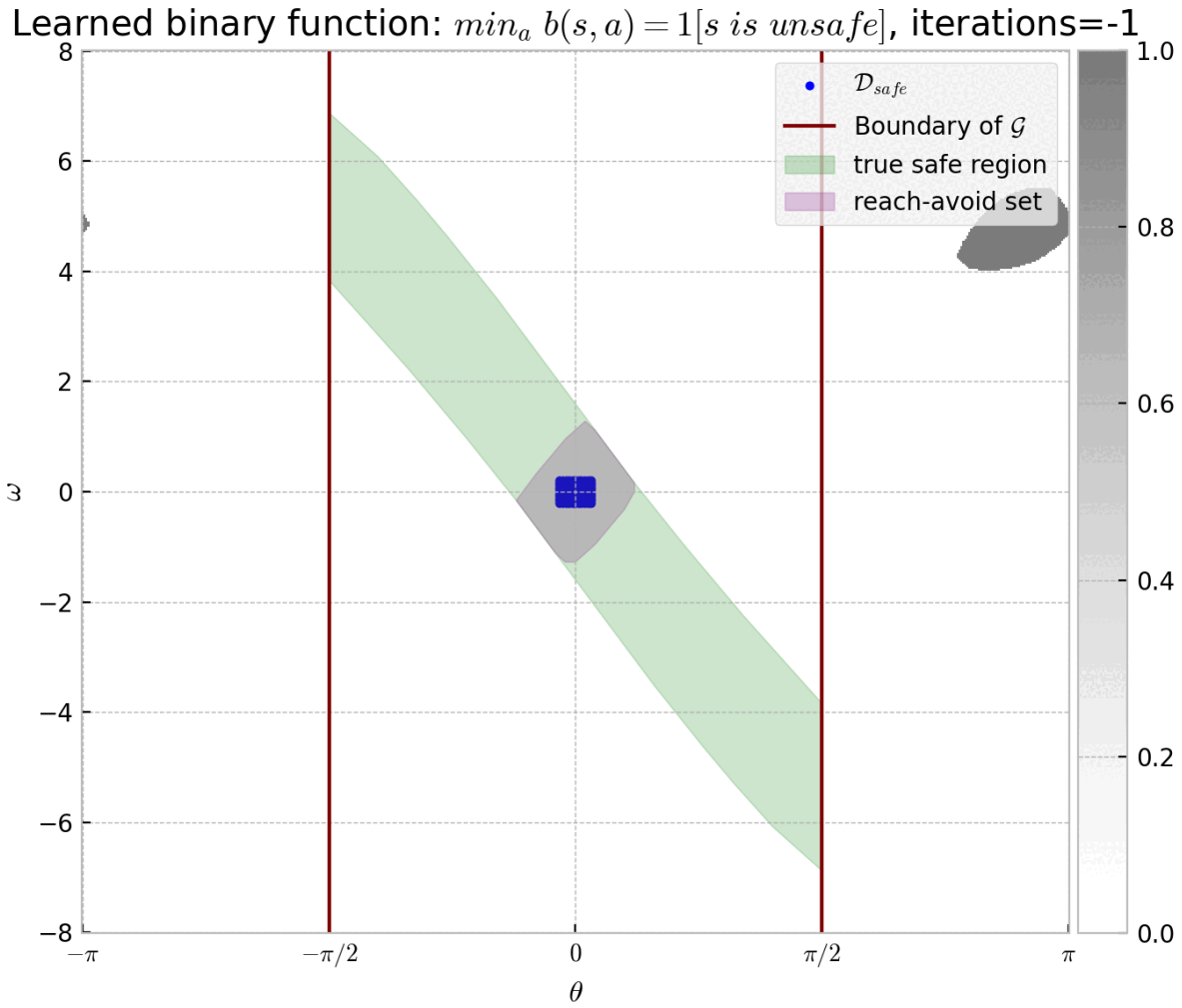


# Numerical Illustration

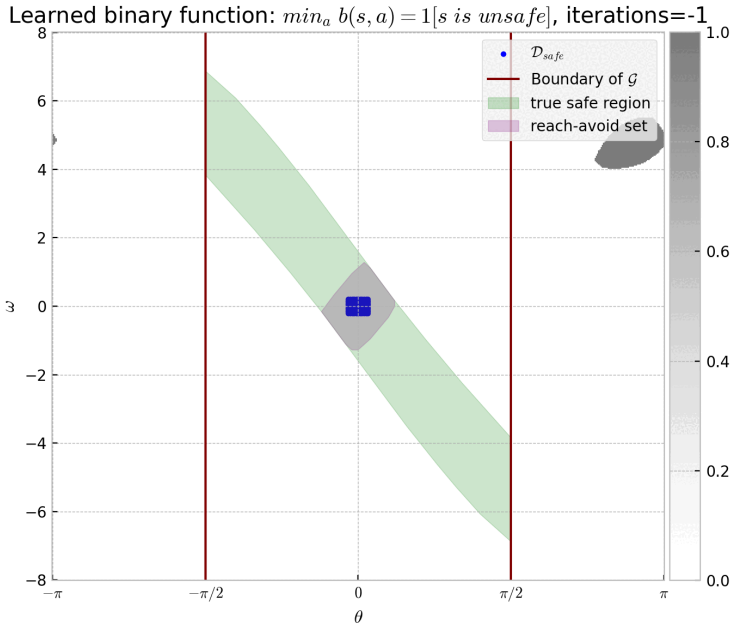
*Control Engineer Favorite's:* Inverted Pendulum

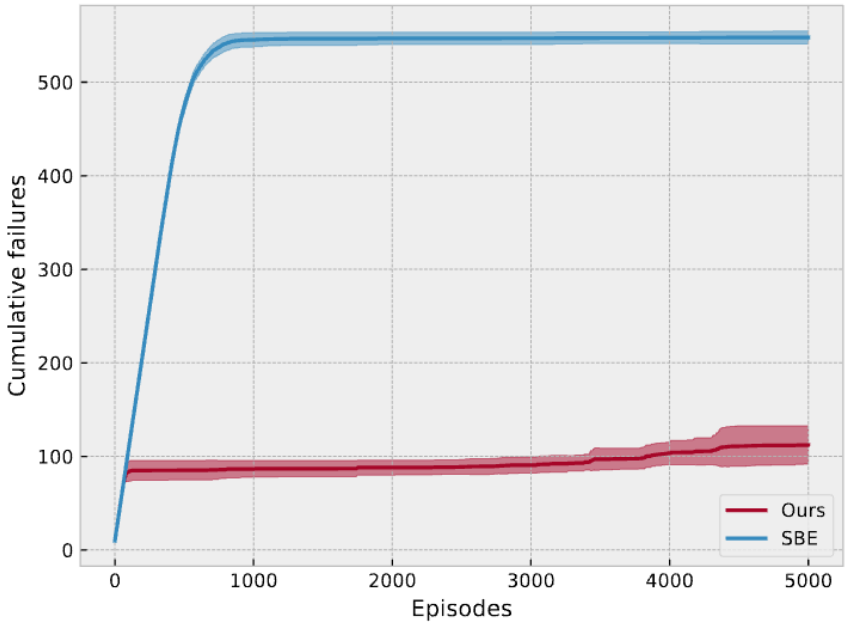


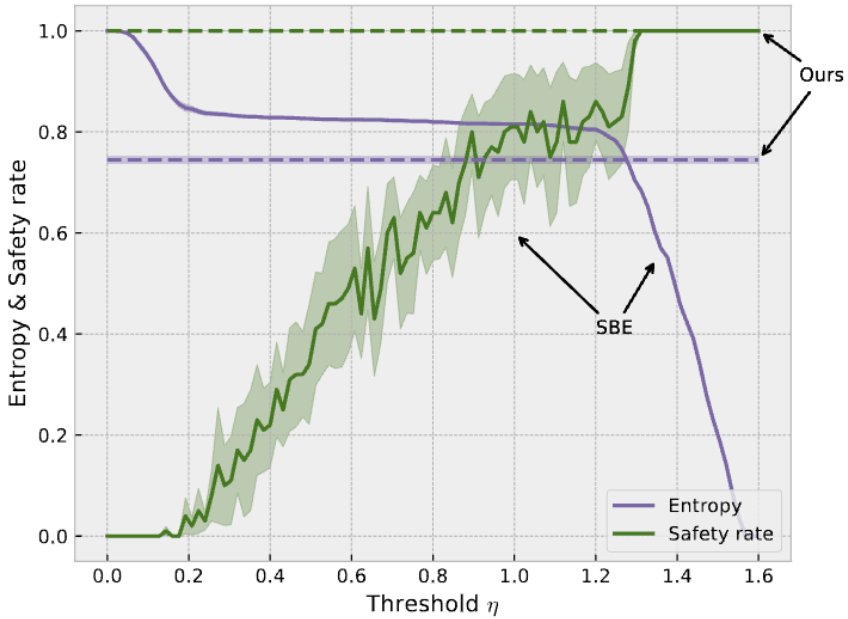



# Numerical Illustration

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

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

[2] T Zheng P You, and E Mallada. Constrained reinforcement learning via dissipative saddle flow dynamics **Asilomar 2023**

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

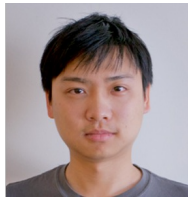
[5] Castellano, Min, Bazerque M, Correct-by-design Safety Critics Using Non-contractive Bellman Operators, **submitted**


Tianqi Zheng

amazon

Agustin Castellano

Johns Hopkins University



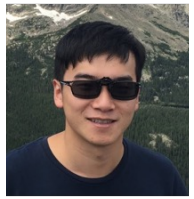
Hancheng Min

Penn


Enrique Mallada

mallada@jhu.edu

http://mallada.ece.jhu.edu

Pengcheng You

北京大学 Peking University

Juan Bazerque

University of Pittsburgh