Reinforcement Learning with Almost Sure Constraints

Enrique Mallada

JOHNS HOPKINS UNIVERSITY

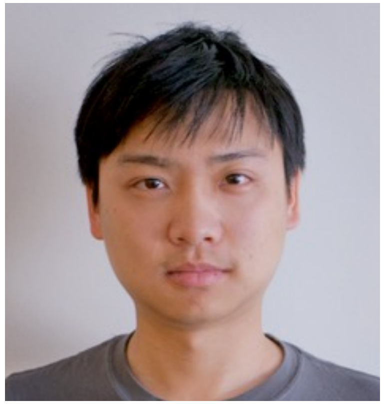
NSF TRIPODS PI Meeting

November 3, 2021

Acknowledgements

Agustin Castellano

JOHNS HOPKINS UNIVERSITY



Hancheng Min

JOHNS HOPKINS UNIVERSITY

Juan Bazerque

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

NSF

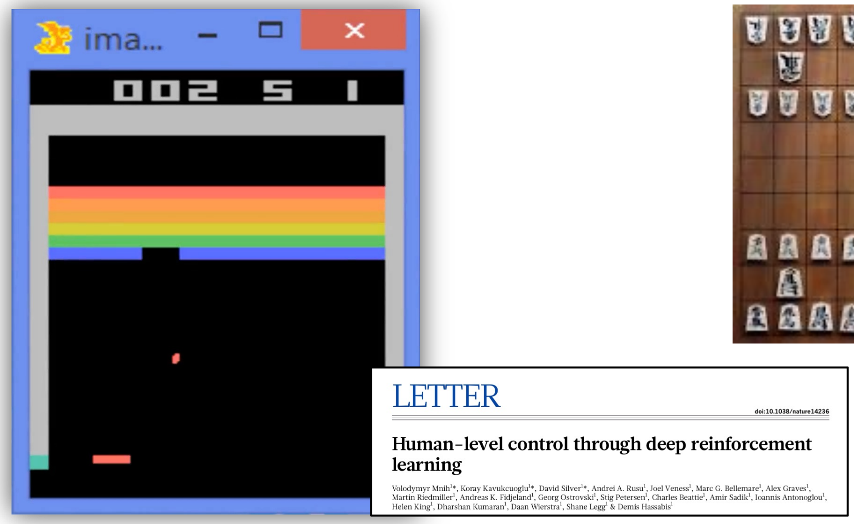
DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

SCIENCE TO SHAPE THE FUTURE OF THE ARMY · ARMY RESEARCH OFFICE · 1951

JOHNS HOPKINS ENVIRONMENT, ENERGY, SUSTAINABILITY & HEALTH INSTITUTE

A World of Success Stories

2017 Google DeepMind's DQN

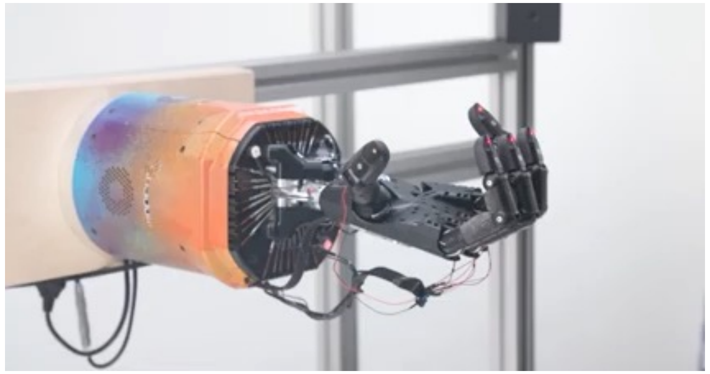


2017 AlphaZero – Chess, Shogi, Go

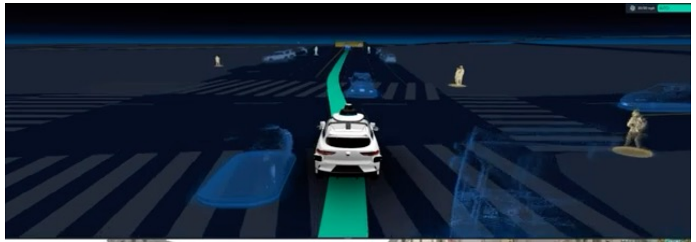
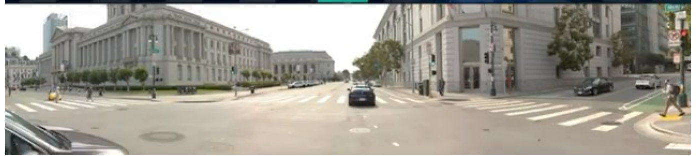
2019 AlphaStar – Starcraft II

Boston Dynamics

OpenAI – Rubik's Cube



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.

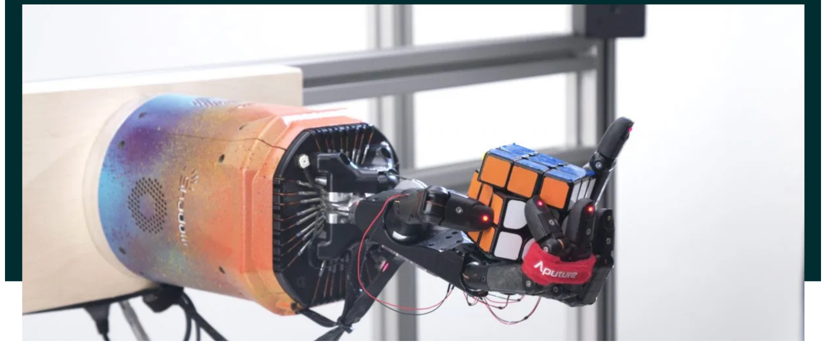
AARIAN MARSHALL BUSINESS 12.07.2020 04:06 PM

Uber Gives Up on the Self-Driving Dream

The ride-hail giant invested more than $1 billion in autonomous vehicles. Now it's selling the unit to Aurora, which makes self-driving tech.

OpenAI disbands its robotics research team

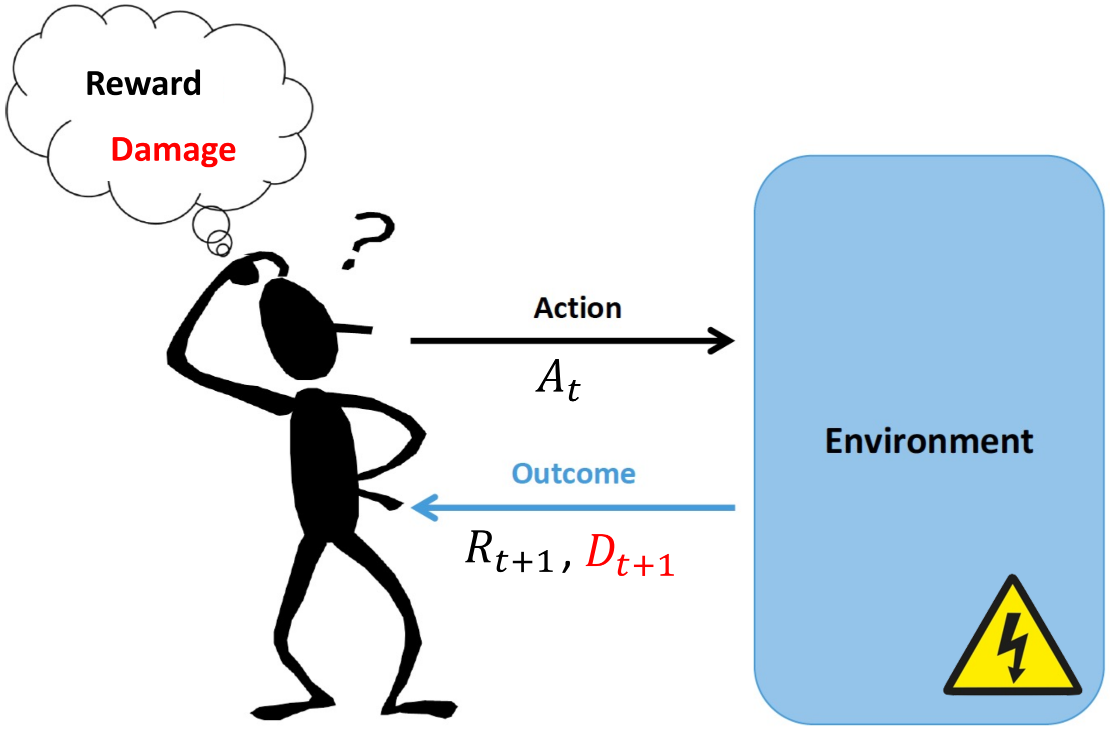
Kyle Wiggers @Kyle_L_Wiggers July 16, 2021 11:24 AM



Self-driving Uber car that hit and killed 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.

Safety Critical Sequential Decision Making



Requirements:

High Priority -> Safety

- Sequential / Online / Real-time
- Limited Failures/Mistakes
- High-probability (or A.S.) Guarantees

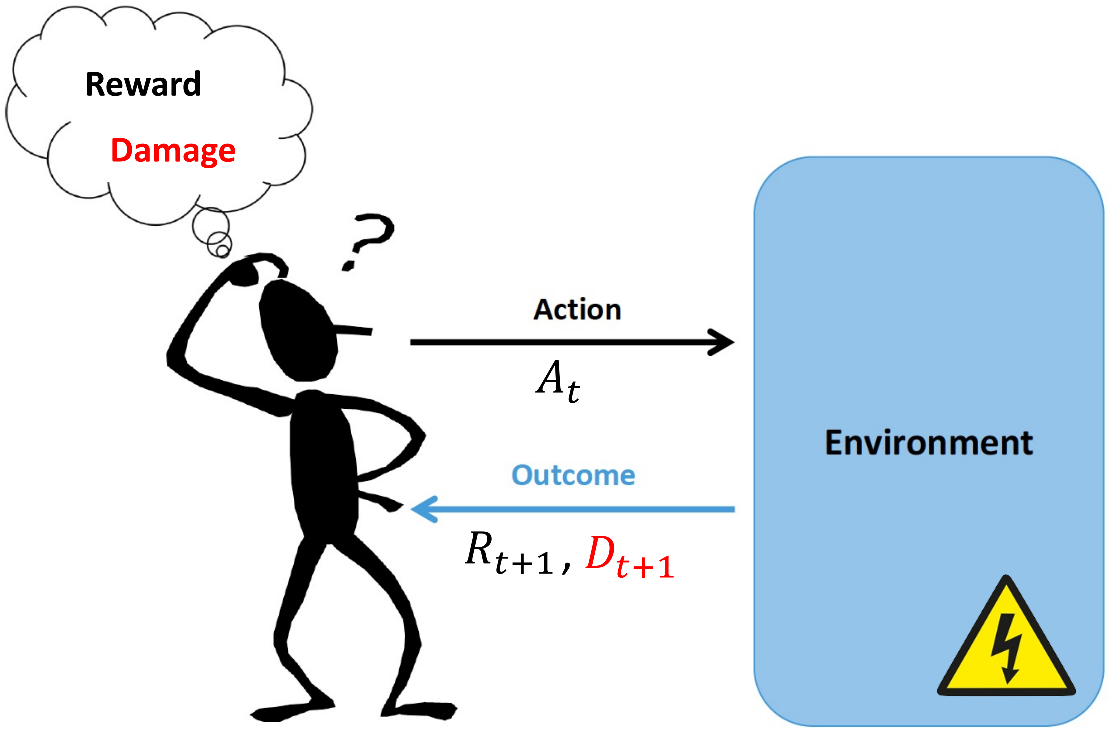
Lower Priority -> Accuracy

- Optimality of the policy
- Full characterization of the safety set?

Key ideas:

- Focus on almost sure **feasibility**, not optimality (Egerstedt et al.,2018)
- Enhanced with **logical** feedback, naturally arising from constraint violations
  - Damage may depend on R_t, or not. May not be directly accessible

Safety Critical Sequential Decision Making



Talk Punchline:

- Can **characterize** all feasible policies ($D_t \equiv 0$) with **finite mistakes**
- **Learning feasible policies** is simpler **than learning** the optimal ones
- Adding **constraints** makes **optimal policies easier to find**

Learning to Act Safely with Limited Exposure and Almost Sure Certainty

Agustin Castellano, Hancheng Min, Juan Bazerque, and Enrique Mallada

ArXiv Preprint, arXiv:2105.08748

Talk Punchline:

- Can **characterize** all feasible policies ($D_t \equiv 0$) with **finite mistakes**
- **Learning feasible policies** is simpler **than learning** the optimal ones
- Adding **constraints** makes **optimal policies easier to find**

Prior Art

- Foundation work on constrained Markov Decision Processes (Altman, 1998)

- **Learning with modelling assumptions**
  - Constrained LQR (Dean et al., 2019)
  - (Achiam et al., 2017)
  - Explore safety for GPs and then optimize rewards (Wachi and Sui, 2020)

- **Model-free Constraints via Primal Dual methods**
  - (Paternain et al., 2019;Ding et al., 2020)
  - Safety guarantees are achieved in the limit

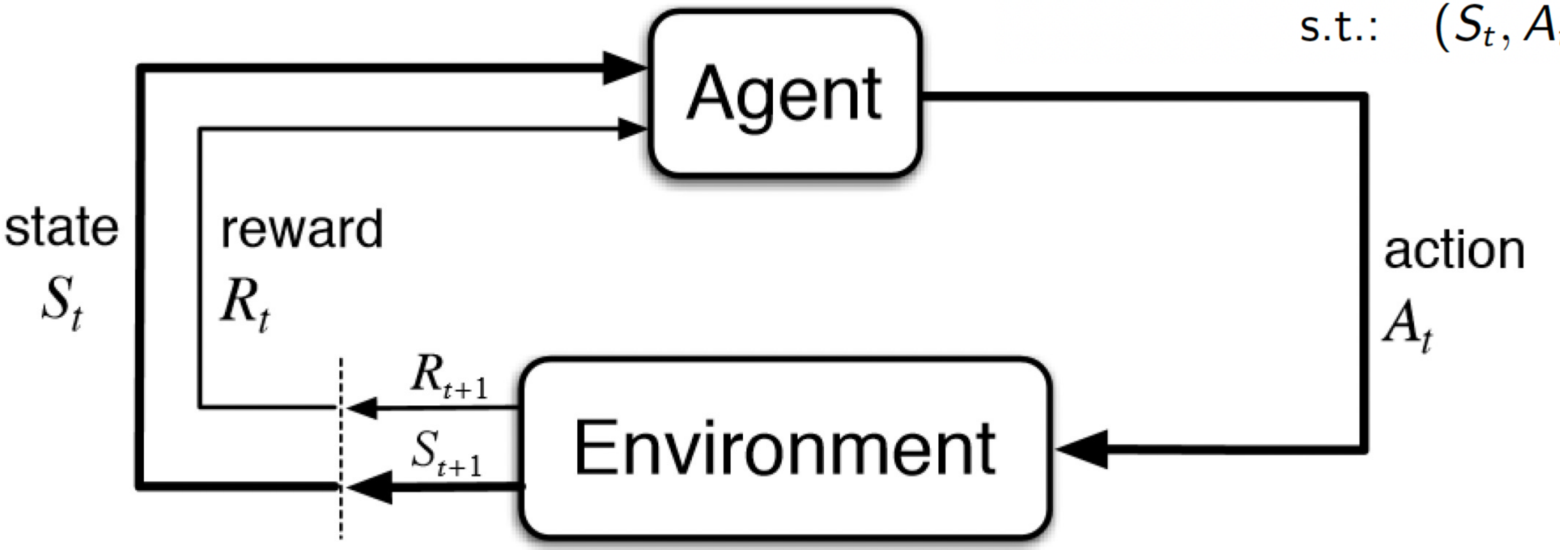
- **Our approach**
  - **Model-free** based learning of **constraints** and optimal **policy**
  - Detect and prevent unsafe actions in **finite time**
  - Learn from constraint violations and rewards together

Reinforcement Learning with Almost Sure Constraints

- Formulated as an MDP with a.s. constraints

$$V^*(s) := \max_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \;\middle|\; S_0 = s \right]$$

$$\text{s.t.:} \quad (S_t, A_t, S_{t+1}) \in \mathcal{F} \quad a.s. \quad \forall t$$



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

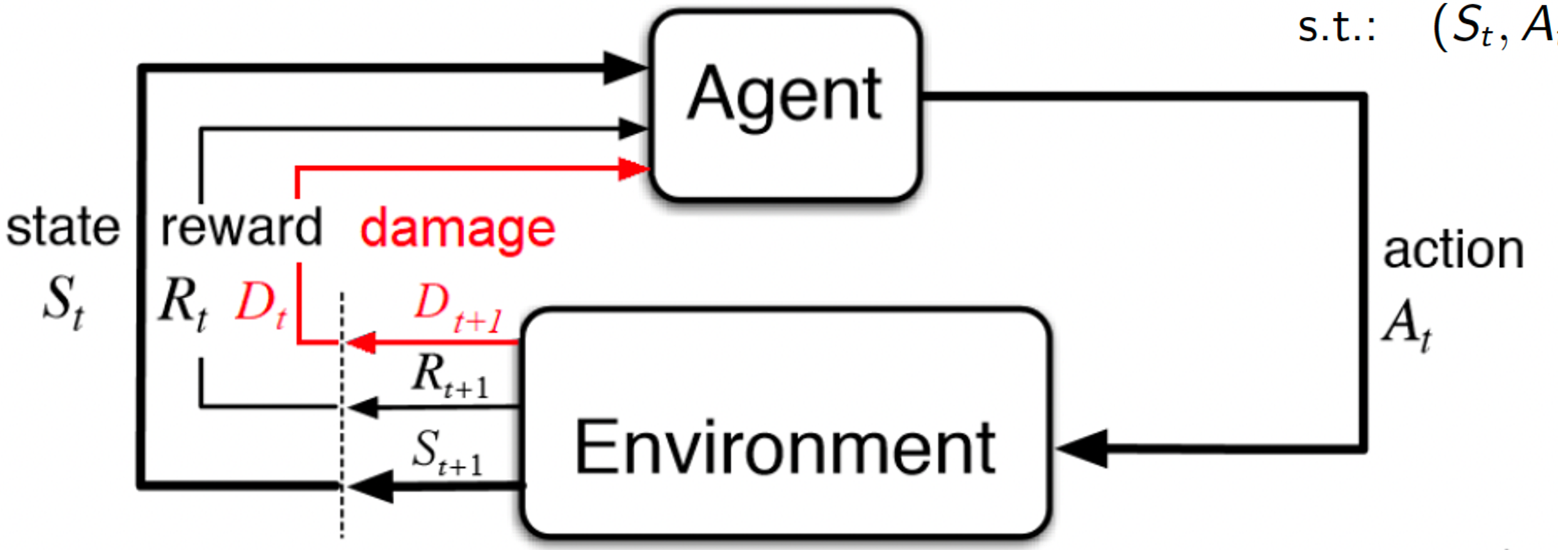
[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Reinforcement Learning with Almost Sure Constraints

- Formulated as an MDP with a.s. constraints

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

$$\text{s.t.:} \quad (S_t, A_t, S_{t+1}) \in \mathcal{F} \quad a.s. \quad \forall t$$



- Constraints not given a priori: Need to learn from experience!
- **Notice:** Model free ➔ Constraint violations are inevitable
- Damage indicator $D_t \in \{0,1\}$ turns on ($D_t = 1$) when constraints are violated

[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Formulation via hard barrier indicator

- Safe RL problem

$$V^*(s) := \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \;\middle|\; S_0 = s\right]$$

$$\text{s.t.:} \quad (S_t, A_t, S_{t+1}) \in \mathcal{F} \quad a.s. \;\; \forall t$$

- Unconstrained formulation achieved using

$$V^{\pi}(s) := \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \left(\gamma^t R_{t+1} - \mathbb{I}_{\{(S_t, A_t, S_{t+1}) \in \mathcal{F}\}}\right) \;\middle|\; S_0 = s\right]$$

where the hard-barrier indicator is given by $\mathbb{I}_{\{.\}} = \begin{cases} 0 & \text{if } \cdot \text{ is true} \\ \infty & \text{if } \cdot \text{ is false} \end{cases}$

[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Hard Barrier Action-Value Functions

Consider the Q-function for a given policy π,

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\left(\gamma^t R_{t+1} - \mathbb{I}_{\{(S_t,A_t,S_{t+1})\in\mathcal{F}\}}\right)\,\middle|\, S_0 = s,\, A_0 = a\right]$$

and define the hard-barrier function

$$B^\pi(s,a) = \mathbb{E}_\pi\left[-\sum_{t=0}^{\infty}\mathbb{I}_{\{(S_t,A_t,S_{t+1})\in\mathcal{F}\}}\,\middle|\, S_0 = s,\, A_0 = a\right]$$

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

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

[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Separation Principle

Theorem

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 π_*

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

Notes:

- As mentioned, the HBF (B^{π}) can be learned in dependently from the rewards
- Any solution to

$$B^*(s,a) := \max_{\pi} B^{\pi}(s,a)$$

  is feasible for the Constrained MDP, in fact $B^{\pi_*} = B^*$.
- We can thus focus on learning B^*!

[1]Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Feasibility Principle

Theorem (Bellman's Optimality for B^*)

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

$$B^*(s,a) = \mathbb{E}\left[-\mathbb{I}_{\{(S_t,A_t,S_{t+1})\in\mathcal{F}\}} + \max_a B^*(S_{t+1},a) \;\middle|\; S_0 = s, A_0 = a\right]$$

Algorithm 3: Barrier-learner

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

- Experienced damage summarized in hard barrier function (HBF)
- Wraps around existing learning algorithms (Q-learning, SARSA)
- Use the HBF to trim the exploration set and avoid repeating unsafe actions, next!

[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Assured Q-Learning with Generative Model

- Generative Model: Query on any (s, a)-pair to sample a transition (s, a, s', d) according to the MDP
- Update barrier function with sampled transition

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

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

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 $P(S_1 = s_t', D_1 = d_t | S_0 = s_t, A_0 = a_t)$

  $B^{(t+1)} \leftarrow \texttt{barrier\_update}(B^{(t)}, s_t, a_t, s_t', d_t)$

end

[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Assured Q-Learning with Generative Model

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**

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

[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Assured Q-Learning with Generative Model

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)

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

Li et al. "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model", NeurIPS, 2020

Assured Q-Learning with Generative Model

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**)

Li et al. "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model", NeurIPS, 2020

Assured Q-Learning with Generative Model

Barrier learner + Learning ϵ-optimal policy

- First, run Barrier Learner for $(L+1)\frac{|S||A|}{\mu}\left(\sum_{k=1}^{|S||A|}\frac{1}{k}\right)\log\frac{1}{\delta/2}$ Iterations.
- Then, learn an ϵ-optimal policy with **the trimmed MDP** using existing algorithms (Li et al. 2020, etc.), with $1-\delta/2$ confidence level

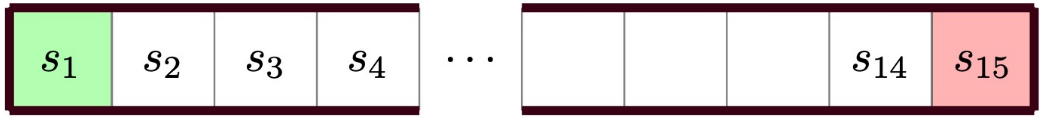
- With at least $1-\delta$ probability, the **Optimal Barrier Function $\boldsymbol{B}^*$** and the **ϵ-optimal <u>safe</u> policy** is learnt, and the entire procedure takes

$$(L+1)\frac{|S||A|}{\mu}\left(\sum_{k=1}^{|S||A|}\frac{1}{k}\right)\log\frac{2}{\delta}+\frac{|S_{safe}||A_{safe}|}{(1-\gamma)^4\varepsilon^2}\log^2\left(\frac{|S_{safe}||A_{safe}|}{(1-\gamma)\varepsilon\delta/2}\right)$$

samples

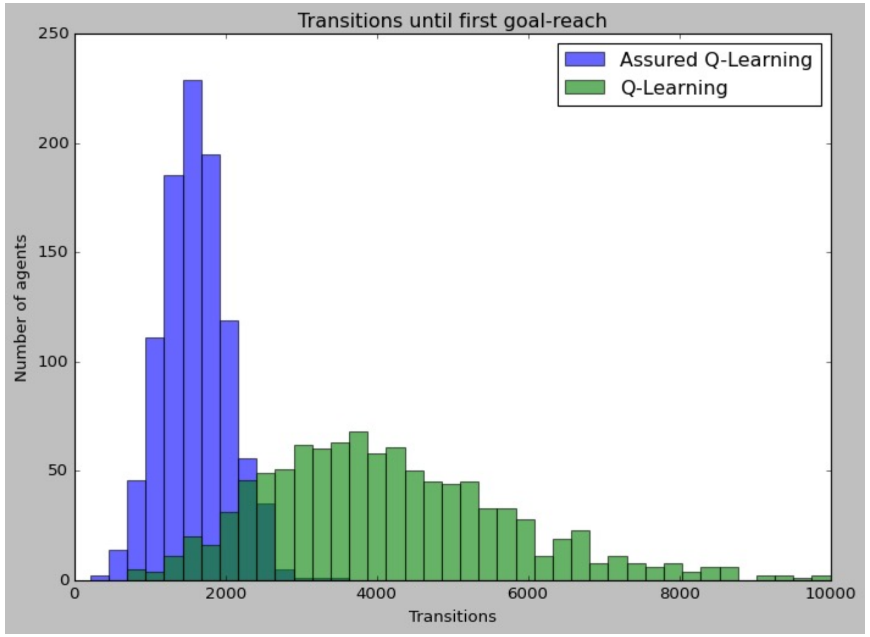
Li et al. "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model", NeurIPS, 2020

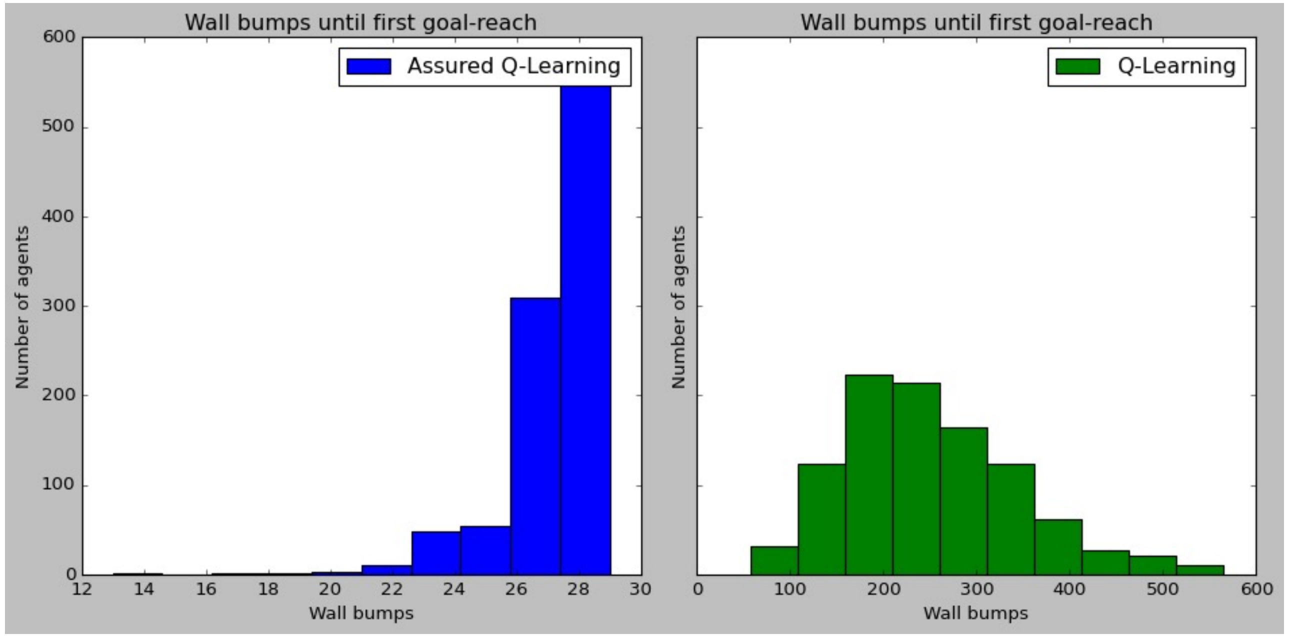
Numerical Experiments



Goal: Reach the end of the aisle without touching the yellow wall

Results





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

[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Numerical Experiments II

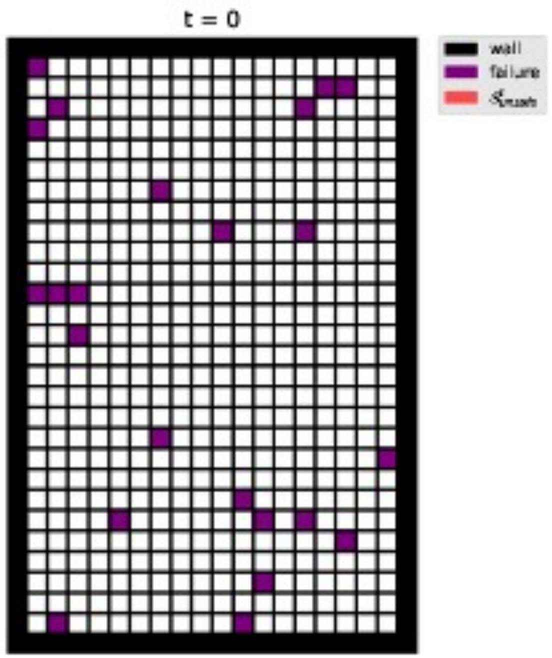
Setup: Rectangular grid, stepping into **holes** gives damage $D_t = 1$.

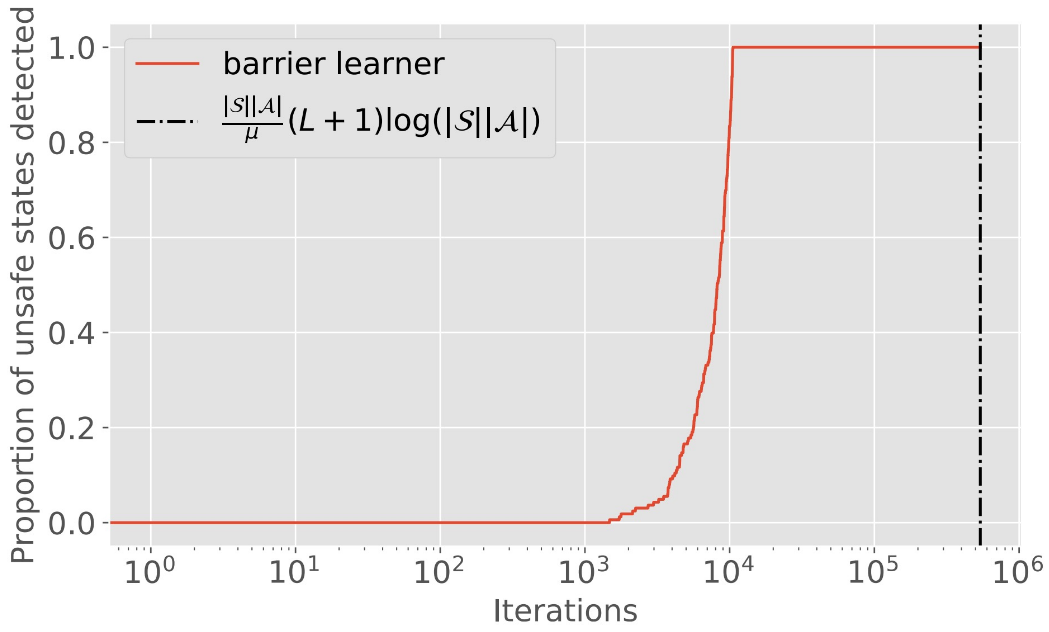
Actions $A = \{up, down, left, right\}$.

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

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

Immediately unsafe states (near **damage**) are identified first.





[1]Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Numerical Experiments II

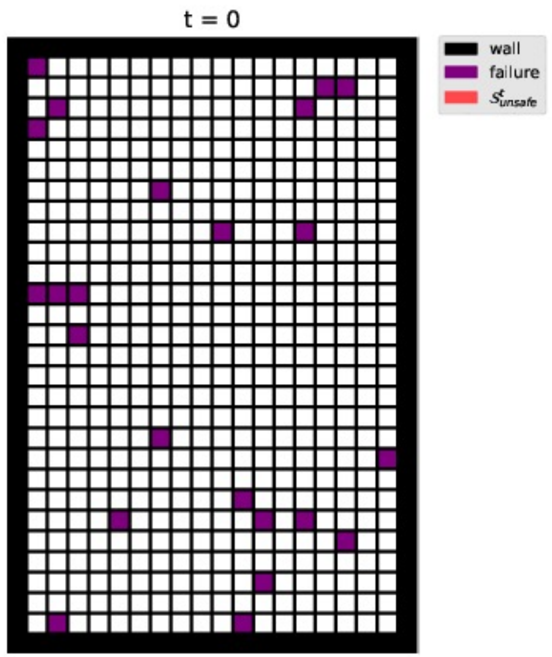
Setup: Rectangular grid, stepping into **holes** gives damage $D_t = 1$.

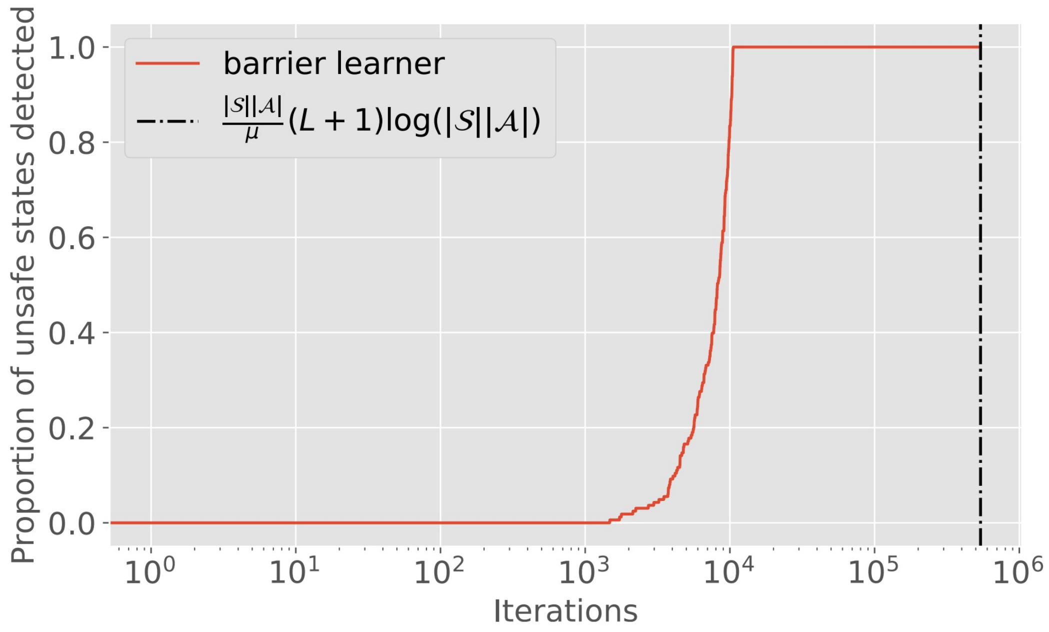
Actions $A = \{up, down, left, right\}$.

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

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

Immediately unsafe states (near **damage**) are identified first.





[1] Castellano, Min, Bazerque and M, "Learning to Act Safely with Limited Exposure and Almost Sure constraints", *under review*

Summary and Future Work

Summary

- Studied safe/constrained reinforcement learning:
  - Focus on safety first, show it can be achieved quickly, and with strong
  - Treat constraints separately, or in parallel
- Motivate the need of additional information, *damage*

Ongoing Work

- Reinforcement Learning with constraints: $\sum_{k=0}^{+\infty} D_{k+1} \leq C \quad a.s.$
- More generally, trajectory dependent constraints

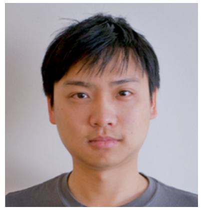
Thanks!

Related Publications:

- Castellano, Bazerque and M, “Learning to be safe, in finite time”, ACC, 2021
- Castellano, Min, Bazerque and M, “Learning to Act Safely with Limited Exposure and Almost Sure constraints”, under review

Agustin Castellano

Johns Hopkins University



Hancheng Min

Johns Hopkins University

Enrique Mallada

mallada@jhu.edu

http://mallada.ece.jhu.edu

Juan Bazerque

Universidad de la República Uruguay