Learning with Constraints

Alejandro Ribeiro

Dept. of Electrical and Systems Engineering

University of Pennsylvania

Email: aribeiro@seas.upenn.edu

Web: alelab.seas.upenn.edu

LinkdIn: https://www.linkedin.com/in/alejandro-ribeiro-penn/

October 26, 2025

- In constrained learning, losses appear as objectives as well as statistical and pointwise constraints

$$P = \ell_0(\Phi_\theta^*) = \underset{\Phi_\theta}{\text{minimum}} \quad \ell_0(\Phi_\theta) = \mathbb{E}\Big[\ell_0(\Phi_\theta(x), y)\Big] \qquad \text{Minimize objective loss}$$

$$\text{subject to} \quad \ell_i(\Phi_\theta) = \mathbb{E}\Big[\ell_i(\Phi_\theta(x), y)\Big] \leq c_i \qquad \text{Statistical loss requirements}$$

$$\ell_i'(\Phi_\theta(x), y) \leq c_i \quad \text{a.e.} \qquad \text{Pointwise loss requirements}$$

- Find the parametric function Φ_θ^* that minimizes the statistical objective loss ℓ_0 while incurring ...

  ... at most c_i units of statistical constraint loss ℓ_i as well as ...

  ... at most c_i units of constraint loss ℓ_i' almost everywhere over the data distribution

Chamon-Ribeiro, *Probably Approximately Correct Constrained Learning*, Neurips 2020, arxiv:2006.05487

Chamon-Paternain-Calvo Fullana-Ribeiro, *Constrained Learning with Non-Convex Losses*, TIT 2022, arxiv:2103.05134

- In constrained learning, losses appear as objectives as well as statistical and pointwise constraints

$$P = \ell_0(\Phi_\theta^*) = \underset{\Phi_\theta}{\text{minimum}} \quad \ell_0(\Phi_\theta) = \mathbb{E}\Big[\ell_0(\Phi_\theta(x), y)\Big] \qquad \text{Minimize objective loss}$$

$$\text{subject to} \quad \ell_i(\Phi_\theta) = \mathbb{E}\Big[\ell_i(\Phi_\theta(x), y)\Big] \leq c_i \qquad \text{Statistical loss requirements}$$

$$\ell_i'(\Phi_\theta(x), y) \leq c_i \quad \text{a.e.} \qquad \text{Pointwise loss requirements}$$

- Find the parametric function Φ_θ^* that minimizes the statistical objective loss ℓ_0 while incurring ...
  - ... at most c_i units of statistical constraint loss ℓ_i as well as ...
  - ... at most c_i units of constraint loss ℓ_i' almost everywhere over the data distribution

Chamon-Ribeiro, *Probably Approximately Correct Constrained Learning*, Neurips 2020, arxiv:2006.05487

Chamon-Paternain-Calvo Fullana-Ribeiro, *Constrained Learning with Non-Convex Losses*, TIT 2022, arxiv:2103.05134

- In constrained learning, losses appear as objectives as well as statistical and pointwise constraints

$$P = \ell_0\big(\Phi_\theta^*\big) = \underset{\Phi_\theta}{\text{minimum}} \quad \ell_0\big(\Phi_\theta\big) = \mathbb{E}\Big[\ell_0\big(\Phi_\theta(x), y\big)\Big] \qquad \text{Minimize objective loss}$$

$$\text{subject to} \quad \ell_i\big(\Phi_\theta\big) = \mathbb{E}\Big[\ell\big(\Phi_\theta(x), y\big)\Big] \leq c \qquad \text{Statistical loss requirements}$$

$$\ell'\big(\Phi_\theta(x), y\big) \leq c \quad \text{a.e.} \qquad \text{Pointwise loss requirements}$$

- Find the parametric function Φ_θ^* that minimizes the statistical objective loss ℓ_0 while incurring ...

  ... at most c units of statistical constraint loss ℓ as well as ...

  ... at most c units of constraint loss ℓ' almost everywhere over the data distribution

Chamon-Ribeiro, *Probably Approximately Correct Constrained Learning*, Neurips 2020, arxiv:2006.05487

Chamon-Paternain-Calvo Fullana-Ribeiro, *Constrained Learning with Non-Convex Losses*, TIT 2022, arxiv:2103.05134

- In constrained reinforcement learning, rewards appear as objectives and constraints

$$P = V_0(\pi^*) = \underset{\pi}{\text{maximum}} \; V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right] \quad \text{Maximize objective reward}$$

$$\text{subject to} \; V_i(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_t,a_t)\right] \geq c_i \quad \text{Subject to reward requirements}$$

- Find the Policy π^* that maximizes the accumulation of objective reward r_0 while accumulating ...

  ... at least c_i units of constraint reward r_i

Paternain-Chamon-Calvo Fullana-Ribeiro, *Constrained Reinforcement Learning has Zero Duality Gap*, 2019, arxiv:1910.13393

Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, arxiv:2102.11941

- In constrained reinforcement learning, rewards appear as objectives and constraints

$$P = V_0(\pi^*) = \underset{\pi}{\text{maximum}} \; V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right] \qquad \text{Maximize objective reward}$$

$$\text{subject to } \mathbf{V}(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r\,(s_t,a_t)\right] \geq c \qquad \text{Subject to reward requirements}$$

- Find the Policy π^* that maximizes the accumulation of objective reward r_0 while accumulating ...

  ... at least c units of constraint reward r

Paternain-Chamon-Calvo Fullana-Ribeiro, *Constrained Reinforcement Learning has Zero Duality Gap*, 2019, arxiv:1910.13393

Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, arxiv:2102.11941

Artificial Intelligence under Requirements

Motivation 1

We often choose to ignore AI's glaring limitations. We not only want to fit data as best as possible.

We also want to be Safe, Robust, Fair, Representative, Truthful...

- Alignment of generative Large Language Models (LLMs) to user preferences

Zhang-Li-Hounie-Bastani-Ding-Ribeiro, *Alignment of Large Language Models with Constrained Learning*, Neurips 2025, arxiv:2505.19387

- Alignment of a language model requires adapting a pre trained LLM to satisfy user requirements

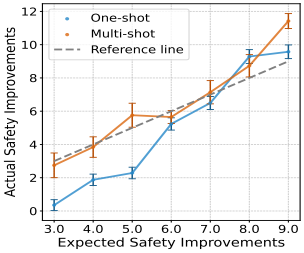
$$P = \underset{\theta}{\text{minimize}} \quad \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\pi_\theta}\left[D_{\text{KL}}(\pi_\theta(\cdot\,|\,\mathbf{x})\,\|\,\pi_{\text{ref}}(\cdot\,|\,\mathbf{x}))\right]\right] \qquad \text{Maximize KL-regularized reward}$$

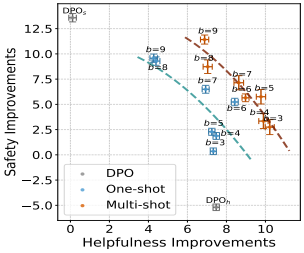
$$\text{subject to } \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\pi_\theta}\left[g_i(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{\mathbf{y}\sim\pi_{\text{ref}}}\left[g_i(\mathbf{x},\mathbf{y})\right]\right] \geq c_i \qquad \text{Subject to utility requirements}$$

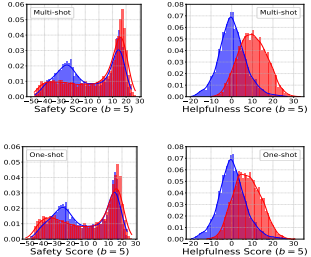
- Policy π_θ that minimizes the KL-divergence to (pretrained) reference model π_{ref} while attaining ...

... an improvement of at least c_i units in user-specified utilities g_i relative to π_{ref}

- Align a pretrained LLM to enhance helpfulness and safety of text generated in response to prompts







- Pareto front of optimality vs helpfulness moves right and up relative to state of the art heuristics

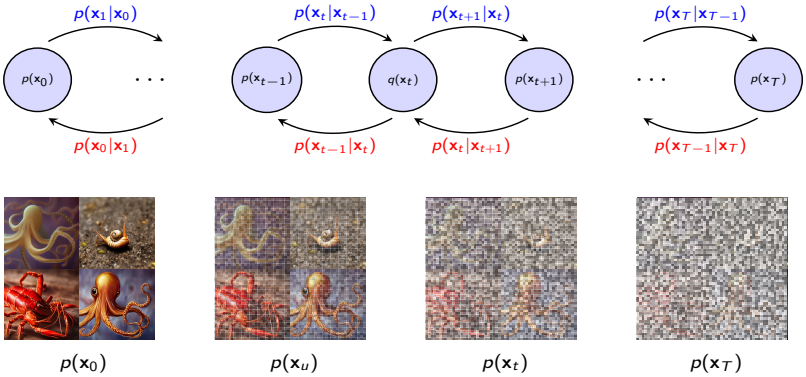
Motivation 2

Some AI problems that a priori do not look like constrained learning problems are often easier to formulate using constraints.

- Composition of a set of distributions parameterized by different generative diffusion models

Khalafi-Hounie-Ding-Ribeiro, *Composition and Alignment of Diffusion Models using Constrained Learning*, Neurips 2025, arxiv:2508.19104

- The backward process is trained to remove noise from from a forward diffusion (noising) process



- Train denoiser $\epsilon_\theta(x_t, t)$ to imitate the data distribution $q(\mathbf{x}_0)$ with the learned distribution $p(x_0; \epsilon_\theta)$

- We want to sample from a composition of a set of diffusion models pretrained with different criteria

  $\Rightarrow$ E.g., models that optimize for different measures of human preferences

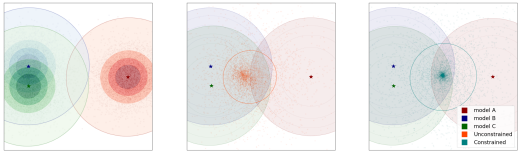
$$P = \underset{\epsilon_\theta, u}{\text{minimize}} \quad u \qquad \text{Minimize allowed divergence threshold}$$

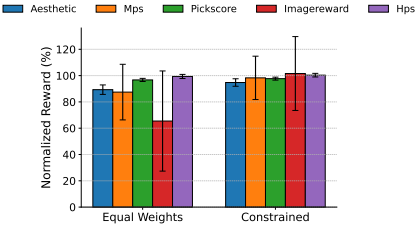
$$\text{subject to} \quad D_{\text{KL}}\Big[\, p(x_0; \epsilon_\theta) \,\|\, q_i \,\Big] \leq u \qquad \text{Keep divergences below threshold}$$

- Find the denoiser ϵ_θ that leads to a distribution $p(x_0; \epsilon_\theta)$ with the smallest upper bound u ...

  ... on the KL divergence to all of the pretrained models q_i

- Compose pretrained diffusion models optimized to four different measures of human preference

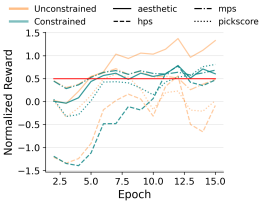


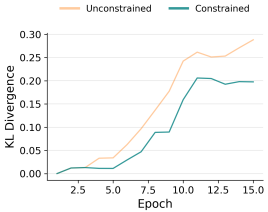


- KL divergence constraints yield samples with good scores on all four measurements

$\Rightarrow$ Unconstrained composition overemphasizes some distributions. Likely because of overlap

- Constrained composition sampling stays close to pre-trained while balancing different rewards





- Unconstrained composition deviates too much from pretrained model and overfits to some rewards

Motivation 3A

AI has transformative potential in physical systems but we must remember that physical systems are designed to satisfy requirements first and to be optimal second.

- Radio resource management in wireless communication and networking

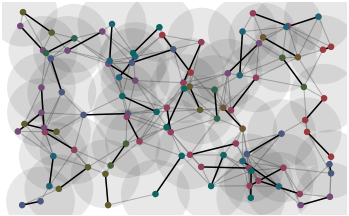
Eisen-Ribeiro, *Optimal Wireless Resource Allocation with Random Edge Graph Neural Networks*, arxiv:1909.01865

Uslu-NaderiAlizadeh-Eisen-Ribeiro, *Fast State-Augmented Learning for Wireless Resource Allocation with Dual Variable Regression*, arxiv:2506.18748

Uslu-Hadou-Bidokhti-Ribeiro, *Generative Diffusion Models for Resource Allocation in Wireless Networks*, arxiv:2504.20277

- Allocate powers p_i in a wireless communication channel across channel state realizations h_{ij}

Communication rate determined by SINR $\Rightarrow \text{SINR}_{it} = \dfrac{h_{ii}p_{it}}{1+\sum_{j\in n(i)} h_{ij}p_{jt}}$



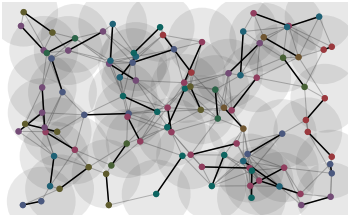
P = maximum$_{p_i \sim \pi}$ Network-wide Sum Rate Utility

subject to Individual Min. Rate QoS

Individual Max. Power Budget

- Allocate powers p_i in a wireless communication channel across channel state realizations h_{ij}

Communication rate determined by SINR $\Rightarrow \text{SINR}_{it} = \dfrac{h_{ii} p_{it}}{1 + \sum_{j \in n(i)} h_{ij} p_{jt}}$



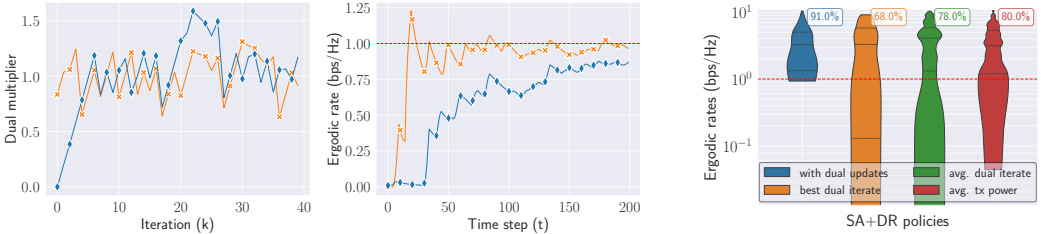
$$P = \underset{p_i \sim \pi}{\text{maximum}} \quad \mathbb{E}_{h, p_{it} \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t \sum_j r(\text{SINR}_{jt}) \right]$$

$$\text{subject to} \quad \mathbb{E}_{h, p_{it} \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t r(\text{SINR}_{jt}) \right] \geq r_{\min}, \quad \text{for all } j,$$

$$\mathbb{E}_{h, p_{it} \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t p_{it} \right] \leq p_{\max}, \quad \text{for all } i$$

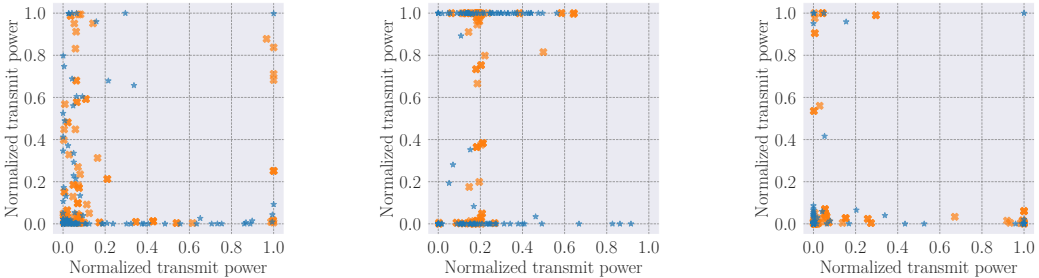
- Optimal resource allocation policy is stochastic $\Rightarrow$ Learn to sample from optimal distributions

  $\Rightarrow$ Augment state with Lagrange multipliers to sample realizations of optimal distribution



- Less constraints are violated and violated constraints are violated by smaller amounts

- Alternatively, train a generative diffusion model to solve the wireless resource allocation problem



- Generate samples from a stationary (optimal) solution distribution in lieu of iterative sampling

Motivation 3B

AI has transformative potential in physical systems but we must remember that physical systems are designed to satisfy requirements first and to be optimal second.

- Approximating solutions of optimal power flow (OPF) on electrical power distribution grids

Damian Owerko, Anna Scaglionne and Alejandro Ribeiro, *Learning Optimal Power Flow with Pointwise Constraints,* Neurips 2025, arxiv:2510.20777

- Find voltages and power allocations that optimize a generation cost objective while satisfying ....

  ... the conservation of power flow and operational constraints on buses and branches

$$
\begin{aligned}
P = \underset{s,v}{\text{minimize}} \quad & \text{Generation cost} \\
\text{subject to} \quad & \text{Power flow conservation,} \\
& \text{Power and voltage operating ranges on nodes,} \\
& \text{Power and reactive power operating ranges on transmission lines,} \\
\text{with} \quad & \text{Power Flow Equation.}
\end{aligned}
$$

- Find voltages and power allocations that optimize a generation cost objective while satisfying ....

  ... the conservation of power flow and operational constraints on buses and branches

$$P = \underset{s,v}{\text{minimize}} \quad \sum_{i=1}^{N} c_{0i} + c_{1i}\text{Re}(s_i) + c_{2i}\text{Re}^2(s_i)$$

$$\text{subject to} \quad s_i - r_i = \sum_{j \in n(i)} f_{ij} - y_i^{S*} |v_i|^2,$$

$$s_{i,\min}^{G} \leq s_i \leq s_{i,\max}^{G} \quad v_{i,\min} \leq |v_i| \leq v_{i,\max},$$

$$|f_{ij}| \leq f_{ij,\max}, \quad |f_{ji}| \leq f_{ji,\max}, \quad \theta_{ij,\min} \leq \angle(v_i v_j^*) \leq \theta_{ij,\max},$$

$$\text{with} \quad f_{ij} = \left(y_{ij} + y_{ij}^{C}\right)^* \left|\frac{v_i}{t_{ij}}\right|^2 - y_{ij}^* \frac{v_i v_j^*}{t_{ij}}, \quad f_{ji} = \left(y_{ij} + y_{ji}^{C}\right)^* |v_j|^2 - y_{ij}^* \frac{v_i^* v_j}{t_{ij}^*}.$$

- Find voltages and power allocations that optimize a generation cost objective while satisfying ....

  ... the conservation of power flow and operational constraints on buses and branches

$$
\begin{aligned}
P \;=\; & \underset{s,v}{\text{minimize}} && C(s) \\
& \text{subject to} && h(\,s,\,v;\,r,\;\,) \;=\; \mathbf{0} \\
& && g(\,s,\,v;\,r,\;\,) \;\leq\; \mathbf{0},
\end{aligned}
$$

- Find voltages and power allocations that optimize a generation cost objective while satisfying ....

  ... the conservation of power flow and operational constraints on buses and branches

$$
\begin{aligned}
P = \quad & \underset{s,v}{\text{minimize}} && C(s) \\
& \text{subject to} && h(s, v; r, \ ) = \mathbf{0} \\
& && g(s, v; r, \ ) \leq \mathbf{0},
\end{aligned}
$$

- Equality constraints represent flow conservation. Inequality constraints represent physical constraints

  $\Rightarrow$ Constraint violations move buses and branches outside of their operating range

- For a distribution of power demand realizations $r \sim \rho$, train a parametric function $\Phi(r; A)$ to...

  ... minimmize the expected cost while satisfying equality and inequality constraints

$$P = \underset{A}{\text{minimum}} \quad \mathbb{E}\Big[\, C\Big(\Phi(r; A)\Big)\,\Big]$$

$$\text{subject to} \quad h\Big(\Phi(r; A); r\Big) = 0 \quad \text{for all } r$$

$$g\Big(\Phi(r; A); r\Big) \leq 0 \quad \text{for all } r$$

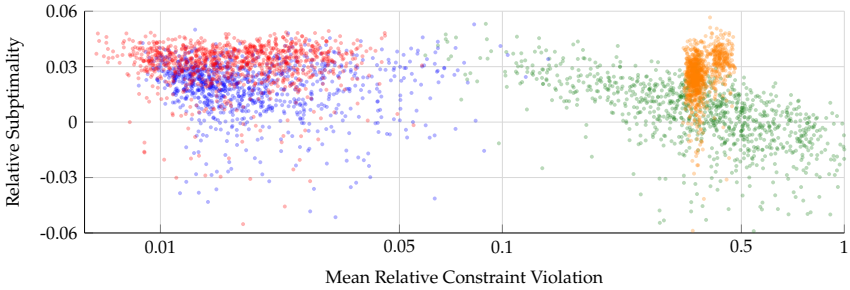
$$P = \underset{A}{\text{minimum}} \quad \mathbb{E}\Big[\, C\Big(\Phi(r; A)\Big)\,\Big]$$

$$\text{subject to} \quad \mathbb{E}\Big[\, h\Big(\Phi(r; A); r\Big)\,\Big] = 0$$

$$\mathbb{E}\Big[\, g\Big(\Phi(r; A); r\Big)\,\Big] \leq 0$$

- An average objective is fine but constraints must be pointwise on individual demand realizations

- Train graph attention to solve OPF in IEEE 300 $\Rightarrow$ Test feasibility of individual demand realizations



- Training with pointwise constraints (red and blue) is the only method with workable constraints

Challenges

16 - 21

- A CRL problem is exactly what the name says it is $\Rightarrow$ Maximize a reward subject to other rewards

$$P = \underset{\pi}{\text{maximum}} \quad V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right]$$

$$\text{subject to} \quad V_i(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_t,a_t)\right] \geq c_i$$

- Policy π that maximizes accumulation of reward r_0 while accumulating at least c_i units of reward r_i

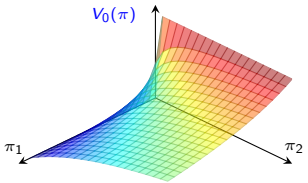
- A CRL problem is exactly what the name says it is $\Rightarrow$ Maximize a reward subject to other rewards

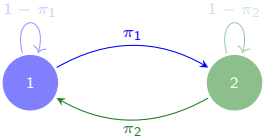
$$P = \underset{\pi}{\text{maximum}} \quad V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right]$$

$$\text{subject to} \quad \mathbf{V}(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbf{r}\,(s_t,a_t)\right] \geq \mathbf{c}$$

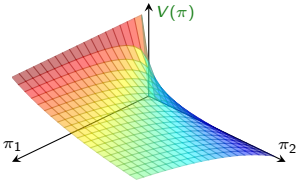
- Policy π that maximizes accumulation of reward r_0 while accumulating at least $\mathbf{c}$ units of reward $\mathbf{r}$

- CRL is challenging to solve because value functions $V_0(\pi)$ and $\mathbf{V}(\pi)$ are not concave on the policy π





$$V_0(\pi) = \frac{\pi_2}{\pi_1 + \pi_2} \qquad V(\pi) = \frac{\pi_1}{\pi_1 + \pi_2}$$



- Set aside this challenge for a moment and attempt to solve CRL in the Lagrangian dual domain

- The Lagrangian is a linear combination of objective and constraints weighted by multiplier $\boldsymbol{\lambda} \geq \mathbf{0}$

$$\mathcal{L}(\pi, \boldsymbol{\lambda}) = V_0(\pi) + \boldsymbol{\lambda}^T \Big(\mathbf{V}(\pi) - \mathbf{c} \Big) = \mathbb{E}_{s,a\sim\pi} \left[\sum_{t=0}^{\infty} \gamma^t \Big(r_0(s_t, a_t) + \boldsymbol{\lambda}^T \mathbf{r}(s_t, a_t) \Big) \right] - \boldsymbol{\lambda}^T \mathbf{c}$$

- The dual function is the maximum of the Lagrangian over the policy variable

$$g(\boldsymbol{\lambda}) = \max_{\pi} \mathcal{L}(\pi, \boldsymbol{\lambda}) = \max_{\pi} \mathbb{E}_{s,a\sim\pi} \left[\sum_{t=0}^{\infty} \gamma^t \Big(r_0(s_t, a_t) + \boldsymbol{\lambda}^T \mathbf{r}(s_t, a_t) \Big) \right] - \boldsymbol{\lambda}^T \mathbf{c}$$

- The dual problem is the minimum of the dual function $\Rightarrow D = g(\boldsymbol{\lambda}^*) = \min_{\boldsymbol{\lambda} \geq \mathbf{0}} g(\boldsymbol{\lambda})$

- Maximizing the Lagrangian is a standard unconstrained reinforcement learning (RL) problem

$$g(\boldsymbol{\lambda}) = \underset{\pi}{\text{maximum}}\ \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r_0(s_t,a_t)+\boldsymbol{\lambda}^T\mathbf{r}(s_t,a_t)\Big)\right]-\boldsymbol{\lambda}^T\mathbf{c}$$

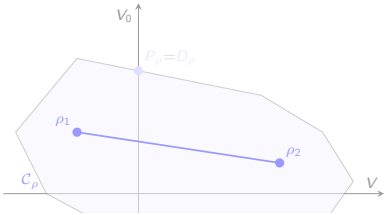
$$:= \underset{\pi}{\text{maximum}}\ \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r_{\boldsymbol{\lambda}}(s_t,a_t)\Big)\right]-\boldsymbol{\lambda}^T\mathbf{c}$$

- A good reason for using the dual optimum $D = g(\boldsymbol{\lambda}^*)$ as a proxy in lieu of the primal CRL problem

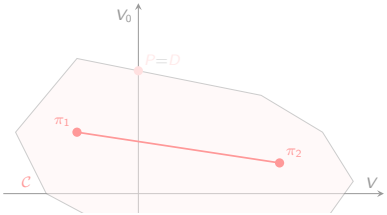
- In general nonconvex problems, dual maxima are strict upper bounds of primal maxima $\Rightarrow P < D$

Strong Duality of Constrained Reinforcement Learning in Policy Space

If a strictly feasible policy exists, $P = D$ even though value functions $V_i(\pi)$ are not concave on π



$V\left[\alpha\rho + (1-\alpha)\rho'\right] = \alpha V(\rho) + (1-\alpha)V(\rho')$

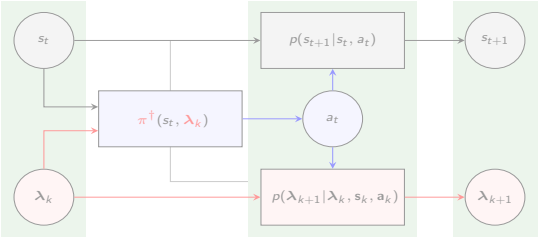


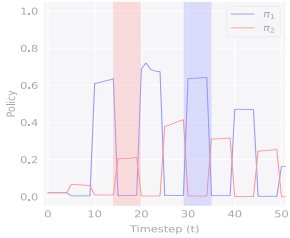
There exist π_α such that $V\left[\pi_\alpha\right] = \alpha V(\pi) + (1-\alpha)V(\pi')$

Paternain-Chamon-Calvo Fullana-Ribeiro, *Constrained Reinforcement Learning has Zero Duality Gap*, 2019, arxiv:1910.13393

State Augmented Constrained Reinforcement Learning

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

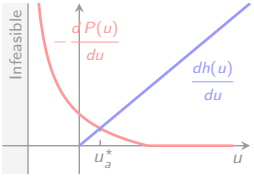


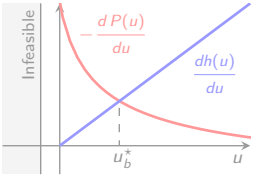


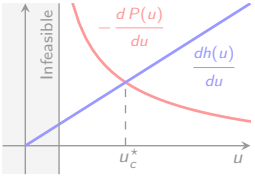
Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, arxiv:2102.11941

Resilient Constrained Reinforcement Learning

Adapt requirements (constraint levels c_i) to equate the marginal costs and benefits of relaxations







Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

Ding-Huan-Ribeiro, *Resilient Constrained Reinforcement Learning*, 2023, arxiv:2312.17194

Strong Duality of Constrained Reinforcement Learning

21 - 28

Theorem (Paternain et al '19)

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

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

Paternain-Chamon-Calvo Fullana-Ribeiro, *Constrained Reinforcement Learning has Zero Duality Gap*, 2019, https://arxiv.org/abs/1910.13393

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

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

- 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

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

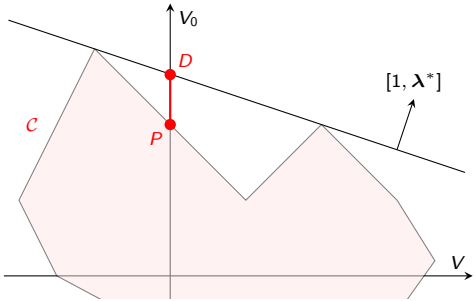
$$P = \underset{\pi}{\text{maximum}} \; V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right] \qquad = P_\rho = \underset{\rho}{\text{maximum}} \; V_0(\rho) := \mathbb{E}_{(s,a)\sim\rho}\left[r_0(s_t,a_t)\right]$$

$$\text{subject to } \mathbf{V}(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbf{r}\,(s_t,a_t)\right] \geq \mathbf{c} \qquad \text{subject to } \mathbf{V}(\rho) := \mathbb{E}_{(s,a)\sim\rho}\left[\mathbf{r}\,(s_t,a_t)\right] \geq \mathbf{c}$$

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

$$P_\rho \;=\; D_\rho \;=\; \underset{\boldsymbol{\lambda}}{\text{minimum}} \quad \underset{\rho}{\text{maximum}} \quad V_0(\rho) \;+\; \boldsymbol{\lambda}^T\Big(\mathbf{V}(\rho) - \mathbf{c}\Big)$$

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



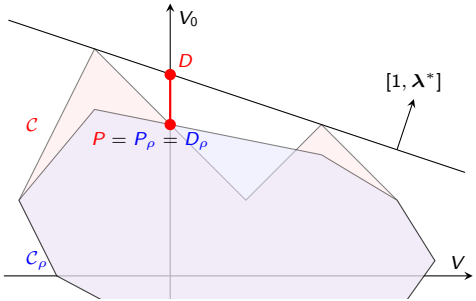
- Epigraph of policy CRL need not be convex

$$\mathcal{C} = \left\{ \begin{bmatrix} V_0(\pi); \mathbf{V}(\pi) \end{bmatrix} \text{ for some } \pi \right\}$$

- Epigraph of occupancy measure CRL is convex

$$\mathcal{C}_\rho = \left\{ \begin{bmatrix} V_0(\rho); \mathbf{V}(\rho) \end{bmatrix} \text{ for some } \rho \right\}$$

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



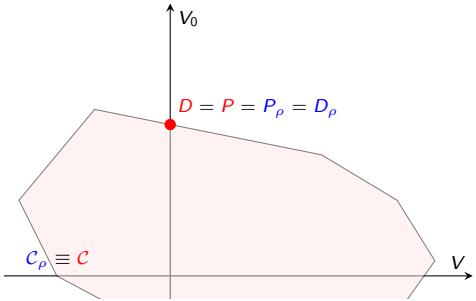
- Epigraph of policy CRL need not be convex

$$\mathcal{C} = \left\{ \begin{bmatrix} V_0(\pi); \mathbf{V}(\pi) \end{bmatrix} \text{ for some } \pi \right\}$$

- Epigraph of occupancy measure CRL is convex

$$\mathcal{C}_\rho = \left\{ \begin{bmatrix} V_0(\rho); \mathbf{V}(\rho) \end{bmatrix} \text{ for some } \rho \right\}$$

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



- Epigraph of policy CRL need not be convex

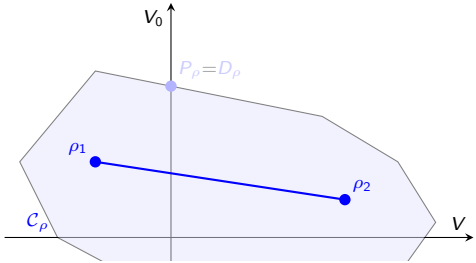
$$\mathcal{C} = \left\{ \begin{bmatrix} V_0(\pi); \mathbf{V}(\pi) \end{bmatrix} \text{ for some } \pi \right\}$$

- Epigraph of occupancy measure CRL is convex

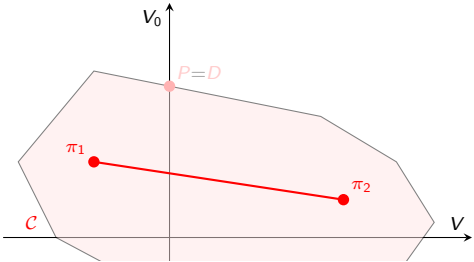
$$\mathcal{C}_\rho = \left\{ \begin{bmatrix} V_0(\rho); \mathbf{V}(\rho) \end{bmatrix} \text{ for some } \rho \right\}$$

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

- 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_1 and π_2 (which will become a headache soon)

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

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

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

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

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

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

Theorem (Paternain et al '19)

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

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

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

Paternain-Chamon-Calvo Fullana-Ribeiro, *Constrained Reinforcement Learning has Zero Duality Gap*, 2019, https://arxiv.org/abs/1910.13393

- CRL problems are not convex when formulated in policy variables

  Even though they are convex (linear) when formulated in occupancy measure variables

- Nevertheless, they have no duality gap $\Rightarrow P = D$. Because their epigraph sets are convex

- If we use ν-universal learning parameterizations CRL problems have $\mathcal{O}(\nu)$ duality gaps

Dual Gradient Descent (DGD)

- Since the duality gap is $\mathcal{O}(\nu)$ (small) we can solve CRL in the parameterized dual domain

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

- For given multiplier $\boldsymbol{\lambda}$, we find the parameter $\boldsymbol{\theta}^\dagger(\boldsymbol{\lambda})$ that maximizes the corresponding Lagrangian

$$\boldsymbol{\theta}^\dagger(\boldsymbol{\lambda}) \in \underset{\theta}{\text{argmax}} \; \mathbb{E}_{s,a \sim \pi_{\boldsymbol{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t \Big(r_0(s_t, a_t) + \boldsymbol{\lambda}^T \mathbf{r}(s_t, a_t) \Big) \right] + \boldsymbol{\lambda}^T \mathbf{c} \equiv \underset{\theta}{\text{argmax}} \; \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda})$$

- Lagrangian maximizers $\theta^\dagger(\boldsymbol{\lambda})$ are unconstrained RL solutions $\Rightarrow r_{\boldsymbol{\lambda}}(s_t, a_t) = r_0(s_t, a_t) + \boldsymbol{\lambda}^T \mathbf{r}(s_t, a_t)$

- Constraint slacks evaluated at Lagrangian maximizers yield dual function gradients $\Rightarrow$ Update $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda}^+ = \left[\boldsymbol{\lambda} - \eta \left(\mathbb{E}_{s,a \sim \pi_{\theta^\dagger(\boldsymbol{\lambda})}} \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t) \right] - \mathbf{c} \right) \right]_+$$

- A set of policy evaluations of unconstrained RL problems. One policy evaluation per constraint

- Since the dual function is convex (they always are), dual gradient descent approaches $\boldsymbol{\lambda}^*$

- Convergence of dual variables still holds if we consider stochastic approximations (policy rollouts)

- Constraint slacks evaluated at Lagrangian maximizers yield dual function gradients $\Rightarrow$ Update $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda}^+ = \left[\boldsymbol{\lambda} - \eta \left(\mathbb{E}_{s,a \sim \pi_{\theta^\dagger(\boldsymbol{\lambda})}} \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t) \right] - \mathbf{c} \right) \right]_+$$

- A set of policy evaluations of unconstrained RL problems. One policy evaluation per constraint

- Since the dual function is convex (they always are), dual gradient descent approaches $\boldsymbol{\lambda}^*$

- Convergence of dual variables still holds if we consider stochastic approximations (policy rollouts)

Theorem (Calvo Fullana et al'23)

The sequence of Lagrangian maximizing policies $\pi(t) = \pi^\dagger(\lambda(t))$ generated by (stochastic) dual gradient descent are:

(i) Asymptotically feasible $\Rightarrow \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{V}\Big(\pi(t)\Big) \geq \mathbf{c}$

(ii) Asymptotically near-optimal $\Rightarrow \lim_{T\to\infty} \mathbb{E}\Big[\frac{1}{T}\sum_{t=0}^{T-1} V\Big(\pi(t)\Big)\Big] \geq P^\star - \frac{\eta B^2}{2}$

(mild conditions apply)

Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, https://arxiv.org/abs/2102.11941

A Tantalizing Conjecture (Time Immemorial)

The sequence of Lagrangian maximizing policies $\pi(t) = \pi^\dagger(\lambda(t))$ generated by (stochastic) dual gradient descent converge to the optimal policy:

(i) Asymptotically feasible $\Rightarrow \lim_{T\to\infty} \mathbf{V}\left[\frac{1}{T}\sum_{t=0}^{T-1} \pi(t) \right] \geq \mathbf{c}$

(ii) Asymptotically near-optimal $\Rightarrow \lim_{T\to\infty} V\left[\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \pi(t) \right] \right] \geq P^\star - \frac{\eta B^2}{2}$

(mild conditions apply)

Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, https://arxiv.org/abs/2102.11941

A False Statement Because Value Functions are not Convex

The sequence of Lagrangian maximizing policies $\pi(t) = \pi^\dagger(\lambda(t))$ generated by (stochastic) dual gradient descent converge to the optimal policy:

(i) Asymptotically feasible $\Rightarrow \lim_{T\to\infty} \mathbf{V}\left[\frac{1}{T}\sum_{t=0}^{T-1} \pi(t) \right] \geq \mathbf{c}$

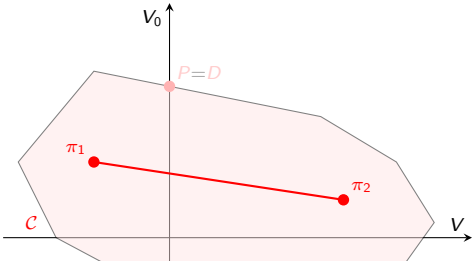
(ii) Asymptotically near-optimal $\Rightarrow \lim_{T\to\infty} V\left[\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \pi(t) \right] \right] \geq P^\star - \frac{\eta B^2}{2}$

(mild conditions apply)

Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, https://arxiv.org/abs/2102.11941

- The epigraph $\mathcal{C}$ is convex in a strange way $\Rightarrow$ policy π_α is not a convex combination of π and π'
- An average of value functions $\frac{1}{T}\sum_{t=1}^{T} V\left[\pi(t)\right]$ is not the value function average $V\left[\frac{1}{T}\sum_{t=1}^{T}\pi(t)\right]$

$$V\left[\alpha\rho + (1-\alpha)\rho'\right] = \alpha V(\rho) + (1-\alpha)V(\rho')$$



There exist π_α such that $V\left[\pi_\alpha\right] = \alpha V(\pi) + (1-\alpha)V(\pi')$

Dual Gradient Descent, Good News and Bad News

- Dual gradient descent alternates between Lagrangian maximization and dual gradient descent steps

- Lagrangian maximization is a standard (unconstrained) reinforcement learning problem

  This is good news. It means that we know how to solve this maximization

- Dual gradient descent does not, alas (poor Yorick), converge to the optimal policy

  $\Rightarrow$ But it does converge in a sense $\Rightarrow$ State Augmented Constrained Reinforcement Learning

Segarra-Eisen-Egan-Ribeiro, *Attributing the Authorship of the Henry VI Plays by Word Adjacency,* Shakespeare Quarterly 67(2) pp, 232-256, 2016

State Augmented Constrained Reinforcement Learning

35 - 44

- Policy π that maximizes accumulation of reward r_0 while accumulating at least c_i units of reward r_i

$$P = \underset{\pi}{\text{maximum}} \quad V_0(\pi) := \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_0(s_t,a_t)\right]$$

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

- Same formulation but without discounting and with ergodic averages (limits of time averages)

- Policy π that maximizes accumulation of reward r_0 while accumulating at least $\mathbf{c}$ units of reward $\mathbf{r}$

$$P = \underset{\pi}{\text{maximum}} \quad V_0(\pi) := \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_0(s_t,a_t)\right]$$

$$\text{subject to} \quad \mathbf{V}(\pi) := \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} \mathbf{r}\,(s_t,a_t)\right] \geq \mathbf{c}$$

- Same formulation but without discounting and with ergodic averages (limits of time averages)

- Lagrangian $\Rightarrow \mathcal{L}(\pi, \boldsymbol{\lambda}) = \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_0(s_t, a_t) + \boldsymbol{\lambda}^T \mathbf{r}(s_t, a_t)\right]$

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

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

Execute policy $\pi^\dagger(\boldsymbol{\lambda}_k)$ for T_0 time steps. Accumulate reward violations on associated multipliers

Theorem (Calvo Fullana et al'23)

Rollout dual gradient descent generates state-action sequences $\Big\{ (s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)) \Big\}_{t \geq 0}$ that are:

(i) Almost surely feasible $\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{r}\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k) \Big) \geq \mathbf{c}$ a.s.

(ii) Near-optimal $\lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} r_0\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k) \Big) \right] \geq P^\star - \frac{\eta B^2}{2}$

(mild conditions apply)

Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, https://arxiv.org/abs/2102.11941

Theorem (Calvo Fullana et al'23)

(i) Almost surely feasible $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{r}\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big) \geq \mathbf{c}$ a.s.

(ii) Near-optimal $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big)\right] \geq P^\star - \frac{\eta B^2}{2}$

- 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.

Theorem (Calvo Fullana et al'23)

(i) Almost surely feasible $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{r}\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big) \geq \mathbf{c}$ a.s.

(ii) Near-optimal $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big)\right] \geq P^\star - \frac{\eta B^2}{2}$

- Alas (poor Yorick), we do not have a claim on the optimal policy $\Rightarrow \pi^\dagger(\boldsymbol{\lambda}_k) \not\to \pi^*$

Theorem (Calvo Fullana et al'23)

(i) Almost surely feasible $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{r}\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big) \geq \mathbf{c}$ a.s.

(ii) Near-optimal $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big)\right] \geq P^\star - \frac{\eta B^2}{2}$

- Alas (poor Yorick), we do not have a claim on the optimal policy $\Rightarrow$ $\frac{1}{K}\sum_{k=1}^{K} \pi^\dagger(\boldsymbol{\lambda}_k) \not\mapsto \pi^*$

Theorem (Calvo Fullana et al'23)

(i) Almost surely feasible $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{r}\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big) \geq \mathbf{c}$ a.s.

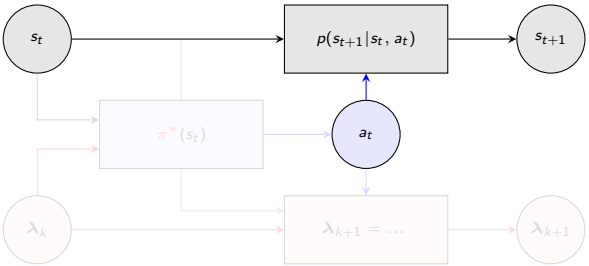
(ii) Near-optimal $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_0\Big(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)\Big)\right] \geq P^\star - \frac{\eta B^2}{2}$

- The sequence $\Big\{(s_t, a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k))\Big\}_{t\geq 0}$ samples actions from the optimal policy (it solves CRL)

  $\Rightarrow$ We just need a way to train a parameterization that generates the sequence $a_t \sim \pi^\dagger(\boldsymbol{\lambda}_k)$

Constrained reinforcement learning is solved by learning policies that maximize Lagrangians

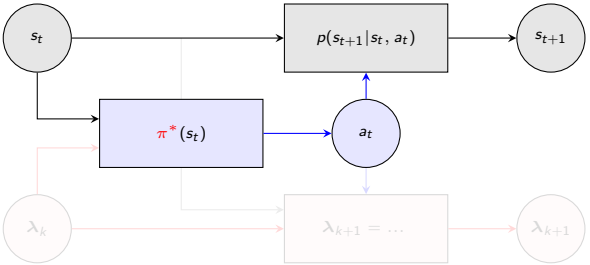
$$\pi^\dagger(\boldsymbol{\lambda}_k) \in \underset{\pi}{\text{argmax}} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_{\boldsymbol{\lambda}_k}(s_t, a_t)\right]$$



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

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

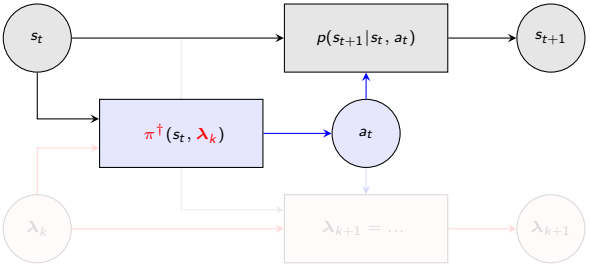
- For a Markov decision process (MDP) we want to choose actions that solve a CRL problem



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

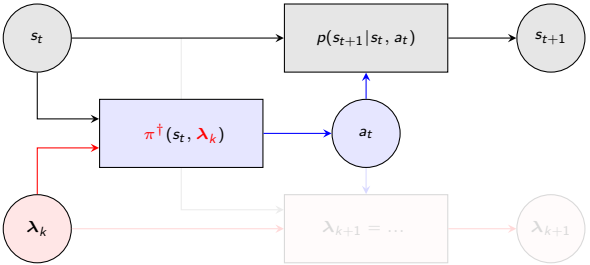
$$\text{subject to } \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} \mathbf{r}\ (s_t, a_t)\right] \geq \mathbf{c}$$

- Requires finding optimal policy π^* $\Rightarrow$ I do not know how to find it operating in policy space



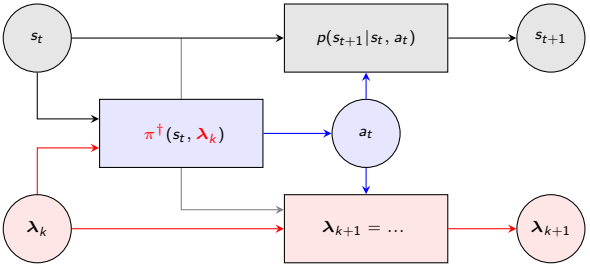
$$\pi^\dagger(s_t, \boldsymbol{\lambda}_k) \in \operatorname*{argmax}_{\pi} \lim_{T\to\infty} \mathbb{E}_{s,a\sim\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_{\boldsymbol{\lambda}_k}(s_t, a_t)\right]$$

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



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

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

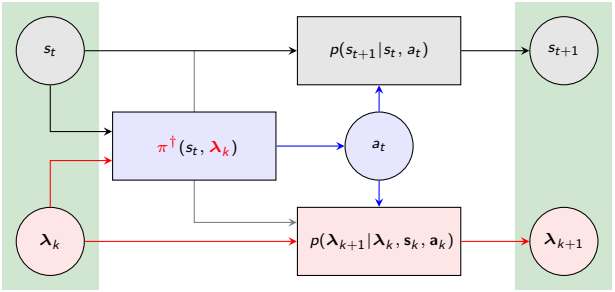


$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0} \sum_{t=kT_0}^{(k+1)T_0-1} \left[\mathbf{r}(s_t, a_t) - \mathbf{c} \right] \right]_+$$

$$\mathbf{s}_k = \left[s_{kT-0:(k+1)T_0-1} \right]$$

$$\mathbf{a}_k = \left[a_{kT-0:(k+1)T_0-1} \right]$$

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



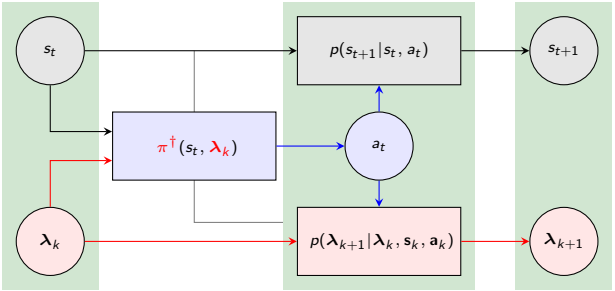
$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0} \sum_{t=kT_0}^{(k+1)T_0-1} \left[\mathbf{r}(s_t, a_t) - \mathbf{c} \right] \right]_+$$

$$\mathbf{s}_k = \left[s_{kT-0:(k+1)T_0-1} \right]$$

$$\mathbf{a}_k = \left[a_{kT-0:(k+1)T_0-1} \right]$$

- 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



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

$$\mathbf{s}_k = \left[s_{kT-0:(k+1)T_0-1} \right]$$

$$\mathbf{a}_k = \left[a_{kT-0:(k+1)T_0-1} \right]$$

- 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

- In practice, policies are functions of learning parameterizations $\Rightarrow$ Choose actions as $a \sim \pi_\phi(s, \boldsymbol{\lambda})$

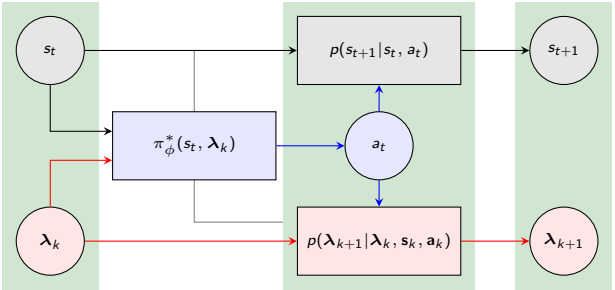
$$\pi_\phi^* \in \underset{\pi_\phi}{\text{argmax}} \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\lambda}} \mathbb{E}_{s,a\sim\pi_\phi} \left[\frac{1}{T} \sum_{t=0}^{T} r_{\boldsymbol{\lambda}_t}(s_t, a_t) \right] \equiv \underset{\pi_\phi}{\text{argmax}} \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\lambda}} \mathbb{E}_{s,a\sim\pi_\phi} \left[\frac{1}{T} \sum_{t=0}^{T} r(s_t, \lambda_t, a_t) \right]$$

- Since this is an state augmented MDP we also need to take expectation over a $\boldsymbol{\lambda}$ distribution

Choosing this distribution presents the usual challenges of off-policy RL

- Learn parameterized policy π^*_ϕ that maximizes the Lagrangian averaged over the dual distribution

Execute policy π^*_ϕ while keeping track of dual variable updates $\Rightarrow$ Generate optimal trajectory



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

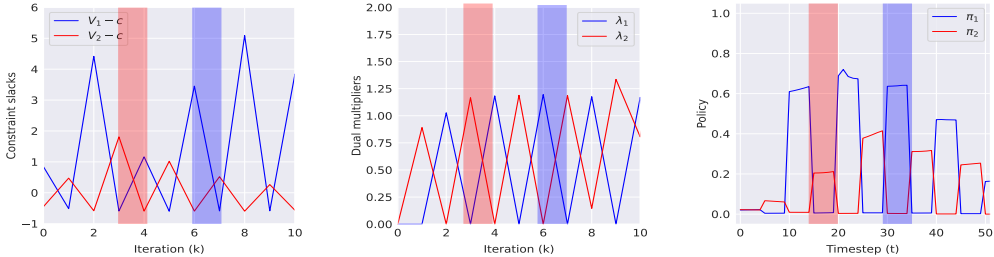
$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \frac{\eta}{T_0} \sum_{t=kT_0}^{(k+1)T_0-1} \left[\mathbf{r}(s_t, a_t) - \mathbf{c} \right] \right]_+$$

$$\mathbf{s}_k = \left[s_{kT-0:(k+1)T_0-1} \right]$$

$$\mathbf{a}_k = \left[a_{kT-0:(k+1)T_0-1} \right]$$

Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, https://arxiv.org/abs/2102.11941

- Constraint slacks oscillate around zero $\Rightarrow$ They spend enough time below zero (feasibility claim)

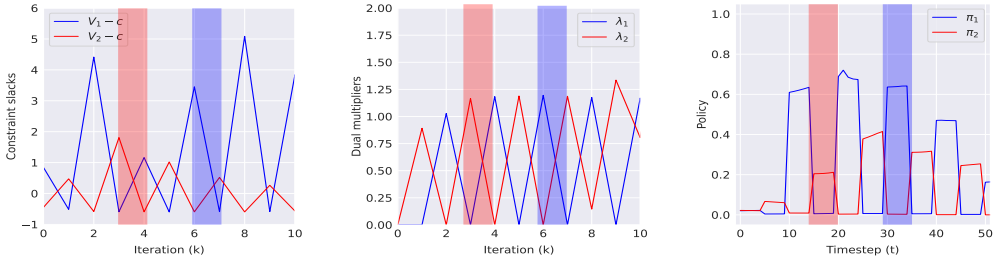


- The slack oscillation is driven by multiplier oscillation which in turn drives policy switching

The multipliers drive the policies $\pi(\boldsymbol{\lambda}_k)$ to switch at the right rate

Uslu-Doostnejad-Ribeiro-NaderiAlizadeh, *Learning to Slice Wi-Fi Networks: A State-Augmented Primal-Dual Approach*, 2024, arxiv:2102.11941

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



- At any given epoch the policies $\pi^\dagger(\boldsymbol{\lambda}_k)$ are not optimal $\Rightarrow$ Their combined action is "optimal"

You want me to take the time average of policies $\Rightarrow$ I can't, because $V(\pi)$ is not convex

Uslu-Doostnejad-Ribeiro-NaderiAlizadeh, *Learning to Slice Wi-Fi Networks: A State-Augmented Primal-Dual Approach*, 2024, arxiv:2405.05748

- To learn solutions of constrained reinforcement learning problems we learn to maximize Lagrangians

  Maximizing Lagrangians is equivalent to solving unconstrained MDPs with modified rewards

  Equivalent to augmenting the MDP's state with dual variables which we update online

- This is not settling for a lesser goal $\Rightarrow$ We are still solving the original CRL problem

  Which we otherwise don't how how to solve except with regularizations that induce suboptimality

Ding-Wei-Zhang-Ribeiro, *Last-Iterate Convergent Policy Gradient Primal-Dual Methods for Constrained MDPs*, 2023, arxiv:2306.11700

Resilient Constrained (Reinforcement) Learning

- Ecological resilience is the ability of an ecosystem to adapt function to withstand varying conditions
- Learning resilience is the ability to adapt specifications to accommodate varying data properties

44 - 49

- Specifications in learning are difficult $\Rightarrow$ Feasible specifications depend on unknown distributions

  $\Rightarrow$ Requirement specifications (c_i) can be relaxed during system design. They are variables

Perturbation function

With constraint relaxation $\mathbf{u}$, the perturbation function $P(\mathbf{u})$ is the solution of the relaxed problem

$$P(\mathbf{u}) = \underset{\pi}{\text{maximum}} \quad V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right]$$

$$\text{subject to} \quad \mathbf{V}(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbf{r}\ (s_t,a_t)\right] \geq \mathbf{c}+\mathbf{u}$$

- Larger relaxations decrease objective loss (a benefit) but increase specification violation (a cost).

- To balance costs and benefits of relaxation we relax constraints in proportion to their difficulty

Resilient Equilibrium

For strictly convex function $h(\mathbf{u})$ we say that relaxation $\mathbf{u}^*$ achieves the resilient equilibrium if

$$\nabla h(\mathbf{u}^\star) \in -\partial P(\mathbf{u}^\star).$$

- At the resilient equilibrium the marginal cost of relaxation equals the marginal benefit of relaxation

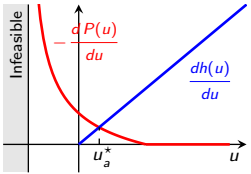
Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

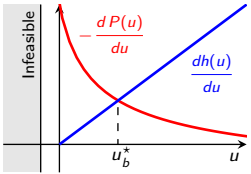
Ding-Huan-Ribeiro, *Resilient Constrained Reinforcement Learning*, 2023, arxiv:2312.17194

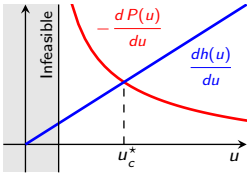
Resilient Equilibrium

For strictly convex function $h(\mathbf{u})$ we say that relaxation $\mathbf{u}^*$ achieves the resilient equilibrium if

$$\nabla h(\mathbf{u}^\star) \in -\partial P(\mathbf{u}^\star).$$







Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

Ding-Huan-Ribeiro, *Resilient Constrained Reinforcement Learning*, 2023, arxiv:2312.17194

- Subdifferentials of perturbation functions are the opposite of corresponding optimal multipliers

Resilient Equilibrium

For strictly convex function $h(\mathbf{u})$ we say that relaxation $\mathbf{u}^*$ achieves the resilient equilibrium if

$$\nabla h(\mathbf{u}^\star) \in \boldsymbol{\lambda}^*(\mathbf{u}^*) = -\partial P(\mathbf{u}^\star).$$

- Resilient constrained learning problems have smaller sample complexity. They generalize better.

  $\Rightarrow$ The optimal multiplier $\boldsymbol{\lambda}^*(\mathbf{u}^*)$ is smaller than the optimal multiplier $\boldsymbol{\lambda}^*(0)$

Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

Resilient Constrained Learning Program

A relaxation $\mathbf{u}^\star$ satisfies the resilient equilibrium if and only if it is a solution of the program

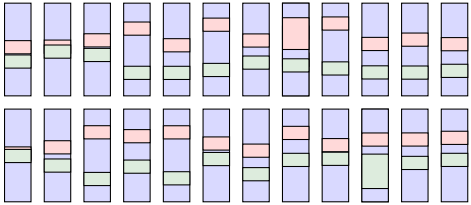
$$P(\mathbf{u}^*) = \max_{\pi} \; V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right] + h(\mathbf{u})$$

$$\text{subject to} \quad \mathbf{V}(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbf{r}\,(s_t,a_t)\right] \geq \mathbf{c}+\mathbf{u}$$

- The resilient equilibrium exist and is unique because we have assumed that $h(\mathbf{u})$ is strictly convex
- Learning resilient solutions is equivalent to a regularized constrained learning problem

Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

- Learn a common model with heterogeneous data distributed among C clients
- Client i loss $\Rightarrow R_i(f_\theta) = \mathbb{E}\Big[\ell\big(f_\theta(\mathbf{x}), y\big)\Big]$
- Average loss $\Rightarrow \bar{R}(f_\theta) = \dfrac{1}{C}\displaystyle\sum_{i=1}^{C} R_i(f_\theta)$



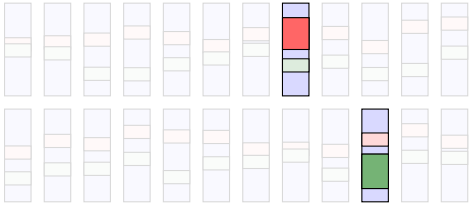
- We seek a model that is best across all clients but is also good (not bad) for each individual client

$$P^\star \;=\; \min_{f_\theta} \quad \bar{R}(f_\theta)$$
$$\text{subject to} \;\; R_i(f_\theta) - \bar{R}(f_\theta) \;\leq\; \epsilon$$

- Minority Samples have few samples in the whole dataset but are significant in some clients

Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

- Learn a common model with heterogeneous data distributed among C clients
- Client i loss $\Rightarrow R_i(f_\theta) = \mathbb{E}\Big[\ell\big(f_\theta(\mathbf{x}), y\big)\Big]$
- Average loss $\Rightarrow \bar{R}(f_\theta) = \dfrac{1}{C}\displaystyle\sum_{i=1}^{C} R_i(f_\theta)$



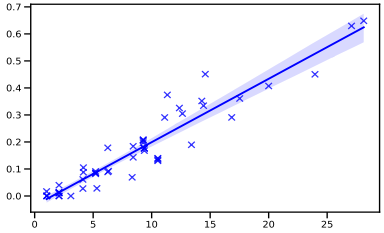
- We seek a model that is best across all clients but is also good (not bad) for each individual client

$$P^\star \;=\; \min_{f_\theta} \quad \bar{R}(f_\theta)$$
$$\text{subject to} \;\; R_i(f_\theta) - \bar{R}(f_\theta) \;\leq\; \epsilon$$

- Minority Samples have few samples in the whole dataset but are significant in some clients

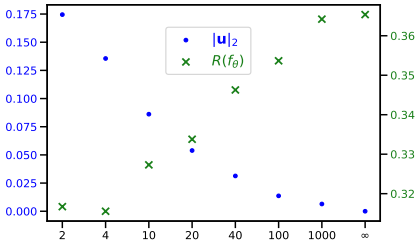
Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

Resilient relaxation $\mathbf{u}^*$ as a function of the percent of entries drawn from the minority class



Clients with more minority samples see larger relaxations as their constraints are more difficult

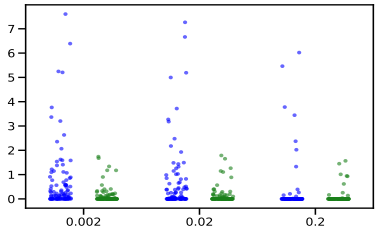
Resilient relaxation and resilient loss for steeper constraint relaxation cost $h(\mathbf{u}) = (\alpha/2)\|\mathbf{u}\|^2$



As we increase α the resilient perturbation decreases and the resilient cost increases

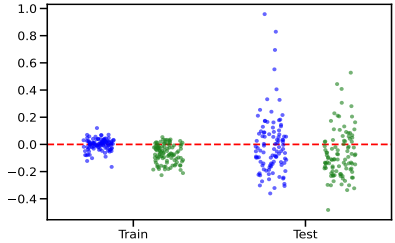
Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, TIT 2022, arxiv:2306.02426

Standard and resilient multipliers of different clients as a function of reference constraint level



Tighter reference constraints do not yield large multipliers. We pay in the form of larger relaxations

Standard and resilient constraint violation of different clients in the training and test sets



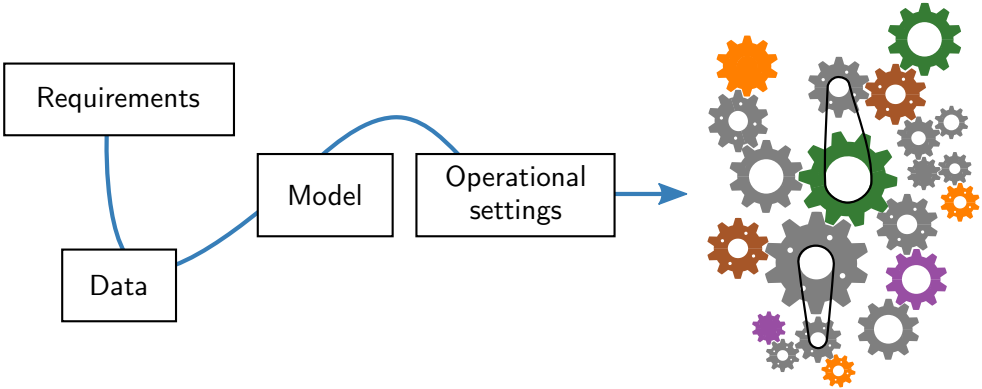
Constraint violations of resilient solutions in the test set are smaller. Closer to test set values

Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

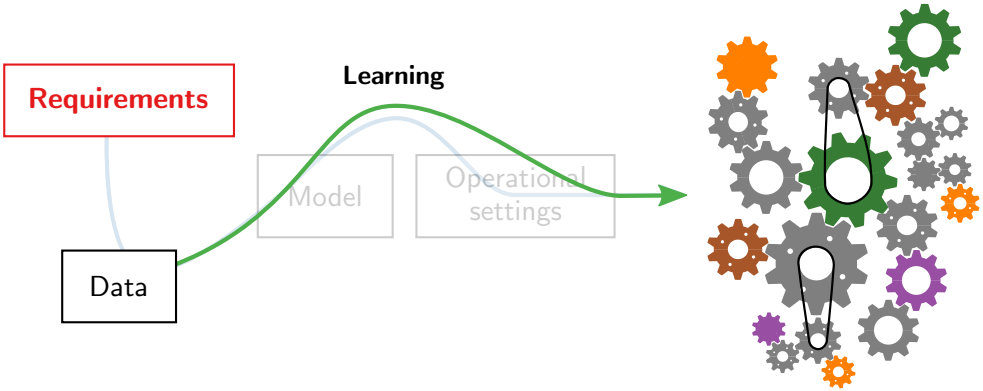
Learning Under Requirements

49 - 51

- Learning can transform systems engineering practice by automating the engineering design cycle



- But it can do so only if we incorporate requirements in the practice of machine learning



- In constrained learning, losses appear as objectives as well as statistical and pointwise constraints

$$P = \ell_0(\Phi_\theta^*) = \underset{\Phi_\theta}{\text{minimum}} \quad \ell_0(\Phi_\theta) = \mathbb{E}\Big[\ell_0(\Phi_\theta(x), y)\Big] \qquad \text{Minimize objective loss}$$

$$\text{subject to} \quad \ell_i(\Phi_\theta) = \mathbb{E}\Big[\ell_i(\Phi_\theta(x), y)\Big] \leq c_i \qquad \text{Statistical loss requirements}$$

$$\ell_i'(\Phi_\theta(x), y) \leq c_i \quad \text{a.e.} \qquad \text{Pointwise loss requirements}$$

- Find the parametric function Φ_θ^* that minimizes the statistical objective loss ℓ_0 while incurring ...

  ... at most c_i units of statistical constraint loss ℓ_i as well as ...

  ... at most c_i units of constraint loss ℓ_i' almost everywhere over the data distribution

- In constrained reinforcement learning, rewards appear as objectives and constraints

$$P = V_0(\pi^*) = \underset{\pi}{\text{maximum}} \quad V_0(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right] \qquad \text{Maximize objective reward}$$

$$\text{subject to} \quad V_i(\pi) := \mathbb{E}_{s,a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_t,a_t)\right] \geq c_i \qquad \text{Subject to reward requirements}$$

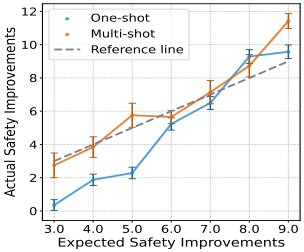
- Find the Policy π^* that maximizes the accumulation of objective reward r_0 while accumulating ...

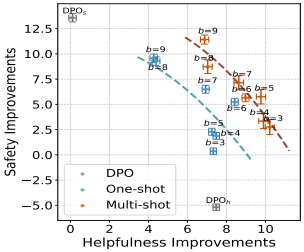
  ... at least c_i units of constraint reward r_i

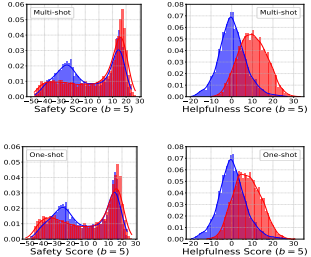
Requirements can be Transformative in Practice

51 - 53

- Align a pretrained LLM to enhance helpfulness and safety of text generated in response to prompts

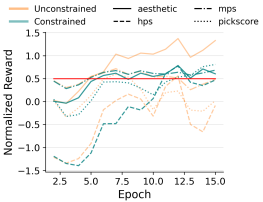


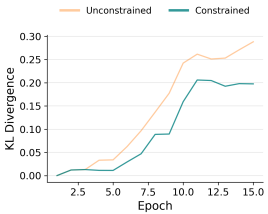




- Pareto front of optimality vs helpfulness moves right and up relative to state of the art heuristics

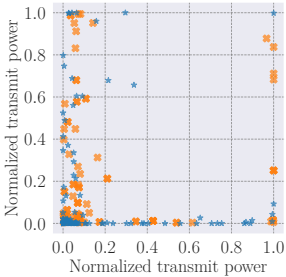
- Constrained composition sampling stays close to pre-trained while balancing different rewards

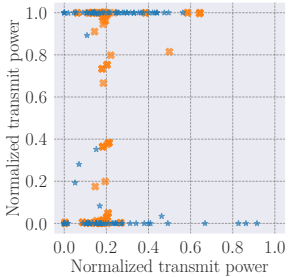


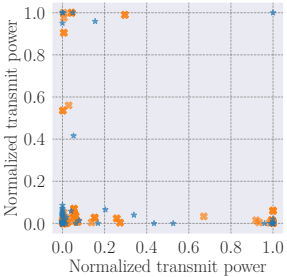


- Unconstrained composition deviates too much from pretrained model and overfits to some rewards

- Optimal resource allocation policy is stochastic $\Rightarrow$ Learn to sample from optimal distributions



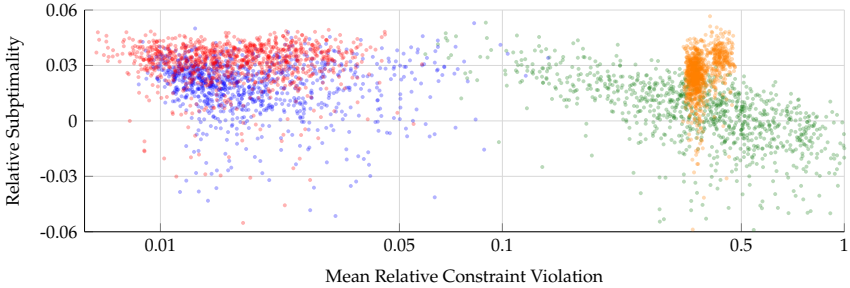




- Less constraints are violated and violated constraints are violated by smaller amounts

- Train a learning parameterization that optimizes cost objective while satisfying ....

  ... the conservation of power flow and operational constraints on buses and branches



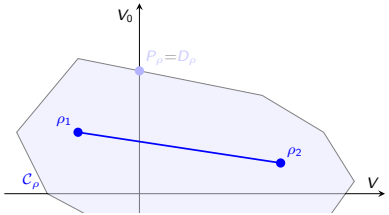
- Training with pointwise constraints (red and blue) is the only method with workable constraints

Requirements in AI Raise Interesting Fundamental Questions

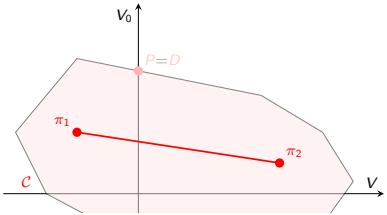
53 - 55

Strong Duality of Constrained Reinforcement Learning in Policy Space

If a strictly feasible policy exists, $P = D$ even though value functions $V_i(\pi)$ are not concave on π



$$V\left[\alpha\rho + (1-\alpha)\rho'\right] = \alpha V(\rho) + (1-\alpha)V(\rho')$$

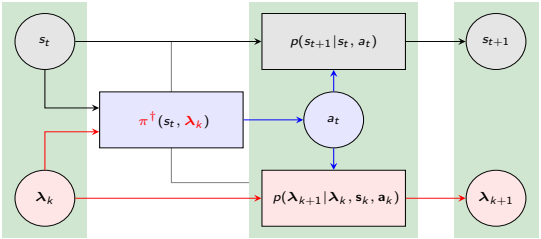


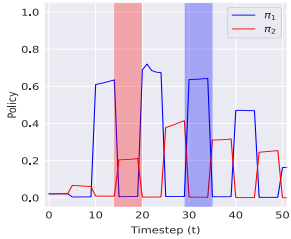
There exist π_α such that $V\left[\pi_\alpha\right] = \alpha V(\pi) + (1-\alpha)V(\pi')$

Paternain-Chamon-Calvo Fullana-Ribeiro, *Constrained Reinforcement Learning has Zero Duality Gap*, 2019, arxiv:1910.13393

State Augmented Constrained Reinforcement Learning

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

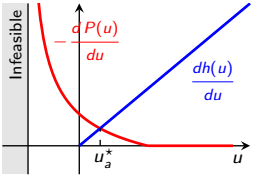


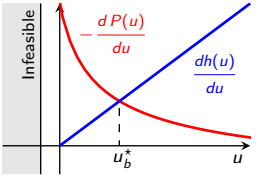


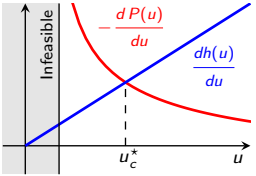
Calvo Fullana-Paternain-Chamon-Ribeiro, *State Augmented Constrained Reinforcement Learning*, 2021, arxiv:2102.11941

Resilient Constrained Reinforcement Learning

Adapt requirements (constraint levels c_i) to equate the marginal costs and benefits of relaxations







Hounie-Ribeiro-Chamon, *Resilient Constrained Learning*, 2023, arxiv:2306.02426

Ding-Huan-Ribeiro, *Resilient Constrained Reinforcement Learning*, 2023, arxiv:2312.17194

Systems Engineering and Artificial Intelligence

55 - 56

Claim 1. Systems Engineering and Artificial Intelligence (AI) are closer disciplines than is often recognized. We use more AI in systems engineering and more systems engineering in AI

Claim 2. Ignoring requirements is poor systems engineering practice. $\Rightarrow$ We can solve limitations of AI and we can expand its reach if we incorporate requirements in AI

Claim 3. Constrained (reinfocement) learning problems are interesting mathematical objects. They are not convex but have small duality gaps