Agenda

I. Constrained supervised learning
   - Constrained learning theory
   - Constrained learning algorithms
   - Resilient constrained learning

Break (10 min)

II. Constrained reinforcement learning
   - Constrained RL duality
   - Constrained RL algorithms

Q&A and discussions

https://luizchamon.com/imprs2024

Universität Stuttgart

IMPRS tutorial
Sep. 19, 2024

supervised and reinforcement learning under requirements

Luiz F. O. Chamon

SimTech Cluster of Excellence

KI institute

Constrained reinforcement learning

Agenda

Constrained reinforcement learning

CMDP duality

CRL algorithms

Agenda

Constrained reinforcement learning

CMDP duality

CRL algorithms

Agenda

Constrained reinforcement learning

CMDP duality

CRL algorithms

Primal-dual algorithm

$$D_\theta^\star = \min_{\lambda \succeq 0} \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{s,a \sim \pi_{\boldsymbol{\theta}}} \left[\frac{1}{T} \sum_{t=0}^{T-1} \gamma^t r_0(s_t, a_t) \right] + \lambda \left(\mathbb{E}_{s,a \sim \pi_{\boldsymbol{\theta}}} \left[\frac{1}{T} \sum_{t=0}^{T-1} \gamma^t r_1(s_t, a_t) \right] - c_1 \right)$$

Primal-dual algorithm

$$D_\theta^\star = \min_{\lambda \succeq 0} \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{s,a \sim \pi_{\boldsymbol{\theta}}} \left[\frac{1}{T} \sum_{t=0}^{T-1} \gamma^t r_0(s_t, a_t) \right] + \lambda \left(\mathbb{E}_{s,a \sim \pi_{\boldsymbol{\theta}}} \left[\frac{1}{T} \sum_{t=0}^{T-1} \gamma^t r_1(s_t, a_t) \right] - c_1 \right)$$

- Maximize the primal (≡ vanilla RL)

$$\boldsymbol{\theta}^\dagger \in \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{s,a \sim \pi_{\boldsymbol{\theta}}} \left[\frac{1}{T} \sum_{t=0}^{T-1} \gamma^t r_{\lambda_k}(s_t, a_t) \right]$$

$$r_{\lambda_k}(s,a) = r_0(s,a) + \lambda_k r_1(s,a)$$

Primal-dual algorithm

$$D^\star_\theta = \min_{\lambda \succeq 0} \max_{\theta \in \Theta} \mathbb{E}_{s,a\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] + \lambda\left(\mathbb{E}_{s,a\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right] - c_1\right)$$

- Maximize the primal (≡ vanilla RL)

$$\theta^\dagger \in \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{s,a\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_{\lambda_k}(s_t,a_t)\right]$$

$$r_{\lambda_k}(s,a) = r_0(s,a) + \lambda_k r_1(s,a)$$

- Update the dual (≡ policy evaluation)

$$\lambda_{k+1} = \left[\lambda_k - \eta\left(\mathbb{E}_{s,a\sim\pi_{\theta^\dagger}}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right] - c_1\right)\right]_+$$

In practice...

$$D^\star_\theta = \min_{\lambda \succeq 0} \max_{\theta \in \Theta} \mathbb{E}_{s,a\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] + \lambda\left(\mathbb{E}_{s,a\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right] - c_1\right)$$

- Maximize the primal (≡ vanilla RL): $\{s_t, a_t\} \sim \pi_{\theta_k}$

$$\theta_{k+1} = \theta_k + \eta\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_{\lambda_k}(s_t,a_t)\right]\nabla_\theta \log\left(\pi_\theta(a_0|s_0)\right)$$

- Update the dual (≡ policy evaluation): $\{s_t, a_t\} \sim \pi_{\theta_k}$

$$\lambda_{k+1} = \left[\lambda_k - \eta\left(\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t) - c_1\right)\right]_+$$

Dual CRL

Theorem

Suppose $\theta^\dagger$ is a ρ-approximate solution of the regularized RL problem:

$$\theta^\dagger \approx \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{s,a\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_\lambda(s_t,a_t)\right].$$

Then, after $K = \left\lceil \frac{\|\lambda^\star\|^2}{2\eta\nu} \right\rceil + 1$ dual iterations with step size $\eta \le \frac{1-\gamma}{mB}$,

the iterates (θ_K, λ_K) are such that

$$\left|P^\star - L\left(\theta_K, \lambda_K\right)\right| \le \frac{1+\|\lambda^\star_\nu\|_1}{1-\gamma}B\nu + \rho$$

[Paternain, Chamon, Calvo-Fullana, and Ribeiro, NeurIPS'19; Calvo-Fullana, Paternain, Chamon, and Ribeiro, IEEE TAC'24]

Dual CRL

Theorem

$$\left|P^\star - L\left(\theta_K, \lambda_K\right)\right| \le \frac{1+\|\lambda^\star_\nu\|_1}{1-\gamma}B\nu + \rho$$

Theorem

The state-action sequence $\left\{s_t, a_t \sim \pi^\dagger(\lambda_k)\right\}$ generated by dual CRL is ($\rho = \nu = 0$)

(i) almost surely feasible: $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} r_i(s_t,a_t) \ge c_i$ a.s., for all i

(ii) near-optimal: $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t,a_t)\right] \ge P^\star - \frac{\eta B^2}{2}$

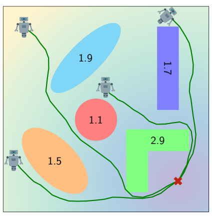
i.e., is a *solution* of the CRL problem (in fact, it is *stronger*: constraints are satisfied a.s.)

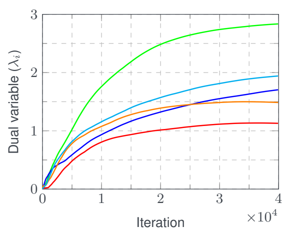
[Paternain, Chamon, Calvo-Fullana, and Ribeiro, NeurIPS'19; Calvo-Fullana, Paternain, Chamon, and Ribeiro, IEEE TAC'24]

Safe navigation

Problem

Find a control policy that navigates the environment effectively and safely



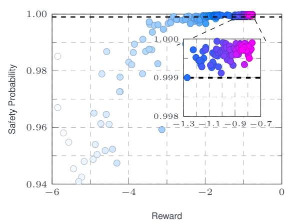


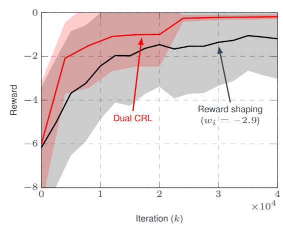
[Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Safe navigation

Problem

Find a control policy that navigates the environment effectively and safely



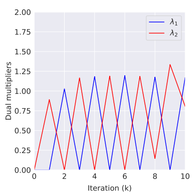


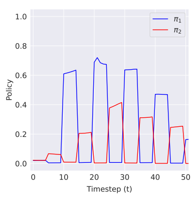
[Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

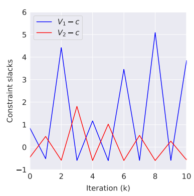
Wireless resource allocation

Problem

Allocate the least transmit power to m device pairs to achieve a communication rate







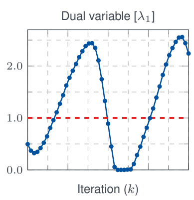
- The dual variables oscillate ⇒ the policy switch ⇒ constraint slacks to oscillate (feasible *on average*)

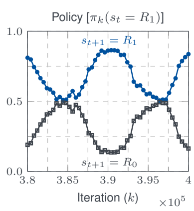
[Uslu, Doostnejad, Ribeiro, NaderiAlizadeh, arxiv:2405.05748]

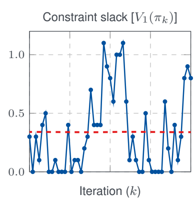
Monitoring task

Problem

Find a policy that maximizes the time in R_0 while monitoring R_1 and R_2 at least $1/3$ of the time each







- The dual variables oscillate ⇒ the policy switch ⇒ constraint slacks to oscillate (feasible *on average*)

[Calvo-Fullana, Paternain, Chamon, and Ribeiro, IEEE TAC'24]

What dual CRL cannot do

Theorem

$$\left|P^\star - L\left(\boldsymbol{\theta}_K, \lambda_T\right)\right| \leq \frac{1 + \|\boldsymbol{\lambda}_\nu^\star\|_1}{1-\gamma} B\nu + \rho$$

Theorem
The state-action sequence $\left\{s_t, a_t \sim \pi^\dagger(\lambda_k)\right\}$ generated by dual CRL is ($\rho = \nu = 0$)

(i) almost surely feasible: $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} r_i(s_t, a_t) \geq c_i$ a.s., for all i

(ii) near-optimal: $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} r_0(s_t, a_t)\right] \geq P^\star - \frac{\eta B^2}{2}$

i.e., is a *solution* of the CRL problem.

$\Rightarrow$ Cannot *effectively* obtain an optimal policy $\pi^\star$ from the sequence of Lagrangian maximizers $\pi^\dagger(\lambda_k)$

[Paternain, Chamon, Calvo-Fullana, and Ribeiro, NeurIPS'19; Calvo-Fullana, Paternain, Chamon, and Ribeiro, IEEE TAC'24]

Primal recovery

- General issue with duality
  - (Primal-)dual methods: $\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$, but $f(\boldsymbol{\theta}_k) \not\to f(\boldsymbol{\theta}^\star)$

✔ Convex optimization $\Rightarrow$ dual averaging

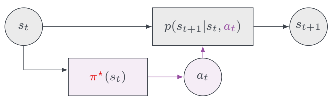
- $f\left(\frac{1}{K}\sum_{k=0}^{K-1} \theta_k\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} f(\theta_k)$ for all K (convexity) $\Rightarrow \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\theta}_k \to \boldsymbol{\theta}^\star$

✘ Non-convex optimization $\Rightarrow$ randomization

- $\boldsymbol{\theta}^\dagger \sim$ Uniform$(\boldsymbol{\theta}_k) \Rightarrow \mathbb{E}\left[f(\boldsymbol{\theta}^\dagger)\right] = \frac{1}{K}\sum_{k=1}^{K} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$

(requires memorizing the whole training sequence)

What we CANNOT do

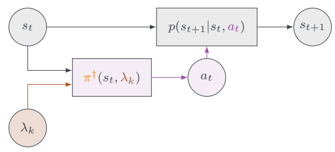


✘ We do not know how to find an optimal policy $\pi^\star$ in the policy space

$$\pi^\star \in \operatorname*{argmax}_{\pi \in \mathcal{P}(\mathcal{S})} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t, a_t)\right]$$
$$\text{subject to } \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_1(s_t, a_t)\right] \geq c_1$$

[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

What we CAN do

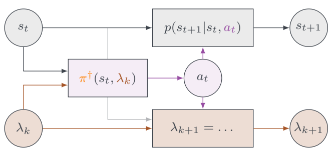


✔ Find Lagrangian maximizing policies $\pi^\dagger(\lambda_k) \Rightarrow$ unconstrained RL problem with reward $r_{\lambda_k}(s, a)$

$$\pi^\dagger(\lambda_k) \in \operatorname*{argmax}_{\pi \in \mathcal{P}(\mathcal{S})} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_{\lambda_k}(s_t, a_t)\right]$$

[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

What we CAN do



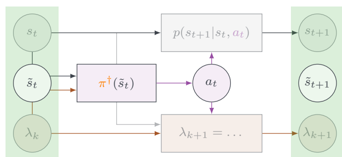
✔ Find Lagrangian maximizing policies $\pi^\dagger(\lambda_k) \Rightarrow$ unconstrained RL problem with reward $r_{\lambda_k}(s, a)$

✔ Update λ_k to generate a sequence of $\pi^\dagger(\lambda_k)$ that are "samples" from $\pi^\star$

$$\lambda_{k+1} = \left[\lambda_k - \eta\left(\mathbb{E}_{s,a\sim\pi^\dagger(\lambda_k)}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_1(s_t, a_t)\right] - c_1\right)\right]_+$$

[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

State-augmented CRL



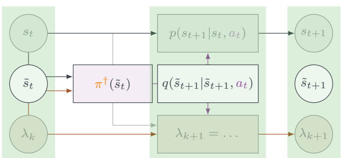
✔ Find Lagrangian maximizing policies $\pi^\dagger(\lambda_k) \Rightarrow$ unconstrained RL problem with reward $r_{\lambda_k}(s, a)$

✔ Update λ_k to generate a sequence of $\pi^\dagger(\lambda_k)$ that are "samples" from $\pi^\star$

$\Rightarrow$ equivalent to an MDP with (augmented) states $\tilde{s} = (s, \lambda)$

[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

State-augmented CRL



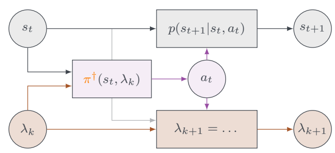
✔ Find Lagrangian maximizing policies $\pi^\dagger(\lambda_k) \Rightarrow$ unconstrained RL problem with reward $r_{\lambda_k}(s, a)$

✔ Update λ_k to generate a sequence of $\pi^\dagger(\lambda_k)$ that are "samples" from $\pi^\star$

$\Rightarrow$ equivalent to an MDP with (augmented) states $\tilde{s} = (s, \lambda)$
and (augmented) transition kernel that includes the dual variables updates

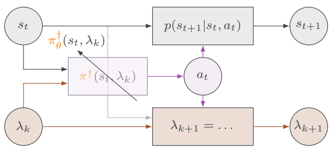
[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

State-augmented CRL in practice

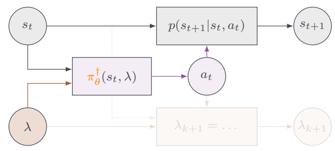


[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

State-augmented CRL in practice

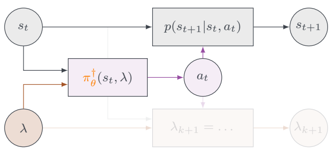


State-augmented CRL in practice



- **During training:** Learn a family of policies $\pi_\theta^\dagger(s, \lambda)$ that maximizes the Lagrangian for all (fixed) λ

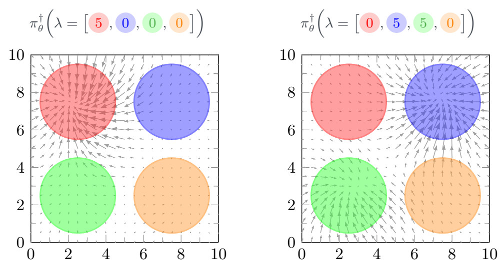
State-augmented CRL in practice



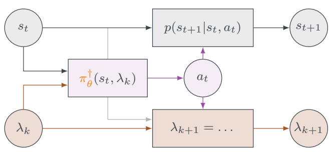
- **During training:** Learn a family of policies $\pi_\theta^\dagger(s, \lambda)$ that maximizes the Lagrangian for all (fixed) λ

$$\pi_\theta^\dagger(\lambda) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \; \mathbb{E}_{\lambda \sim \mathfrak{m}} \left[\lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} r_\lambda(s_t, a_t) \right] \right]$$

Monitoring task



State-augmented CRL in practice

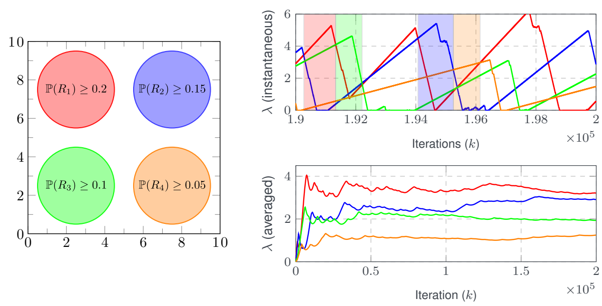


- **During training:** $\pi_\theta^\dagger(\lambda) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} r_\lambda(s_t, a_t) \right]$, for all λ

- **During deployment:** Execute $a_t \sim \pi_\theta^\dagger(\lambda_k)$ for T_0 iterations and update λ_k

$$\lambda_{k+1} = \left[\lambda_k - \frac{\eta}{T_0} \sum_{t=kT_0}^{(k+1)T_0-1} \Big(r_1(s_t, a_t) - c_1 \Big) \right]_+$$

Monitoring task



Solving CRL

A-CRL:
$$\begin{cases} \textbf{Training:} & \pi_\theta^\dagger(\lambda) \in \operatorname{argmax}_{\theta\in\Theta} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} r_\lambda(s_t, a_t) \right], \quad \text{for all } \lambda \\ \textbf{Deployment:} & \lambda_{k+1} = \left[\lambda_k - \frac{\eta}{T_0} \sum_{t=kT_0}^{(k+1)T_0-1} \Big(r_1(s_t, a_t) - c_1 \Big) \right]_+, \quad a_t \sim \pi_\theta^\dagger(\lambda_k) \end{cases}$$

- A-CRL solves (P-CRL) by generating state-action sequences $\{(s_t, a_t)\}$ that are (i) almost surely feasible and (ii) $O(\eta)$-optimal [Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

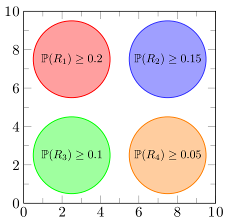
Solving CRL

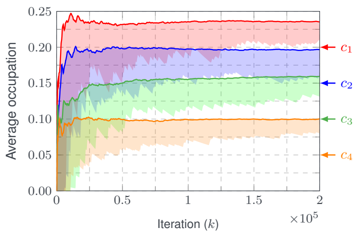
A-CRL:
$$\begin{cases} \textbf{Training:} & \pi_\theta^\dagger(\lambda) \in \operatorname{argmax}_{\theta\in\Theta} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} r_\lambda(s_t, a_t) \right], \quad \text{for all } \lambda \\ \textbf{Deployment:} & \lambda_{k+1} = \left[\lambda_k - \frac{\eta}{T_0} \sum_{t=kT_0}^{(k+1)T_0-1} \Big(r_1(s_t, a_t) - c_1 \Big) \right]_+, \quad a_t \sim \pi_\theta^\dagger(\lambda_k) \end{cases}$$

- A-CRL solves (P-CRL) by generating state-action sequences $\{(s_t, a_t)\}$ that are (i) almost surely feasible and (ii) $O(\eta)$-optimal [Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

- But A-CRL does not find a feasible and $\mathcal{O}(\eta)$-optimal policy $\pi^\star$
  - $\Rightarrow$ It finds a policy $\pi_\theta^\dagger$ on an augmented MDP (s, λ) that generates the same trajectories as dual CRL on the original MDP (s)

Monitoring task



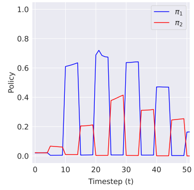


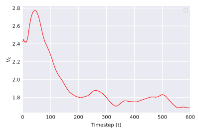
[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

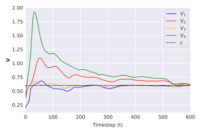
Wireless resource allocation

Problem

Allocate the least transmit power to m device pairs to achieve a communication rate







[Uslu, Doostnejad, Ribeiro, NaderiAlizadeh, arxiv:2405.05748]

Summary

- **Constrained RL is the a tool for decision making under requirements**
- **Constrained RL is hard...**
- **...but possible. How?**

Summary

- **Constrained RL is the a tool for decision making under requirements**

CRL is a natural way of specifying complex behaviors that cannot be handled by unconstrained RL
$\Rightarrow$ (P-RL) $\subsetneq$ (P-CRL)
e.g., safety [Paternain et al., IEEE TAC'23], wireless resource allocation [Eisen et al., IEEE TSP'19; Chowdhury et al., Asilomar'23], monitoring [Calvo-Fullana et al., IEEE TAC'24]

- **Constrained RL is hard...**
- **...but possible. How?**

Summary

- **Constrained RL is the a tool for decision making under requirements**

CRL is a natural way of specifying complex behaviors that cannot be handled by unconstrained RL
$\Rightarrow$ (P-RL) $\subsetneq$ (P-CRL)
e.g., safety [Paternain et al., IEEE TAC'23], wireless resource allocation [Eisen et al., IEEE TSP'19; Chowdhury et al., Asilomar'23], monitoring [Calvo-Fullana et al., IEEE TAC'24]

- **Constrained RL is hard...**

CRL is strongly dual (despite non-convexity), but that is not always enough to obtain feasible solutions
$\Rightarrow$ ~~primal-dual methods~~

- **...but possible. How?**

Summary

- **Constrained RL is the a tool for decision making under requirements**

CRL is a natural way of specifying complex behaviors that cannot be handled by unconstrained RL
$\Rightarrow$ (P-RL) $\subsetneq$ (P-CRL)
e.g., safety [Paternain et al., IEEE TAC'23], wireless resource allocation [Eisen et al., IEEE TSP'19; Chowdhury et al., Asilomar'23], monitoring [Calvo-Fullana et al., IEEE TAC'24]

- **Constrained RL is hard...**

CRL is strongly dual (despite non-convexity), but that is not always enough to obtain feasible solutions
$\Rightarrow$ ~~primal-dual methods~~

- **...but possible. How?**

When combined with a *systematic state augmentation* technique,
we can use policies that solve (P-RL) to solve (P-CRL)

Agenda

I. Constrained supervised learning
   - Constrained learning theory
   - Constrained learning algorithms
   - Resilient constrained learning

Break (10 min)

II. Constrained reinforcement learning
   - Constrained RL duality
   - Constrained RL algorithms

Q&A and discussions

https://luizchamon.com/imprs2024

Survey: Universität Stuttgart

forms.gle/Ja1Ej1Yyyh7BXUMj9

www.luizchamon.com/imprs2024

SimTech Cluster of Excellence

KI institute

IMPRS tutorial
Sep. 19, 2024

supervised and reinforcement learning under requirements