Reinforcement Learning with Almost Sure Constraints

Enrique Mallada

JOHNS HOPKINS UNIVERSITY



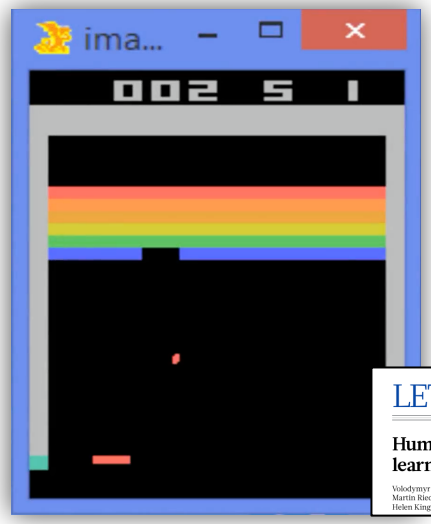
A. Castellano H. Min J. Bazerque

Workshop on Control and Machine Learning: Challenges and Progress

Oct 12, 2023

A World of Success Stories

2017 Google DeepMind's DQN



LETTER

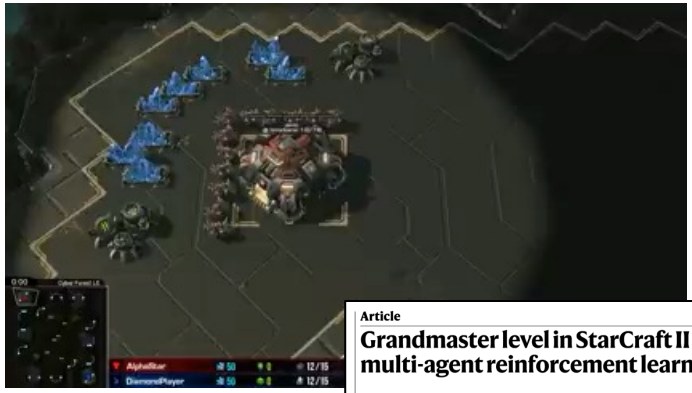
doi:10.1038/nature14236

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

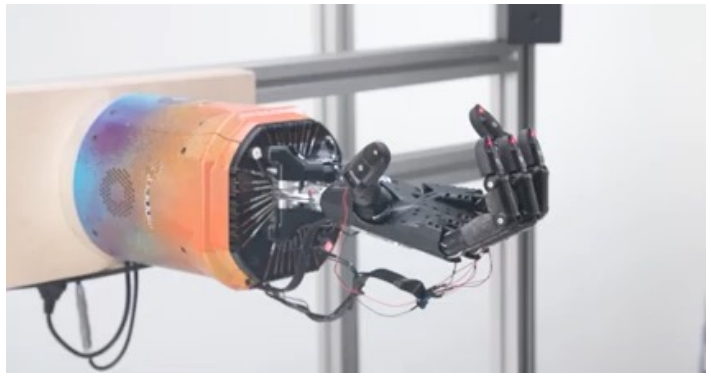
Received: 30 August 2019

Accepted: 10 October 2019

Published online: 30 October 2019

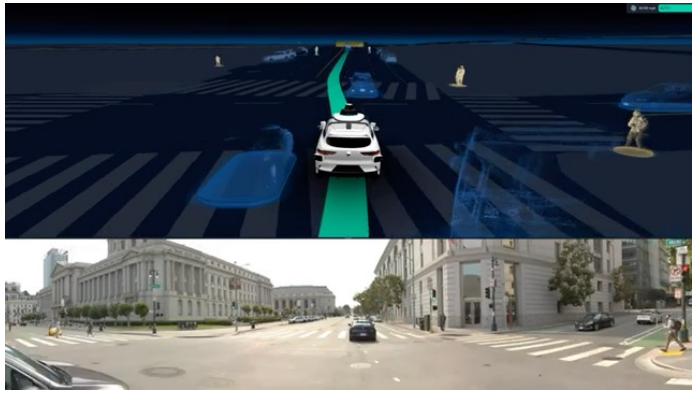
Oriol Vinyals[1,2*], Igor Babuschkin[1,2], Wojciech M. Czarnecki[1,2], Michaël Mathieu[1,2], Andrew Dudzik[1,2], Junyoung Chung[1,2], David H. Choi[1,2], Richard Powell[1,2], Timo Ewalds[1,2], Petko Georgiev[1,2], Junhyuk Oh[1,2], Dan Horgan[1,2], Manuel Kroiss[1,2], Ivo Danihelka[1,2], Aja Huang[1,2], 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.

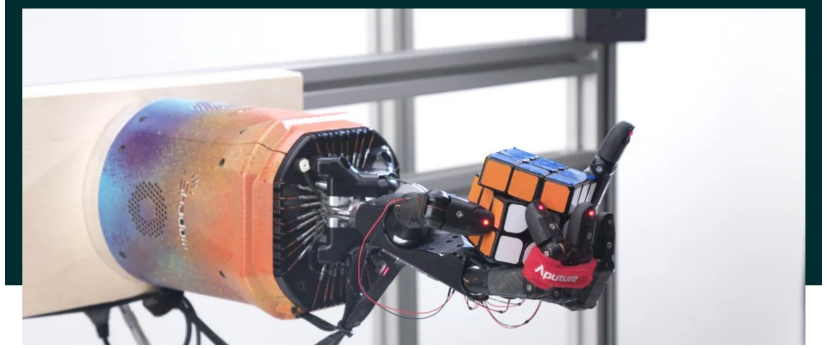
AARIAN MARSHALL BUSINESS 12.07.2020 04:06 PM

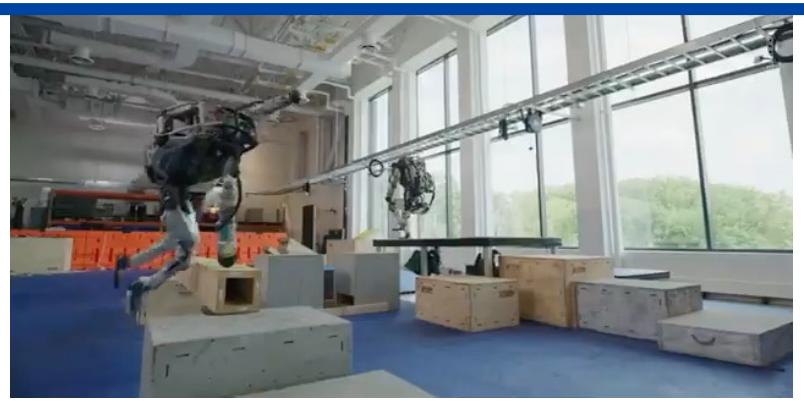
Uber Gives Up on the Self-Driving Dream

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

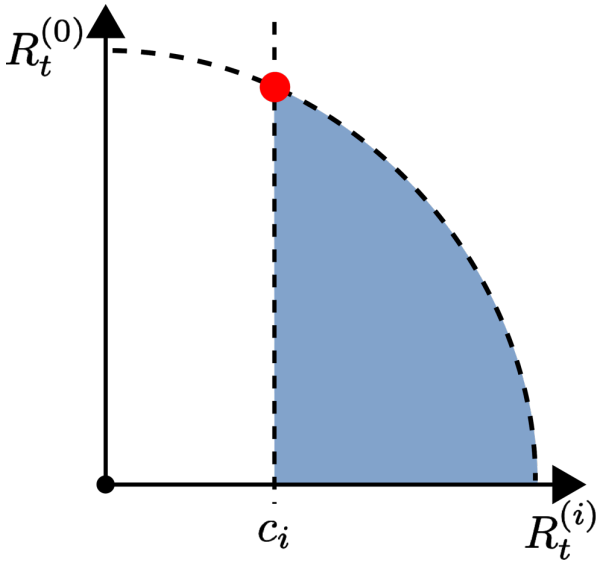
woman did not recognize that pedestrians jaywalk

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

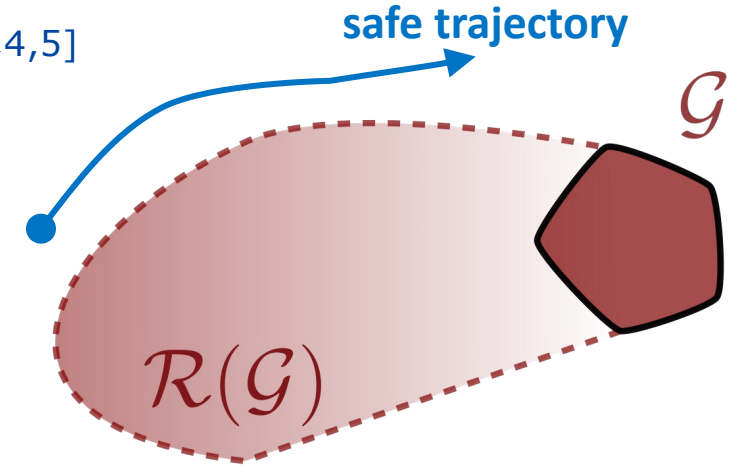
- Physical systems must meet multiple objectives
  - Need to trade off between the different goals
  - Constrained RL allows to explore the Pareto Front [1,2]

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



- 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

Safety-critical Constraints in Dynamical Systems

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

Safety-critical Constraints in Dynamical Systems

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	Size	Control Invariant?
Reachability Theory[1-2]	No	Hard	Maximum CIS	Yes
Control Barrier Functions[3-4]	No	Hard	Subset of M-CIS	Yes
Safety Critics[5-7]	Yes	Soft/Approx.	Maximum CIS as $\gamma \to 1$	No

Safety-critical Constraints in Dynamical Systems

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	Size	Control Invariant?
Reachability Theory[1-2]	No	Hard	Maximum CIS	Yes
Control Barrier Functions[3-4]	No	Hard	Subset of M-CIS	Yes
Safety Critics[5-7]	Yes	Soft/Approx.	Maximum CIS as $\gamma \to 1$	No
Ours	Yes	Hard	M-CIS and Subsets	Yes

[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)]

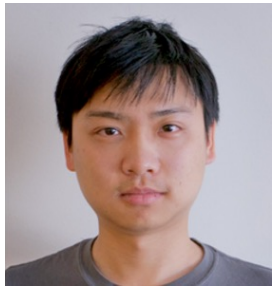
Learning to Act Safely with Limited Exposure and Almost Sure Certainty

Agustin Castellano, Hancheng Min, Juan Bazerque, Enrique Mallada

arXiv > eess > arXiv:2105.08748

Agustin Castellano

JOHNS HOPKINS UNIVERSITY



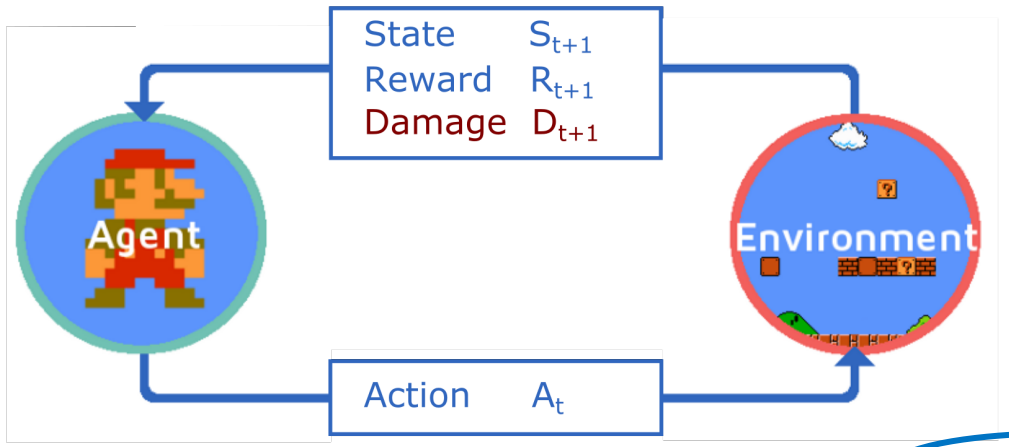
Hancheng Min

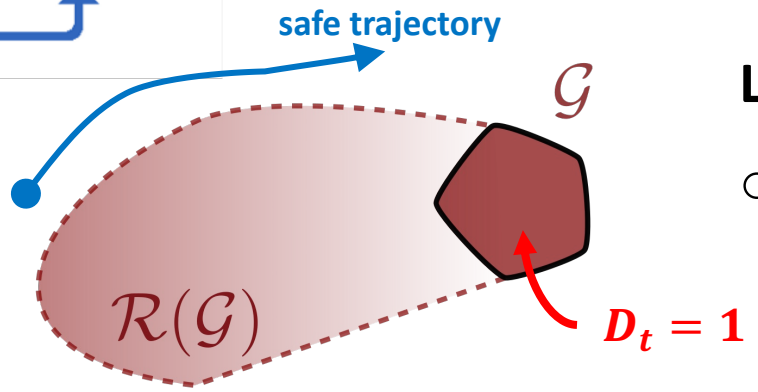
JOHNS HOPKINS UNIVERSITY

Juan Bazerque

University of Pittsburgh

Learning for Safety-critical Sequential Decision Making





Requirements:

High Priority -> Safety

- Limited Failures/Mistakes
- Hard Constraints/ A.S. Guarantees

Lower Priority -> Accuracy

- Optimality of the policy

Methodology:

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

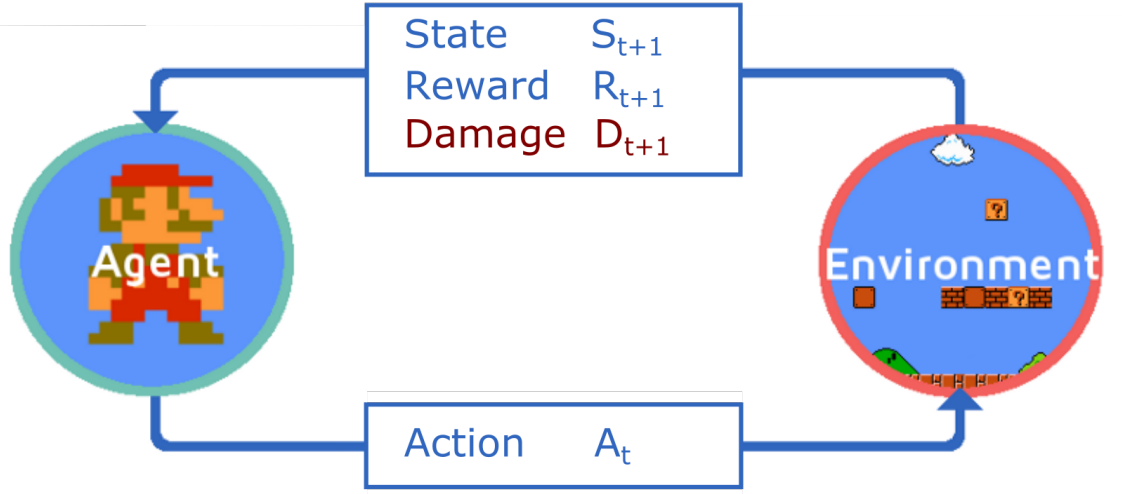
$$S_t \in \mathcal{G} \Leftrightarrow D_t = 1$$

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

Recap: RL 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
- Constraints not given a priori: Need to learn from experience!
- **Notice:** Model free ➔ Constraint violations are inevitable

Outline

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

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:

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

$$\begin{matrix} 0 & if\ D_{t+1} = 0 \\ -\infty & if\ D_{t+1} = 1 \end{matrix}$$

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,

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

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 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:

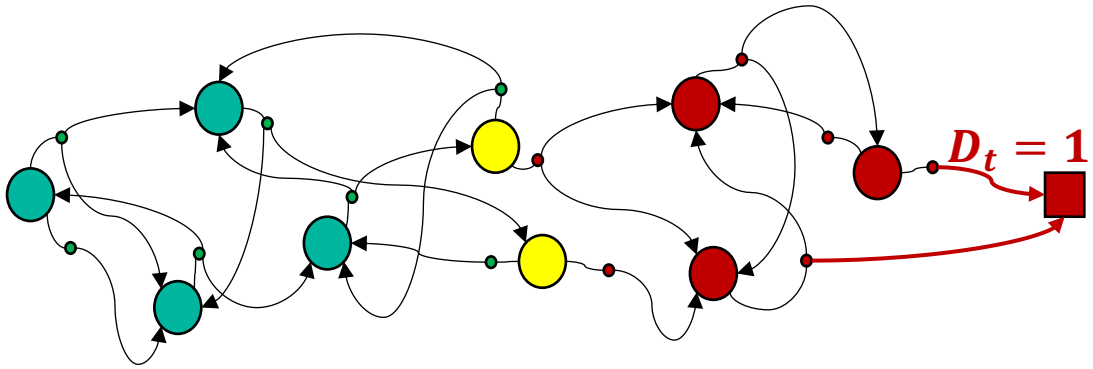
$$B^*(s,a) = \mathbb{E}\left[-\log(1 - D_{t+1}) + \max_{a'} B^*(S_{t+1}, a') \mid S_0 = s, A_0 = a \right]$$

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

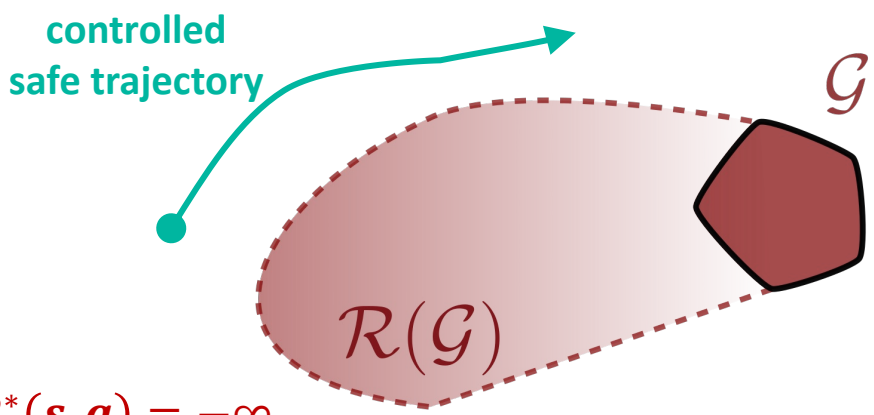
$B^*(s,a) \in \{0, -\infty\}$ summarizes safety information of the entire MDP

- $B^*(s,a) = 0$ if ∃ safe 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

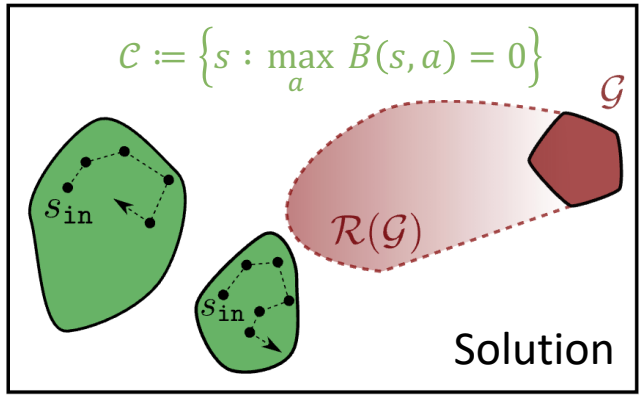
Theorem (Safety Bellman Equation for B^*)

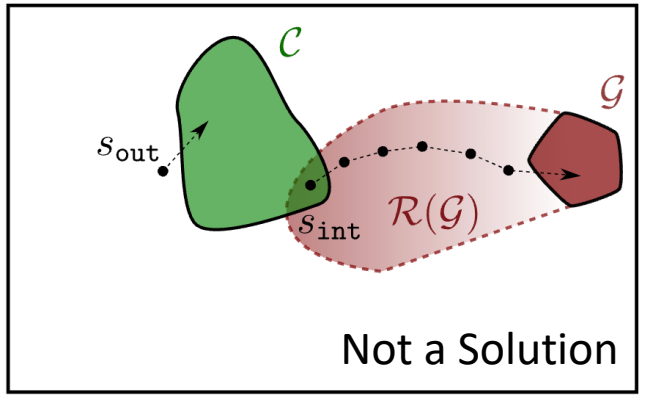
Let $B^*(s,a) := \max_\pi B^\pi(s,a)$, then the following holds:

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

Understanding 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 $\mathcal{C} := \left\{ s : \max_a \tilde{B}(s,a) = 0 \right\}$ 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^*
Initialize $B^{(0)}(s, a) = 0, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$
for $t = 0, 1, \cdots$ **do**
  Draw $(s_t, a_t) \sim \text{Unif}(\{(s, a) : B^{(t)}(s, a) \neq -\infty\})$
  Sample transition (s_t, a_t, s'_t, d_t) according to $P(S_1 = s'_t, D_1 = d_t | S_0 = s_t, A_0 = a_t)$
  $B^{(t+1)} \leftarrow$ `barrier_update`$(B^{(t)}, s_t, a_t, s'_t, d_t)$
end

Initially, all (s, a)-pairs are "safe"

Draw (s, a)-pair uniformly among those considered to be "safe" at time t

Update barrier function

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

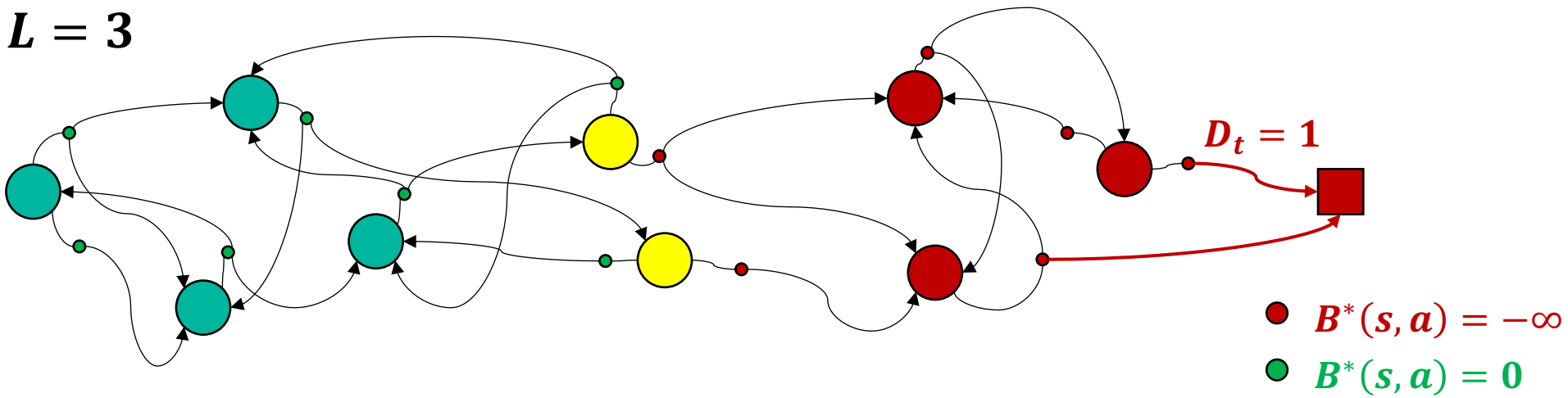
$$\mu = min\{p(s', d|s, a): p(s', d|s, a) \neq 0\}$$

- 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

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



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

- **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

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

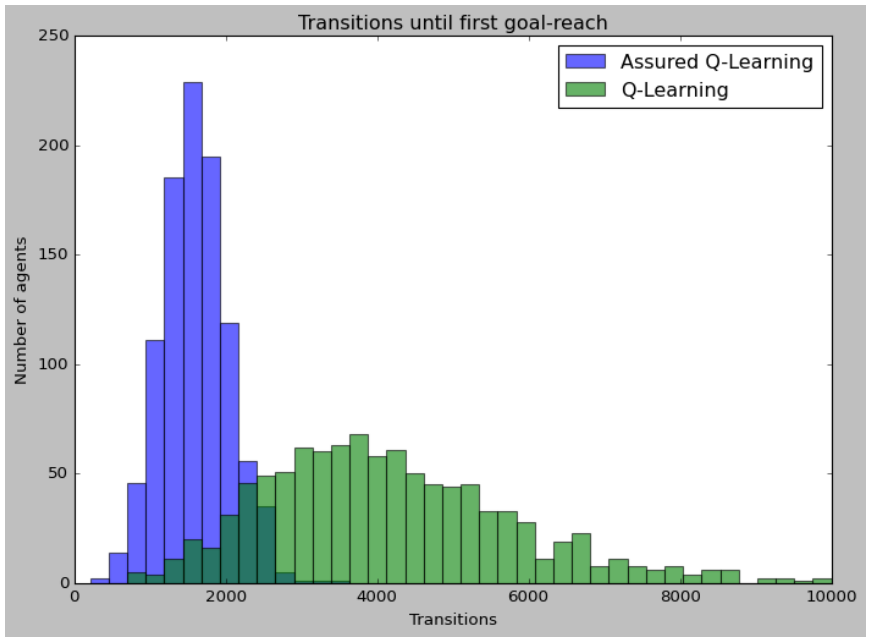
samples (**Trimming the MDP by learning the barrier**)

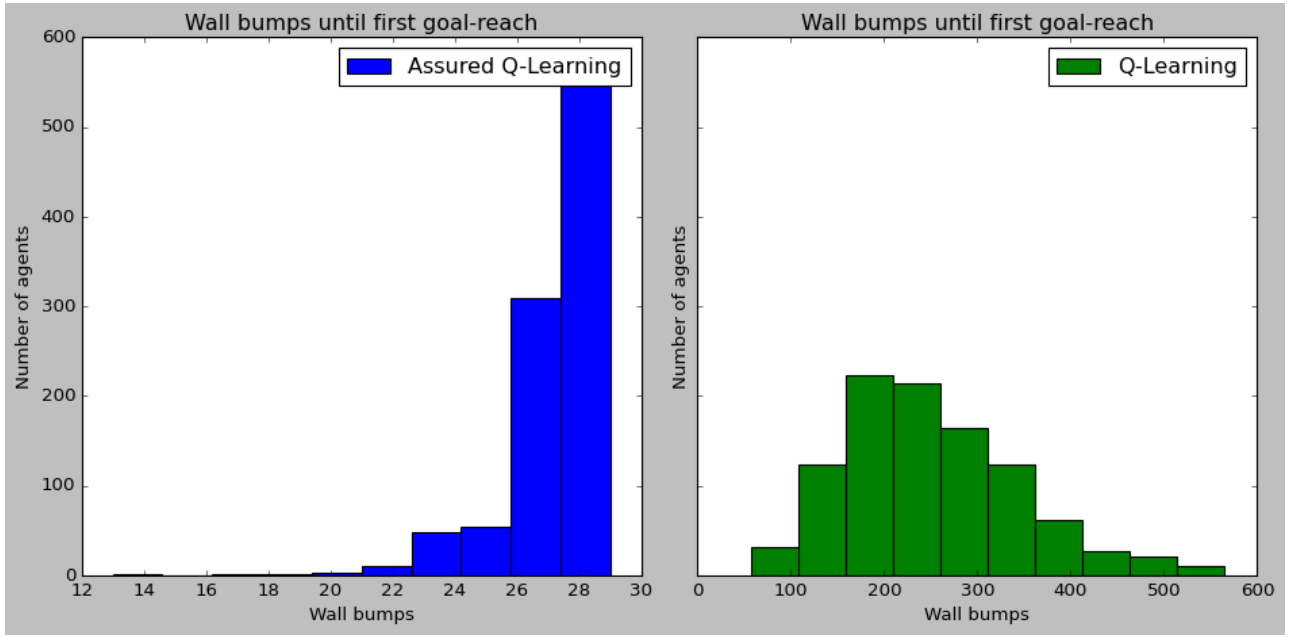
Numerical Experiments

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

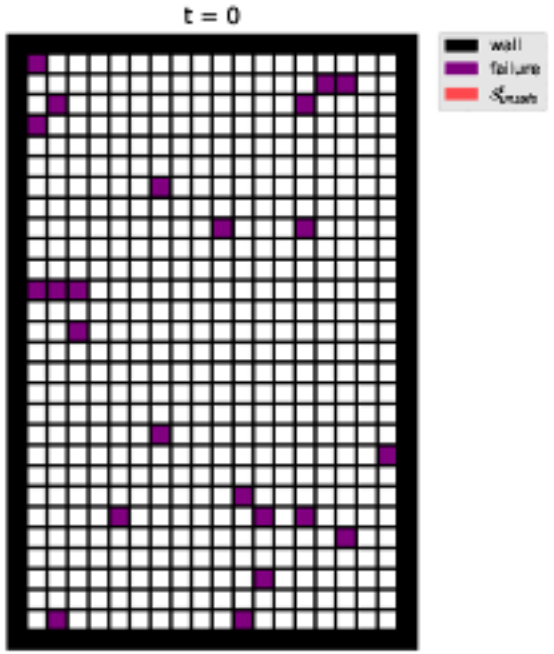
Numerical Experiments II

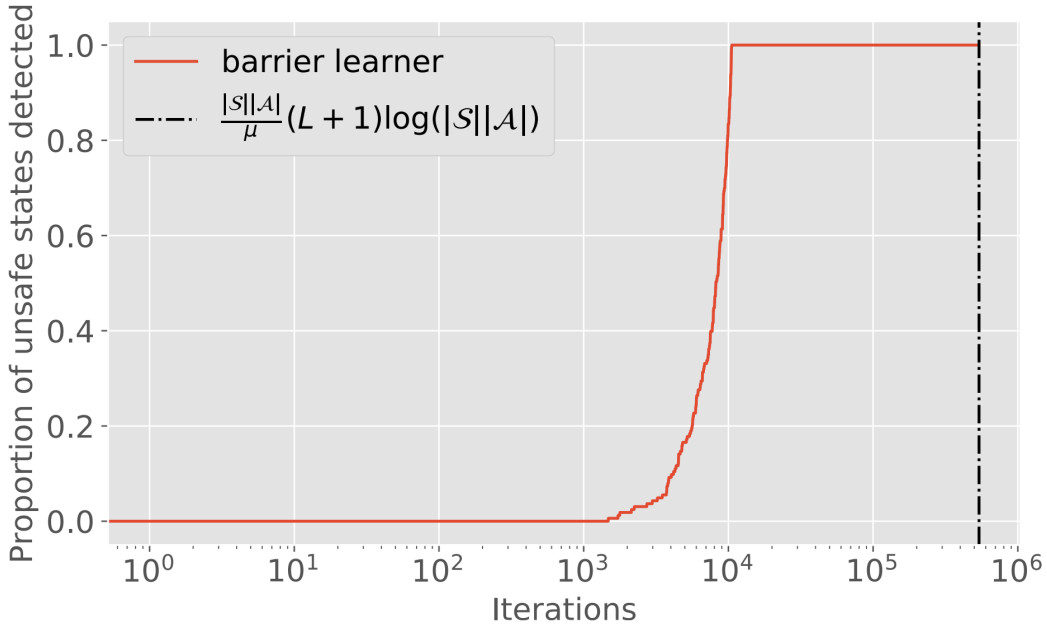
Setup: Rectangular grid, stepping into **holes** gives damage $D_t = 1$.

Actions $A = \{up, down, left, right\}$.

With every action, small probability to move to a random adjacent state.

Result: Barrier-learner identifies **all** the state space as unsafe.





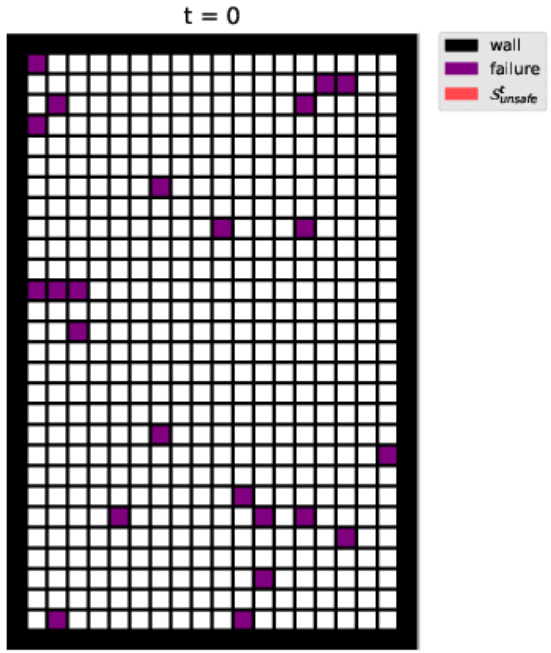
Numerical Experiments II

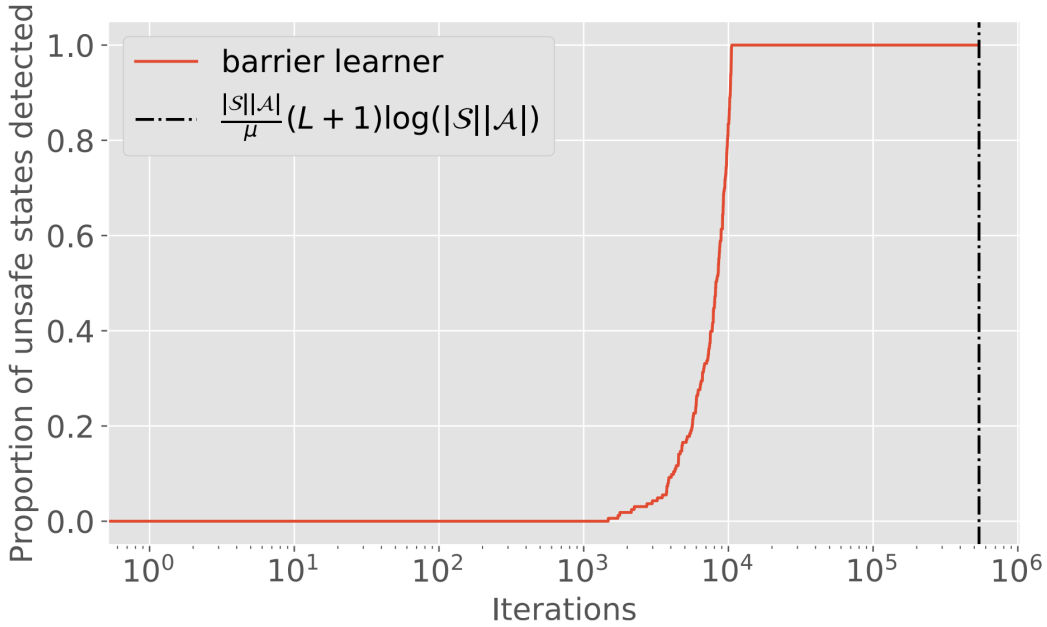
Setup: Rectangular grid, stepping into **holes** gives damage $D_t = 1$.

Actions $A = \{up, down, left, right\}$.

With every action, small probability to move to a random adjacent state.

Result: Barrier-learner identifies **all** the state space as unsafe.





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

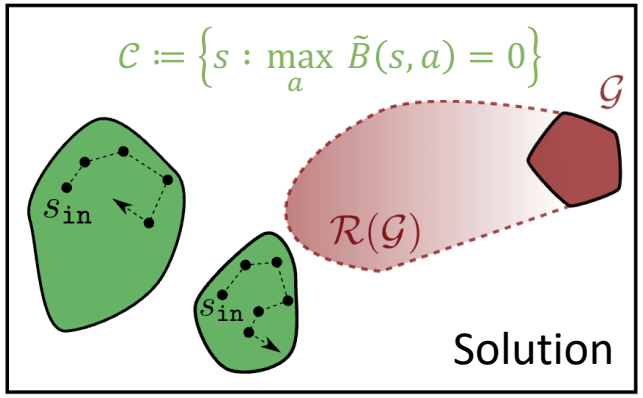
Theorem (Safety Bellman Equation for B^*)

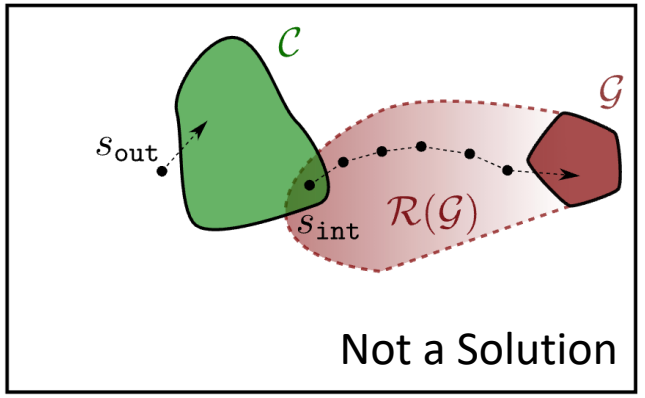
Let $B^*(s,a) := \max_\pi B^\pi(s,a)$, then the following holds:

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

Understanding 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 $\mathcal{C} := \left\{s : \max_a \tilde{B}(s,a) = 0\right\}$ 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 $\mathcal{R}(\mathcal{G})$

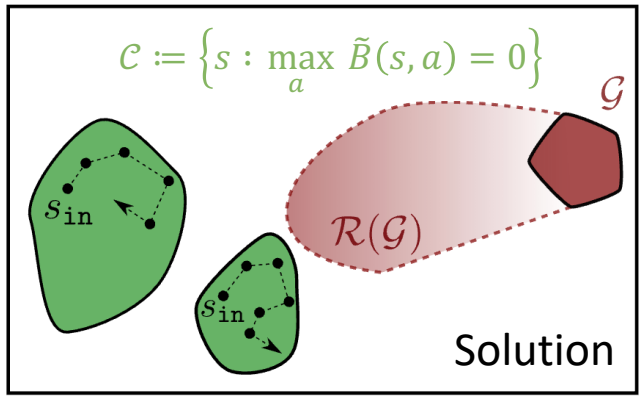
One-Sided Safety Bellman Equation

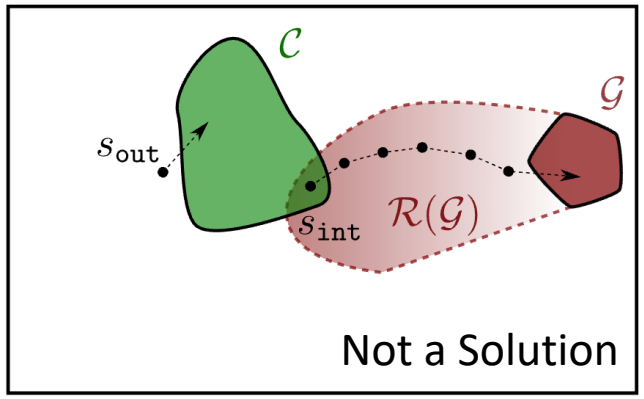
Theorem (One-Sided Safety Bellman Equation)

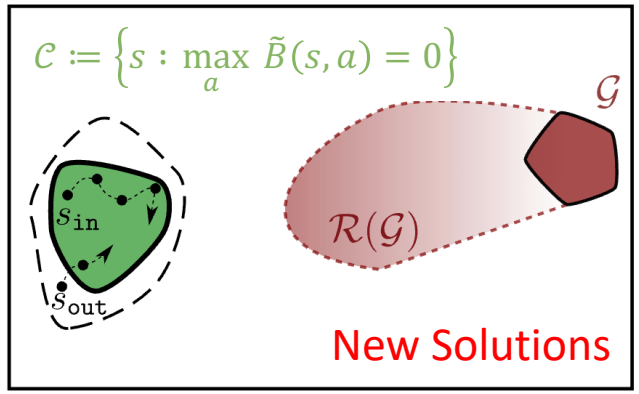
Let $\tilde{B}(s, a)$ be a solution of the following set of inequalities:

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

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







Learning CIS Using Deep Neural Nets

Algorithm Summary

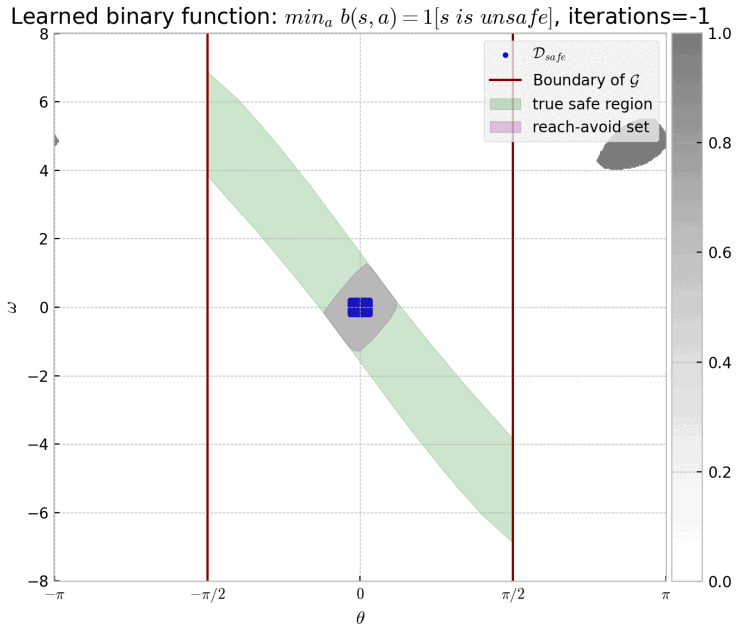
- Uses *axiomatic data* $(\mathrm{s}, \mathrm{a}, \mathrm{d}, \mathrm{s}') \in \mathcal{D}_{safe}$ known to be safe
- *Initialize* $\hat{b}^{\theta}(s,a) = 0$, where $\hat{b}(s,a) = 1 - e^{B(s,a)}$ (all presumed safe)
- *At each iteration*, take N episodes starting from $\mathcal{D}_{safe}$
  - Behavioral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

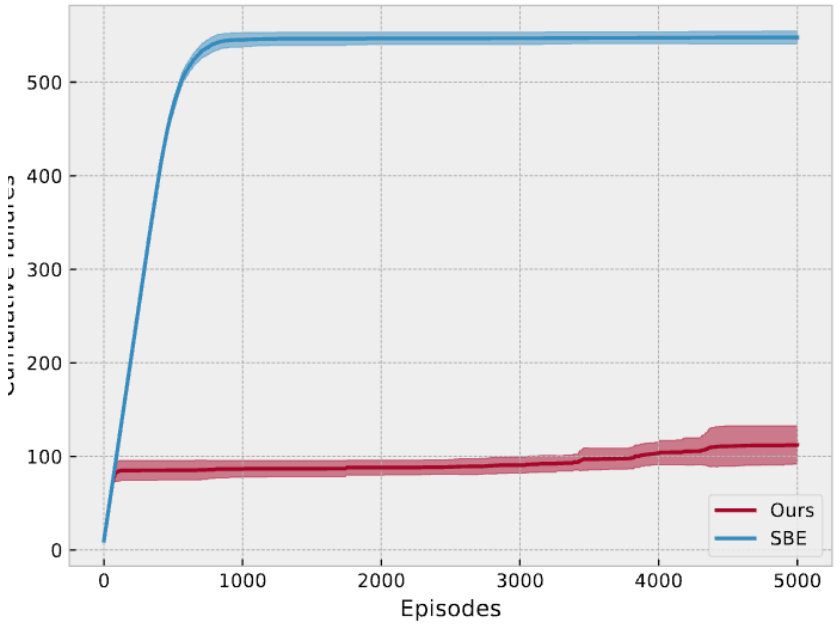
$$\pi^{\theta}(a|s) = \begin{cases} 0 & \text{if } \hat{b}^{\theta}(s,a) = 1 \\ 1/\sum_{a' \in \mathcal{A}} \mathbb{1}\{\hat{b}^{\theta}(s,a') = 0\} & \text{if } \hat{b}^{\theta}(s,a) = 0 \end{cases}$$

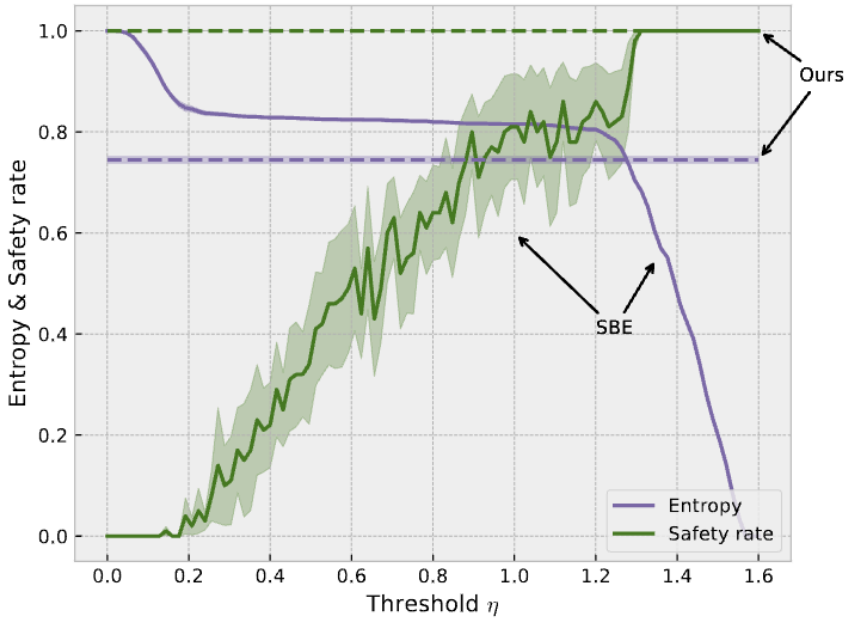
  - Train NN using SGD *until fully fitting the data*
  - Start a new iteration, and repeat

Numerical Illustration

Control Engineer Favorite's: Inverted Pendulum







SBE = Fisac's '19 Safety Critic

Summary and future work

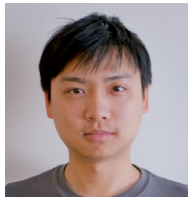
- Reinforcement Learning for Safety-Critical Systems
- Treat constraints separately or in parallel (Barrier Learner)
- Can **characterize** all feasible policies ($D_t \equiv 0$) with **finite mistakes**
- Requires learning using Bellman equations with non-unique solutions
- **Takeaways:**
  - **Learning 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**

Agustin Castellano
JOHNS HOPKINS UNIVERSITY



Hancheng Min
Penn University of Pennsylvania

Enrique Mallada
mallada@jhu.edu
http://mallada.ece.jhu.edu

Juan Bazerque
University of Pittsburgh