Universitat Pompeu Fabra Barcelona · Universität Stuttgart · Rensselaer · Penn

AAAI tutorial
Feb. 20, 2023

supervised and reinforcement learning under requirements

Miguel Calvo-Fullana
Universitat Pompeu Fabra, Spain

Luiz F. O. Chamon
Universität Stuttgart, Germany

Santiago Paternain
Rensselaer Polytechnic Institute, USA

Alejandro Ribeiro
University of Pennsylvania, USA

Constrained reinforcement learning

Agenda

Constrained reinforcement learning

CMDP duality

Primal-Dual algorithms, state augmentation, guarantees

Agenda

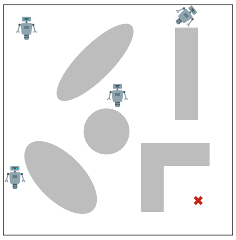
Constrained reinforcement learning

CMDP duality

Primal-Dual algorithms, state augmentation, guarantees

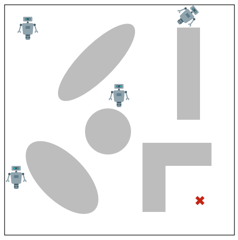
Safe navigation

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



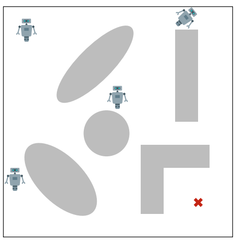
Safe navigation

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



Safe navigation

Problem
~~Safely~~ find a control policy that navigates the environment effectively and safely



Safe navigation

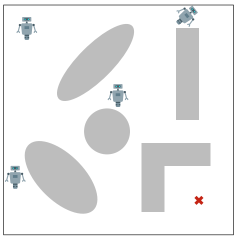
Problem
Find a control policy that navigates the environment effectively and safely

- CBFs, artificial potentials, MPC
  [Koditschek et al., AAM'90; Mayne et al., Autom.'00; Wieland et al., IFAC'07. . .]
  - ✖ knowledge of dynamical system
- System identification
  [Deistler et al., Autom.'95; Tsiamis et al., CDC'19; Dean et al., FCM'19. . .]
  - ✖ "consistency" guarantees for linear systems

Safe navigation

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

- CBFs, artificial potentials, MPC
  [Koditschek et al., AAM'90; Mayne et al., Autom.'00; Wieland et al., IFAC'07...]
  ✖ knowledge of dynamical system
- System identification
  [Deistler et al., Autom.'95; Tsiamis et al., CDC'19; Dean et al., FCM'19...]
  ✖ "consistency" guarantees for linear systems
- RL
  [Bertsekas & Tsitsiklis'96; Sutton & Barto'18; Bertsekas'19...]



Reinforcement learning

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

Reinforcement learning

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

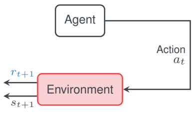
$$\Pr\left(s_{t+1} \mid \{s_u, a_u\}_{u \leq t}\right) = \Pr\left(s_{t+1} \mid s_t, a_t\right) = p(s_{t+1} \mid s_t, a_t)$$

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

Reinforcement learning

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

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

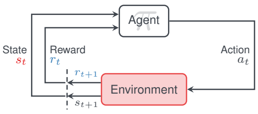


- MDP: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), p (transition kernel), $r : \mathcal{S} \times \mathcal{A} \to [0, B]$ (reward)

Reinforcement learning

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

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



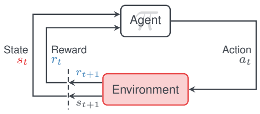
$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \ V(\pi) \triangleq \mathbb{E}_{s,a \sim \pi}\left[\frac{1}{T}\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right] \qquad \text{(P-RL)}$$

- 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

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



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

- (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

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

- 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...]

Safe navigation

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

$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \ V(\pi)$$

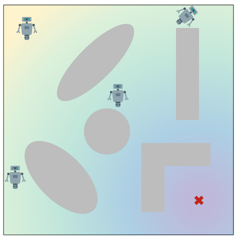
$$r(s, a) =$$

Safe navigation

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

$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; V(\pi)$$

$$r(s,a) = \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0}$$

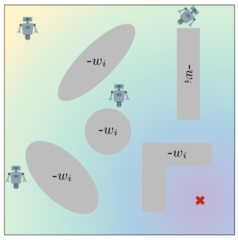


Safe navigation

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

$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; V(\pi)$$

$$r(s,a) = \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0} + \sum_{i=1}^{5} \underbrace{w_i \mathbb{I}(s_t \in \mathcal{O}_i)}_{r_i}$$

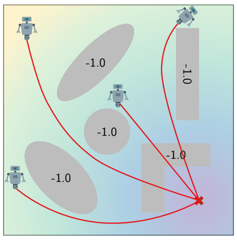


Safe navigation

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

$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; V(\pi)$$

$$r(s,a) = \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0} + \sum_{i=1}^{5} \underbrace{w_i \mathbb{I}(s_t \in \mathcal{O}_i)}_{r_i}$$

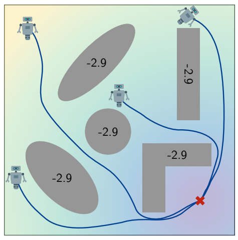


Safe navigation

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

$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; V(\pi)$$

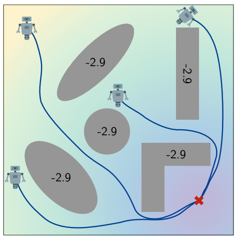
$$r(s,a) = \underbrace{-\|s - s_{\text{goal}}\|^2}_{r_0} + \sum_{i=1}^{5} \underbrace{w_i \mathbb{I}(s_t \in \mathcal{O}_i)}_{r_i}$$



Safe navigation

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

- CBFs, artificial potentials, MPC
  [Koditschek et al., AAM'90; Mayne et al., Autom.'00; Wieland et al., IFAC'07...]
  ✖ knowledge of dynamical system
- System identification
  [Deistler et al., Autom.'95; Tsiamis et al., CDC'19; Dean et al., FCM'19...]
  ✖ "consistency" guarantees for linear systems
- RL with reward shaping
  [Bertsekas & Tsitsiklis'96; Sutton & Barto'18; Bertsekas'19...]
  ✖ weak guarantee



Safe navigation

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

$$\begin{aligned} &\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} && \text{Task reward} \\ &\text{subject to} && \Pr\left(\text{Not colliding with } \mathcal{O}_i\right) \geq 1 - \delta, \quad i = 1, 2, \dots \end{aligned}$$

Safe navigation

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

$$\begin{aligned} &\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} && V_0(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t, a_t)\right] \\ &\text{subject to} && \Pr\left(\text{Not colliding with } \mathcal{O}_i\right) \geq 1 - \delta, \quad i = 1, 2, \dots \end{aligned}$$

Safe navigation

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

$$\begin{aligned} &\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} && V_0(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t, a_t)\right] \\ &\text{subject to} && \Pr\left(\bigcap_{t=0}^{T-1} \{s_t \notin \mathcal{O}_i\} \;\middle|\; \pi\right) \geq 1 - \delta_i, \quad i = 1, 2, \dots \end{aligned}$$

- Probabilistic version of control invariant sets

Safe navigation

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

$$\begin{aligned}
&\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} && V_0(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t,a_t)\right] \\
&\text{subject to} && V_i(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{\mathbb{I}(s_t \notin \mathcal{O}_i)}_{r_i}\right] \geq 1 - \frac{\delta_i}{T}, \quad i = 1, 2, \dots
\end{aligned}$$

- Probabilistic version of control invariant sets
- Constraint tightening: $\Pr\left(\bigcap_{t=0}^{T-1} \mathcal{E}_t\right) \geq 1 - \delta \Longleftarrow \sum_{t=0}^{T-1} \Pr(\mathcal{E}_t) \geq T - \delta$

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

Constrained RL

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

- 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. . .]

CRL methods

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

- 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

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

- 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

Primal-Dual algorithms, state augmentation, guarantees

Duality

DUAL

↕

PRIMAL

Duality

DUAL

↕

PRIMAL

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

Duality

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

DUAL

↕

PRIMAL

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

Duality

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

$$\updownarrow$$

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

- $D^\star = \min_{\lambda\succeq 0}\max_{\pi\in\mathcal{P}(\mathcal{S})} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1}\gamma^t\big(r_0(s_t,a_t)+\lambda r_1(s_t,a_t)\big)\right]$
- No hyperparameters to be tuned in the problem $\Rightarrow$ Domain Independent
- Equivalent to solving a sequence a unconstrained RL problems $\Rightarrow$ Tractable

Duality

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

$$\updownarrow$$

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

- Approximation guarantees?
- In general, $D^\star \geq P^\star$

Duality

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

$$\updownarrow$$

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

- Approximation guarantees?
- In general, $D^\star \geq P^\star$
- But in some cases, $D^\star = P^\star$ (strong duality) [e.g., convex optimization]

Duality

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

$$\updownarrow$$

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

- Approximation guarantees?
- In general, $D^\star \geq P^\star$
- ~~But in some cases, $D^\star = P^\star$ (strong duality) [e.g., convex optimization]~~

Strong duality of CRL

Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)
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 $D^\star = P^\star$.

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

Strong duality of CRL

Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)
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 $D^\star = P^\star$.

- **Non-proof:** There is an equivalent linear program

(P-CRL) $\equiv$ LP : $\rho_\pi(s,a) = \dfrac{1-\gamma}{1-\gamma^T}\displaystyle\sum_{t=0}^{T-1}\gamma^t \Pr_\pi\big(s_t = s, a_t = a\big) \longleftrightarrow \pi(a|s) = \dfrac{\rho_\pi(s,a)}{\int_{\mathcal{A}}\rho_\pi(s,a)da}$

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

$$\begin{array}{ll} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} & V_0(\pi) \\ \text{subject to} & V_i(\pi) \geq c_i \end{array} \quad\equiv\quad \begin{array}{ll} \underset{\rho\in\mathcal{P}}{\text{maximize}} & \mathbb{E}_{(s,a)\sim\rho}\big[r_0(s,a)\big] \\ \text{subject to} & \mathbb{E}_{(s,a)\sim\rho}\big[r_i(s,a)\big] \geq \bar{c}_i \end{array}$$

(strongly dual)

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

Strong duality of CRL

Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)
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 $D^\star = P^\star$.

- **Non-proof:** There is an equivalent linear program

$\not\Leftarrow$ (P-CRL) $\equiv$ LP : $\rho_\pi(s,a) = \dfrac{1-\gamma}{1-\gamma^T}\displaystyle\sum_{t=0}^{T-1}\gamma^t \Pr_\pi\big(s_t = s, a_t = a\big) \longleftrightarrow \pi(a|s) = \dfrac{\rho_\pi(s,a)}{\int_{\mathcal{A}}\rho_\pi(s,a)da}$

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

$$\begin{array}{ll} \underset{\pi\in\mathcal{P}(\mathcal{S})}{\text{maximize}} & V_0(\pi) \\ \text{subject to} & V_i(\pi) \geq c_i \end{array} \quad\equiv\quad \begin{array}{ll} \underset{\rho\in\mathcal{P}}{\text{maximize}} & \mathbb{E}_{(s,a)\sim\rho}\big[r_0(s,a)\big] \\ \text{subject to} & \mathbb{E}_{(s,a)\sim\rho}\big[r_i(s,a)\big] \geq \bar{c}_i \end{array}$$

~~(strongly dual)~~ $\not\Leftarrow$ (strongly dual)

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

Counterexample (1)

- Consider the following equivalent optimization problems

$$\begin{array}{rl} P^\star = \max\limits_{x} & -x \\ \text{subject to} & x^2 - 1 \geq 0 \\ & x \geq 0 \end{array} \quad\equiv\quad \begin{array}{rl} P^\star_{LP} = \max\limits_{x} & -x \\ \text{subject to} & x - 1 \geq 0 \end{array}$$

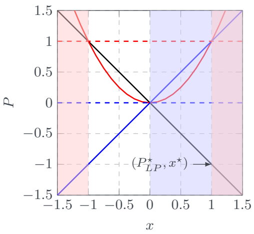
- They have the same objective and the same feasible set $x \geq 1 \Rightarrow$ Equivalent problems

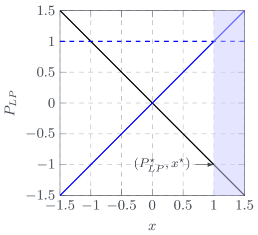
$$x^\star = 1, \qquad P^\star = P^\star_{LP} = -1$$

- Problem P_{LP} is convex (Linear Program) $\Rightarrow$ Zero duality gap
- Problem P is not convex $\Rightarrow$ Zero duality gap?

Counterexample (2)

$$P^\star = \max_x \; -x \quad \text{subject to} \quad x^2 - 1 \geq 0, \quad x \geq 0 \qquad \equiv \qquad P_{LP}^\star = \max_x \; -x \quad \text{subject to} \quad x - 1 \geq 0$$





Counterexample (3)

- Let us solve the dual problem of the LP first

$$P_{LP}^\star = \max_x \; -x \qquad = -1 \qquad \text{subject to} \quad x - 1 \geq 0$$

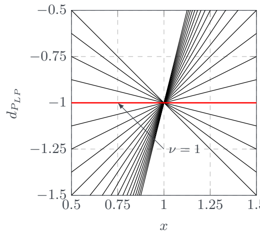
- The dual function is ($\nu \geq 0$)

$$d_{P_{LP}}(\nu) = \max_x -x + \nu(x-1) = \begin{cases} -1 & \nu = 1 \\ \infty & \text{if } \nu \neq 1 \end{cases}$$

- The solution to the dual problem is

$$D_{LP}^\star = \min_{\nu \geq 0} d_{P_{LP}}(\nu) = -1$$

- We have $D_{LP}^\star = P_{LP}^\star \Rightarrow$ no duality gap



Counterexample (4)

- Let us solve the dual problem of the non-convex problem

$$P^\star = \max_x \; -x = \qquad -1 \qquad \text{subject to} \quad x^2 - 1 \geq 0, \quad x \geq 0$$

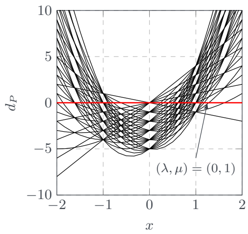
- The dual function is ($\lambda, \mu \geq 0$)

$$d_P(\lambda, \mu) = \max_x -x + \lambda(x^2 - 1) + \mu x = \begin{cases} 0 & \text{if } \lambda = 0, \mu = 1 \\ \infty & \text{otherwise} \end{cases}$$

- The solution to the dual problem is

$$D_P^\star = \min_{\lambda, \mu \geq 0} d_P(\lambda, \mu) = 0$$

- We have $D_{LP}^\star \neq P_{LP}^\star \Rightarrow$ There is duality gap



Proof outline

- The proof of the result is based on geometric arguments

$$P^\star \triangleq \max_{\pi \in \mathcal{P}(\mathcal{S})} \; V_0(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_0(s_t, a_t)\right]$$
$$\text{subject to} \quad V_i(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t)\right] \geq c_i, i = 1, \ldots, m.$$

- Define the epigraph set

$$\mathcal{C} = \left\{\xi \in \mathbb{R}^{m+1} \mid \exists \pi \text{ s.t.} V_i(\pi) \geq \xi_i \text{ for all } i = 0, \ldots, m\right\}$$

  - The set $\mathcal{C}$ is convex $\Rightarrow$ Zero duality gap follows the same arguments as in convex optimization
  - The supporting hyper-plane at $(P^\star, \mathbf{0})$ is defined by the optimal Lagrange multipliers

$$P^\star + \sum_{i=1}^m \lambda_i \mathbf{0} \geq \xi_0 + \sum_{i=1}^m \lambda_i \xi_i \geq V_0(\pi) + \sum_{i=1}^m \lambda_i V_i(\pi)$$

  - This implies strong duality $P^\star \geq D^\star$

Dual CRL

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

DUAL

- $D^\star = P^\star$ (strong duality) [despite non-convexity]

Dual CRL

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

DUAL

- $D^\star = P^\star$ (strong duality) [despite non-convexity]
- Infinite dimensionality of $\mathcal{P}(\mathcal{S})$

Dual CRL

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

DUAL

- $D^\star = P^\star$ (strong duality) [despite non-convexity]
- ~~Infinite dimensionality of $\mathcal{P}(\mathcal{S})$~~ Finite dimensional parametrization π_θ

Dual CRL

Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)

Let π_θ be ν-universal, i.e.,

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

Then,

$$\left|P^\star - D_\theta^\star\right| \leq \frac{1 + \|\lambda_\nu^\star\|_1}{1 - \gamma} B\nu$$

Alternative: $|P_\theta^\star - D_\theta^\star|$ can be bounded using ν-universality only over $\pi \in \overline{\text{conv}}\left(\{\pi_\theta | \theta \in \Theta\}\right)$

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

Dual CRL

Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)

Let $\pi_{\boldsymbol{\theta}}$ be ν-universal, i.e.,

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

Then,

$$\left|P^\star - D_\theta^\star\right| \leq \frac{1 + \|\lambda_\nu^\star\|_1}{1-\gamma} B\nu$$

Alternative: $|P_\theta^\star - D_\theta^\star|$ can be bounded using ν-universality only over $\pi \in \overline{\operatorname{conv}}\left(\{\pi_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \Theta\}\right)$

Sources of error

parametrization richness (ν) requirements difficulty ($\lambda_\nu^\star$) horizon (γ)

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

Dual CRL

Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)

Let $\pi_{\boldsymbol{\theta}}$ be ν-universal, i.e.,

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

Then,

$$\left|P^\star - D_\theta^\star\right| \leq \frac{1 + \|\lambda_\nu^\star\|_1}{1-\gamma} B\nu$$

Alternative: $|P_\theta^\star - D_\theta^\star|$ can be bounded using ν-universality only over $\pi \in \overline{\operatorname{conv}}\left(\{\pi_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \Theta\}\right)$

Sources of error

parametrization richness (ν) requirements difficulty ($\lambda_\nu^\star$) horizon (γ)

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

Dual CRL

Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)

Let $\pi_{\boldsymbol{\theta}}$ be ν-universal, i.e.,

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

Then,

$$\left|P^\star - D_\theta^\star\right| \leq \frac{1 + \|\lambda_\nu^\star\|_1}{1-\gamma} B\nu$$

Alternative: $|P_\theta^\star - D_\theta^\star|$ can be bounded using ν-universality only over $\pi \in \overline{\operatorname{conv}}\left(\{\pi_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \Theta\}\right)$

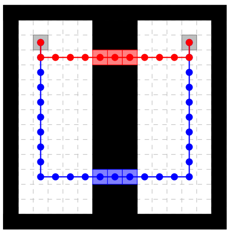
Sources of error

parametrization richness (ν) requirements difficulty ($\lambda_\nu^\star$) horizon (γ)

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

Grid world example

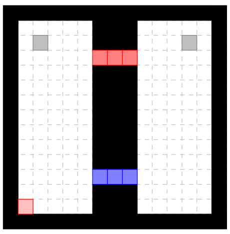
- Consider a grid world with a safe and an unsafe bridge
  - Only two potentially optimal policies depending on the cost of crossing each bridge

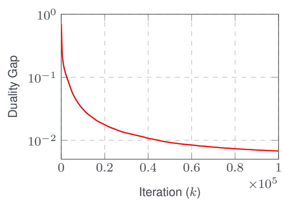


[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19]

Grid world example

- Consider a grid world with a safe and an unsafe bridge
  - Only two potentially optimal policies depending on the cost of crossing each bridge

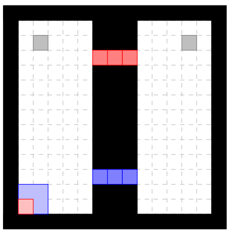


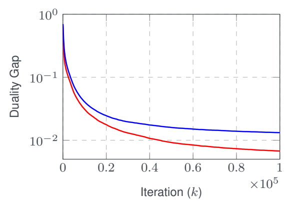


[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19]

Grid world example

- Consider a grid world with a safe and an unsafe bridge
  - Only two potentially optimal policies depending on the cost of crossing each bridge

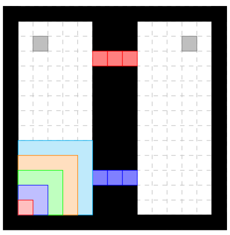


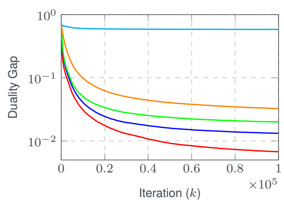


[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19]

Grid world example

- Consider a grid world with a safe and an unsafe bridge
  - Only two potentially optimal policies depending on the cost of crossing each bridge





[Paternain, Chamon, Calvo-Fullana, Ribeiro, NeurIPS'19]

Dual CRL

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

- $D^\star = P^\star$ (strong duality) [despite non-convexity]

- Infinite dimensionality of $\mathcal{P}(\mathcal{S})$ Finite dimensional parametrization $\pi_{\boldsymbol{\theta}}$

  $\pi_{\boldsymbol{\theta}}$ is ν-universal $\Rightarrow \left|P^\star - D_\theta^\star\right| \leq O(\nu)$

Agenda

Constrained reinforcement learning

CMDP duality

Primal-Dual algorithms, state augmentation, guarantees

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 ($\equiv$ vanilla RL)

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

$$r_\lambda(s,a) = r_0(s,a) + \lambda r_1(s,a)$$

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 ($\equiv$ vanilla RL)

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

$$r_\lambda(s,a) = r_0(s,a) + \lambda r_1(s,a)$$

- Update the dual ($\equiv$ policy evaluation)

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

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 ($\equiv$ vanilla RL)

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

$$r_\lambda(s,a) = r_0(s,a) + \lambda r_1(s,a)$$

- Update the dual ($\equiv$ policy evaluation)

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

In practice...

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

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

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

- Update the dual ($\equiv$ policy evaluation): $\{s_t, a_t\} \sim \pi_{\boldsymbol{\theta}_{k+1}}$

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

Dual CRL

Theorem

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

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

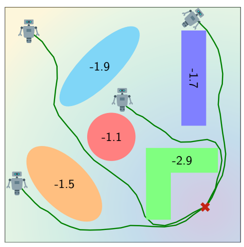
Then, after $K = \left\lceil \frac{\|\boldsymbol{\lambda}^\star\|^2}{2\eta\nu} \right\rceil + 1$ dual iterations with step size $\eta \leq \frac{1-\gamma}{mB}$,

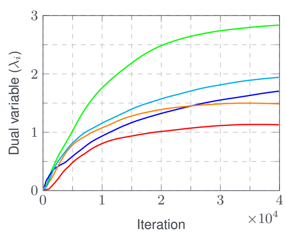
the iterates $\left(\boldsymbol{\theta}^{(T)}, \boldsymbol{\lambda}^{(T)}\right)$ are such that

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

[Paternain, C., Calvo-Fullana, and Ribeiro, NeurIPS'19; C. and Ribeiro, NeurIPS'20; C., Paternain, Calvo-Fullana, and Ribeiro, IEEE TIT'23]

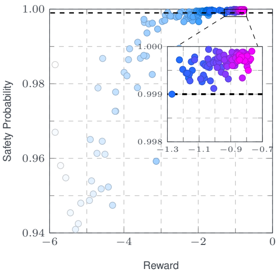
Safe navigation

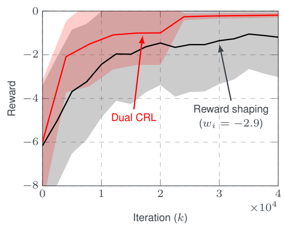




[Paternain, Calvo-Fullana, Chamon, Ribeiro, IEEE TAC'23]

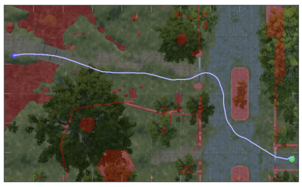
Safe navigation





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

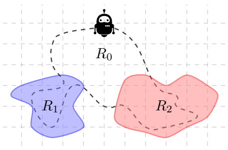
Safe navigation



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

Monitoring task

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

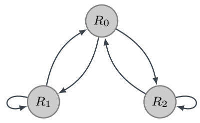


[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

Monitoring task

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

$$\max_{\pi \in \mathcal{P}(\mathcal{S})} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\big(s_t \in R_0\big)\right]$$
$$\text{s. to } \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\big(s_t \in R_i\big)\right] \geq \frac{1}{3}$$

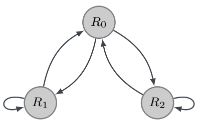


[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

Monitoring task

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

$$\max_{\pi \in \mathcal{P}(\mathcal{S})} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\big(s_t \in R_0\big)\right]$$
$$\text{s. to } \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\big(s_t \in R_i\big)\right] \geq \frac{1}{3}$$



- ✓ $\pi^\star =$ draw actions uniformly at random

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

Monitoring task

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

$$\max_{\pi \in \mathcal{P}(\mathcal{S})} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\big(s_t \in R_0\big)\right]$$
$$\text{s. to } \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\big(s_t \in R_i\big)\right] \geq \frac{1}{3}$$

$$\max_{\pi \in \mathcal{P}(\mathcal{S})} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_{\boldsymbol{\lambda}}(s_t)\right]$$
$$r_{\boldsymbol{\lambda}}(s) = \mathbb{I}\big(s \in R_0\big) + \lambda_1 \mathbb{I}\big(s \in R_1\big) + \lambda_2 \mathbb{I}\big(s \in R_2\big)$$

- ✓ $\pi^\star =$ draw actions uniformly at random

- ✓ $\lambda_1 = \lambda_2 = 1$: all $\pi \in \mathcal{P}(\mathcal{S})$ are optimal
- ✗ $\lambda_1, \lambda_2 < 1$: $\pi^\star$ s.t. $\Pr\left[s \in R_0\right] = 1/2$
- ✗ $\lambda_i > 1$ and $\lambda_i > \lambda_j$: $\pi^\star$ s.t. $\Pr\left[s \in R_i\right] = 1$

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

So CRL is hard?

- There are tasks that CRL can tackle and RL cannot

$$\begin{array}{ll} \max\limits_{\pi \in \mathcal{P}(\mathcal{S})} & V_0(\pi) \\ \text{subject to} & V_i(\pi) \geq c_i \end{array} \quad \supsetneq \quad \max_{\pi \in \mathcal{P}(\mathcal{S})} V(\pi)$$

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

So CRL is hard?

- There are tasks that CRL can tackle and RL cannot

$$\begin{array}{ll} \max\limits_{\pi \in \mathcal{P}(\mathcal{S})} & V_0(\pi) \\ \text{subject to} & V_i(\pi) \geq c_i \end{array} \quad \supsetneq \quad \max_{\pi \in \mathcal{P}(\mathcal{S})} V(\pi)$$

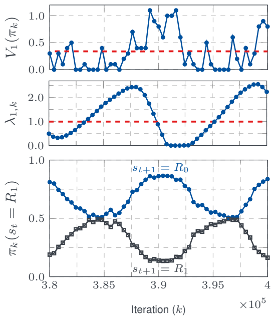
- Dual CRL cannot solve all CRL problems

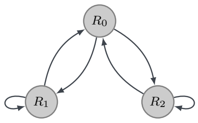
Theorem (Paternain, Chamon, Calvo-Fullana, Ribeiro'19)
If $\pi_{\boldsymbol{\theta}}$ is ν-universal, then $|P^\star - D_\theta^\star| \leq O(\nu)$.

$$\implies \exists\, \boldsymbol{\theta}^\dagger \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ V_0(\pi_{\boldsymbol{\theta}}) + \sum_{i=1}^m \lambda_i^\star V_i(\pi_{\boldsymbol{\theta}})$$ that is approximately feasible.

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

So CRL is hard?





Primal recovery

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

Primal recovery

- General issue with duality
  - (Primal-)dual methods: $f(\boldsymbol{\theta}_k) \not\to f(\boldsymbol{\theta}^\star)$ but $\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$
- Convex optimization ⇒ dual averaging
  - Convexity: $f\left(\frac{1}{K}\sum_{k=0}^{K-1}\theta_k\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} f(\theta_k)$ for all K ⇒ $\boldsymbol{\theta}^\star = \lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{\theta}_k$

Primal recovery

- General issue with duality
  - (Primal-)dual methods: $f(\boldsymbol{\theta}_k) \not\to f(\boldsymbol{\theta}^\star)$ but $\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$
- Convex optimization ⇒ dual averaging
  - Convexity: $f\left(\frac{1}{K}\sum_{k=0}^{K-1}\theta_k\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} f(\theta_k)$ for all K ⇒ $\boldsymbol{\theta}^\star = \lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{\theta}_k$
- Non-convex optimization ⇒ randomization
  - $\boldsymbol{\theta}^\dagger \sim$ Uniform$(\boldsymbol{\theta}_k) \Rightarrow \mathbb{E}\left[f(\boldsymbol{\theta}^\dagger)\right] = \frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$

Intuition

$$\begin{cases} \boldsymbol{\theta}_k \in \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} V_{\lambda_k}(\pi_{\boldsymbol{\theta}}), \quad V_\lambda(\pi) = V_0(\pi) + \lambda V_1(\pi) \\ \lambda_{k+1} = \left[\lambda_k - \eta\big(V_1(\pi_{\boldsymbol{\theta}_k}) - c_1\big)\right]_+ \end{cases}$$

- Only the ergodic average of (approximate) dual ascent iterates converges

$$V_i(\pi_{\boldsymbol{\theta}_k}) \not\to V_i(\pi_{\boldsymbol{\theta}^\star}) \quad \text{but} \quad \frac{1}{K}\sum_{k=0}^{K-1} V_i(\pi_{\boldsymbol{\theta}_k}) \to V_i(\pi_{\boldsymbol{\theta}^\star})$$

⇒ ~~Randomization:~~ $\theta^\dagger \sim$ Uniform$(\boldsymbol{\theta}_k) \Rightarrow \mathbb{E}\left[V_i(\pi_{\boldsymbol{\theta}^\dagger})\right] = \frac{1}{K}\sum_{k=0}^{K-1} V_i(\pi_{\boldsymbol{\theta}^\dagger}) \to V_i(\pi_{\boldsymbol{\theta}^\star})$

Intuition

$$\begin{cases} \boldsymbol{\theta}_k \in \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} V_{\lambda_k}(\pi_{\boldsymbol{\theta}}), \quad V_\lambda(\pi) = V_0(\pi) + \lambda V_1(\pi) \\ \lambda_{k+1} = \left[\lambda_k - \eta\big(V_1(\pi_{\boldsymbol{\theta}_k}) - c_1\big)\right]_+ \end{cases}$$

- Only the ergodic average of (approximate) dual ascent iterates converges

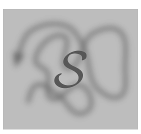
$$V_i(\pi_{\boldsymbol{\theta}_k}) \not\to V_i(\pi_{\boldsymbol{\theta}^\star}) \quad \text{but} \quad \frac{1}{K}\sum_{k=0}^{K-1} V_i(\pi_{\boldsymbol{\theta}_k}) \to V_i(\pi_{\boldsymbol{\theta}^\star})$$

- Value function is an ergodic average: $V(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t, a_t)\right]$

State augmentation

- Construct a new MDP based on *known* state space $\mathcal{M}$ and transition kernel q:

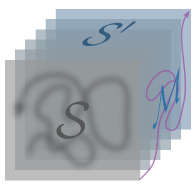
$$\text{MDP}' = \begin{cases} \text{State space:} & \mathcal{S}\times\mathcal{M} = \mathcal{S}' \Rightarrow s' = [s, m] \text{ for } s \in \mathcal{S} \text{ and } m \in \mathcal{M} \\ \text{Action space:} & \mathcal{A} \\ \text{Transition kernel:} & p(s_{t+1}|s_t, a)q(m_{t+1}|m_t, s_t, a) = p'(s'_{t+1}|s'_t, a) \end{cases}$$



State augmentation

- Construct a new MDP based on *known* state space $\mathcal{M}$ and transition kernel q:

$$\text{MDP}' = \begin{cases} \text{State space:} & \mathcal{S}\times\mathcal{M} = \mathcal{S}' \Rightarrow s' = [s, m] \text{ for } s \in \mathcal{S} \text{ and } m \in \mathcal{M} \\ \text{Action space:} & \mathcal{A} \\ \text{Transition kernel:} & p(s_{t+1}|s_t, a)q(m_{t+1}|m_t, s_t, a) = p'(s'_{t+1}|s'_t, a) \end{cases}$$

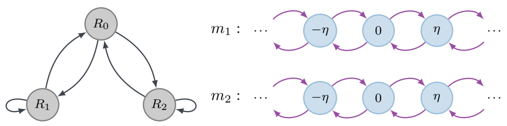


State augmentation

- Construct a new MDP based on *known* state space $\mathcal{M}$ and transition kernel q:

$$\text{MDP}' = \begin{cases} \text{State space:} & \mathcal{S}\times\mathcal{M} = \mathcal{S}' \Rightarrow s' = [s, m] \text{ for } s \in \mathcal{S} \text{ and } m \in \mathcal{M} \\ \text{Action space:} & \mathcal{A} \\ \text{Transition kernel:} & p(s_{t+1}|s_t, a)q(m_{t+1}|m_t, s_t, a) = p'(s'_{t+1}|s'_t, a) \end{cases}$$

  - e.g., $\mathcal{M} = \mathbb{R}^2$ and $m_{i,t+1} = m_{i,t} + \eta\big[\mathbb{I}(s_t = R_i) - \mathbb{I}(s_t \neq R_i)\big]$



State augmentation

- Construct a new MDP based on *known* state space $\mathcal{M}$ and transition kernel q:

$$\text{MDP}' = \begin{cases} \text{State space:} & \mathcal{S}\times\mathcal{M} = \mathcal{S}' \Rightarrow s' = [s, m] \text{ for } s \in \mathcal{S} \text{ and } m \in \mathcal{M} \\ \text{Action space:} & \mathcal{A} \\ \text{Transition kernel:} & p(s_{t+1}|s_t, a)q(m_{t+1}|m_t, s_t, a) = p'(s'_{t+1}|s'_t, a) \end{cases}$$

- In general, it is not clear…
  - …*how many* and *which* states to augment ($\mathcal{M}$)
  - …*what dynamics* these states should follow (q)

…to guarantee optimality and feasibility

Intuition: State-augmented CRL

$$\begin{cases} \boldsymbol{\theta}_k \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ V_{\lambda_k}(\pi_{\boldsymbol{\theta}}), \quad V_\lambda(\pi) = V_0(\pi) + \lambda V_1(\pi) \\ \lambda_{k+1} = \Big[\lambda_k - \eta\big(V_1(\pi_{\boldsymbol{\theta}_k}) - c_1\big)\Big]_+ \end{cases}$$

- Only the ergodic average of (approximate) dual ascent iterates converges

$$V_i(\pi_{\boldsymbol{\theta}_k}) \not\to V_i(\pi_{\boldsymbol{\theta}^\star}) \quad \text{but} \quad \frac{1}{K}\sum_{k=0}^{K-1} V_i(\pi_{\boldsymbol{\theta}_k}) \to V_i(\pi_{\boldsymbol{\theta}^\star})$$

- Value function is an ergodic average: $V(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t, a_t)\right]$

Intuition: State-augmented CRL

$$\begin{cases} \boldsymbol{\theta}_k \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ V_{\lambda_k}(\pi_{\boldsymbol{\theta}}), \quad V_\lambda(\pi) = V_0(\pi) + \lambda V_1(\pi) \\ \lambda_{k+1} = \Big[\lambda_k - \eta\big(V_1(\pi_{\boldsymbol{\theta}_k}) - c_1\big)\Big]_+ \end{cases}$$

- Only the ergodic average of (approximate) dual ascent iterates converges

$$V_i(\pi_{\boldsymbol{\theta}_k}) \not\to V_i(\pi_{\boldsymbol{\theta}^\star}) \quad \text{but} \quad \frac{1}{K}\sum_{k=0}^{K-1} V_i(\pi_{\boldsymbol{\theta}_k}) \to V_i(\pi_{\boldsymbol{\theta}^\star})$$

- Value function is an ergodic average: $V(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t, a_t)\right]$

Intuition: State-augmented CRL

Offline $\quad \boldsymbol{\theta}_k \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ V_{\lambda_k}(\pi_{\boldsymbol{\theta}}), \quad V_\lambda(\pi) = V_0(\pi) + \lambda V_1(\pi)$

$\lambda_{k+1} = \Big[\lambda_k - \eta\big(V_1(\pi_{\boldsymbol{\theta}_k}) - c_1\big)\Big]_+$

- Only the ergodic average of (approximate) dual ascent iterates converges

$$V_i(\pi_{\boldsymbol{\theta}_k}) \not\to V_i(\pi_{\boldsymbol{\theta}^\star}) \quad \text{but} \quad \frac{1}{K}\sum_{k=0}^{K-1} V_i(\pi_{\boldsymbol{\theta}_k}) \to V_i(\pi_{\boldsymbol{\theta}^\star})$$

- Value function is an ergodic average: $V(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t, a_t)\right]$

Intuition: State-augmented CRL

Offline $\quad \boldsymbol{\theta}_k \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ V_{\lambda_k}(\pi_{\boldsymbol{\theta}}), \quad V_\lambda(\pi) = V_0(\pi) + \lambda V_1(\pi)$

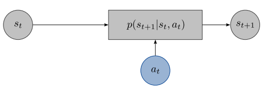
Online $\quad \lambda_{k+1} = \Big[\lambda_k - \eta\big(V_1(\pi_{\boldsymbol{\theta}_k}) - c_1\big)\Big]_+$

- Only the ergodic average of (approximate) dual ascent iterates converges

$$V_i(\pi_{\boldsymbol{\theta}_k}) \not\to V_i(\pi_{\boldsymbol{\theta}^\star}) \quad \text{but} \quad \frac{1}{K}\sum_{k=0}^{K-1} V_i(\pi_{\boldsymbol{\theta}_k}) \to V_i(\pi_{\boldsymbol{\theta}^\star})$$

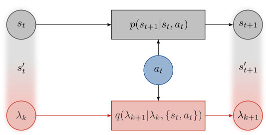
- Value function is an ergodic average: $V(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t, a_t)\right]$

State-augmented CRL



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

State-augmented CRL

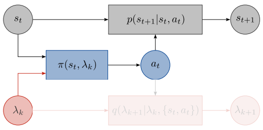


State space: $\mathcal{M} = \{\lambda\} \Rightarrow s' = (s, \lambda)$

Dynamics: $\lambda_{i,k+1} = \Big[\lambda_{i,k} - \eta\big(V_i(\pi) - c_i\big)\Big]_+$

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

State-augmented CRL



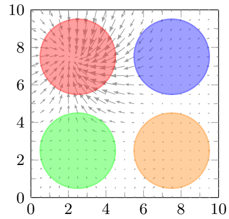
State space: $\mathcal{M} = \{\lambda\} \Rightarrow s' = (s,\lambda)$

Dynamics: $\lambda_{i,k+1} = \left[\lambda_{i,k} - \eta\left(V_i(\pi) - c_i\right)\right]_+$

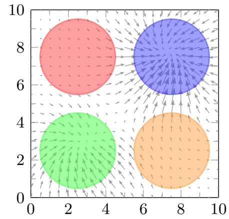
- **Training** (offline)
  - Train policy against $r(s',a) = r_0(s,a) + \sum_{i=1}^m \lambda_i r_1(s,a)$ with static λ (no dynamics)

$$\equiv \pi^\dagger(\lambda) \in \underset{\pi \in \mathcal{P}(\mathcal{S})}{\operatorname{argmax}}\ V_0(\pi) + \sum_{i=1}^m \lambda_i \left(V_i(\pi) - c_i\right)$$

Monitoring task

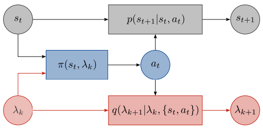


Policy learned for
$\lambda = [5, 0, 0, 0]$



Policy learned for
$\lambda = [0, 5, 5, 0]$

State-augmented CRL



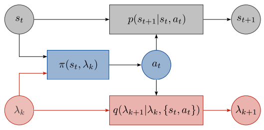
State space: $\mathcal{M} = \{\lambda\} \Rightarrow s' = (s,\lambda)$

Dynamics: $\lambda_{i,k+1} = \left[\lambda_{i,k} - \eta\left(V_i(\pi) - c_i\right)\right]_+$

- **Training** (offline) $\Rightarrow \pi^\dagger(\lambda) \approx \underset{\pi \in \mathcal{P}(\mathcal{S})}{\operatorname{argmax}}\ V_0(\pi) + \sum_{i=1}^m \lambda_i V_i(\pi)$
- **Execution** (online)
  - Execute $\pi^\dagger(\cdot|s,\lambda_k)$ for fixed horizon T_0 and use stochastic approximation of λ-dynamics

$$\lambda_{i,k+1} = \left[\lambda_{i,k} - \eta\left(\frac{1}{T_0}\sum_{\tau=0}^{T_0-1} r_{i,\tau} - c_i\right)\right]_+$$

State-augmented CRL

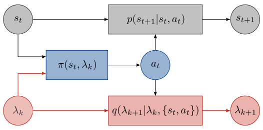


State space: $\mathcal{M} = \{\lambda\} \Rightarrow s' = (s,\lambda)$

Dynamics: $\lambda_{i,k+1} = \left[\lambda_{i,k} - \eta\left(V_i(\pi) - c_i\right)\right]_+$

- It is systematic: no *ad hoc* state augmentation

State-augmented CRL

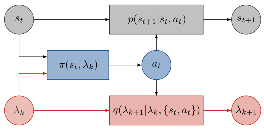


State space: $\mathcal{M} = \{\lambda\} \Rightarrow s' = (s,\lambda)$

Dynamics: $\lambda_{i,k+1} = \left[\lambda_{i,k} - \eta\left(V_i(\pi) - c_i\right)\right]_+$

- It is systematic: no *ad hoc* state augmentation
- Accommodates online modifications of requirements: trained policy does not depend on c

State-augmented CRL



State space: $\mathcal{M} = \{\lambda\} \Rightarrow s' = (s,\lambda)$

Dynamics: $\lambda_{i,k+1} = \left[\lambda_{i,k} - \eta\left(V_i(\pi) - c_i\right)\right]_+$

- It is systematic: no *ad hoc* state augmentation
- Accommodates online modifications of requirements: trained policy does not depend on c
- It works

State-augmented CRL

Theorem (Calvo-Fullana, Paternain, Chamon, Ribeiro'23)

State-augmented CRL generates *state-action sequences* $\{(s_t,a_t)\}$ that are almost surely feasible

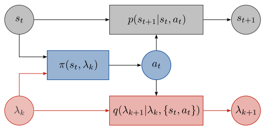
$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} r_i(s_t,a_t) \geq c_i \text{ a.s., for all } i,$$

and near-optimal

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

(mild conditions apply)

State-augmented CRL

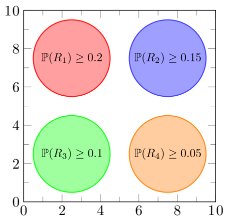


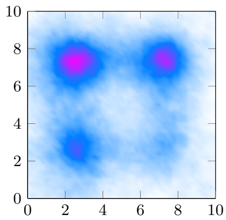
State space: $\mathcal{M} = \{\lambda\} \Rightarrow s' = (s,\lambda)$

Dynamics: $\lambda_{i,k+1} = \left[\lambda_{i,k} - \eta\left(V_i(\pi) - c_i\right)\right]_+$

- It is systematic: no *ad hoc* state augmentation
- Accommodates online modifications of requirements: trained policy does not depend on c
- It works
  - Does not find a policy $\Rightarrow$ generates trajectories during execution that solve (P-CRL)

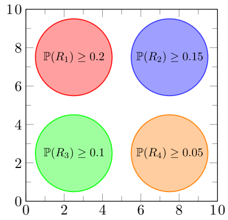
Monitoring task

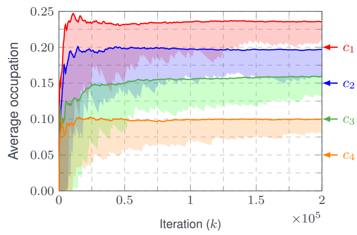




[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

Monitoring task





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

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 precludes fine tuning of rewards, e.g., safety
  [Paternain et al., IEEE TAC'23]
- **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 precludes fine tuning of rewards, e.g., safety
  [Paternain et al., IEEE TAC'23]
- **Constrained RL is hard...**
  Although strong duality holds for CRL (despite non-convexity),
  that is not always enough to obtain feasible solutions $\Rightarrow$ (P-RL) $\subsetneq$ (P-CRL)
- **...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 precludes fine tuning of rewards, e.g., safety
  [Paternain et al., IEEE TAC'23]
- **Constrained RL is hard...**
  Although strong duality holds for CRL (despite non-convexity),
  that is not always enough to obtain feasible solutions $\Rightarrow$ (P-RL) $\subsetneq$ (P-CRL)
- **...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

II. Robustness-constrained learning

Break (30 min)

III. Constrained reinforcement learning

`https://luizchamon.com/aaai`

Universitat Pompeu Fabra Barcelona · Universität Stuttgart · Rensselaer · Penn University of Pennsylvania

Survey:

`www.luizchamon.com/aaai`

AAAI tutorial
Feb. 20, 2023

supervised and reinforcement learning under requirements