Universitat Pompeu Fabra Barcelona · Universität Stuttgart · Rensselaer · Penn

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

L4DC tutorial
July 15, 2024

supervised and reinforcement learning under requirements

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

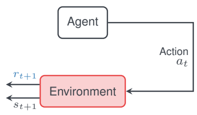
$$\Pr\left(s_{t+1} \mid \{s_u, a_u\}_{u \leq t}\right) = \Pr\left(s_{t+1} \mid s_t, a_t\right) = p(s_{t+1} \mid s_t, a_t)$$

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

Reinforcement learning

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

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

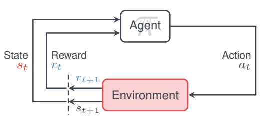


- MDP: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), p (transition kernel), $r : \mathcal{S} \times \mathcal{A} \to [0, B]$ (reward)

Reinforcement learning

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

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



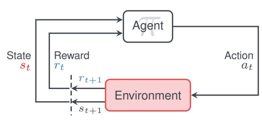
$$\underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \; V(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right] \tag{P-RL}$$

- 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

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



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

- (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

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

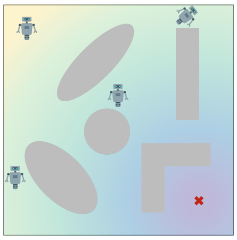
$$r(s,a) =$$

Safe navigation

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

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

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

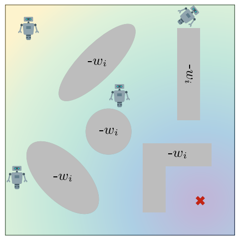


Safe navigation

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

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

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

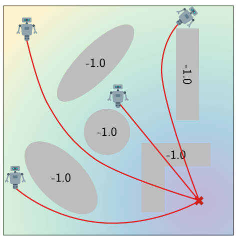


Safe navigation

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

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

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

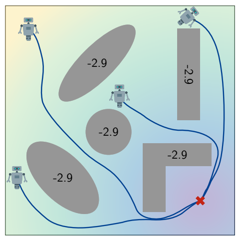


Safe navigation

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

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

Safe navigation

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

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

Safe navigation

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

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

- Probabilistic version of control invariant sets

Safe navigation

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

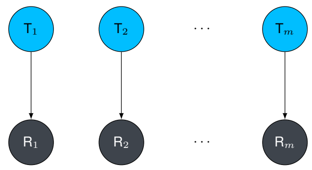
$$\begin{aligned} \underset{\pi \in \mathcal{P}(\mathcal{S})}{\text{maximize}} \quad & V_0(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0(s_t, a_t)\right] \\ \text{subject to} \quad & V_i(\pi) \triangleq \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{\mathbb{I}(s_t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

- 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]

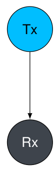
Wireless resource allocation

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



Wireless resource allocation

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



Input (p) $\leftarrow$ Action

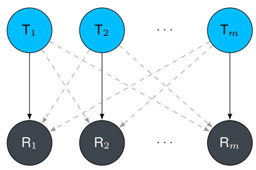
Channel (h) $\leftarrow$ State

Output $=$ Input $\times$ Channel $+$ Noise

$\Rightarrow$ Rate(p, h) $\leftarrow$ Reward

Wireless resource allocation

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

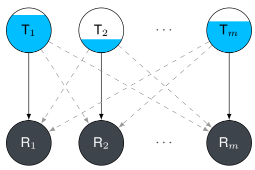


$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \sum_{i=1}^{m} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} p_i(h_t)\right] \\ \text{s. to} \quad & \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \text{Rate}_i\big(\boldsymbol{p}(h_t), h_t\big)\right] \geq c_i \end{aligned}$$

[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

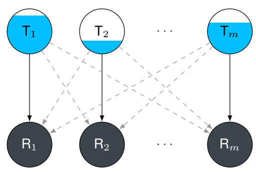


$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \sum_{i=1}^{m} \Pr\left[\bigcap_{t=0}^{T-1} \{b_{i,t} = 0\}\right] \\ \text{s. to} \quad & \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \text{Rate}_i\big(\boldsymbol{p}(h_t), h_t\big)\right] \geq c_i \end{aligned}$$

[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



$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \sum_{i=1}^{m} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\left(b_{i,t} = 0\right)\right] \\ \text{s. to} \quad & \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \text{Rate}_i\big(\boldsymbol{p}(h_t), h_t\big)\right] \geq c_i \end{aligned}$$

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

Constrained RL

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

- 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...]

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 $\approx$ penalty methods
  - ✗ Manual, time-consuming, domain-dependent
  - ✗ Trade-offs, training plateaux
- Prior knowledge $\approx$ 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 $\approx$ penalty methods
- Prior knowledge $\approx$ 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

Strong Duality of CRL

- Define the dual problem as

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

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

Assume that there exist a strictly feasible policy $\pi^\dagger$ such that $V(\pi^\dagger) < c$. Then, the constrained reinforcement learning problem has zero duality gap $P = D$

- There is some sort of hidden convexity in CRL problems $\Rightarrow$ Occupancy measure reformulation

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

Occupancy Measure Reformulation

- The occupancy measure of policy π is the accumulated probability of visiting each state action pair

$$\rho_\pi(s,a) = (1-\gamma)\sum_{t=0}^{T-1}\gamma^t \mathbb{P}_\pi\big(s_t = s, a_t = a\big) \;\Rightarrow\; \pi(a|s) = \rho_\pi(s,a) \times \left[\int_{\mathcal{A}} \rho_\pi(s,a)\, da\right]^{-1}$$

- The value functions $V_i(\pi)$ can be rewritten as expectations with respect to the occupancy measure

$$V_i(\rho) = \mathbb{E}_{(s,a)\sim\rho}\Big[r_i(s,a)\Big] = \int_{\mathcal{S}\times\mathcal{A}} r(s,a)\, \rho_\pi(s,a)\, da\, ds$$

- Thus, value functions $V_i(\rho)$ are linear with respect to the occupancy measure variable

A Non-Proof of Strong Duality

- CRL is a nonconvex program in policy variables but a linear program on occupancy measure variables

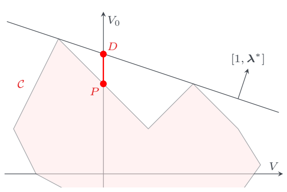
$$\begin{aligned} P = \max_{\pi} \quad & V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right] & = P_\rho = \max_{\rho} \quad & V_0(\rho) := \mathbb{E}_{(s,a)\sim\rho}\Big[r_0(s,a)\Big] \\ \text{subject to} \quad & V(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \geq c & \text{subject to} \quad & V(\rho) := \mathbb{E}_{(s,a)\sim\rho}\Big[r(s,a)\Big] \geq c \end{aligned}$$

- CRL formulated in terms of occupancy measure variables has no duality gap because it is an LP

$$P_\rho = D_\rho = \min_{\boldsymbol{\lambda}} \max_{\rho} V_0(\rho) + \boldsymbol{\lambda}^T\Big(V(\rho) - c\Big)$$

- Primal equivalence $\neq$ dual equivalency $\Rightarrow$ CRL with policy variables may still have a duality gap

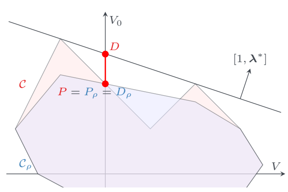
A Proof Sketch of Strong Duality



- Epigraph of policy CRL need not be convex

$$\mathcal{C} = \Big\{\Big[V_0(\pi);\, V(\pi)\Big] \text{ for some } \pi\Big\}$$

A Proof Sketch of Strong Duality



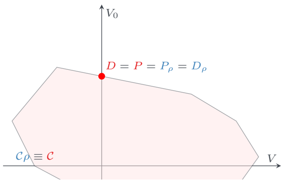
- Epigraph of policy CRL need not be convex

$$\mathcal{C} = \Big\{\Big[V_0(\pi);\, V(\pi)\Big] \text{ for some } \pi\Big\}$$

- Epigraph of occupancy measure CRL is convex

$$\mathcal{C}_\rho = \Big\{\Big[V_0(\rho);\, V(\rho)\Big] \text{ for some } \rho\Big\}$$

A Proof Sketch of Strong Duality



- Epigraph of policy CRL need not be convex

$$\mathcal{C} \;=\; \Big\{ \Big[\, V_0(\pi);\; V(\pi) \,\Big] \text{ for some } \pi \Big\}$$

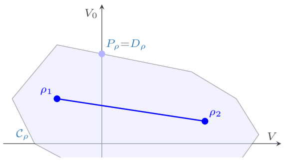
- Epigraph of occupancy measure CRL is convex

$$\mathcal{C}_\rho = \Big\{ \Big[\, V_0(\rho);\; V(\rho) \,\Big] \text{ for some } \rho \Big\}$$

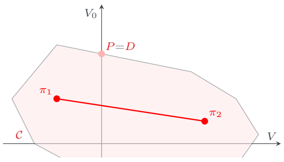
- These two sets are the same $\Rightarrow \mathcal{C}_\rho \equiv \mathcal{C}$

Epigraphs are Convex in Different Ways

- The epigraphs $\mathcal{C}_\rho$ and $\mathcal{C}$ of occupancy measure and policy CRL are convex in different ways



$V\Big[\alpha\rho_1 + (1-\alpha)\rho_2\Big] = \alpha V(\rho_1) + (1-\alpha) V(\rho_2)$



There exist π_α such that $V\Big[\pi_\alpha\Big] = \alpha V(\pi_1) + (1-\alpha) V(\pi_2)$

- The policy π_α is not a convex combination of π and π' challenges convergence of dual methods

Learning Parameterization

- Strong duality, $D = P$, despite having value functions $V_0(\pi)$ and $V(\pi)$ that are not concave on π

$$P \;=\; D \;=\; \min_{\boldsymbol{\lambda}\geq 0}\;\max_{\pi}\; \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r_0(s_t,a_t) + \boldsymbol{\lambda}^T r(s_t,a_t)\Big)\right] + \lambda^T c$$

- In practice, policies are functions of learning parameterizations $\Rightarrow$ Choose actions as $a \sim \pi_\theta$

$$D_\theta \;=\; \min_{\boldsymbol{\lambda}\geq 0}\;\max_{\pi_\theta}\; \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r_0(s_t,a_t) + \lambda^T r(s_t,a_t)\Big)\right] + \lambda^T c$$

- Induces a duality gap because standard learning parameterizations are not convex

Duality Gap in Parameterized CRL

- The learning parameterization is ν-universal $\Rightarrow \min_\theta\ \max_s \int_{\mathcal{A}} \Big|\, \pi(a|s) - \pi_\theta(a|s) \,\Big|\, da \leq \nu$ for all π

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

The difference between the CRL parameterized dual D_θ and the CRL primal P is bounded by

$$\Big|\, P - D_\theta \,\Big| \leq \Big(1 + \|\lambda^\star\|_1\Big)\frac{B\nu}{1-\gamma}$$

- Duality gap depends on parameterization richness relative to discount factor and constraint difficulty

[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

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

Primal-dual algorithm

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

- Maximize the primal ($\equiv$ vanilla RL)

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

Primal-dual algorithm

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

- Maximize the primal ($\equiv$ vanilla RL)

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

- 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

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

- Maximize the primal (≡ vanilla RL)

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

- Update the dual (≡ policy evaluation)

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

In practice...

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

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

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

- Update the dual (≡ policy evaluation): $\{s_t, a_t\} \sim \pi_{\theta_{k+1}}$

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

Dual CRL

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

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

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

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

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

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

Dual gradient descent claims

Theorem (Calvo-Fullana et al'23)
The generated state-action sequences $\left(s_t, a_t \sim \pi^\dagger(\lambda_k)\right)$ are:

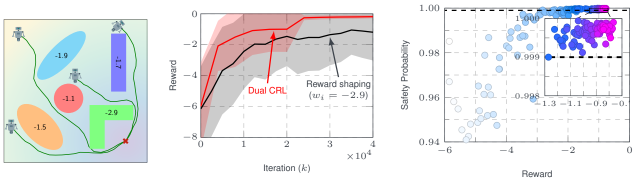
(i) Almost surely feasible: $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} r_i(s_t,a_t) \ge c_i$ a.s., for all i

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

- The time average of the rewards of the sequence generated by rollout dual descent converges
  This sequence is a "solution" of the CRL problem. Stronger, in fact. Constraints satisfied a.s.

Safe navigation

- Reach a target destination while avoiding collisions with a number of obstacles (w.h.p)

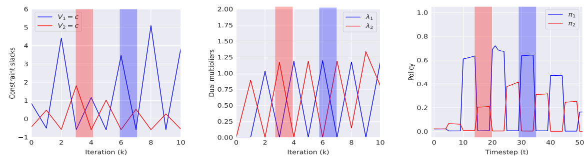


- Policy learned in dual domain outperforms optimal reward shaping policy (obstacle heterogeneity)

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

Wireless resource allocation

- Constraint slacks oscillate around zero ⇒ They spend enough time below zero (feasibility claim)



- The slack oscillation is driven by multiplier oscillation which in turn drives policy switching
  The multipliers drive the policies to switch at the right rate

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

Dual gradient descent does not claim

Theorem (Calvo-Fullana et al'23)
The generated state-action sequences $\left(s_t, a_t \sim \pi^\dagger(\lambda_k)\right)$ are:

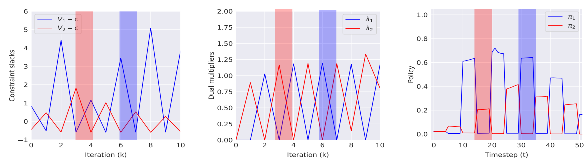
(i) Almost surely feasible: $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} r_i(s_t,a_t) \ge c_i$ a.s., for all i

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

- No claim on optimal policy $\pi^\star$ ⇒ Generate policies $\pi^\dagger(\lambda_k)$ that are samples of near optimal policies

Optimal policy recovery

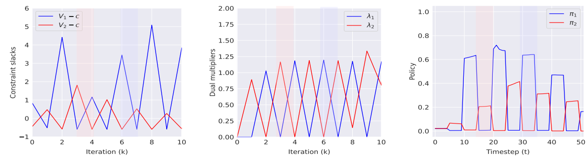
- DGD learns to allocate different users at different points in time with the right amount of power



- At any given epoch the policies $\pi^\dagger(\lambda_k)$ are not optimal ⇒ Their combined action is "optimal"
  Would want to take the time average of policies ⇒ Can't because $V_i(\pi)$ is not convex

Optimal policy recovery

- DGD learns to allocate different users at different points in time with the right amount of power



- At any given epoch the policies $\pi^\dagger(\lambda_k)$ are not optimal $\Rightarrow$ Their combined action is "optimal"

  Would want to take the time average of policies $\Rightarrow$ Can't because $V_i(\pi)$ is not convex

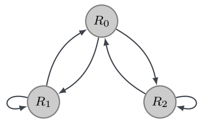
Cannot recover a near optimal policy $\pi^\star$ from sequence of Lagrangian maximizing policies $\pi^\dagger(\lambda_k)$

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

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

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

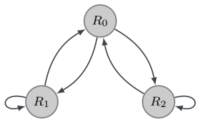


Monitoring task

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

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

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



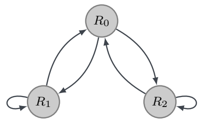
Monitoring task

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

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

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

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



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

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

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

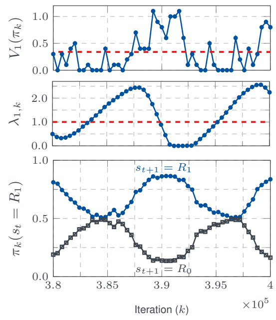
✔ $\pi^\star$ = draw actions uniformly at random

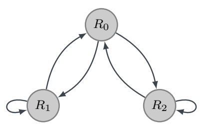
$$\max_{\pi \in \mathcal{P}(\mathcal{S})} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_{\boldsymbol{\lambda}}(s_t)\right]$$

$$r_{\boldsymbol{\lambda}}(s) = \mathbb{I}\left(s \in R_0\right) + \lambda_1 \mathbb{I}\left(s \in R_1\right) + \lambda_2 \mathbb{I}\left(s \in R_2\right)$$

✔ $\lambda_1 = \lambda_2 = 1$: 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. $\Pr\left[s \in R_i\right] = 1$

Monitoring task





Primal recovery

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

Primal recovery

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

✔ Convex optimization $\Rightarrow$ 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 $K \Rightarrow \boldsymbol{\theta}^\star = \lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{\theta}_k$

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 $K \Rightarrow \boldsymbol{\theta}^\star = \lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{\theta}_k$
- ❌ Non-convex optimization ⇒ randomization
  - $\boldsymbol{\theta}^\dagger \sim \text{Uniform}(\boldsymbol{\theta}_k) \Rightarrow \mathbb{E}\left[f(\boldsymbol{\theta}^\dagger)\right] = \frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$

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 $K \Rightarrow \boldsymbol{\theta}^\star = \lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{\theta}_k$
- ❌ Non-convex optimization ⇒ randomization
  - $\boldsymbol{\theta}^\dagger \sim \text{Uniform}(\boldsymbol{\theta}_k) \Rightarrow \mathbb{E}\left[f(\boldsymbol{\theta}^\dagger)\right] = \frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{\theta}_k) \to f(\boldsymbol{\theta}^\star)$
  - Must memorize the complete training sequence of policies

So CRL is hard?

- There are tasks that CRL can tackle and RL cannot

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

- Regularized RL is unable to represent all CRL problems (cannot really "solve" them)
- How can we solve CRL?

So CRL is hard?

- There are tasks that CRL can tackle and RL cannot

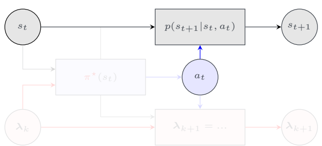
$$\begin{array}{ll} \max_{\pi\in\mathcal{P}(\mathcal{S})} & V_0(\pi) \\ \text{subject to} & V_i(\pi) \geq c_i \end{array} \quad \supsetneq \quad \max_{\pi\in\mathcal{P}(\mathcal{S})} V(\pi)$$

- Regularized RL is unable to represent all CRL problems (cannot really "solve" them)
- How can we solve CRL?

  To solve CRL we augment the state with Lagrange multipliers and learn to maximize Lagrangians

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

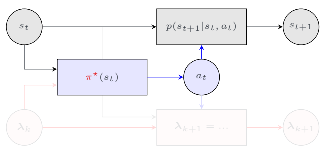
State-augmented CRL



$$\begin{array}{ll} \pi^\star = \operatorname*{argmax}_{\pi} & \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_0(s_t,a_t)\right] \\ \text{subject to} & \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} \mathbf{r}(s_t,a_t)\right] \geq \mathbf{c} \end{array}$$

- For a Markov decision process (MDP) we want to choose actions that solve a CRL problem

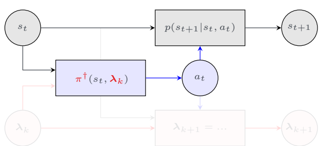
State-augmented CRL



$$\begin{array}{ll} \pi^\star = \operatorname*{argmax}_{\pi} & \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_0(s_t,a_t)\right] \\ \text{subject to} & \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} \mathbf{r}(s_t,a_t)\right] \geq \mathbf{c} \end{array}$$

- Requires finding optimal policy $\pi^\star$ ⇒ We do not know how to find it operating in policy space

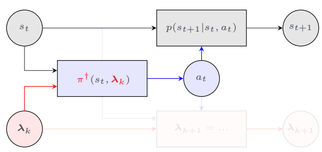
State-augmented CRL



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

- Find Lagrangian maximizing policies $\pi^\dagger(\boldsymbol{\lambda}_k)$ ⇒ Solve unconstrained RL with rewards $r_{\boldsymbol{\lambda}_k}(s_t, a_t)$

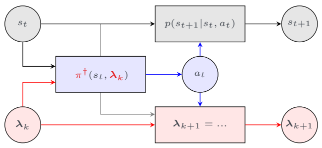
State-augmented CRL



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

- Needs dual variable $\boldsymbol{\lambda}_k$ as input. Also need to update $\boldsymbol{\lambda}_k$ to accumulate constraint violations

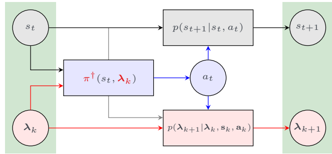
State-augmented CRL



$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$$

- Needs dual variable $\boldsymbol{\lambda}_k$ as input. Also need to update $\boldsymbol{\lambda}_k$ to accumulate constraint violations

State-augmented CRL

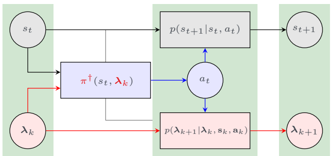


$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$$

- This is equivalent to defining an augmented MDP with (augmented) state $\tilde{S}_t = (s_t, \boldsymbol{\lambda}_t)$

  And an augmented transition probability kernel that included the dual variable updates

State-augmented CRL

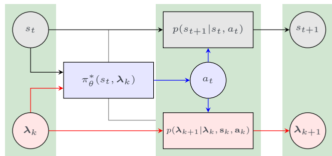


$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$$

- This is equivalent to defining an augmented MDP with (augmented) state $\tilde{S}_t = (s_t, \boldsymbol{\lambda}_t)$

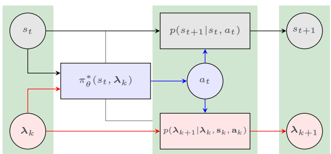
  And an augmented transition probability kernel that included the dual variable updates

Parameterized state-augmented CRL



- In practice, policies are functions of learning parameterizations ⇒ Choose actions as $a \sim \pi_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$
- During training:
  ⇒ Learn policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ that maximizes the Lagrangian averaged over the dual distribution
- During deployment:
  ⇒ Execute policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ while keeping track of dual variable updates λ_k

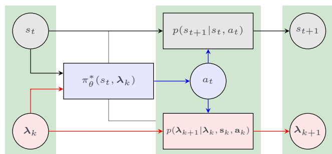
Parameterized state-augmented CRL



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

- In practice, policies are functions of learning parameterizations ⇒ Choose actions as $a \sim \pi_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$
- During training:
  ⇒ Learn policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ that maximizes the Lagrangian averaged over the dual distribution
- During deployment:
  ⇒ Execute policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ while keeping track of dual variable updates λ_k

Parameterized state-augmented CRL

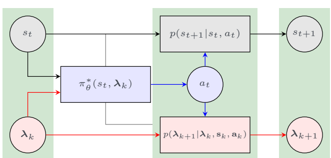


$$a_t \sim \pi^*_{\boldsymbol{\theta}}(s_t,\boldsymbol{\lambda}_k)$$

$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$$

- In practice, policies are functions of learning parameterizations ⇒ Choose actions as $a \sim \pi_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$
- During training:
  ⇒ Learn policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ that maximizes the Lagrangian averaged over the dual distribution
- During deployment:
  ⇒ Execute policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ while keeping track of dual variable updates λ_k

Parameterized state-augmented CRL



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

$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$$

- In practice, policies are functions of learning parameterizations ⇒ Choose actions as $a \sim \pi_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$
- During training:
  ⇒ Learn policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ that maximizes the Lagrangian averaged over the dual distribution
- During deployment:
  ⇒ Execute policy $\pi^*_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ while keeping track of dual variable updates λ_k

Solving CRL

(S1) At epoch k, choose policy ⇒ $\pi^\dagger(\boldsymbol{\lambda}_k) \in \operatorname*{argmax}_{\pi} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_{\boldsymbol{\lambda}}(s_t,a_t)\right]$

(S2) Choose actions $a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)$ between times kT_0 and $(k+1)T_0 - 1$

(S3) Update multiplier ⇒ $\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$

- The algorithm (S1)-(S3) solves CRL in the sense that it generates a state-action sequence (s_t, a_t) that is almost surely feasible and $\mathcal{O}(\eta)$-optimal in expectation
  ⇒ Dual gradient descent "solves" it in the sense of generating the state-action sequence
  ⇒ State-augmented CRL finds an augmented policy $\pi_\theta(s, \lambda)$ to generate them
- We would like a policy $\pi_\theta(s)$ that is feasible and $\mathcal{O}(\eta)$-optimal in expectation
  ⇒ The non-concavity of value functions means that this might not be posible

Solving CRL

(S1) At epoch k, choose policy $\Rightarrow \pi^\dagger(\boldsymbol{\lambda}_k) \in \operatorname{argmax}_{\pi} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_{\boldsymbol{\lambda}}(s_t,a_t)\right]$

(S2) Choose actions $a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)$ between times kT_0 and $(k+1)T_0 - 1$

(S3) Update multiplier $\Rightarrow \boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$

- The algorithm (S1)-(S3) solves CRL in the sense that it generates a state-action sequence (s_t, a_t) that is almost surely feasible and $\mathcal{O}(\eta)$-optimal in expectation
  - $\Rightarrow$ Dual gradient descent "solves" it in the sense of generating the state-action sequence
  - $\Rightarrow$ State-augmented CRL finds an augmented policy $\pi_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ to generate them
- We would like a policy $\pi_{\boldsymbol{\theta}}(s)$ that is feasible and $\mathcal{O}(\eta)$-optimal in expectation
  - $\Rightarrow$ The non-concavity of value functions means that this might not be posible

Solving CRL

(S1) At epoch k, choose policy $\Rightarrow \pi^\dagger(\boldsymbol{\lambda}_k) \in \operatorname{argmax}_{\pi} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_{\boldsymbol{\lambda}}(s_t,a_t)\right]$

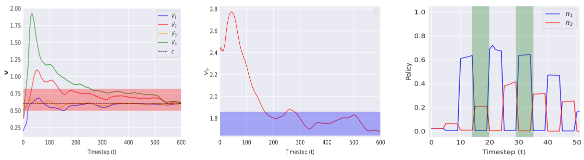
(S2) Choose actions $a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)$ between times kT_0 and $(k+1)T_0 - 1$

(S3) Update multiplier $\Rightarrow \boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\left[\mathbf{r}(s_t,a_t) - \mathbf{c}\right]\right]_+$

- The algorithm (S1)-(S3) solves CRL in the sense that it generates a state-action sequence (s_t, a_t) that is almost surely feasible and $\mathcal{O}(\eta)$-optimal in expectation
  - $\Rightarrow$ Dual gradient descent "solves" it in the sense of generating the state-action sequence
  - $\Rightarrow$ State-augmented CRL finds an augmented policy $\pi_{\boldsymbol{\theta}}(s, \boldsymbol{\lambda})$ to generate them
- We would like a policy $\pi_{\boldsymbol{\theta}}(s)$ that is feasible and $\mathcal{O}(\eta)$-optimal in expectation
  - $\Rightarrow$ The non-concavity of value functions means that this might not be posible

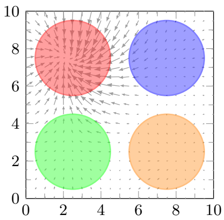
Wireless resource allocation

- State augmented CRL learns policies that satisfy constraints and minimize objective on average

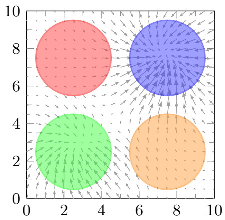


- Even though we still have policy switching $\Rightarrow$ Multipliers drive policies to switch at the right rate

Monitoring task



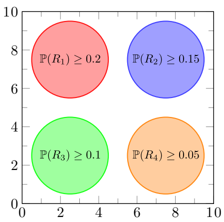
Policy learned for
$\lambda = [5, 0, 0, 0]$

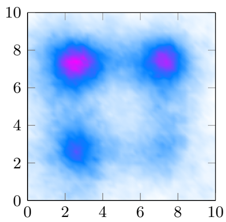


Policy learned for
$\lambda = [0, 5, 5, 0]$

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

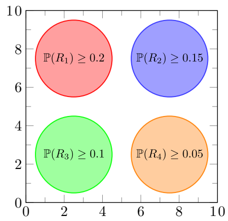
Monitoring task

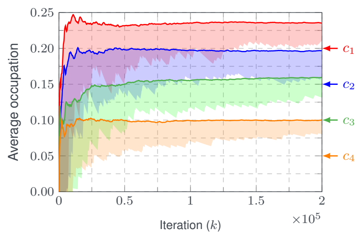




[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

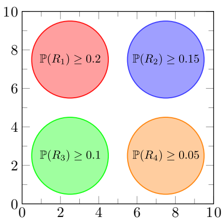
Monitoring task

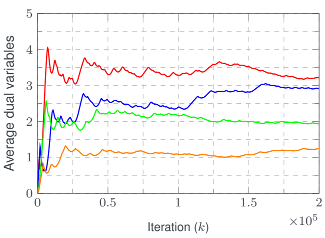




[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

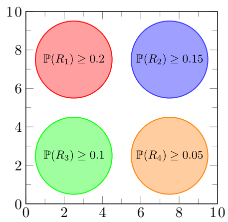
Monitoring task

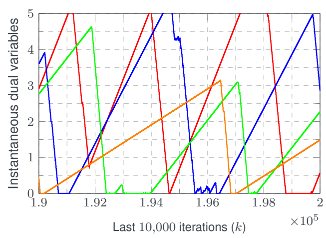




[Calvo-Fullana, Paternain, Chamon, Ribeiro, IEEE TAC'23]

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
  - Constrained learning theory
  - Resilient constrained learning
  - Robust learning

Break (30 min)

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

https://luizchamon.com/l4dc

Universitat Pompeu Fabra Barcelona
Universität Stuttgart
Rensselaer
Penn

Survey:

www.luizchamon.com/l4dc

L4DC tutorial
July 15, 2024

supervised and reinforcement learning under requirements