Reinforcement Learning with Almost Sure Constraints

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

JOHNS HOPKINS UNIVERSITY

ITA Workshop
San Diego, CA

May 27, 2022

[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)]

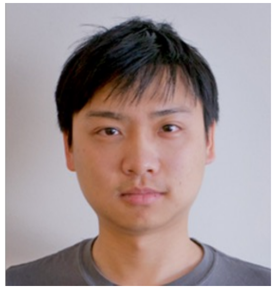
Learning to Act Safely with Limited Exposure and Almost Sure Certainty

Agustin Castellano, Hancheng Min, Juan Bazerque, Enrique Mallada

arXiv > eess > arXiv:2105.08748

Agustin Castellano

Johns Hopkins University



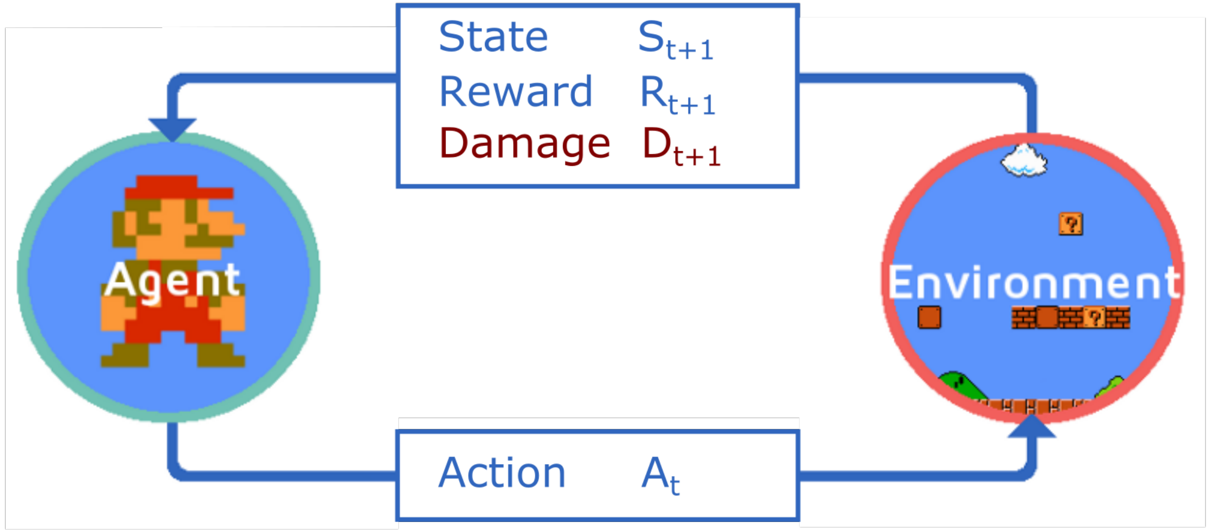
Hancheng Min

Johns Hopkins University

Juan Bazerque

Universidad de la República Uruguay

Learning for Safety-critical Sequential Decision Making



Requirements:

High Priority -> Safety

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

Lower Priority -> Accuracy

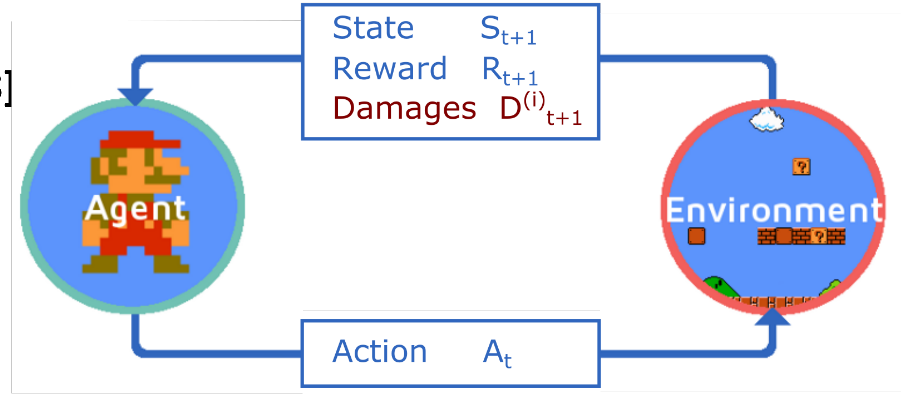
- Optimality of the policy

Key ideas:

- Focus on almost sure **feasibility**, not optimality (Egerstedt et al.,2018)
- Enhanced with **logical** feedback, naturally arising from constraint violations

Background

- **Constrained Markov Decision Processes (CMDPs)** **[Altman'98]**



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

$$\text{s.t.:} \quad C_i^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t D^{(i)}_{t+1} | S_0 = s\right] \leq c_i \quad i = 1, \ldots, m$$

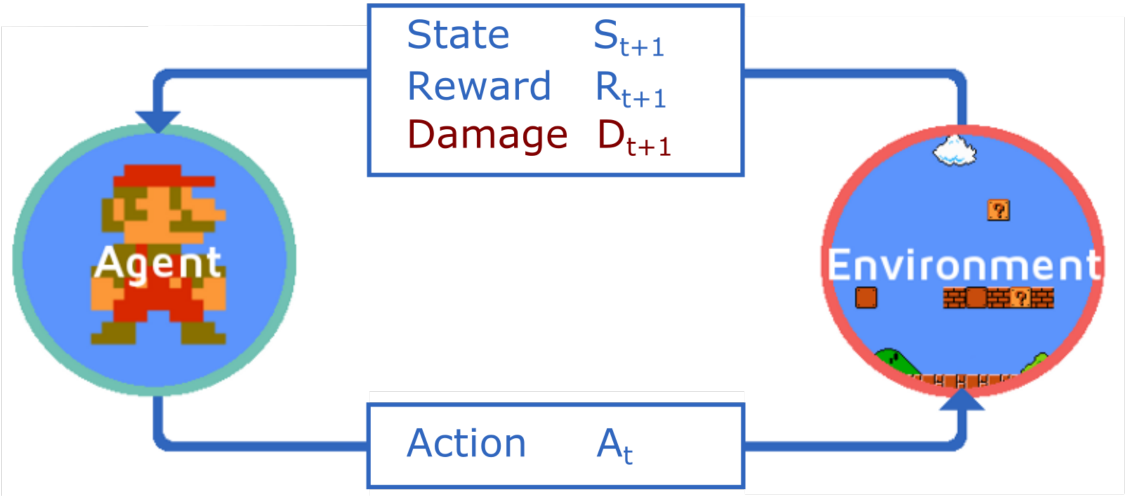
- Solvable if MDP is "known" (Linear Program).
- ∃ stationary optimal solution $\pi^*(a|s)$

- **What to do if MDP is "unknown"? Examples of Model-based and Model-free methods**
- (MB) Learn transitions and reward/constraint signals, solve for a (near) optimal policy.
  [Aria HZ et al'20], [Bai et al'20], [Wang et al 20], [Chen et al'21]
- (MF) Primal or Primal-dual methods.
  [Chow et al'17], [Tessler et al'19], [Paternain et al'19], [Ding et al'20], [Stooke et al. '20], [Xu et al'21]

Reinforcement Learning with Almost Sure Constraints

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

$$\text{s.t.: } \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t D_{t+1} \mid S_0 = s \right] \le c \iff D_{t+1} = 0 \text{ almost surely } \forall t$$



- 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

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:

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

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_*

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

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

Optimal Hard Barrier Action-Value Function

Theorem (Bellman Equation for B^*)

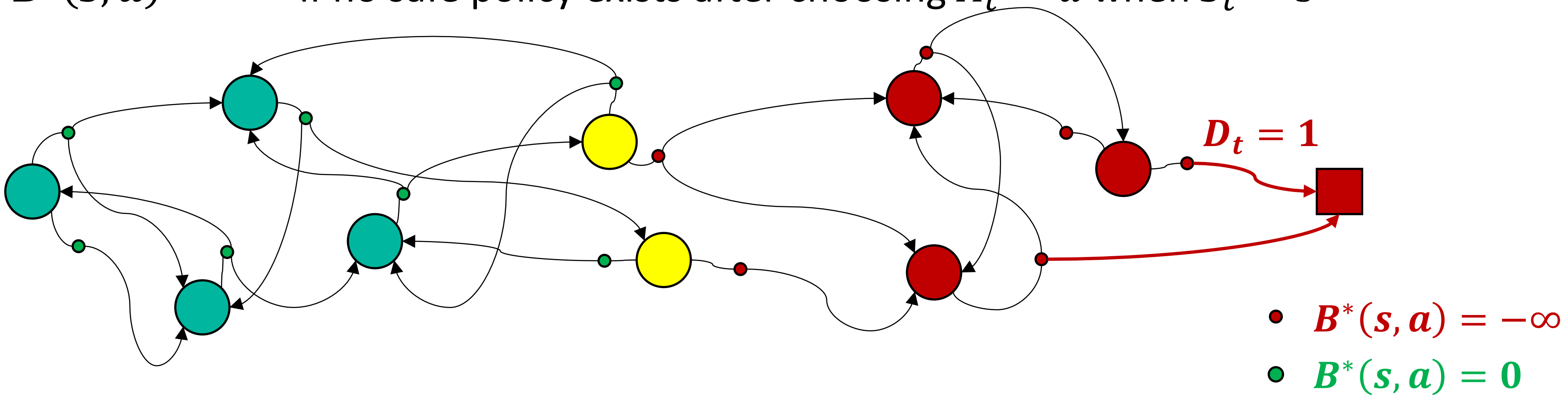
Let $B^*(s,a) := \max_\pi B^\pi(s,a)$, then the following holds:

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

Understanding $\boldsymbol{B^*(s,a)}$:

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

- $\boldsymbol{B^*(s,a)} = \mathbf{0}$ if $\exists$ safe π after choosing $A_t = a$ when $S_t = s$
- $\boldsymbol{B^*(s,a)} = -\infty$ if no safe policy exists after choosing $A_t = a$ when $S_t = s$



Learning the barrier…

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 HBF to trim 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 $B^{(0)}(s, a) = 0, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$ — Initially, all (s, a)-pairs are "safe"
for $t = 0, 1, \cdots$ **do**
  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 $P(S_1 = s'_t, D_1 = d_t | S_0 = s_t, A_0 = a_t)$
  $B^{(t+1)} \leftarrow \texttt{barrier\_update}(B^{(t)}, s_t, a_t, s'_t, d_t)$ — 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

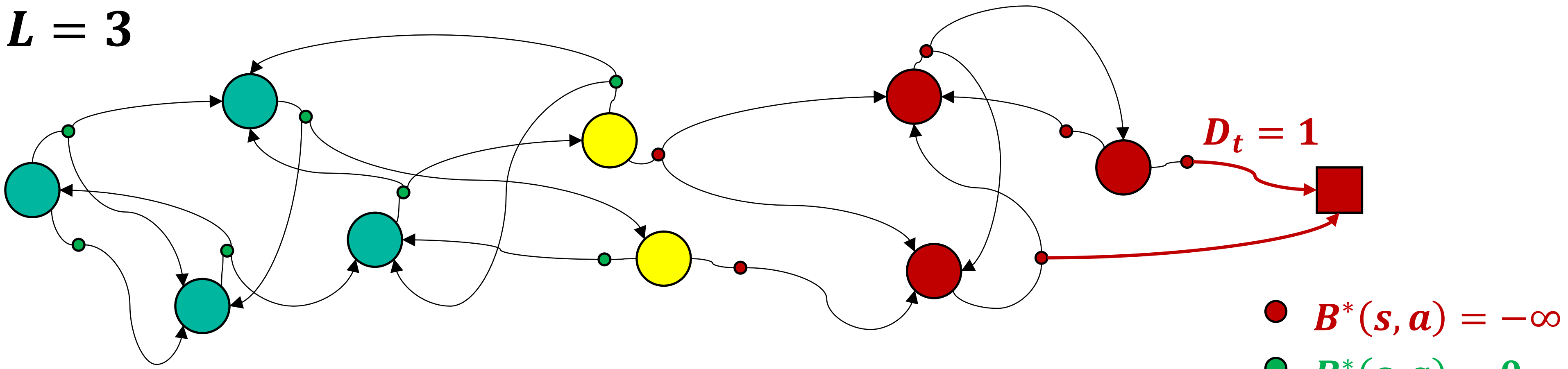
$$\mu = min\{p(s',d|s,a): p(s',d|s,a) \neq 0\}$$

- 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

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



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

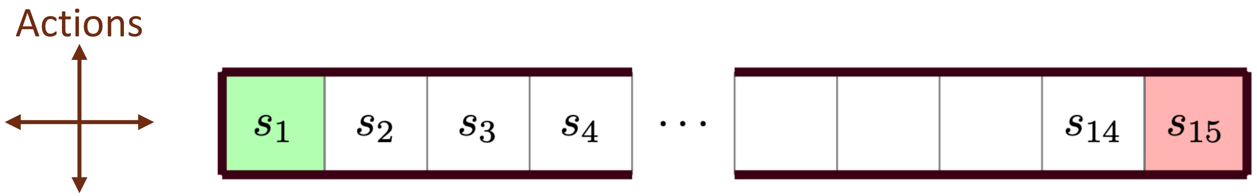
$$N' = \frac{|S_{safe}||A_{safe}|}{(1-\gamma)^4\varepsilon^2}\log^2\left(\frac{|S_{safe}||A_{safe}|}{(1-\gamma)\varepsilon\delta}\right)$$

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

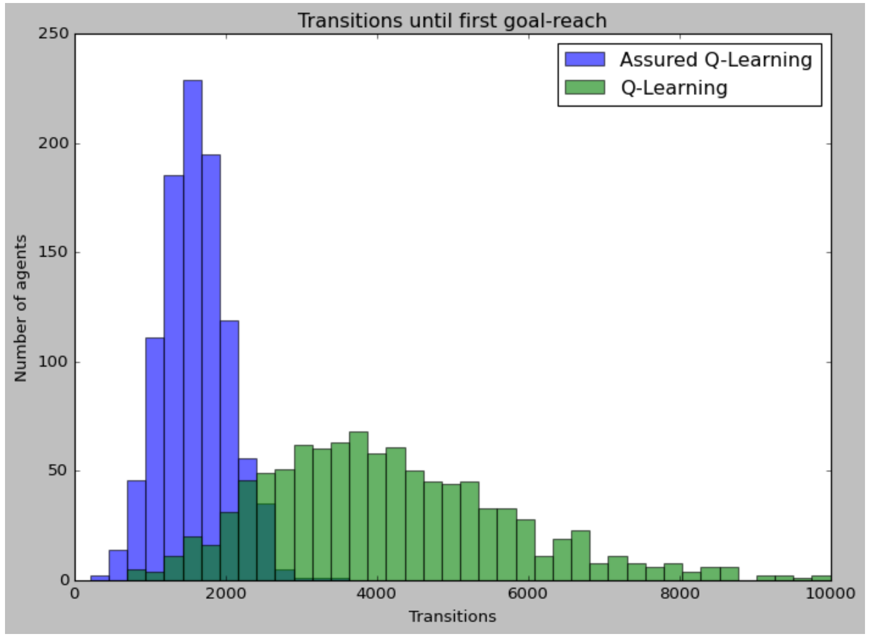
Numerical Experiments

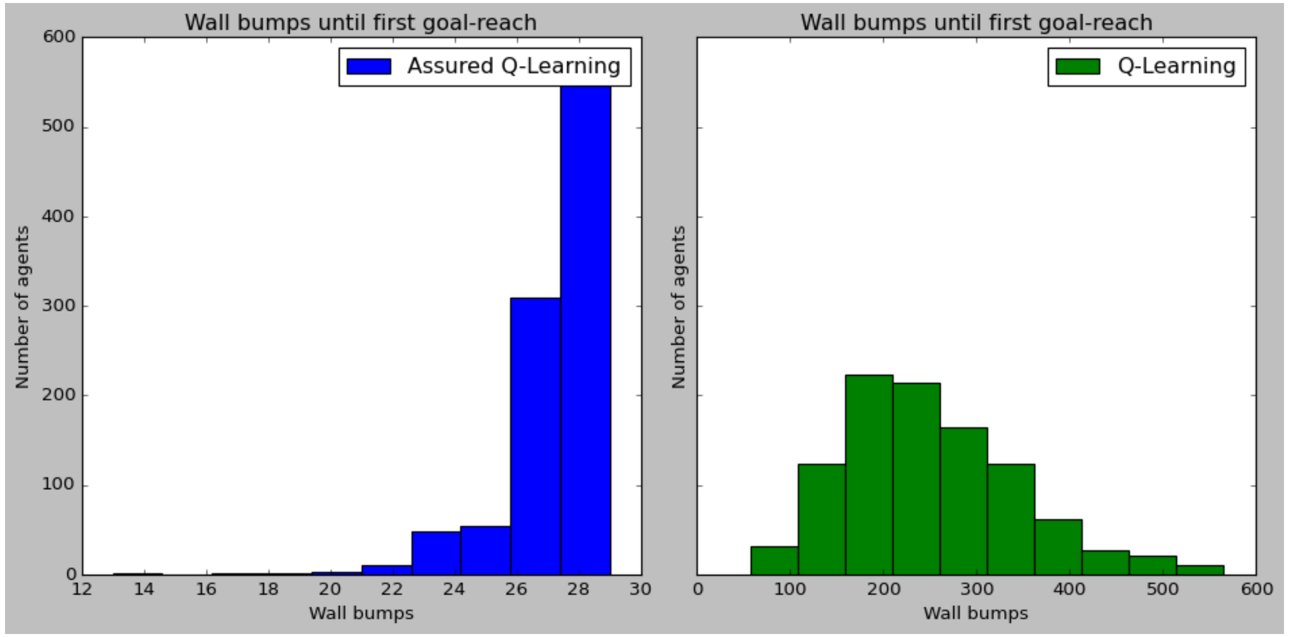
Goal: Reach the end of the aisle ($R_{t+1} = 10$)

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

Almost sure RL with positive budget (Δ)

- Almost Sure RL with positive budget

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

$$\text{s.t: } P_\pi \left(\sum_{t=0}^{\infty} D_{t+1} \leq \Delta \mid S_0 = s \right) = 1$$

Π_H: history-dependent policies

$h_t = (S_0, A_0, R_1, D_1, \ldots, S_t)$; $\pi(a|h_t)$

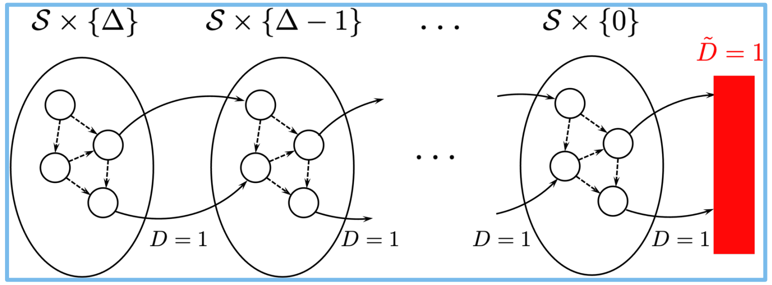
- Current budget at time t:

$$K_t = \Delta - \sum_{\ell=0}^{t-1} D_{\ell+1} \quad \forall t \geq 1$$

"How much more damage I can sustain and still be feasible"

- Augmented MDP $\tilde{\mathcal{M}}$

$$\tilde{S}_t = (S_t, K_t)\,, \qquad \tilde{D}_{t+1} = \mathbf{1}\{K_t - D_{t+1} < 0\}\,.$$



- <u>Equivalent</u> problem:

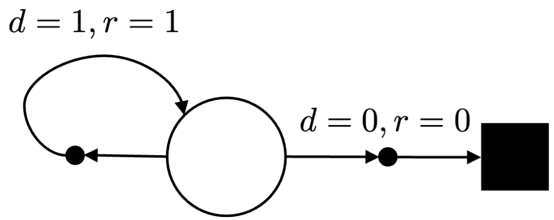
$$\max_{\tilde{\pi} \in \tilde{\Pi}_H} \mathbb{E}_{\tilde{\pi}, \tilde{\mathcal{M}}} \left[\sum_{t=0}^{\infty} R_{t+1} \,\middle|\, (S_0, K_0) = (s, \Delta) \right]$$

$$\text{s.t: } P_{\tilde{\pi}} \left(\tilde{D}_{t+1} = 0 \right) = 1 \qquad \forall t \geq 0$$

Fits previous formulation! →

- Could learn $B^*(s, k, a)$
- Separation & Feasibility Principles
- Potential drawback: working in higher dimensions?

Experiment: comparing constraints



Goal

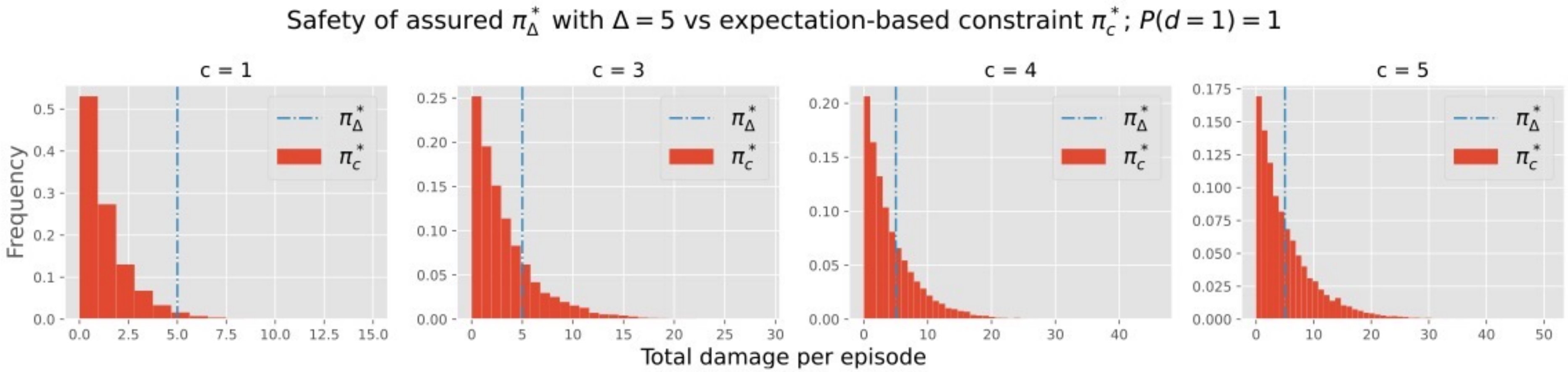
$$\max_{\pi} \quad \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} R_{t+1}\right]$$

1) Proposed constraint

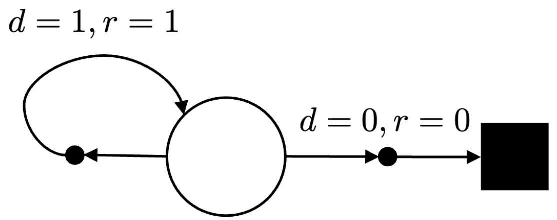
$$\mathbb{P}_{\pi_\Delta}\left(\sum_{t=0}^{\infty} D_{t+1} \leq \Delta \,\middle|\, S_0 = s\right) = 1$$

2) Classic CMDP constraint

$$\mathbb{E}_{\pi_c}\left[\sum_{t=0}^{\infty} D_{t+1}\right] \leq c$$



Experiment: comparing constraints



Goal

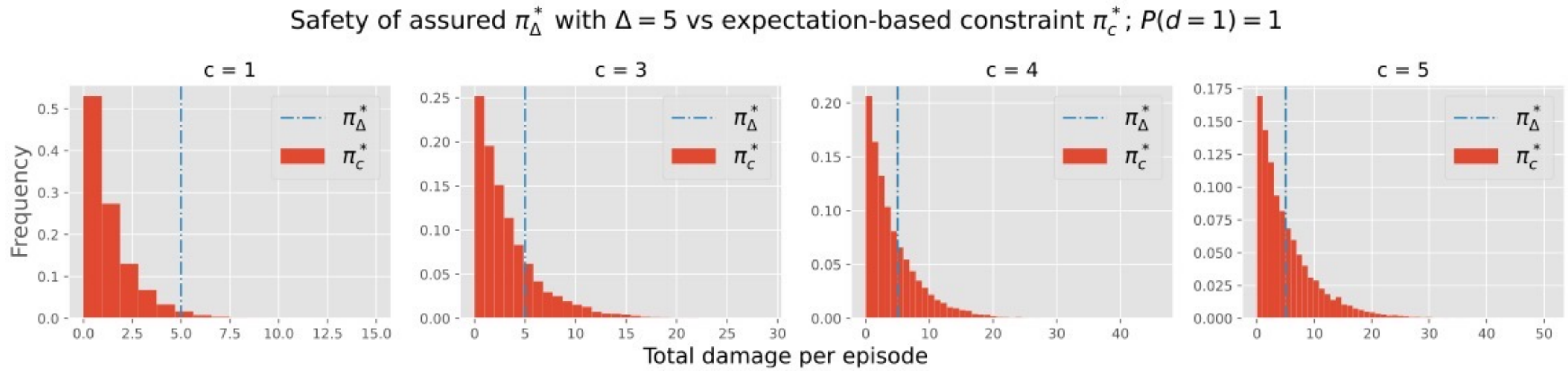
$$\max_{\pi} \quad \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} R_{t+1}\right]$$

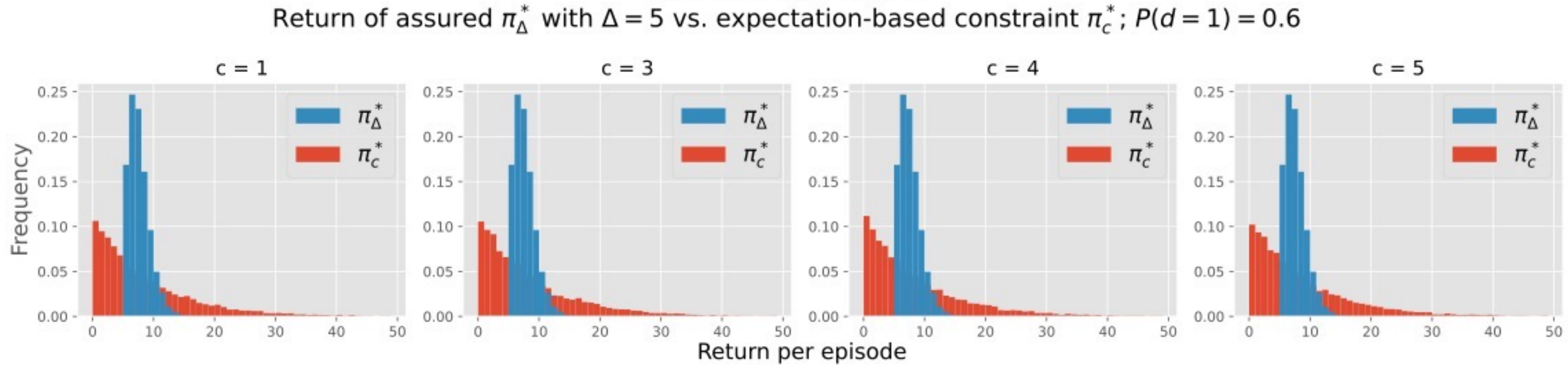
1) Proposed constraint

$$\mathbb{P}_{\pi_\Delta}\left(\sum_{t=0}^{\infty} D_{t+1} \le \Delta \;\middle|\; S_0 = s\right) = 1$$

2) Classic CMDP constraint

$$\mathbb{E}_{\pi_c}\left[\sum_{t=0}^{\infty} D_{t+1}\right] \le c$$





Summary and future work

Summary

- 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**
- **Take aways:**
  - **Learning feasible policies** is simpler **than learning** the optimal ones
  - Adding **constraints** makes **optimal policies easier to find**

Future work:

- Theory: Extensions to continue state and action spaces
- Application: Deep RL with almost sure constraints

Thanks!

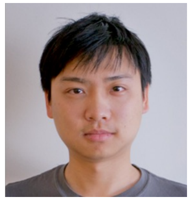
Related Publications:

[L4DC 22] Castellano, Min, Bazerque, M, *Reinforcement Learning with Almost Sure Constraints,* **Learning for Dynamics and Control (L4DC) Conference, 2022**

[arXiv 21] Castellano, Min, Bazerque, M, *Learning to Act Safely with Limited Exposure and Almost Sure Certainty,* **submitted to IEEE TAC, 2021, under review**, preprint arXiv:2105.08748

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