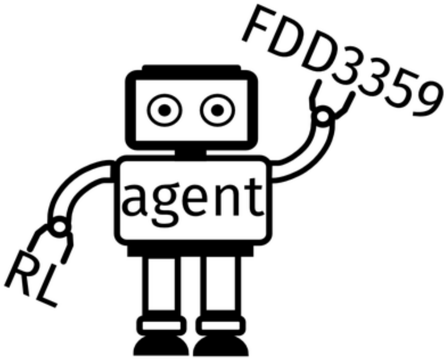
KTH ROYAL INSTITUTE OF TECHNOLOGY

Reinforcement Learning

Temporal Logic Constrained RL

FDD3359

Alexis Linard



Temporal Logic Constrained RL

- Goal
  - Get to know the basics of safe reinforcement learning with shielding
  - Get to know the concepts of Linear (LTL) and Signal Temporal Logic (STL)
  - Understand how shielding avoids the agent taking unsafe actions
  - Use TL rewards in RL

- Acknowledgements:
  - Some figures taken from literature (cited along the slides)
  - Tutorial on Safe RL of Berkenkamp and Krause
  - Some figures taken from Alexandre Donzé's lecture notes on STL

[1] F. Berkenkamp and A. Krause. "Tutorial on Safe Reinforcement Learning". Lecture notes. ETH Zürich. 2018. https://las.inf.ethz.ch/files/ewrl18_SafeRL_tutorial.pdf
[2] A. Donzé. "On Signal Temporal Logic". Lecture notes. University of California, Berkley. 2014. https://people.eecs.berkeley.edu/~sseshia/fmee/lectures/EECS294-98_Spring2014_STL_Lecture.pdf

Intended Learning Outcomes

By the end of this, you should be able to:

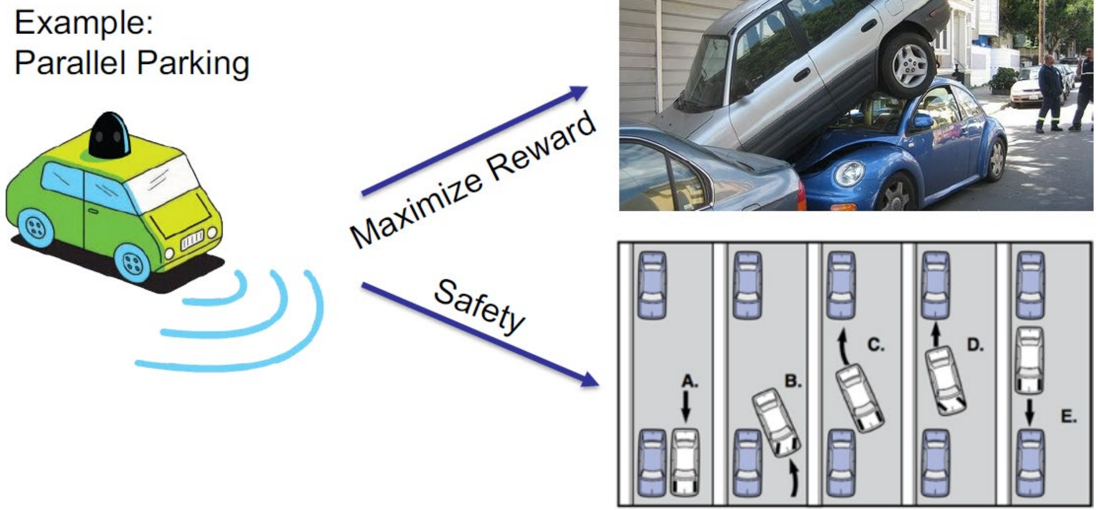
- Apply shielding
- Define specifications using Linear Temporal Logic
- Define specifications using Signal Temporal Logic and use quantitative semantics as reward in the RL framework

Reinforcement Learning – Limits

- Good at learning optimal policy/converging to local maximum of reward

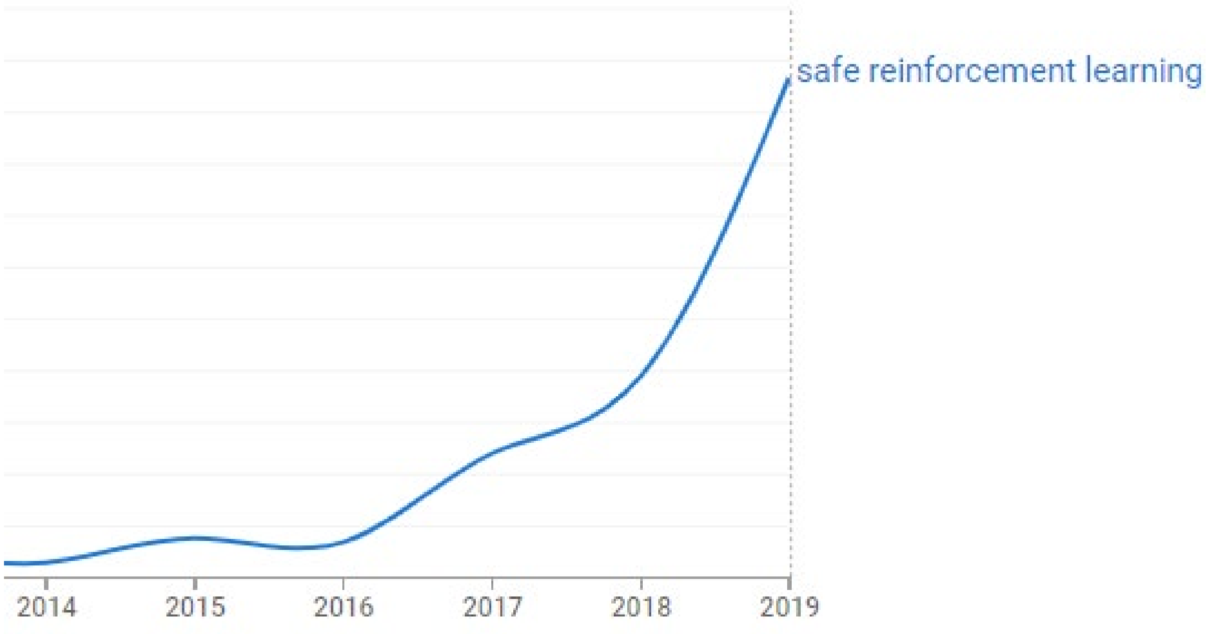
Reinforcement Learning – Limits

- Good at learning optimal policy/converging to local maximum of reward
- Bad a guaranteeing safety

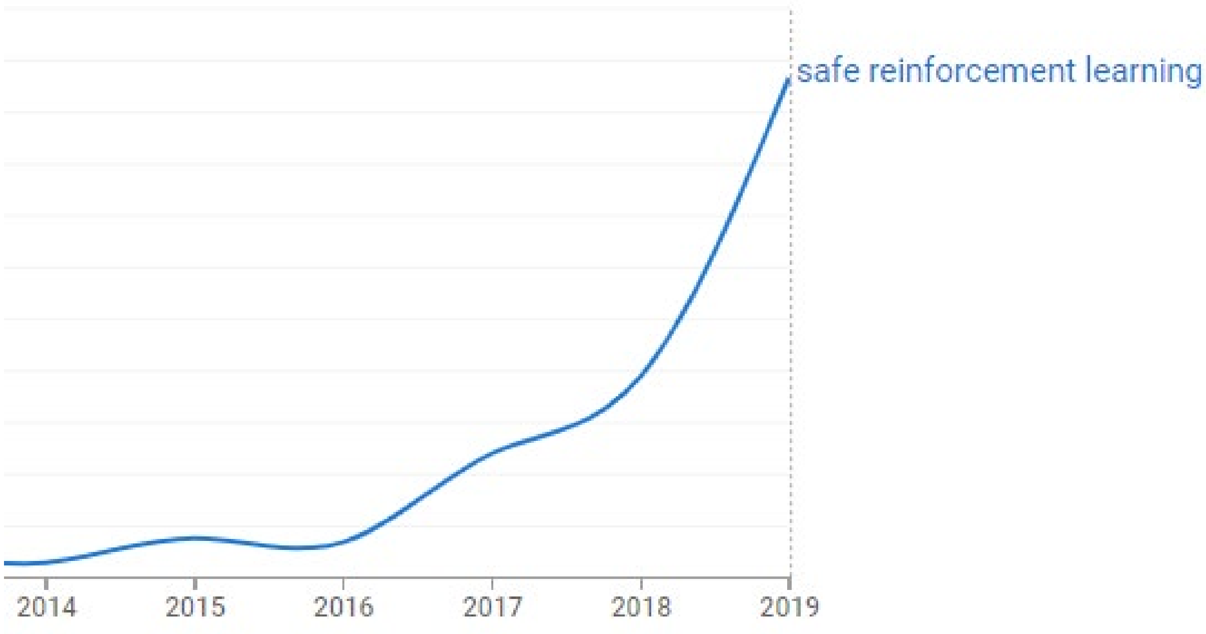


Inspired by Carl Hildebrandt
https://1funny.com/wp-content/uploads/2009/04/bad-parking.jpg
https://www.drivingtips.com/minimizing-risks-of-accidents-while-driving/

Safe Reinforcement Learning



Safe Reinforcement Learning



Disclaimer: we don't cover the whole literature, just a selection that matches our research.

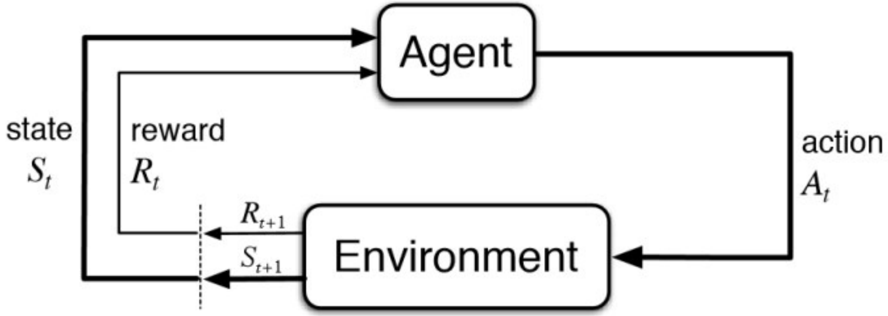
Safe Reinforcement Learning

- Safe RL
  - Optimization Criterion
    - Worst Case Criterion
      - Inherent Uncertainty
        - Heger (1994b,a)
        - Gaskett (2003)
      - Parameter Uncertainty
        - Nilim and El Ghaoui (2005)
        - Tamar et al. (2013)
    - Risk-Sensitive Criterion
      - Exponential Functions
        - Howard and Matheson (1972)
        - Borkar (2001, 2002)
        - Basu et al. (2008)
      - Weighted Sum of Return and Risk
        - Mihatsch and Neuneier (2002)
        - Sato et al. (2002)
        - Geibel and Wysotzki (2005)
    - Constrained Criterion
      - Moldovan and Abbeel (2011, 2012a)
      - Castro et al. (2012)
      - Kadota et al. (2006)
    - Other Optimization Criteria
      - Morimura et al. (2010a,b)
      - Luenberger (2013)
      - Castro et al. (2012)
  - Exploration Process
    - External Knowledge
      - Providing Initial Knowledge
        - Driessens and Džeroski (2004)
        - Martín H. and Lope (2009)
        - Song et al. (2012)
      - Deriving a Policy from Demonstrations
        - Abbeel et al. (2010)
        - Tang et al. (2010)
      - Teacher Advice
        - Ask for Help
          - Clouse (1997)
          - García and Fernández (2012)
          - Geramifard et al. (2013)
        - Teacher Provide Advices
          - Clouse and Utgoff (1992)
          - Thomaz and Breazeal (2006, 2008)
          - Vidal et al. (2013)
        - Other Approaches
          - Rosenstein and Barto (2002, 2004)
          - Kuhlmann et al. (2004)
          - Torrey and Taylor (2012)
    - Risk-directed Exploration
      - Gehring and Precup (2013)
      - Law (2005)

[3] J. Garcıa and F. Fernández. "A comprehensive survey on safe reinforcement learning." Journal of Machine Learning Research 16.1 (2015): 1437-1480.

Safe Reinforcement Learning (for Robotics)

- Learn control parameters such that:
  - the resulting policy satisfies a safety specification φ
  - the system stays safe during the learning process



Safe Reinforcement Learning (for Robotics)

- Shielding
- Including Temporal Logics rewards

Safe Reinforcement Learning (for Robotics)

- **Shielding**
- Including Temporal Logics rewards

Linear Temporal Logic (LTL)

- To define safety specifications
- Temporal logics specify patterns that timed behaviors of systems may or may not satisfy
- The most intuitive is Linear Temporal Logic (LTL), dealing with discrete sequences of states.
- Based on logic operators:
  - $\wedge$
  - $\neg$
  - $\vee$
- Based on temporal operators:
  - $\mathcal{N}$ *(also written* $\bigcirc$) "Next"
  - $\mathcal{G}$ *(also written* $\square$) "Always"
  - $\mathcal{U}$ "Until"
  - $\mathcal{F}$ *(also written* $\lozenge$ *)* "Eventually"

LTL Semantics

- An LTL formula φ is evaluated on a sequence, e.g., $w = a\ a\ a\ a\ b\ b\ b\ a\ a\ a$ ...
- At each step of w, we can define a truth value of φ, noted $\chi^{\varphi}(w, i)$
- LTL atoms $\pi \in AP$ are represented by symbols: $a\ b\ \ldots$
- We say that $w \vDash \varphi \leftrightarrow \chi^{\varphi}(w, 0) = 1$

$$
\begin{array}{rccccccccccc}
i = & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \ldots \\
w = & a & a & a & a & b & b & b & a & a & a & \ldots \\
\chi^{a}(w, i) = & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & \ldots \\
\chi^{b}(w, i) = & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & \ldots
\end{array}
$$

LTL Semantics

- Temporal operators are evaluated at each step wrt the future of sequences

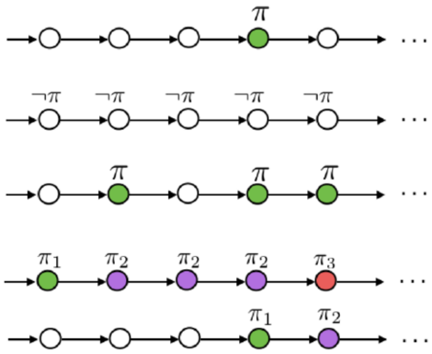
$\varphi ::= \pi \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \mathcal{N}\varphi \mid \varphi_1 \mathcal{U} \varphi_2$,
where $\pi \in AP$

$\mathcal{F}\varphi \leftrightarrow \bot\, \mathcal{U}\, \varphi$
$\mathcal{G}\varphi \leftrightarrow \neg\mathcal{F}\neg\varphi$

$$
\begin{aligned}
w \vDash \varphi_1 \wedge \varphi_2 &\leftrightarrow w \vDash \varphi_1 \text{ and } w \vDash \varphi_2 \\
w \vDash \neg\varphi &\leftrightarrow w \nvDash \varphi \\
w \vDash \mathcal{N}\varphi &\leftrightarrow w^2 \vDash \varphi \\
w \vDash \varphi_1 \mathcal{U} \varphi_2 &\leftrightarrow \exists j \geq 1, w^j \vDash \varphi_2 \\
&\quad and\ w^i \vDash \varphi_1, \forall\, 1 \leq i < j
\end{aligned}
$$

- TL patterns:

Reachability	$\mathcal{F}\pi$
Safety	$\mathcal{G}\neg\pi$
Surveillance	$\mathcal{G}\mathcal{F}\pi$
Sequencing	$\pi_1 \mathcal{U} (\pi_2 \mathcal{U} \pi_3)$
	$\mathcal{F}(\pi_1 \wedge \mathcal{F}\pi_2)$
Response	$\mathcal{G}(request \Rightarrow \mathcal{F}\, response)$



LTL Semantics

- Temporal operators are evaluated at each step wrt the future of sequences

$i =$	0	1	2	3	4	5	6	7	8	9	…
$w =$	a	a	a	a	b	b	b	a	a	a	…
$\chi^{\mathcal{N}b}(w,i) =$	0	0	0	1	1	1	0	0	0	?	…
$\chi^{\mathcal{G}\,a}(w,i) =$	0	0	0	0	0	0	0	1?	1?	1?	…
$\chi^{\mathcal{F}\,b}(w,i) =$	1	1	1	1	1	1	1	0?	0?	0?	…
$\chi^{a\mathcal{U}b}(w,i) =$	1	1	1	0	0	0	0	0?	0?	0?	…

LTL exercise

5 minutes in the breakout rooms:

- One of you – click on share screen and select whiteboard
- All of you – write as you wish
- When the last minute countdown starts, take a screenshot
- When you get back to the main room, share the screen with the screenshot (multiple sharing will be enabled).

Write LTL properties of a traffic light:

1. Red and green are never on simultaneously.
2. Whenever there is red, it will stay red until there is yellow and then it will stay yellow until there is green.

Btw, have you ever heard of…

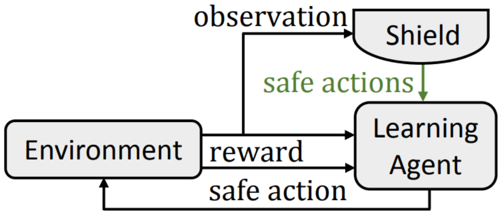
- …model checking?

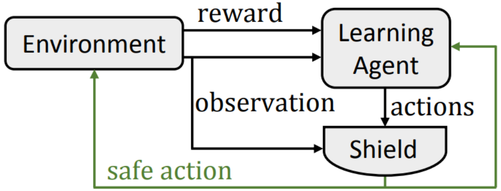
$$\boxed{\text{System } \mathcal{M}} \longrightarrow aaaabbbaa\ldots \longrightarrow \boxed{\text{Property } \varphi} \longrightarrow 111000\ldots$$

- Model checking consists of proving that $\mathcal{M} \vDash \varphi$
  - Formally, $\mathcal{M} \vDash \varphi \leftrightarrow \forall w \in traces(\mathcal{M}),\ w \vDash \varphi$

Safe RL via Shielding

- Generate a set of system specifications and an abstraction of the agent's environment expressed as temporal logic.
- Synthesize a reactive system (shield) which enforces the safety properties of the systems specifications.
- Modify the learning loop by placing the shield in 1 of 2 places:
  - Before the learning agent, thus removing any unsafe actions.
  - After the learning agent, thus monitoring the selected actions and correcting them only if an unsafe action is chosen.





[4] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu (2018). “Safe Reinforcement Learning via Shielding”. In AAAI-18: 32nd AAAI Conference on Artificial Intelligence (pp. 2669-2678).

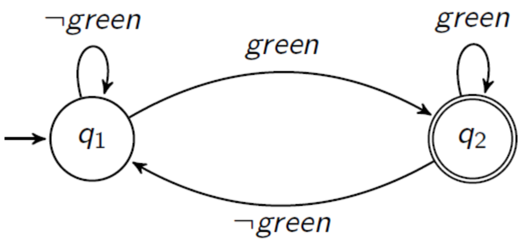
Shield Synthesis

- System specifications are given as temporal logic.
- Convert the safety specification into an automaton φ_s in which only safe states F may be visited: $\varphi_s = (Q, q_0, \Sigma, \delta, F)$
- Convert the environment abstraction (often modeled as a MDP) into an automaton: $\varphi_M = (Q, q_0, \Sigma, \delta, F)$
- Use reactive synthesis to enforce φ_s by solving a safety game built from φ_s and φ_M which is won if the system only ever visits safe states F.

[5] R. Bloem, B. Könighofer, R. Könighofer, and C. Wang. (2015, April). “Shield synthesis”. In International Conference on Tools and Algorithms for the Construction and Analysis of Systems (pp. 533-548

Shield Synthesis

- System specifications are given as temporal logic.
- Convert the safety specification into a Büchi Automaton (BA) $\varphi_s = (Q, q_0, \Sigma, \delta, F)$
  - Every LTL formula can be algorithmically translated into a language equivalent BA
  - An accepting run is a run that intersects F infinitely many times
  - An input word is accepted if there exists an accepting run over it
  - An example BA for $\mathcal{F}\ green$:



- Convert the environment abstraction (often modeled as a MDP) into an automaton: $\varphi_M = (Q, q_0, \Sigma, \delta, F)$
- Use reactive synthesis to enforce φ_s by solving a safety game built from φ_s and φ_M which is won if the system only ever visits safe states F.

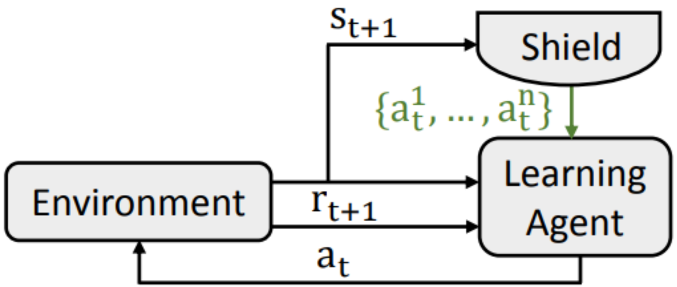
Shield Synthesis

- System specifications are given as temporal logic.
- Convert the safety specification into an automaton φ_s in which only safe states F may be visited: $\varphi_s = (Q, q_0, \Sigma, \delta, F)$
- Convert the environment abstraction (often modeled as a MDP) into an automaton: $\varphi_M = (Q, q_0, \Sigma, \delta, F)$
- Use reactive synthesis to enforce φ_s by solving a safety game built from φ_s and φ_M which is won if the system only ever visits safe states F.

[5] R. Bloem, B. Könighofer, R. Könighofer, and C. Wang. (2015, April). “Shield synthesis”. In International Conference on Tools and Algorithms for the Construction and Analysis of Systems (pp. 533-548

Safe RL via Shielding (Preemptive Shielding)

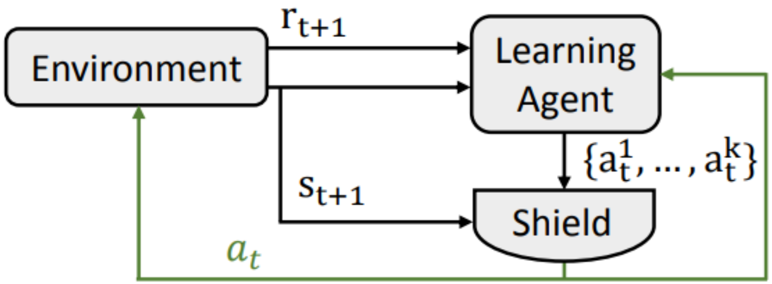
- Transforms the original MDP $\mathcal{M}$ into a new MDP $\mathcal{M}' = (\mathcal{S}', \mathcal{A}', \mathcal{R}', \mathbb{P}')$ with the unsafe actions at each state removed.
- $\mathcal{S}'$ is the product of the original MDP and the state space of the shield
- For each $s \in \mathcal{S}'$ create a new subset $\mathcal{A}'_s \subseteq \mathcal{A}_s$



[4] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu (2018). "Safe Reinforcement Learning via Shielding". In AAAI-18: 32nd AAAI Conference on Artificial Intelligence (pp. 2669-2678).

Safe RL via Shielding (Post-Posed Shielding)

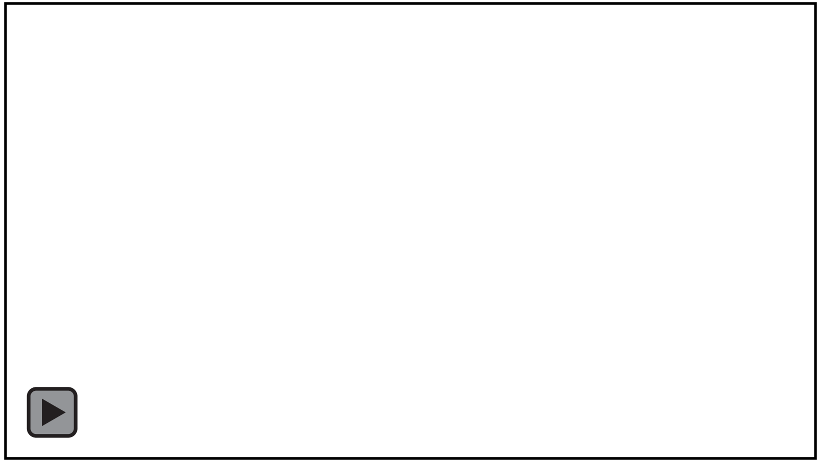
- Allows fixed policy
- Learning algorithm only sees state of the MDP (without shield)
- Shielding is transparent



[4] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu (2018). "Safe Reinforcement Learning via Shielding". In AAAI-18: 32nd AAAI Conference on Artificial Intelligence (pp. 2669-2678).

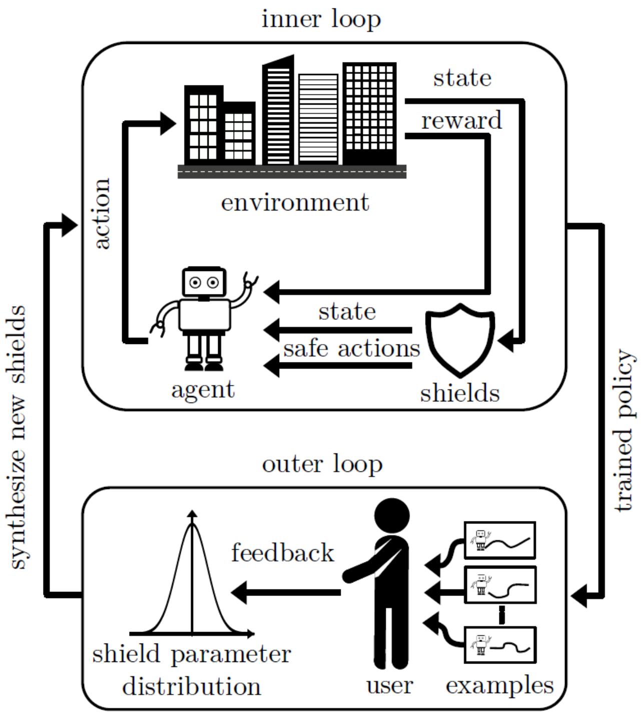
Along the lines of shielding

- Shielded decision-making in MDPs [6]
- Probabilistic Shielding [7]



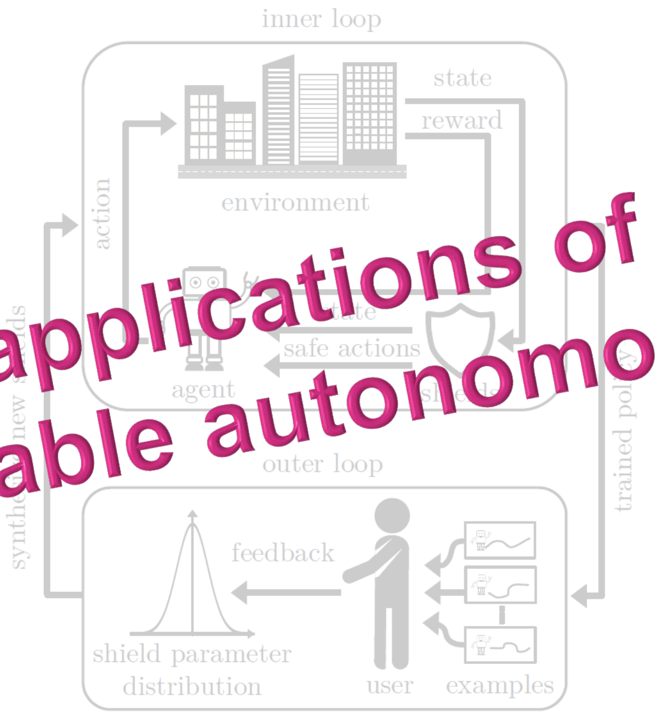
[6] N. Jansen, B. Könighofer, S. Junges, and R. Bloem, (2018). “Shielded decision-making in MDPs”. arXiv preprint arXiv:1807.06096.
[7] N. Jansen, B. Könighofer, S. Junges, A. Serban, and R. Bloem. (2018). “Safe Reinforcement Learning via Probabilistic Shields”. arXiv, arXiv-1807.

Along the lines of shielding



[7B] Daniel Marta, Christian Pek, Gaspar Isaac, Melsión, Jana Tumova, Iolanda Leite, Human-Feedback Shield Synthesis for Perceived Safety in Deep Reinforcement Learning. IEEE Robotics and Automation Letters 7.1 (2021): 406-413.

Along the lines of shielding



[8] Daniel Marta, Christian Pek, Gaspar Isaac, Melsión, Jana Tumova, Iolanda Leite, Human-Feedback Shield Synthesis for Perceived Safety in Deep Reinforcement Learning. IEEE Robotics and Automation Letters 7.1 (2021): 406-413.

Safe Reinforcement Learning (for Robotics)

- Shielding
- **Including Temporal Logics reward**

RL with TL rewards

- While LTL only comes up with *qualitative* semantics

- LTL

$\varphi ::= \pi \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \mathcal{N}\varphi \mid \varphi_1 \mathcal{U} \varphi_2$,
where $\pi \in AP$

[9] P. Kapoor, A. Balakrishnan, and J. V. Deshmukh (2020). "Model-based Reinforcement Learning from Signal Temporal Logic Specifications". arXiv preprint arXiv:2011.04950.
[10] X. Li, C. I. Vasile, and C. Belta (2017). "Reinforcement learning with temporal logic rewards". In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3834-3839).

RL with TL rewards

- While LTL only comes up with *qualitative* semantics, there are other temporal logics coming up with ***quantitative*** semantics!

- LTL

$\varphi ::= \pi \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \mathcal{N}\varphi \mid \varphi_1 \mathcal{U} \varphi_2$,
where $\pi \in AP$

- TLTL

$\varphi ::= f(s) < c \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \mathcal{N}\varphi \mid \varphi_1 \mathcal{U} \varphi_2$
where $f: \mathbb{R}^n \to \mathbb{R}$

- STL

$\varphi ::= \mu \mid \top \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \varphi_1 \mathcal{U}_{[a,b]} \varphi_2$,
where μ is an atomic predicate of the form $\mu = f(x_1[t], \dots, x_n[t]) > 0$
and $[a, b]$, $a, b \in \mathbb{R}$ is the time interval

[9] P. Kapoor, A. Balakrishnan, and J. V. Deshmukh (2020). "Model-based Reinforcement Learning from Signal Temporal Logic Specifications". arXiv preprint arXiv:2011.04950.
[10] X. Li, C. I. Vasile, and C. Belta (2017). "Reinforcement learning with temporal logic rewards". In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3834-3839).

Signal Temporal Logic

- STL is continuous time and continuous space

$\varphi ::= \mu \mid \top \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \varphi_1 \mathcal{U}_{[a,b]} \varphi_2$

Assume atomic predicates of the form $\mu = f(x_1[t], \dots, x_n[t]) > 0$

The satisfaction of φ by an n-dimensional signal $\mathfrak{x} = (x_1, \dots, x_n)$ at time t :

$$
\begin{aligned}
(\mathfrak{x}, t) &\vDash \mu && \leftrightarrow f(x_1[t], \dots, x_n[t]) > 0 \\
(\mathfrak{x}, t) &\vDash \varphi_1 \wedge \varphi_2 && \leftrightarrow (\mathfrak{x}, t) \vDash \varphi_1 \text{ and } (\mathfrak{x}, t) \vDash \varphi_2 \\
(\mathfrak{x}, t) &\vDash \neg\varphi && \leftrightarrow (\mathfrak{x}, t) \nvDash \varphi \\
(\mathfrak{x}, t) &\vDash \varphi_1 \mathcal{U}_{[a,b]} \varphi_2 && \leftrightarrow \exists t \in [t+a, t+b] \quad \mathit{such\ that}\ (\mathfrak{x}, t') \vDash \varphi_2 \\
& && \qquad \mathit{and}\ \forall t'' \in [t, t'], (\mathfrak{x}, t'') \vDash \varphi_1
\end{aligned}
$$

$$\lozenge\, \mathcal{U}_{[a,b]} \varphi = \top\, \mathcal{U}_{[a,b]} \varphi$$

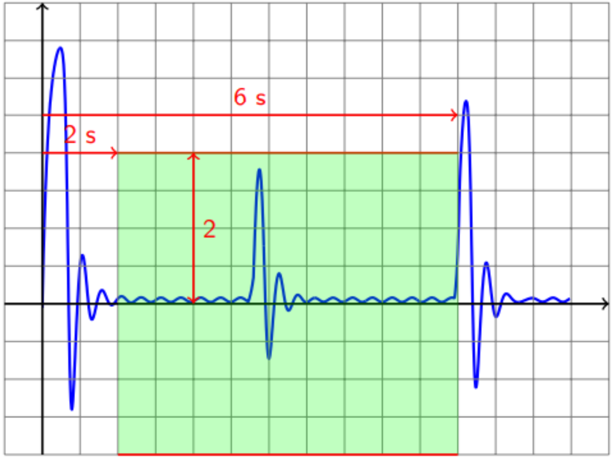
$$\square\, \mathcal{U}_{[a,b]} \varphi = \neg \lozenge\, \mathcal{U}_{[a,b]} \neg\varphi$$

[11] O. Maler and D. Nickovic, “Monitoring temporal properties of continuous signals,” in Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems. Springer, 2004, pp. 152–166.
[2] A. Donzé. “On Signal Temporal Logic”. Lecture notes. University of California, Berkley. 2014. https://people.eecs.berkeley.edu/~sseshia/fmee/lectures/EECS294-98_Spring2014_STL_Lecture.pdf

Signal Temporal Logic: Semantics

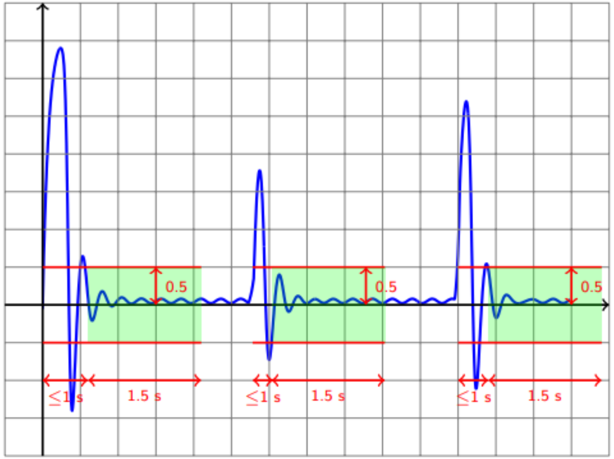
Between 2s and 6s the signal is between -2 and 2

$$\varphi := \mathsf{G}_{[2,6]}\ (|x[t]| < 2)$$



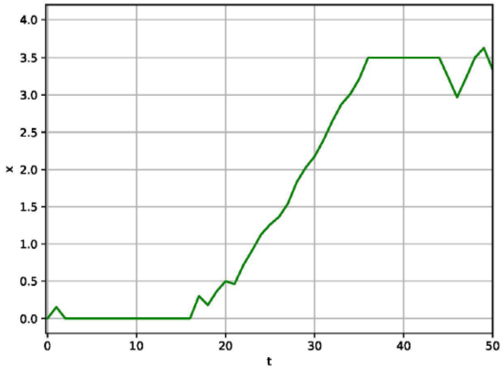
Always $|x| > 0.5 \Rightarrow$ *after 1 s,* $|x|$ *settles under 0.5 for 1.5 s*

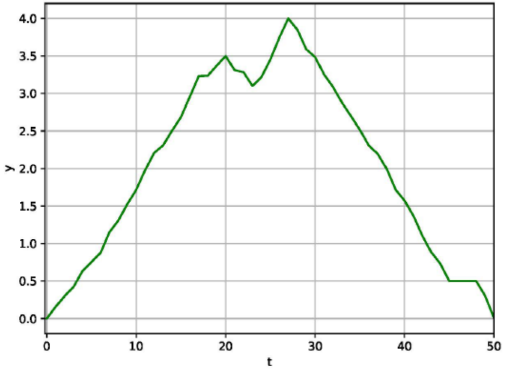
$$\varphi := \mathsf{G}(x[t] > .5 \rightarrow \mathsf{F}_{[0,.6]}\ (\mathsf{G}_{[0,1.5]}\ x[t] < 0.5))$$

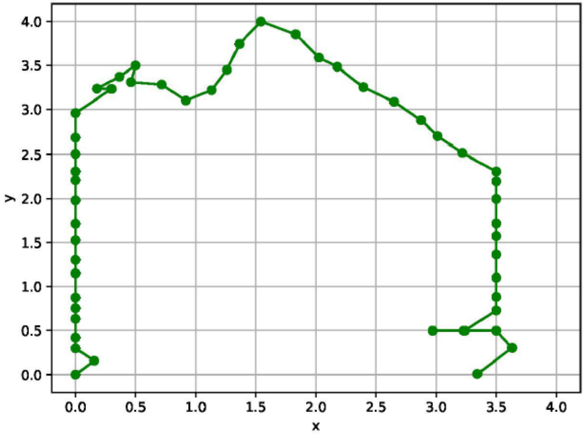


[11] O. Maler and D. Nickovic, "Monitoring temporal properties of continuous signals," in Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems. Springer, 2004, pp. 152–166.
[2] A. Donzé. "On Signal Temporal Logic". Lecture notes. University of California, Berkley. 2014. https://people.eecs.berkeley.edu/~sseshia/fmee/lectures/EECS294-98_Spring2014_STL_Lecture.pdf

Signal Temporal Logic: Semantics

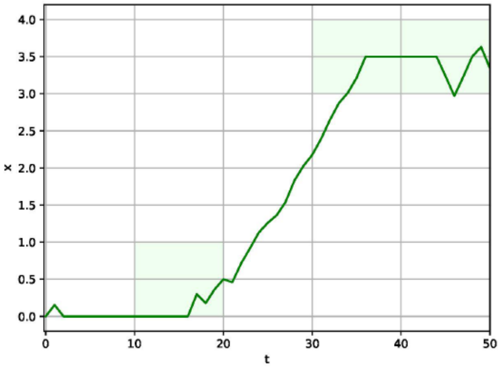




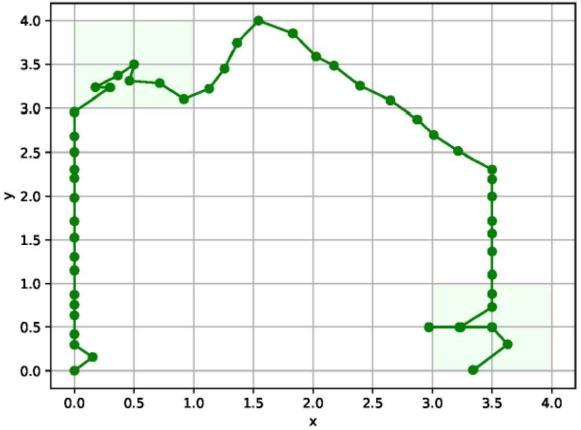


[11] O. Maler and D. Nickovic, “Monitoring temporal properties of continuous signals,” in Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems. Springer, 2004, pp. 152–166.
[2] A. Donzé. “On Signal Temporal Logic”. Lecture notes. University of California, Berkley. 2014. https://people.eecs.berkeley.edu/~sseshia/fmee/lectures/EECS294-98_Spring2014_STL_Lecture.pdf

Signal Temporal Logic: Semantics



$$\varphi = \Diamond_{[10,20]}\,(0 < x < 1 \wedge 3 < y < 4)$$
$$\wedge\ \Diamond_{[30,50]}\,(3 < x < 4 \wedge 0 < y < 1)$$

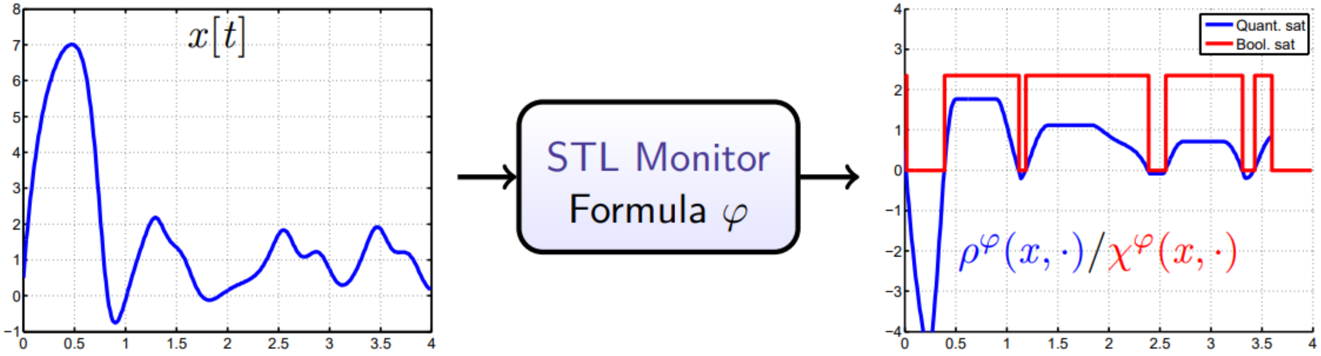


[11] O. Maler and D. Nickovic, "Monitoring temporal properties of continuous signals," in Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems. Springer, 2004, pp. 152–166.
[2] A. Donzé. "On Signal Temporal Logic". Lecture notes. University of California, Berkley. 2014. https://people.eecs.berkeley.edu/~sseshia/fmee/lectures/EECS294-98_Spring2014_STL_Lecture.pdf

Signal Temporal Logic: Quantitative Measures

$$
\begin{aligned}
\rho\left(f(\mathbf{x}) > 0, \mathbf{x}, \tau\right) &= f(\mathbf{x}(\tau)) \\
\rho\left(\neg\varphi, \mathbf{x}, \tau\right) &= -\rho(\varphi, \mathbf{x}, \tau) \\
\rho\left(\varphi_1 \wedge \varphi_2, \mathbf{x}, \tau\right) &= \min\left(\rho(\varphi_1, \mathbf{x}, \tau), \rho(\varphi_2, \mathbf{x}, \tau)\right) \\
\rho\left(\square_I \varphi, \mathbf{x}, \tau\right) &= \inf_{\tau' \in \tau + I} \rho(\varphi, \mathbf{x}, \tau') \\
\rho\left(\diamond_I \varphi, \mathbf{x}, \tau\right) &= \sup_{\tau' \in \tau + I} \rho(\varphi, \mathbf{x}, \tau') \\
\rho\left(\varphi \mathbf{U}_I \psi, \mathbf{x}, \tau\right) &= \sup_{\tau_1 \in \tau + I} \min\left(\rho(\psi, \mathbf{x}, \tau_1), \inf_{\tau_2 \in (\tau, \tau_1)} \rho(\varphi, \mathbf{x}, \tau_2)\right)
\end{aligned}
$$

Robustness of φ on a signal $\boldsymbol{x}$



[12] Donzé, A., Ferrere, T., & Maler, O. (2013, July). Efficient robust monitoring for STL. In International Conference on Computer Aided Verification (pp. 264-279).

STL Robustness Exercise

5 minutes in the breakout rooms:

- One of you – click on share screen and select whiteboard
- All of you – write as you wish
- When the last minute countdown starts, take a screenshot
- When you get back to the main room, share the screen with the screenshot (multiple sharing will be enabled).

You have 2 little tasks:

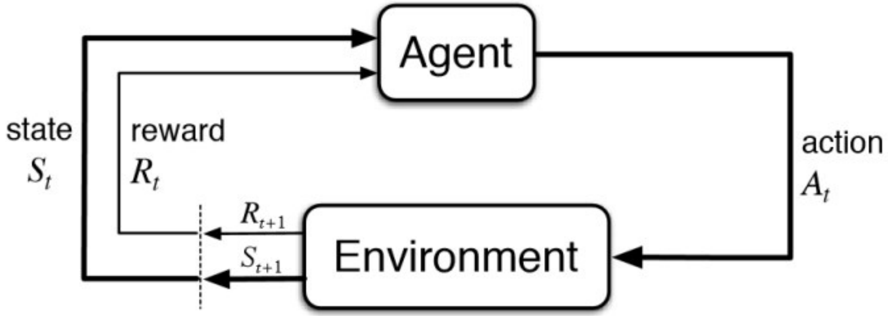
- Compute $\rho(\Box_{[0,10]}\,(x > 2), \boldsymbol{x}, 0)$
- Compute $\rho\left(\lozenge_{[2,5]}\,(x > 2), \boldsymbol{x}, 0\right)$

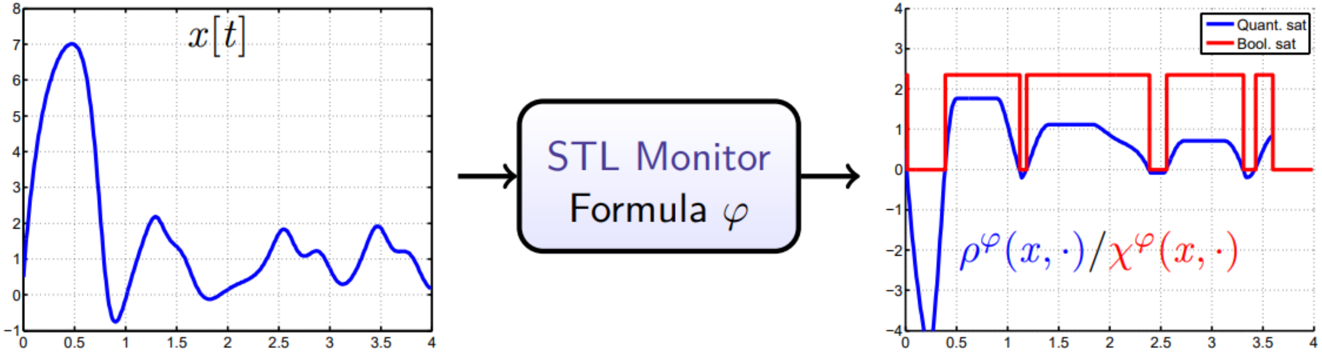
$$
\begin{array}{lcl}
\rho\left(f(\mathbf{x}) > 0, \mathbf{x}, \tau\right) & = & f(\mathbf{x}(\tau)) \\
\rho\left(\neg\varphi, \mathbf{x}, \tau\right) & = & -\rho(\varphi, \mathbf{x}, \tau) \\
\rho\left(\varphi_1 \wedge \varphi_2, \mathbf{x}, \tau\right) & = & \min\left(\rho(\varphi_1, \mathbf{x}, \tau), \rho(\varphi_2, \mathbf{x}, \tau)\right) \\
\rho\left(\Box_I \varphi, \mathbf{x}, \tau\right) & = & \inf\limits_{\tau' \in \tau + I} \rho(\varphi, \mathbf{x}, \tau') \\
\rho\left(\Diamond_I \varphi, \mathbf{x}, \tau\right) & = & \sup\limits_{\tau' \in \tau + I} \rho(\varphi, \mathbf{x}, \tau') \\
\rho\left(\varphi \mathbf{U}_I \psi, \mathbf{x}, \tau\right) & = & \sup\limits_{\tau_1 \in \tau + I} \min\left(\rho(\psi, \mathbf{x}, \tau_1), \inf\limits_{\tau_2 \in (\tau, \tau_1)} \rho(\varphi, \mathbf{x}, \tau_2)\right)
\end{array}
$$

Robustness of φ on a signal $\boldsymbol{x}$

$$\boldsymbol{x} = [1.3, 1.9, 2.1, 2.2, 2.1, 2.4, 2.3, 2.2, 2.1, 1.9, 1.9, 1.8, 1.7]$$

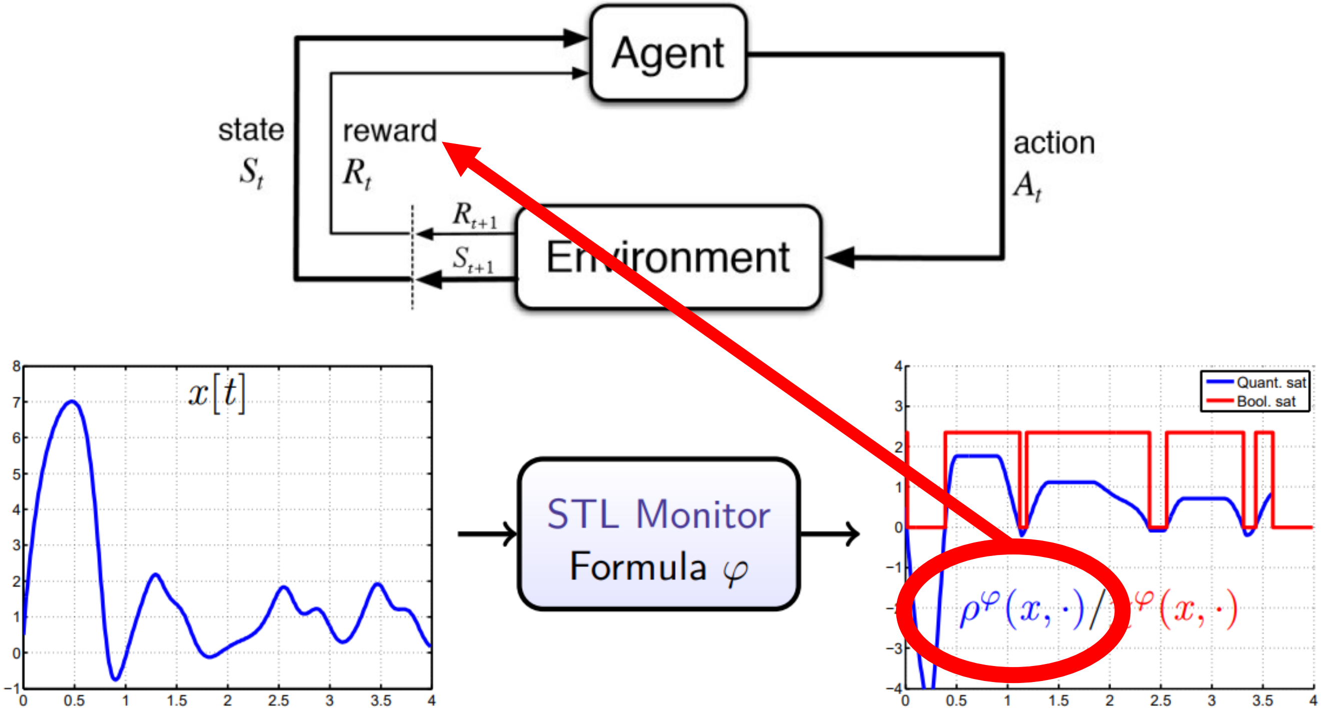
RL with TL rewards





[9] P. Kapoor, A. Balakrishnan, and J. V. Deshmukh (2020). "Model-based Reinforcement Learning from Signal Temporal Logic Specifications". arXiv preprint arXiv:2011.04950.

RL with TL rewards



[9] P. Kapoor, A. Balakrishnan, and J. V. Deshmukh (2020). "Model-based Reinforcement Learning from Signal Temporal Logic Specifications". arXiv preprint arXiv:2011.04950.

RL with TL rewards

- Use the robustness as reward
- Any TL equipped with quantitative semantics can make it!
- For instance, TLTL

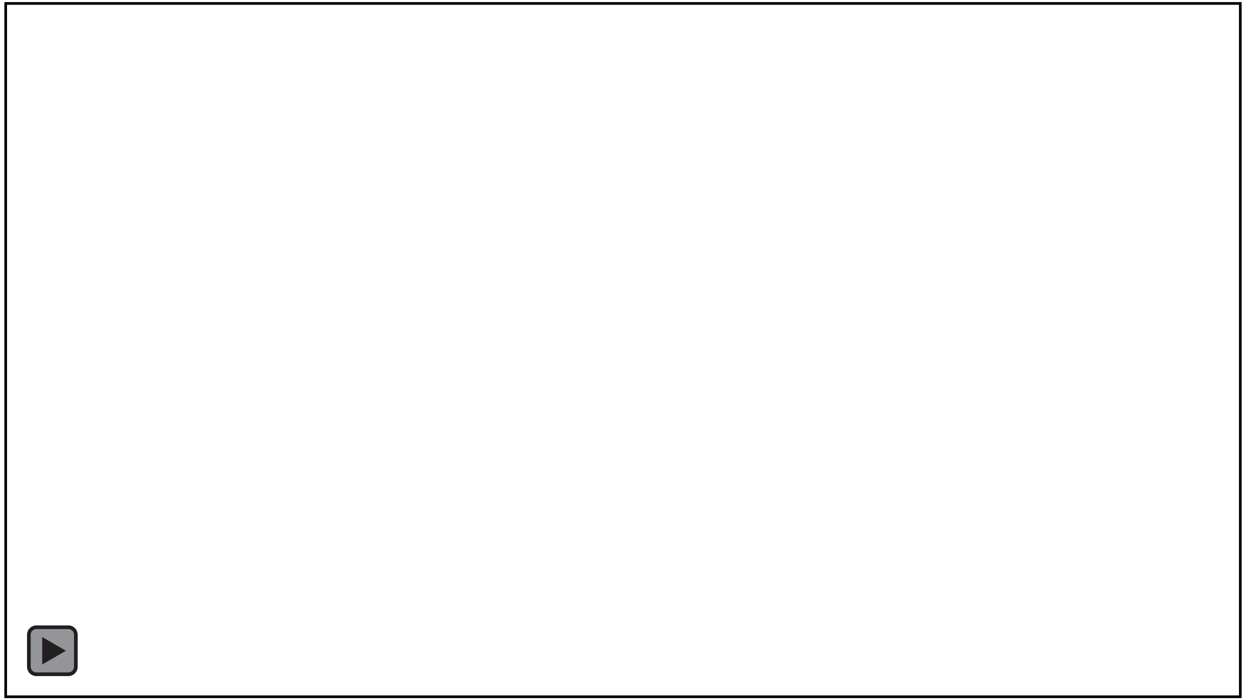
$\varphi ::= f(s) < c \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \mathcal{N}\varphi \mid \varphi_1 \mathcal{U} \varphi_2$
where $f: \mathbb{R}^n \to \mathbb{R}$

Let s_t be the state at time t and $s_{t:t+k}$ be a sequence of states from time t to $t+k$:
$s_{t:t+k} \vDash f(s) < c \quad \leftrightarrow \quad f(s_t) < c$

$$
\begin{aligned}
\rho(s_{t:t+k}, \top) &= \rho_{max}, \\
\rho(s_{t:t+k}, f(s_t) < c) &= c - f(s_t), \\
\rho(s_{t:t+k}, \neg\phi) &= -\rho(s_{t:t+k}, \phi), \\
\rho(s_{t:t+k}, \phi \Rightarrow \psi) &= \max(-\rho(s_{t:t+k}, \phi), \rho(s_{t:t+k}, \psi)) \\
\rho(s_{t:t+k}, \phi_1 \wedge \phi_2) &= \min(\rho(s_{t:t+k}, \phi_1), \rho(s_{t:t+k}, \phi_2)), \\
\rho(s_{t:t+k}, \phi_1 \vee \phi_2) &= \max(\rho(s_{t:t+k}, \phi_1), \rho(s_{t:t+k}, \phi_2)), \\
\rho(s_{t:t+k}, \bigcirc\phi) &= \rho(s_{t+1:t+k}, \phi) \ (k > 0), \\
\rho(s_{t:t+k}, \Box\phi) &= \min_{t' \in [t, t+k)} (\rho(s_{t':t+k}, \phi)), \\
\rho(s_{t:t+k}, \Diamond\phi) &= \max_{t' \in [t, t+k)} (\rho(s_{t':t+k}, \phi)), \\
\rho(s_{t:t+k}, \phi \,\mathcal{U}\, \psi) &= \max_{t' \in [t, t+k)} (\min(\rho(s_{t':t+k}, \psi), \\
&\qquad \min_{t'' \in [t, t')} \rho(s_{t'':t'}, \phi))),
\end{aligned}
$$

[10] X. Li, C. I. Vasile, and C. Belta (2017). "Reinforcement learning with temporal logic rewards". In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3834-3839).1

RL with TL rewards



[10] X. Li, C. I. Vasile, and C. Belta (2017). "Reinforcement learning with temporal logic rewards". In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3834-3839).

More on the topic

[13] Z. Xu and U. Topcu (2019). "Transfer of Temporal Logic Formulas in Reinforcement Learning".
In IJCAI: proceedings of the conference (Vol. 28, p. 4010).

Talk of Ufuk Topcu, University of Texas, USA, at the RL-CONFORM workshop on “Verifiable reinforcement learning systems”
https://youtu.be/dMz14KdGtGs

Conclusion

We covered:

- Shielding in RL, and shield synthesis for perceived safety
- Formal specifications using Linear Temporal Logic and Signal Temporal Logic
- RL with temporal logic rewards

References

[1] F. Berkenkamp and A. Krause. "Tutorial on Safe Reinforcement Learning". Lecture notes. ETH Zürich. 2018. https://las.inf.ethz.ch/files/ewrl18_SafeRL_tutorial.pdf

[2] A. Donzé. "On Signal Temporal Logic". Lecture notes. University of California, Berkley. 2014. https://people.eecs.berkeley.edu/~sseshia/fmee/lectures/EECS294-98_Spring2014_STL_Lecture.pdf

[3] J. Garcıa and F. Fernández. "A comprehensive survey on safe reinforcement learning." Journal of Machine Learning Research 16.1 (2015): 1437-1480.

[4] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu (2018). "Safe Reinforcement Learning via Shielding". In AAAI-18: 32nd AAAI Conference on Artificial Intelligence (pp. 2669-2678).

[5] R. Bloem, B. Könighofer, R. Könighofer, and C. Wang. (2015, April). "Shield synthesis". In International Conference on Tools and Algorithms for the Construction and Analysis of Systems (pp. 533-548)

[6] N. Jansen, B. Könighofer, S. Junges, and R. Bloem, (2018). "Shielded decision-making in MDPs". arXiv preprint arXiv:1807.06096.

References

[7] N. Jansen, B. Könighofer, S. Junges, A. Serban, and R. Bloem. (2018).
“Safe Reinforcement Learning via Probabilistic Shields”. arXiv, arXiv-1807.

[8] Daniel Marta, Christian Pek, Gaspar Isaac Melsión, Jana Tumova, Iolanda Leite, Human-Feedback Shield Synthesis for Perceived Safety in Deep Reinforcement Learning. IEEE Robotics and Automation Letters 7.1 (2021): 406-413.

[9] P. Kapoor, A. Balakrishnan, and J. V. Deshmukh (2020).
"Model-based Reinforcement Learning from Signal Temporal Logic Specifications". arXiv preprint arXiv:2011.04950.

[10] X. Li, C. I. Vasile, and C. Belta (2017). "Reinforcement learning with temporal logic rewards".
In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3834-3839).

[11] O. Maler and D. Nickovic, “Monitoring temporal properties of continuous signals,”
in Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems. Springer, 2004, pp. 152–166.

[12] Donzé, A., Ferrere, T., & Maler, O. (2013, July). Efficient robust monitoring for STL.
In International Conference on Computer Aided Verification (pp. 264-279).

[13] Z. Xu and U. Topcu (2019). "Transfer of Temporal Logic Formulas in Reinforcement Learning".
In IJCAI: proceedings of the conference (Vol. 28, p. 4010).

Talk of Ufuk Topcu, University of Texas, USA, at the RL-CONFORM workshop on “Verifiable reinforcement learning systems”
https://youtu.be/dMz14KdGtGs