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/sgm

Luiz F. O. Chamon

supervised and reinforcement learning under requirements

SIMPAS group meeting
Apr. 7th, 2025

Constrained reinforcement learning

Agenda

Constrained reinforcement learning

CMDP duality

CRL algorithms

Reinforcement learning

- Model-free framework for decision-making in Markovian settings

Reinforcement learning

- Model-free framework for decision-making in Markovian settings

$$\mathbb{P}\left(s_{t+1} \mid \{s_u, a_u\}_{u\le t}\right) = \mathbb{P}\left(s_{t+1} \mid s_t, a_t\right) = p(s_{t+1} \mid s_t, a_t)$$

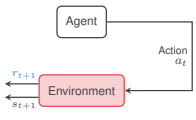
Environment

- MDP: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), p (transition kernel)

Reinforcement learning

- Model-free framework for decision-making in Markovian settings

$$\mathbb{P}\left(s_{t+1} \mid \{s_u, a_u\}_{u\le t}\right) = \mathbb{P}\left(s_{t+1} \mid s_t, a_t\right) = p(s_{t+1} \mid s_t, a_t)$$

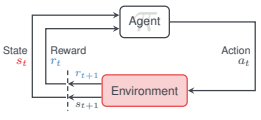


- MDP: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), p (transition kernel), $r : \mathcal{S} \times \mathcal{A} \to [0, B]$ (reward)

Reinforcement learning

- Model-free framework for decision-making in Markovian settings

$$\mathbb{P}\left(s_{t+1} \mid \{s_u, a_u\}_{u\le t}\right) = \mathbb{P}\left(s_{t+1} \mid s_t, a_t\right) = p(s_{t+1} \mid s_t, a_t)$$



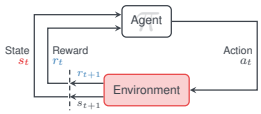
$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}}\ V(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T}\gamma^t r(s_t, a_t)\right] \quad \text{(P-RL)}$$

- MDP: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), p (transition kernel), $r : \mathcal{S} \times \mathcal{A} \to [0, B]$ (reward)
- $\mathcal{P}(\mathcal{S})$: space of probability measures parameterized by $\mathcal{S}$
- T (horizon) (possibly $T \to \infty$) and $\gamma < 1$ (discount factor) (possibly $\gamma = 1$)

Reinforcement learning

- Model-free framework for decision-making in Markovian settings

$$\mathbb{P}\left(s_{t+1} \mid \{s_u, a_u\}_{u \le t}\right) = \mathbb{P}\left(s_{t+1} \mid s_t, a_t\right) = p(s_{t+1} \mid s_t, a_t)$$



$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; V(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right] \quad \text{(P-RL)}$$

- (P-RL) can be solved using policy gradient and/or Q-learning type algorithms

[W'92, WD'92, BT'96, KT'00, JFEPF'14, HKSC'15, NFPIY'15, AJFR'17, PP'18, SB'18, B'19, KCP'19…]

Constrained RL

$$\begin{aligned} \underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & V_0(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \gamma^t r_0(s_t, a_t)\right] \\ \text{subject to} \quad & V_i(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \gamma^t r_i(s_t, a_t)\right] \ge c_i, \quad i = 1, \dots, m \end{aligned} \quad \text{(P-CRL)}$$

- MDP: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), p (transition kernel), $r_i : \mathcal{S} \times \mathcal{A} \to [0, B]$ (reward)
- $\mathcal{P}(\mathcal{S})$: space of probability measures parameterized by $\mathcal{S}$
- T (horizon) (possibly $T \to \infty$) and $\gamma < 1$ (discount factor) (possibly $\gamma = 1$)

[Altman'99; Achiam et al., ICML'17; Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23…]

Safe navigation

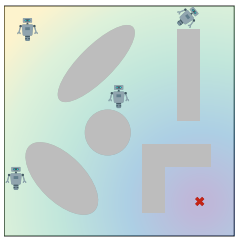
Problem
Find a control policy that navigates the environment effectively and safely

$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\right]$$

Safe navigation

Problem
Find a control policy that navigates the environment effectively and safely

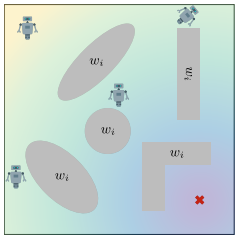
$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0}\right]$$



Safe navigation

Problem
Find a control policy that navigates the environment effectively and safely

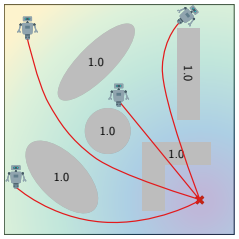
$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0} - \sum_{i=1}^{5} w_i \underbrace{\mathbb{I}(s_t \in \mathcal{O}_i)}_{r_i}\right]$$



Safe navigation

Problem
Find a control policy that navigates the environment effectively and safely

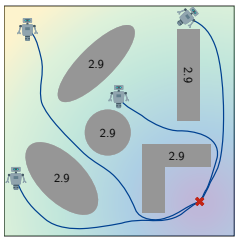
$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0} - \sum_{i=1}^{5} w_i \underbrace{\mathbb{I}(s_t \in \mathcal{O}_i)}_{r_i}\right]$$



Safe navigation

Problem
Find a control policy that navigates the environment effectively and safely

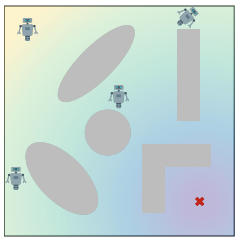
$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0} - \sum_{i=1}^{5} w_i \underbrace{\mathbb{I}(s_t \in \mathcal{O}_i)}_{r_i}\right]$$



Safe navigation

Problem
Find a control policy that navigates the environment effectively and safely

$$\begin{aligned} \underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t, a_t)\right] \\ \text{subject to} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{\mathbb{I}(s_t \notin \mathcal{O}_i)}_{r_i}\right] \ge 1 - \frac{\delta_i}{T} \end{aligned}$$



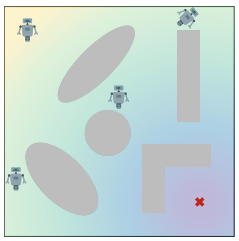
[Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Safe navigation

Problem

Find a control policy that navigates the environment effectively and safely

$$\begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t,a_t)\right] \\ \text{subject to} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{\mathbb{I}(s_t\notin\mathcal{O}_i)}_{r_i}\right] \geq 1-\frac{\delta_i}{T} \end{aligned}$$



- Safety guarantee:

$$\sum_{t=0}^{T-1}\mathbb{P}(\mathcal{E}_t)\geq T-\delta \implies \mathbb{P}\left(\bigcap_{t=0}^{T-1}\mathcal{E}_t\right)\geq 1-\delta$$

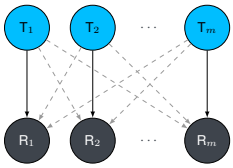
[Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Wireless resource allocation

Problem

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

$$\max_{\pi\in\mathcal{P}(\mathcal{S})} \mathbb{E}_{\boldsymbol{h},\boldsymbol{p}\sim\pi(\boldsymbol{h})}\left[\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{-\sum_{i=1}^{m} p_{i,t}}_{r_0} - \sum_{i=1}^{m} w_i \underbrace{\text{Rate}_i(\boldsymbol{p}_t,\boldsymbol{h}_t)}_{r_i}\right]$$



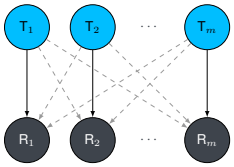
[Eisen, Zhang, Chamon, Lee, and Ribeiro, IEEE TSP'19]

Wireless resource allocation

Problem

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

$$\begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & \mathbb{E}_{\boldsymbol{h},\boldsymbol{p}\sim\pi(\boldsymbol{h})}\left[\frac{1}{T}\sum_{t=0}^{T-1}-\sum_{i=1}^{m} p_{i,t}\right] \\ \text{s. to} \quad & \mathbb{E}_{\boldsymbol{h},\boldsymbol{p}\sim\pi(\boldsymbol{h})}\left[\frac{1}{T}\sum_{t=0}^{T-1}\text{Rate}_i(\boldsymbol{p}_t,\boldsymbol{h}_t)\right]\geq c_i \end{aligned}$$



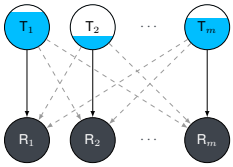
[Chowdhury, Paternain, Verma, Swami, Segarra, Asilomar'23]

Wireless resource allocation

Problem

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

$$\begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & \mathbb{E}_{(\boldsymbol{h},\boldsymbol{b}),\boldsymbol{p}\sim\pi(\boldsymbol{h},\boldsymbol{b})}\left[\frac{1}{T}\sum_{t=0}^{T-1}-\sum_{i=1}^{m}\mathbb{I}(b_{i,t}=0)\right] \\ \text{s. to} \quad & \mathbb{E}_{(\boldsymbol{h},\boldsymbol{b}),\boldsymbol{p}\sim\pi(\boldsymbol{h},\boldsymbol{b})}\left[\frac{1}{T}\sum_{t=0}^{T-1}\text{Rate}_i(\boldsymbol{p}_t,\boldsymbol{h}_t)\right]\geq c_i \end{aligned}$$



[Chowdhury, Paternain, Verma, Swami, Segarra, Asilomar'23]

Constrained RL

$$\begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & V_0(\pi)\triangleq\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] \\ \text{subject to} \quad & V_i(\pi)\triangleq\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_i(s_t,a_t)\right]\geq c_i,\quad i=1,\dots,m \end{aligned} \qquad \text{(P-CRL)}$$

- MDP: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), p (transition kernel), $r_i:\mathcal{S}\times\mathcal{A}\to[0,B]$ (reward)
- $\mathcal{P}(\mathcal{S})$: space of probability measures parameterized by $\mathcal{S}$
- T (horizon) (possibly $T\to\infty$) and $\gamma<1$ (discount factor) (possibly $\gamma=1$)

[Altman'99; Achiam et al., ICML'17; Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23...]

RL $\subsetneq$ CRL

Proposition

There exist environments in which every task cannot be unambiguously described by a reward

- There are tasks that CRL can tackle and RL cannot

$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}}\ V(\pi) \quad \subsetneq \quad \begin{aligned}\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & V_0(\pi)\\ \text{subject to} \quad & V_i(\pi)\geq c_i\end{aligned}$$

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

RL $\subsetneq$ CRL

Proposition

There exist ~~environments~~ in which every ~~task~~ cannot be ~~unambiguously described by a reward~~

(MDPs) (occupation measure) (induced by a unique π^* that maximizes a reward)

- There are tasks that CRL can tackle and RL cannot

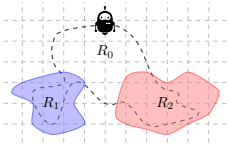
$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}}\ V(\pi) \quad \subsetneq \quad \begin{aligned}\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & V_0(\pi)\\ \text{subject to} \quad & V_i(\pi)\geq c_i\end{aligned}$$

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

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

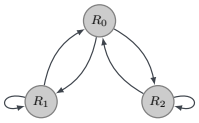


[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'24]

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

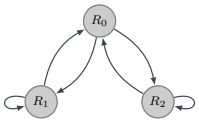


✅ $\pi^\star$ = draw actions uniformly at random

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



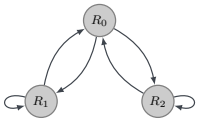
$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \ \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t)\right]$$

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

❌ $r(R_0) > r(R_1), r(R_2)$
$\pi^\dagger$ s.t. $\mathbb{P}\left[s \in R_0\right] = 1/2$



$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \ \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t)\right]$$

Monitoring task

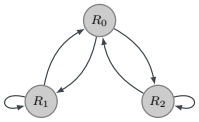
Problem

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

❌ ~~$r(R_0) > r(R_1), r(R_2)$~~

❌ $r(R_1) > r(R_0), r(R_2)$
$\pi^\dagger$ s.t. $\mathbb{P}\left[s \in R_1\right] = 1$

❌ $r(R_2) > r(R_0), r(R_1)$
$\pi^\dagger$ s.t. $\mathbb{P}\left[s \in R_2\right] = 1$



$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \ \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t)\right]$$

Monitoring task

Problem

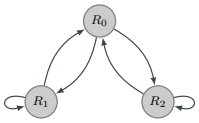
Find a policy that maximizes the time in R_0 while monitoring R_1 and R_2 at least $1/3$ of the time each

❌ ~~$r(R_0) > r(R_1), r(R_2)$~~

❌ ~~$r(R_1) > r(R_0), r(R_2)$~~

❌ ~~$r(R_2) > r(R_0), r(R_1)$~~

✅ $r(R_0) = r(R_1) = r(R_2)$
all $\pi \in \mathcal{P}(\mathcal{S})$ are optimal



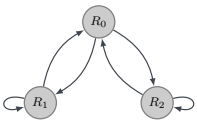
$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \ \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t)\right]$$

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

$$\begin{aligned}
\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\left(s_t \in R_0\right)\right] \\
\text{s. to} \quad & \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\left(s_t \in R_1\right)\right] \geq \frac{1}{3} \\
& \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\left(s_t \in R_2\right)\right] \geq \frac{1}{3}
\end{aligned}$$



RL $\subsetneq$ CRL

Proposition

There exist ~~environments~~ (MDPs) in which every ~~task~~ (occupation measure) cannot be ~~unambiguously described by a reward~~ (induced by a unique $\pi^\star$ that maximizes a reward)

- There are tasks that CRL can tackle and RL cannot

$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \ V(\pi) \quad \subsetneq \quad \begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & V_0(\pi) \\ \text{subject to} \quad & V_i(\pi) \geq c_i \end{aligned}$$

⇒ Regularized RL cannot solve *all* CRL problems

RL $\subsetneq$ CRL

Proposition

There exist ~~environments~~ (MDPs) in which every ~~task~~ (occupation measure) cannot be ~~unambiguously described by a reward~~ (induced by a unique $\pi^\star$ that maximizes a reward)

- There are tasks that CRL can tackle and RL cannot

$$\underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \ V(\pi) \quad \subsetneq \quad \begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & V_0(\pi) \\ \text{subject to} \quad & V_i(\pi) \geq c_i \end{aligned}$$

⇒ Regularized RL cannot solve *all* CRL problems

- How can we tackle CRL problems?

CRL methods

$$\begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] \\ \text{subject to} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_i(s_t,a_t)\right] \geq c_i \end{aligned}$$

- Reward shaping ≈ penalty methods
  - Manual, time-consuming, domain-dependent
  - Trade-offs, training plateaux
- Prior knowledge ≈ projection methods
  e.g., safe exploration [Berkenkamp et al., NeurIPS'17, Dalal et al., arXiv'18]
  - Requires set of safe actions or safe policies
  - Intractable projections
- Linearization and convex surrogates
  e.g., CPO [Achiam et al., ICML'17]
  - No approximation guarantee
  - Approximate problem may be infeasible

CRL methods

$$\begin{aligned} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] \\ \text{subject to} \quad & \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_i(s_t,a_t)\right] \geq c_i \end{aligned}$$

- Reward shaping ≈ penalty methods
- Prior knowledge ≈ projection methods
  e.g., safe exploration [Berkenkamp et al., NeurIPS'17, Dalal et al., arXiv'18]
- Linearization and convex surrogates
  e.g., CPO [Achiam et al., ICML'17]
- Duality
  [Bhatnagar et al., JOTA'12; Tesler et al., ICRL'19; PCCR, NeurIPS'19; Ding et al., NeurIPS'20; PCCR, IEEE TAC'23 . . .]
  - Domain independent
  - Tractable
  - Approximation guarantee [non-convexity]

Agenda

Constrained reinforcement learning

CMDP duality

CRL algorithms

CMDP duality

$$D^\star = \min_{\lambda\geq 0}\max_{\pi\in\mathcal{P}(\mathcal{S})} \overbrace{\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] + \lambda\,\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right]}^{L(\pi,\lambda)}$$

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

- Domain independent ⇐ No hyperparameters tuning
- Tractable ⇐ Equivalent to solving a sequence of unconstrained RL problems
- Approximation guarantee ⇐ ~~$D^\star = P^\star$ (strong duality)~~ ~~[e.g., convex optimization]~~

CMDP duality

$$D^\star = \min_{\lambda\geq 0}\max_{\pi\in\mathcal{P}(\mathcal{S})} \overbrace{\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] + \lambda\,\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right]}^{L(\pi,\lambda)}$$

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

Theorem
If there exists $\pi^\dagger \in \mathcal{P}(\mathcal{S})$ such that $V_i(\pi^\dagger) > c_i$ for all $i = 1,\dots,m$, then $D^\star = P^\star$ (strong duality).

- There is some sort of hidden convexity in CRL ⇒ Occupation measure

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Occupation measure

- The occupation measure of a policy π is the (averaged) probability of visiting each state-action pair

$$\rho_\pi(s,a) = \frac{1-\gamma}{1-\gamma^T}\sum_{t=0}^{T-1}\gamma^t\,\mathbb{P}_{s,a\sim\pi}\big(s_t = s, a_t = a\big) \longleftrightarrow \pi(a|s) = \frac{\rho_\pi(s,a)}{\int_{\mathcal{A}}\rho_\pi(s,a)da}$$

- The value functions $V_i(\pi)$ can be written as an expectation with respect to the ρ_π

$$\begin{aligned} \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{T-1}\gamma^t r_i(s_t,a_t)\right] = V_i(\pi) \propto V(\rho_\pi) &= \mathbb{E}_{(s,a)\sim\rho_\pi}\left[r_i(s,a)\right] \\ &= \int_{\mathcal{S}\times\mathcal{A}} r_i(s,a)\rho_\pi(s,a)\,dsda \end{aligned}$$

  ⇒ The value functions $V_i(\rho_\pi)$ are linear with respect to the occupation measure ρ_π

A non-proof of strong duality

$$\begin{aligned} P^\star = \max_{\pi\in\mathcal{P}(\mathcal{S})} \quad & V_0(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] \\ \text{s. to} \quad & V_1(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right] \geq c \end{aligned} \quad \overset{\equiv}{\underset{(P^\star = P^\star_\rho)}{\uparrow}} \quad \begin{aligned} P^\star_\rho = \max_{\rho\in\mathcal{R}} \quad & V_0(\rho) = \int r_i(s,a)\rho(s,a)dsda \\ \text{s. to} \quad & V_1(\rho) = \int r_1(s,a)\rho(s,a)dsda \geq c \end{aligned}$$

- CRL is *non-convex* in policy space, but *linear* in occupation measure space

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

A non-proof of strong duality

$$\begin{aligned} P^\star = \max_{\pi\in\mathcal{P}(\mathcal{S})} \quad & V_0(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] \\ \text{s. to} \quad & V_1(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right] \geq c \end{aligned} \quad \overset{\equiv}{\underset{(P^\star = P^\star_\rho)}{\uparrow}} \quad \begin{aligned} P^\star_\rho = \max_{\rho\in\mathcal{R}} \quad & V_0(\rho) = \int r_i(s,a)\rho(s,a)dsda \\ \text{s. to} \quad & V_1(\rho) = \int r_1(s,a)\rho(s,a)dsda \geq c \end{aligned}$$

(strongly dual)

- CRL is *non-convex* in policy space, but *linear* in occupation measure space
- CRL in occupation measure space has *no duality gap* (LP)

$$P^\star_\rho = D^\star_\rho = \min_{\lambda\geq 0}\max_{\rho\in\mathcal{R}} V_0(\rho) + \lambda\big(V_1(\rho) - c\big)$$

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

A non-proof of strong duality

$$P^\star = \max_{\pi \in \mathcal{P}(\mathcal{S})} \; V_0(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{T-1} \gamma^t r_0(s_t,a_t)\right] \qquad P^\star_\rho = \max_{\rho\in\mathcal{R}} \; V_0(\rho) = \int r_i(s,a)\rho(s,a)dsda$$

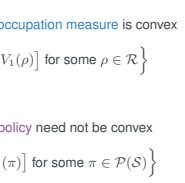
$$\text{s.to} \quad V_1(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{T-1} \gamma^t r_1(s_t,a_t)\right] \geq c \quad \equiv \quad \text{s.to} \quad V_1(\rho) = \int r_1(s,a)\rho(s,a)dsda \geq c$$

~~(strongly dual)~~ ~~$\not\Leftrightarrow$~~ $(P^\star = P^\star_\rho)$ + (strongly dual)

- CRL is *non-convex* in policy space, but *linear* in occupation measure space
- CRL in occupation measure space has *no duality gap* (LP)

$$P^\star_\rho = D^\star_\rho = \min_{\lambda\geq 0} \max_{\rho\in\mathcal{R}} V_0(\rho) + \lambda\big(V_1(\rho) - c\big)$$

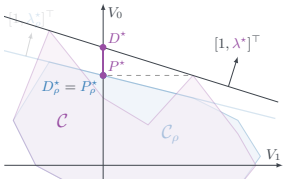
A ~~non~~-proof of strong duality



- Epigraph of CRL in occupation measure is convex

$$\mathcal{C}_\rho = \Big\{ \big[V_0(\rho); V_1(\rho)\big] \text{ for some } \rho \in \mathcal{R} \Big\}$$

A ~~non~~-proof of strong duality



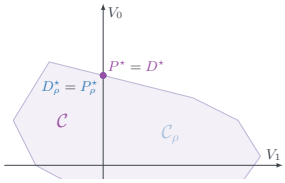
- Epigraph of CRL in occupation measure is convex

$$\mathcal{C}_\rho = \Big\{ \big[V_0(\rho); V_1(\rho)\big] \text{ for some } \rho \in \mathcal{R} \Big\}$$

- Epigraph of CRL in policy need not be convex

$$\mathcal{C} = \Big\{ \big[V_0(\pi); V_1(\pi)\big] \text{ for some } \pi \in \mathcal{P}(\mathcal{S}) \Big\}$$

A ~~non~~-proof of strong duality



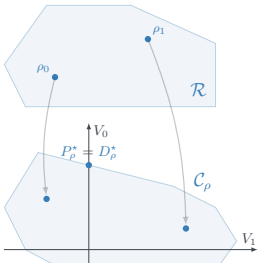
- Epigraph of CRL in occupation measure is convex

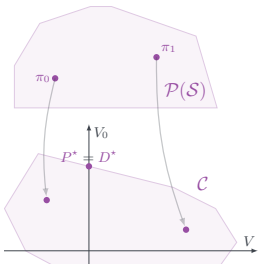
$$\mathcal{C}_\rho = \Big\{ \big[V_0(\rho); V_1(\rho)\big] \text{ for some } \rho \in \mathcal{R} \Big\}$$

- Epigraph of CRL in policy need not be convex

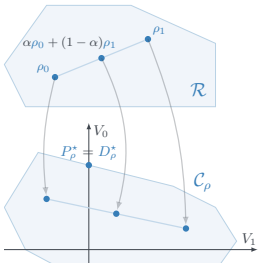
$$\mathcal{C} = \Big\{ \big[V_0(\pi); V_1(\pi)\big] \text{ for some } \pi \in \mathcal{P}(\mathcal{S}) \Big\}$$

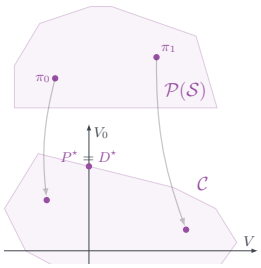
Epigraphs are "convex" in different ways





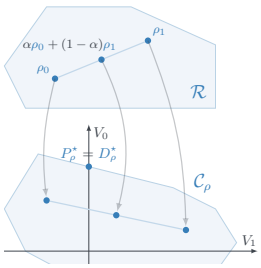
Epigraphs are "convex" in different ways

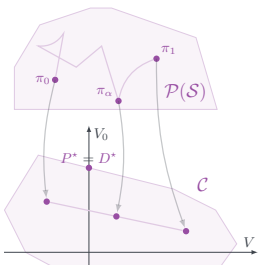




$$V\big(\alpha\rho_1 + (1-\alpha)\rho_2\big) = \alpha V(\rho_1) + (1-\alpha)V(\rho_2)$$

Epigraphs are "convex" in different ways





$V\big(\alpha\rho_1 + (1-\alpha)\rho_2\big) = \alpha V(\rho_1) + (1-\alpha)V(\rho_2)$ $\quad$ $\exists\ \pi_\alpha$ such that $V(\pi_\alpha) = \alpha V(\pi_1) + (1-\alpha)V(\pi_2)$

Strong duality in practice

$$P^\star = D^\star = \min_{\lambda\geq 0} \max_{\pi\in\mathcal{P}(\mathcal{S})} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] + \lambda\,\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right]$$

- Strong duality in policy space $\mathcal{P}(\mathcal{S})$ despite $V_0(\pi)$ and $V(\pi)$ being non-convex

Strong duality in practice

$$P^\star = D^\star = \min_{\lambda \succeq 0} \max_{\pi \in \mathcal{P}(\mathcal{S})} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_0(s_t,a_t)\right] + \lambda\,\mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t r_1(s_t,a_t)\right]$$

$$\updownarrow \Delta$$

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

- Strong duality in policy space $\mathcal{P}(\mathcal{S})$ despite $V_0(\pi)$ and $V(\pi)$ being non-convex
- But in practice, policies are parameterized (π_θ)
  $\Rightarrow$ Introduces a duality gap Δ because standard parametrizations are not convex

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Duality gap of parametrized CRL

Theorem
Let π_θ be ν-universal, i.e.,

$$\min_{\theta\in\Theta} \max_{s\in\mathcal{S}} \int_{\mathcal{A}} \left|\pi(a|s) - \pi_\theta(a|s)\right| da \le \nu, \text{ for all } \pi \in \mathcal{P}(\mathcal{S}).$$

Then,

$$\left|P^\star - D^\star_\theta\right| = \Delta \le \frac{1 + \|\boldsymbol{\lambda}^\star_\nu\|_1}{1-\gamma} B\nu$$

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Duality gap of parametrized CRL

Theorem
Let π_θ be ν-universal, i.e.,

$$\min_{\theta\in\Theta} \max_{s\in\mathcal{S}} \int_{\mathcal{A}} \left|\pi(a|s) - \pi_\theta(a|s)\right| da \le \nu, \text{ for all } \pi \in \mathcal{P}(\mathcal{S}).$$

Then,

$$\left|P^\star - D^\star_\theta\right| = \Delta \le \frac{1 + \|\boldsymbol{\lambda}^\star_\nu\|_1}{1-\gamma} B\nu$$

Sources of error

parametrization richness (ν) requirements difficulty ($\lambda^\star_\nu$) horizon (γ)

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Duality gap of parametrized CRL

Theorem
Let π_θ be ν-universal, i.e.,

$$\min_{\theta\in\Theta} \max_{s\in\mathcal{S}} \int_{\mathcal{A}} \left|\pi(a|s) - \pi_\theta(a|s)\right| da \le \nu, \text{ for all } \pi \in \mathcal{P}(\mathcal{S}).$$

Then,

$$\left|P^\star - D^\star_\theta\right| = \Delta \le \frac{1 + \|\boldsymbol{\lambda}^\star_\nu\|_1}{1-\gamma} B\nu$$

Sources of error

parametrization richness (ν) requirements difficulty ($\lambda^\star_\nu$) horizon (γ)

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Duality gap of parametrized CRL

Theorem
Let π_θ be ν-universal, i.e.,

$$\min_{\theta\in\Theta} \max_{s\in\mathcal{S}} \int_{\mathcal{A}} \left|\pi(a|s) - \pi_\theta(a|s)\right| da \le \nu, \text{ for all } \pi \in \mathcal{P}(\mathcal{S}).$$

Then,

$$\left|P^\star - D^\star_\theta\right| = \Delta \le \frac{1 + \|\boldsymbol{\lambda}^\star_\nu\|_1}{1-\gamma} B\nu$$

Sources of error

parametrization richness (ν) requirements difficulty ($\lambda^\star_\nu$) horizon (γ)

[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19; Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Agenda

Constrained reinforcement learning

CMDP duality

CRL algorithms

Primal-dual algorithm

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

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 ($\equiv$ vanilla RL)

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

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

Primal-dual algorithm

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

- Maximize the primal ($\equiv$ 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 ($\equiv$ 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...

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

- Maximize the primal ($\equiv$ vanilla RL): $\{s_t, a_t\} \sim \pi_{\boldsymbol{\theta}_k}$

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

- Update the dual ($\equiv$ policy evaluation): $\{s_t, a_t\} \sim \pi_{\boldsymbol{\theta}_k}$

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

Dual CRL

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

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

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

the iterates $(\boldsymbol{\theta}_K, \boldsymbol{\lambda}_K)$ are such that

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

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

Dual CRL

Theorem

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

Theorem
The state-action sequence $\left\{s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\right\}$ 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

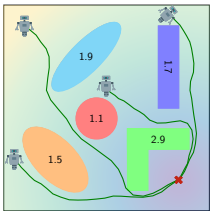
(ii) near-optimal: $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t,a_t)\right] \geq P^\star - \frac{\eta B^2}{2}$

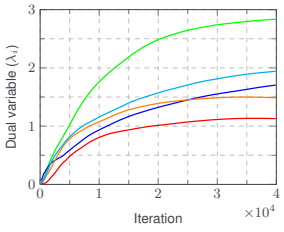
i.e., is a *solution* of the CRL problem (in fact, it is *stronger*: constraints are satisfied a.s.)

[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

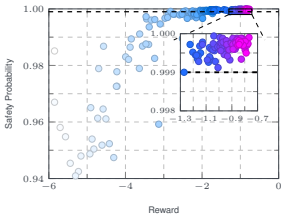


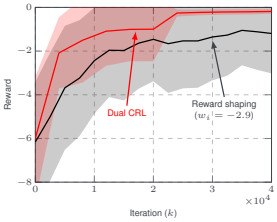


[Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

Safe navigation

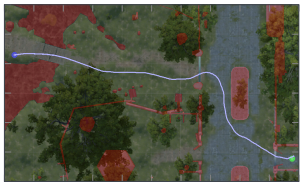
Problem
Find a control policy that navigates the environment effectively and safely





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

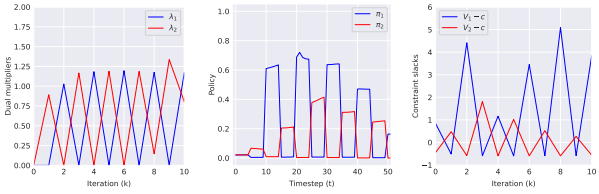
Safe navigation



[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

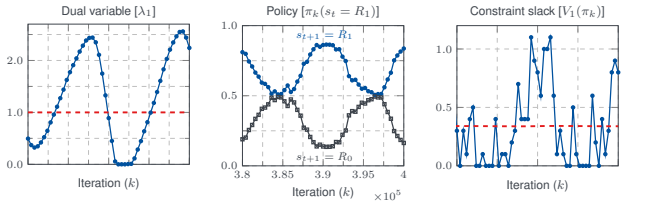


- The dual variables oscillate $\Rightarrow$ the policy switch $\Rightarrow$ constraint slacks to oscillate (feasible *on average*)

[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*)

What dual CRL cannot do

Theorem

$$\left|P^\star - L\left(\theta_K, \lambda_T\right)\right| \leq \frac{1 + \|\lambda_v^*\|_1}{1-\gamma} B\nu + \rho$$

Theorem
The state-action sequence $\{s_t, a_t \sim \pi^\dagger(\lambda_k)\}$ 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.

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

Primal recovery

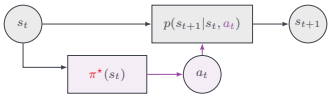
- General issue with duality
  - (Primal-)dual methods: $\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$, but $f(\boldsymbol{\theta}_k) \not\to f(\boldsymbol{\theta}^\star)$

✅ Convex optimization ⇒ dual averaging
  - $f\left(\frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{\theta}_k\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k)$ for all K (convexity) ⇒ $\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_k \to \boldsymbol{\theta}^\star$

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

(requires memorizing the whole training sequence)

What we CANNOT do

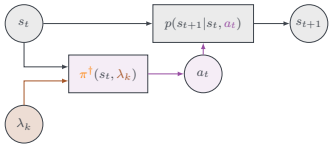


❌ We do not know how to find an optimal policy $\pi^\star$ in the policy space

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

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

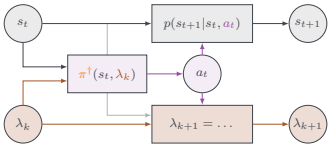
What we CAN do



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

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

What we CAN do

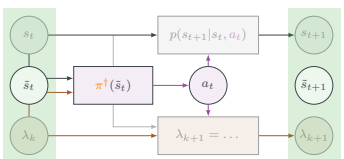


✅ Find Lagrangian maximizing policies $\pi^\dagger(\lambda_k)$ ⇒ unconstrained RL problem with reward $r_{\lambda_k}(s, a)$

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

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

State-augmented CRL

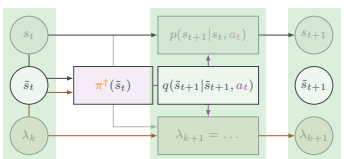


✅ Find Lagrangian maximizing policies $\pi^\dagger(\lambda_k)$ ⇒ unconstrained RL problem with reward $r_{\lambda_k}(s, a)$

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

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

State-augmented CRL

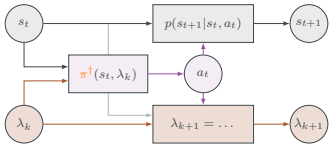


✅ Find Lagrangian maximizing policies $\pi^\dagger(\lambda_k)$ ⇒ unconstrained RL problem with reward $r_{\lambda_k}(s, a)$

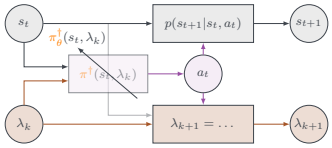
✅ Update λ_k to generate a sequence of $\pi^\dagger(\lambda_k)$ that are "samples" from $\pi^\star$

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

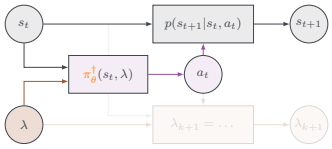
State-augmented CRL in practice



State-augmented CRL in practice

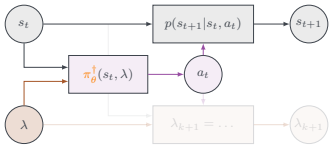


State-augmented CRL in practice



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

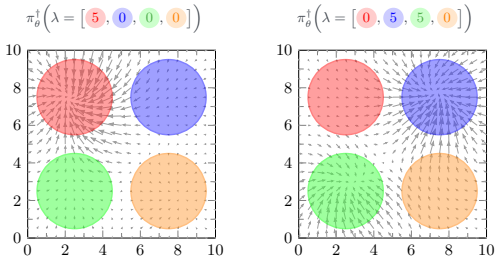
State-augmented CRL in practice



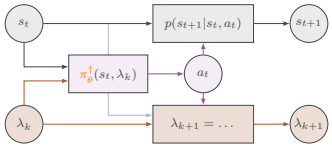
- **During training:** Learn a family of policies $\pi^\dagger_\theta(s, \lambda)$ that maximizes the Lagrangian for all (fixed) λ

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

Monitoring task



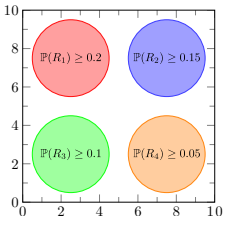
State-augmented CRL in practice

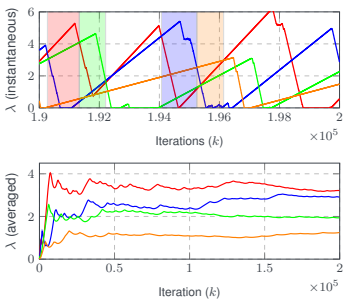


- **During training:** $\pi^\dagger_\theta(\lambda) \in \operatorname*{argmax}_{\theta\in\Theta} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_\lambda(s_t, a_t)\right]$, for all λ
- **During deployment:** Execute $a_t \sim \pi^\dagger_\theta(\lambda_k)$ for T_0 iterations and update λ_k

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

Monitoring task





Solving CRL

A-CRL:

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

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

- 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

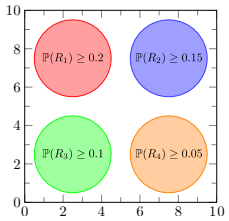
A-CRL:

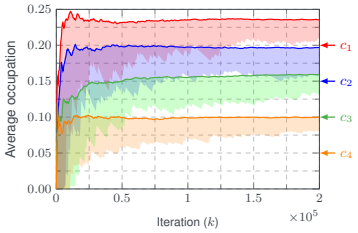
$$\begin{cases} \textbf{Training:} & \pi_\theta^\dagger(\lambda) \in \operatorname{argmax}_{\theta\in\Theta} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_\lambda(s_t,a_t)\right], \quad \text{for all } \lambda \\ \textbf{Deployment:} & \lambda_{k+1} = \left[\lambda_k -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

- 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 $\pi^\star$
  - ⇒ 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)

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

Monitoring task



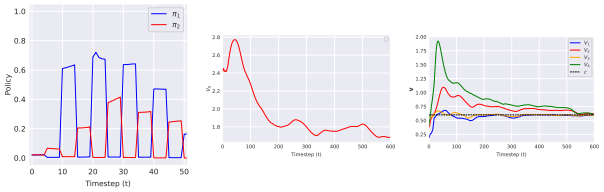


[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

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
  ⇒ (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
  ⇒ (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
  ⇒ ~~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
  ⇒ (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
  ⇒ ~~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/sgm

INSTITUT POLYTECHNIQUE DE PARIS

Luiz F. O. Chamon

www.luizchamon.com
luiz.chamon@polytechnique.edu

supervised and reinforcement learning under requirements