Learning Safe Policies Via Primal-Dual Methods

Santiago Paternain, Miguel Calvo-Fullana, Luiz F.O. Chamon and Alejandro Ribeiro
Electrical and Systems Engineering, University of Pennsylvania
Email: {spater,cfullana,luizf,aribeiro}@seas.upenn.edu

58th IEEE Conference on Decision and Control
December 13th 2019, Nice, France

- Recent years of Reinforcement Learning have shown big success
  - ⇒ Able to deal with complex systems without need of modeling
  - ⇒ Easy to specify ⇒ just requires a reward signal
- Not enough ⇒ We need to be able to work with constraints
  - ⇒ In general we might be interested in performing several goals
  - ⇒ Or satisfy operation constraints
  - ⇒ In general engineering problems come in the form of specifications
- In this work we consider safety constraints ⇒ Non-convex problem
  - ⇒ We propose two relaxations to solve the problem
  - ⇒ The relaxed problem is as easy to solve as unconstrained RL
  - ⇒ The relaxations do not modify the performance much

- Markov Decision Process with state-action space $\mathcal{S} \times \mathcal{A} \subset \mathbb{R}^n \times \mathbb{R}^p$
- Where the transition probabilities satisfy the Markov property

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

- At each time-step the agent receives reward $r_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Consider a family of distributions π_θ parameterized by $\theta \in \mathbb{R}^d$
- We want to select the parameters that maximize the expected return

$$\max_{\theta \in \mathbb{R}^d} \mathbb{E}_{s,a \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t r_0(s_t, a_t) \right]$$

- We desire to learn policies that satisfy certain safety constraints

Safe Policies

- We say that a policy π_θ is $1-\delta_i$ safe for a set $\mathcal{S}_i \subset \mathcal{S}$ if

$$\mathbb{P}\left(\bigcap_{t=0}^{\infty}\{s_t \in \mathcal{S}_i\} \,\middle|\, \pi_\theta\right) \geq 1-\delta_i$$

- The goal is to maximize the return while remaining safe

$$\begin{aligned} \max_{\theta\in\mathbb{R}^d} \quad & \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r_0(s_t,a_t)\right] \\ \text{subject to} \quad & \mathbb{P}\left(\bigcap_{t=0}^{\infty}\{s_t \in \mathcal{S}_i\} \,\middle|\, \pi_\theta\right) \geq 1-\delta_i,\, i=1,\ldots,m. \end{aligned}$$

- The first challenge is that the problem is non-convex
  - $\Rightarrow$ We can solve a convex relaxation by solving the dual instead
- The second challenge is in computing the dual itself
  - $\Rightarrow$ Less obvious but the probability constraints make this difficult
  - $\Rightarrow$ So we will relax these constraints as well
- We try to answer how much is lost in these relaxations

- The goal is to maximize the return while remaining safe

$$\begin{aligned} \max_{\theta \in \mathbb{R}^d} \quad & \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_0(s_t, a_t)\right] \\ \text{subject to} \quad & \mathbb{P}\left(\bigcap_{t=0}^{\infty} \{s_t \in \mathcal{S}_i\} \,\middle|\, \pi_\theta\right) \geq 1 - \delta_i, i = 1, \ldots, m. \end{aligned}$$

- Define multipliers $\lambda \in \mathbb{R}_+^m$ and write the Lagrangian as

$$\mathcal{L}(\theta, \lambda) = \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_0(s_t, a_t)\right] + \sum_{i=1}^{m} \lambda_i \left(\mathbb{P}\left(\bigcap_{t=0}^{\infty} \{s_t \in \mathcal{S}_i\} \,\middle|\, \pi_\theta\right) - (1 - \delta_i)\right)$$

- The dual function $d(\lambda) = \max_{\theta \in \mathbb{R}^n} \mathcal{L}(\theta, \lambda)$ is convex on λ
  - $\Rightarrow$ Solving $\min_{\lambda \in \mathbb{R}_+^m} d(\lambda)$ is easy
  - $\Rightarrow$ Only provides an upper bound on the original problem
  - $\Rightarrow$ Challenge: How can we compute the maximization?

- The maximization of the Lagrangian relaxation is challenging

$$\min_{\lambda \in \mathbb{R}_+^m} \max_{\theta \in \mathbb{R}^n} \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_0(s_t, a_t)\right] + \sum_{i=1}^{m} \lambda_i \left(\mathbb{P}\left(\bigcap_{t=0}^{\infty} \{s_t \in \mathcal{S}_i\} \,\middle|\, \pi_\theta\right) - (1-\delta_i)\right)$$

- We propose to relax the probabilistic constraints in the following way

$$\min_{\lambda \in \mathbb{R}_+^m} \max_{\theta \in \mathbb{R}^n} \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_0(s_t, a_t)\right] + \sum_{i=1}^{m} \lambda_i \left(\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}\left(s_t \in \mathcal{S}_i\right)\right] - \frac{c_i}{1-\gamma}\right)$$

- Defining $r_\lambda(s,a) = r_0 + \sum_{i=1}^{m} \lambda_i (\mathbb{1}(s \in \mathcal{S}_i) - c_i)$

$$D_\theta^\star := \min_{\lambda \in \mathbb{R}_+^m} \max_{\theta \in \mathbb{R}^n} \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_\lambda(s_t, a_t)\right] \qquad \text{(DI)}$$

- The maximization can be solved using any RL algorithm
  - $\Rightarrow$ Solving the problem is as easy as solving an unconstrained RL problem
- We will see that not much is lost in these relaxations

- We propose to relax the probabilistic constraints as follows

$$\mathbb{P}\left(\bigcap_{t=0}^{\infty}\{s_t \in \mathcal{S}_i\} \,\middle|\, \pi_\theta\right) \geq 1-\delta_i \Rightarrow \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}\left(s_t \in \mathcal{S}_i\right)\right] \geq \frac{1-\delta_i+\nu_i}{1-\gamma}$$

- Any policy that is $1-\delta_i$ safe satisfies the relaxation with $\nu_i = 0$

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}\left(s_t \in \mathcal{S}_i\right)\right] = \sum_{t=0}^{\infty}\gamma^t \mathbb{P}\left(s_t \in \mathcal{S}_i\right) \geq \frac{1-\delta_i}{1-\gamma}$$

- Any policy that satisfies the relaxation with $\nu_i > 0$
  - $\Rightarrow$ Can be shown to be safe until a time horizon
  - $\Rightarrow$ Time horizon depends on how close is ν_i to δ_i

Theorem (Paternain et al'19)

Suppose there exists a policy $\pi_{\tilde{\theta}}$ and time horizons T_i such that $\pi_{\tilde{\theta}}$ is $(1-\gamma^{T_i}(1-\gamma)\delta_i)$-safe for the sets $\mathcal{S}_i$ with $i = 1, \ldots, m$. Then, the relaxation with $\nu_i = \delta_i(1-\gamma^{T_i}(1-\gamma))$ yields a $1-\delta_i$ safe policy for the sets $\mathcal{S}_i$ up to time T_i.

- The existence of a safer policy guarantees that
  - $\Rightarrow$ It is possible to tighten the constraint by increasing ν_i
  - $\Rightarrow$ Obtain a policy with the desired safety until a given time horizon
- We also have an analogous result for episodic problems
- We have not lost much in terms of safety with the relaxation

- Define $r_i(s_t, a_t) = \mathbb{1}(s_t \in \mathcal{S}_i)$ and $c_i = (1 - \delta_i + \nu_i)$
- The relaxation proposed induces the following optimization problem

  $\Rightarrow$ Maximize the expected return while satisfying a set of constraints

$$P_\theta^\star \triangleq \max_{\theta \in \mathbb{R}^d} \quad V_0(\pi_\theta) \triangleq \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_0(s_t, a_t)\right]$$

$$\text{subject to} \quad V_i(\pi_\theta) \triangleq \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t)\right] - \frac{c_i}{1-\gamma} \geq 0, i = 1, \ldots, m. \tag{PI}$$

- The dual of this problem yields the relaxation that we said we can solve
- Defining $r_\lambda(s, a) = r_0(s, a) + \sum_{i=1}^{m} \lambda_i(r_i(s, a)) - c_i)$

$$D_\theta^\star := \min_{\lambda \in \mathbb{R}_+^m} \max_{\theta \in \mathbb{R}^n} \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_\lambda(s_t, a_t)\right] \tag{DI}$$

- We are left to characterize the loss of optimality in this relaxation

Small Duality Gap

- π_θ is an ϵ-universal parameterization of functions $\pi \in \mathcal{P}(\mathcal{S})$ if

$$\max_{s\in\mathcal{S}} \int_{\mathcal{A}} |\pi(a|s) - \pi_\theta(a|s)| \; da \leq \epsilon$$

Theorem (Paternain et al'19)

Suppose that r_i is bounded for all $i = 0, \ldots, m$ by constants $B_{r_i} > 0$ and define and $B_r = \max_{i=1\ldots m} B_{r_i}$. Let $\lambda_\epsilon^\star$ be the solution to the following problem

$$\lambda_\epsilon^\star \triangleq \min_{\lambda\in\mathbb{R}_+^m} \max_{\pi\in\mathcal{P}(\mathcal{S})} V_0(\pi) + \sum_{i=1}^m \lambda_i \left(V_i(\pi) - B_r \frac{\epsilon}{1-\gamma} \right).$$

If the parametrization π_θ is an $\epsilon-$universal parametrization of functions $\pi \in \mathcal{P}(\mathcal{S})$ and Slater's condition holds for (PI)*, it follows that*

$$P_\theta^\star \geq D_\theta^\star \geq P_\theta^\star - \left(B_{r_0} + \left\|\lambda_\epsilon^\star\right\|_1 B_r\right) \frac{\epsilon}{1-\gamma},$$

where $P_\theta^\star$ is the optimal value of (PI)*, and $D_\theta^\star$ the value of problem* (DI)*.*

- The better the parameterization the smaller is ϵ

$$P_\theta^\star \geq D_\theta^\star \geq P_\theta^\star - \left(B_{r_0} + \|\lambda_\epsilon^\star\|_1 B_r\right) \frac{\epsilon}{1-\gamma},$$

$\Rightarrow$ The closer we are from solving (PI) by solving (DI)

- The two relaxations introduced are such that

$\Rightarrow$ We can still guarantee safety if a safer policy exists

$\Rightarrow$ The loss in optimality can be made arbitrarily small

$\Rightarrow$ We constructed a formulation that allows us to solve the problem

$\Rightarrow$ Not harder to solve than unconstrained Reinforcement Learning

- The proposed relaxations yields the following problem

$$D_\theta^\star := \min_{\lambda \in \mathbb{R}_+^m} \max_{\theta \in \mathbb{R}^n} \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_\lambda(s_t, a_t)\right] \quad \text{(DI)}$$

- Where we have defined $r_\lambda(s,a) = r_0 + \sum_{i=1}^m \lambda_i (r_i(s,a)) - c_i)$
- Solving the maximization is not harder than solving a RL problem
- If we have $\theta^\star(\lambda) := \operatorname{argmax}_\theta \mathcal{L}_\theta(\theta, \lambda)$
- Let us define the dual function associated to the CRL problem

$$d_\theta(\lambda) = \max_\theta \mathcal{L}_\theta(\theta, \lambda)$$

- The dual function is the point-wise maximum of linear functions
  - $\Rightarrow$ It is a convex function $\Rightarrow$ Easy to solve with SGD
  - $\Rightarrow$ Danskin's Theorem guarantees that $\nabla d_\theta(\lambda) = V(\theta^\star(\lambda))$
  - $\Rightarrow$ Gradient of the dual function solves the problem (DI)

- Policy Gradient algorithms solve RL problems ⇒ Can compute $\theta^\star(\lambda)$

$$\theta_{k+1} = \theta_k + \eta_\theta \nabla_\theta \mathcal{L}_\theta(\theta_k, \lambda_k)$$

- In parallel we can run the dual step

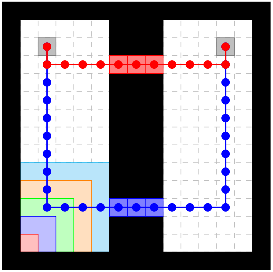
$$\lambda_{k+1} = \left[\lambda_k - \eta_\lambda \nabla_\lambda \mathcal{L}(\theta_k, \lambda_k)\right]_+$$

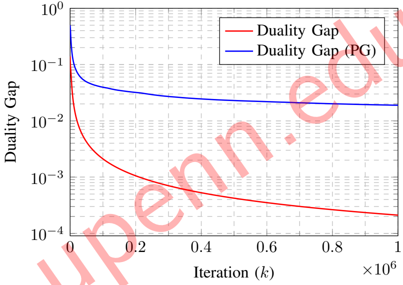
- Typically one needs to chose $\eta_\lambda \ll \eta_\theta$ so λ is approximately constant

Theorem (Paternain et al'19)

If policy gradient finds a solution $\theta^\dagger(\lambda_k)$ that is β-suboptimal, $\mathcal{L}(\theta^\dagger(\lambda_k), \lambda_k) + \beta \geq \mathcal{L}(\theta^\star(\lambda_k), \lambda_k)$ Then the primal-dual algorithm converges in $K \leq \|\lambda_0 - \lambda_\theta^\star\|^2 / (2\eta\varepsilon)$ iterations to a neighborhood of $D_\theta^\star$

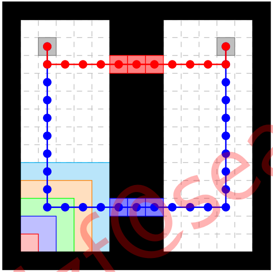
$$d_\theta(\lambda_k) \leq D_\theta^\star + O(\eta, \beta, \varepsilon)$$

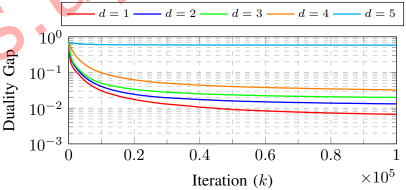




- We consider a gridworld $\Rightarrow$ Agent must navigate from left to right
  - $\Rightarrow$ Red bridge is unsafe while blue bridge is safe
  - $\Rightarrow$ Constrains the agent to not cross the unsafe bridge with 99%
- In this problem we can compute the global primal minimizer
  - $\Rightarrow$ This allows us to explicitly characterize the duality gap.
- Duality gap effectively vanishes for exact minimization
- Duality gap goes to a neighborhood for a single policy gradient step.

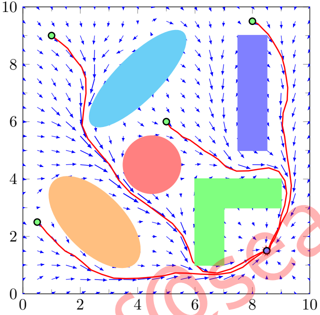
- The effect of parametrization on the duality gap is such that
  - ⇒ Duality gap increases with parametrization coarseness
  - ⇒ Theoretical duality gap depended on its richness

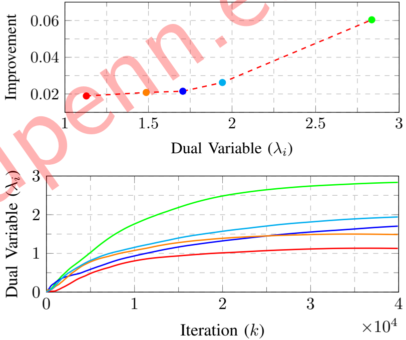




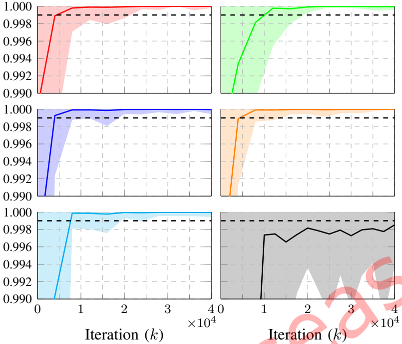
Numerical Examples

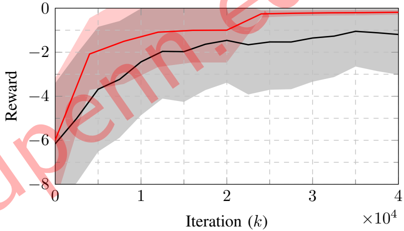
- Consider now safe navigation in an obstacle-ridden environment





- Constrained Reinforcement Learning learns to avoid obstacles
  - $\Rightarrow$ The value of each obstacle is given by the value of its dual variable





- Safety is satisfied for all obstacles and reward is maximized
- Compared with a naive approach (black curves)
  - ⇒ Set the weights to the min/max values of the dual variables
  - ⇒ CRL outperforms and methodologically satisfies the constraints

- We need to be able to work with constraints
  - $\Rightarrow$ In this work we considered safety constraints
- We proposed two relaxations to compute safe policies
  - $\Rightarrow$ Safe policies can be achieved if a safer policy exists
  - $\Rightarrow$ The relaxation of the dual problem yields small duality gap
  - $\Rightarrow$ The gap depends of the how rich the parameterization is
- The relaxations yield a problem formulation that can be solved
  - $\Rightarrow$ Using for instance Primal-Dual methods
  - $\Rightarrow$ As easy as solving unconstrained RL problems

- Let us consider a non-parametric policy $\pi \in \mathcal{P}(\mathcal{S})$
  - $\Rightarrow$ Where $\mathcal{P}(\mathcal{S})$ is the space of probability measures on $(\mathcal{A}, \mathcal{B}(\mathcal{A}))$
- In this case the Constrained Reinforcement Learning Problem is

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

- Problem (PII) upper bounds the parametric problem $\Rightarrow P_\theta^\star \leq P^\star$
  - $\Rightarrow$ Not solvable, however it is important for theoretical results
  - $\Rightarrow$ Also holds that $D_\theta^\star \leq P^\star$. Can we provide a lower bound for $D_\theta^\star$?

Theorem

Suppose that r_i is bounded for all $i = 0, \dots, m$ and that Slater's condition holds for (PII)*. Then, strong duality holds for* (PII)*, i.e.,* $P^\star = D^\star$.

- Idea of the proof:
  - $\Rightarrow$ Let us define the perturbation function associated to (PII)

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

  - $\Rightarrow$ If $P(\xi)$ is concave $\Rightarrow$ Then zero duality holds (Fenchel-Moreau)

- Define the occupation measure $\rho_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_\pi^t(s,a)$
- Construct the following problem equivalent to ($\tilde{\text{PI}}$)

$$\begin{aligned} P(\xi) = \max_{\rho_\pi \in \mathcal{R}} \quad & \int_{\mathcal{S}\times\mathcal{A}} r_0(s,a)\, d\rho_\pi \\ \text{subject to} \quad & \int_{\mathcal{S}\times\mathcal{A}} r_0(s,a)\, d\rho_\pi \geq c_i + \xi_i,\ i = 1,\ldots,m. \end{aligned} \qquad (\tilde{\text{PI}}')$$

- The set $\mathcal{R}$ is a convex set (Borkar'88)
- Then ($\tilde{\text{PI}}'$) is a convex optimization problem
  - ⇒ Its perturbation function is concave