Constrained Reinforcement Learning Has Zero Duality Gap

Santiago Paternain, Luiz F. O. Chamon, Miguel Calvo-Fullana and Alejandro Ribeiro
University of Pennsylvania, Philadelphia, USA

Why Constrained Reinforcement Learning?

- We want agents to perform multiple tasks with some success level
  - ⇒ We can have m reward signals $r_i(s,a)$ with $i=1,\ldots,m$
  - ⇒ We want them all to be larger than some value c_i
- Physical systems are subject to different restrictions
  - ⇒ Level of battery being larger than some value
  - ⇒ Avoiding obstacles or unsafe portions of the state space
- Most approaches to tackle this problem are either
  - ⇒ Integrating prior-knowledge
  - ⇒ Manual selection of Lagrange multipliers
  - ⇒ Primal-Dual methods

Constrained Reinforcement Learning Framework

- 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 $m+1$ rewards $r_i : \times\mathcal{S}\times\mathcal{A}\to\mathbb{R}$
- 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 while satisfying a set of constraints

$$P_\theta^\star \triangleq \max_{\theta\in\mathbb{R}^d} \; V_0(\theta) \triangleq \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^\infty \gamma^t r_0(s_t,a_t)\right]$$
$$\text{subject to } \; V_i(\pi_\theta) \triangleq \mathbb{E}_{s,a\sim\pi_\theta}\left[\sum_{t=0}^\infty \gamma^t r_i(s_t,a_t)\right] \geq c_i,\, i=1,\ldots,m. \qquad \text{(PI)}$$

- This is the Constrained Reinforcement Learning (CRL) problem
- An approach to solve these problems is to use Primal-Dual methods

Why Primal-Dual methods?

- Why use Primal-Dual methods compared to other approaches?
- Prior domain knowledge
  - ⇒ Project chosen actions to a set that ensures the constraints
  - ✗ Safety is not guaranteed unless similar transitions have been observed
  - ✗ Projection might result in sub-optimal operation
- Manual selection of Lagrange Multipliers
  - ✗ The weight of each constraint needs to be hand tuned
  - ✗ For each set of penalty coefficients there are different solutions
  - ✗ It is domain dependent
  - ✗ Competing resources might lead to training plateaus
- Primal-Dual methods
  - ✓ Can be been used successfully
  - ✓ The dual function is always convex
  - ✓ Deal directly with the constraints is not more complicated
  - ✓ Solving the dual can be shown to not be harder than classic RL

Main Contribution

- Constrained Reinforcement Learning has zero duality gap
- Arbitrarily small gap for rich parametrization of the policies
- Solving the dual problem is as good as solving the original problem

Example: Learning Safe Policies

- In this example we are concerned about safety
- We want to maximize the return while remaining on safe sets $\mathcal{S}_i\subset\mathcal{S}$

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

- With high probability for all $i=1,\ldots,m$
- The previous constraint can be relaxed to be of the form

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

- Any policy that satisfies the previous expression
  - ⇒ Can be shown to be safe until a time horizon
  - ⇒ Time horizon depends on how close is ν to δ

Working on the Dual Domain

- Let us define the dual function associated to the CRL problem

$$d_\theta(\lambda) = \max_\theta \mathcal{L}_\theta(\theta,\lambda) = \max_\theta V_0(\theta) + \sum_{i=1}^m \lambda_i V_i(\theta)$$

- The dual function is the point-wise maximum of linear functions
  - ⇒ It is a convex function ⇒ Easy to solve with SGD
  - ⇒ Danskin's Theorem guarantees that $\nabla d_\theta(\lambda) = V(\theta^*(\lambda))$
- If we have $\theta^*(\lambda) := \operatorname{argmax}_\theta \mathcal{L}_\theta(\theta,\lambda)$
  - ⇒ Gradient of the dual function solves the problem

$$D_\theta^\star \triangleq \min_{\lambda\in\mathbb{R}_+^m} d_\theta(\lambda). \qquad \text{(DI)}$$

- There are some limitations of the dual solution
- It only provides a lower bound on the problem (PI)

$$P_\theta^\star \leq D_\theta^\star$$

- We show that actually the sub-optimality is arbitrarily small
- Solving the primal problem might not be possible
  - ⇒ However it is not more difficult than solving a classic RL problem

Primal-Dual Algorithm

- Dual gradient descent requires the computation of

$$\theta^*(\lambda) = \operatorname*{argmax}_{\theta\in\mathbb{R}^d} \mathcal{L}_\theta(\theta,\lambda)$$

- Notice that the Lagrangian can be written as

$$\mathcal{L}_\theta(\theta,\lambda) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t\left(r_0(s_t,a_t) + \sum_{i=1}^m \lambda_i\left(r_i(s_t,a_t) - c_i(1-\gamma)\right)\right)\right]$$

- Let us define a reward depending on the multipliers

$$r_\lambda(s,a) = r_0(s,a) + \sum_{i=1}^m \lambda(r_i(s,a) - c_i(1-\gamma))$$

- Then the Lagrangian can be written as an expected discounted return

$$\mathcal{L}_\theta(\theta,\lambda) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_\lambda(s_t,a_t)\right]$$

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

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

- In parallel the dual step can be run

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

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

Dual descent convergence

If policy gradient finds a solution $\theta^\dagger(\lambda_k)$ that is β-suboptimal,

$$\mathcal{L}(\theta^\dagger(\lambda_k),\lambda_k) + \beta \geq \mathcal{L}(\theta^*(\lambda_k),\lambda_k)$$

Then the primal-dual algorithm converges to a neighborhood of $D_\theta^\star$

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

in $K \leq \|\lambda_0 - \lambda_\theta^\star\|^2/(2\eta\varepsilon)$ iterations.

- The previous result is only useful if sub-optimality is not large

The non-parametric Constrained Reinforcement Learning Problem

- Let us consider a non-parametric policy $\pi\in\mathcal{P}(\mathcal{S})$
  - ⇒ 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

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

- Problem (PII) upper bounds the parametric problem ⇒ $P_\theta^\star \leq P^\star$
  - ⇒ Not solvable, however it is important for theoretical results
- Define the Dual function associated to (PII)

$$d(\lambda) = \max_\theta \mathcal{L}(\theta,\lambda) = \max_\theta V_0(\theta) + \sum_{i=1}^m \lambda_i U_i(\theta)$$

- Then the dual problem is that of finding the best upper bound for (PII)

$$D^\star \triangleq \min_{\lambda\in\mathbb{R}_+^m} d(\lambda). \qquad \text{(DII)}$$

Zero Duality Gap of Constrained Reinforcement Learning

Theorem: Zero Duality Gap

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

- We follow with the reasoning as to why this result holds
- Let us define the perturbation function associated to (PII)

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

- If $P(\xi)$ is concave ⇒ 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 $(\tilde{\text{PII}})$

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

- The set $\mathcal{R}$ is a convex set (Borkar'88)
- Then $(\tilde{\text{PII}}')$ is a convex optimization problem
  - ⇒ In fact it is linear
  - ⇒ It's perturbation function is concave

Almost Zero Duality Gap for Parametric Problems

- For the problem (PI) we have a duality gap that will depend on the quality of the parametrization
- We say that a parameterization π_θ is an ϵ-universal parameterization of functions $\pi\in\mathcal{P}(\mathcal{S})$ if

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

- This is a requirement on the total variation norm
  - ⇒ Milder than approximation in uniform bound
  - ⇒ Satisfied by RBF networks, RKHS, and deep neural networks

Theorem: Almost Zero Duality Gap for parametric problems

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 folowing min-max problem

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

Then, 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

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

where $P^\star$ is the optimal value of (PII), and $D_\theta^\star$ the value of the parametrized dual problem (DI).

- The better the parameterization the smaller is ϵ
- The closer we are from solving (PII) by solving (DI)
- What about infeasible problems?
  - ⇒ If (PI) is infeasible then $D_\theta^\star = -\infty$
  - ⇒ Right hand side inequality holds trivially
  - ⇒ If infeasible then there is no solution to Problem $(\tilde{\text{PII}})$ with $\xi_i = B_r\epsilon/(1-\gamma)$ because π_θ is an ϵ-parameterization of $\mathcal{P}(\mathcal{S})$
  - ⇒ Then, $\lambda_\epsilon^\star$ is infinity ⇒ Right hand side of the bound holds too

Primal-Dual Convergence

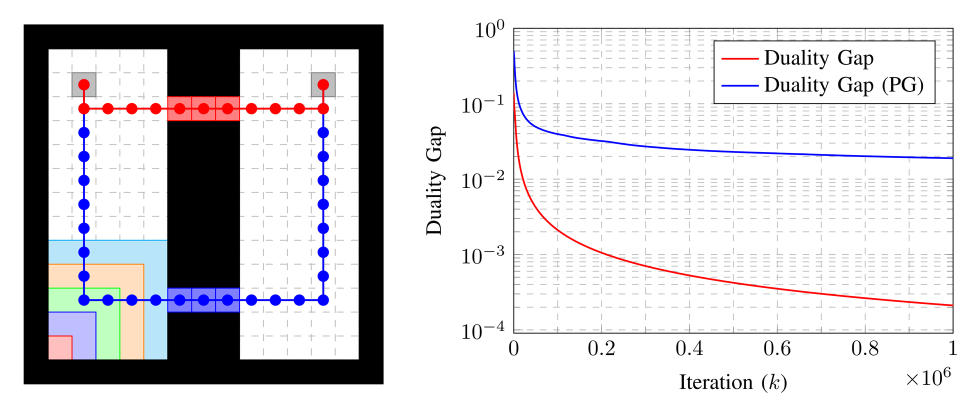
- Combining all the previous results
  - ⇒ Classic convergence of Primal-Dual Algorithm
  - ⇒ Almost zero duality gap
- We can provide a bound on the number of iterations needed to reach a neighborhood of the primal

Theorem: Convergence of Primal-Dual algorithms

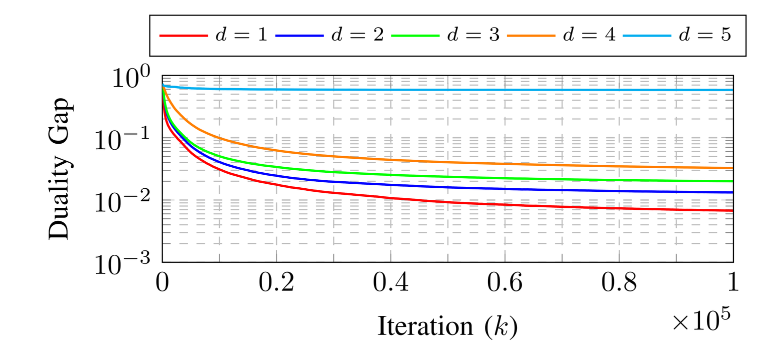
Under the hypothesis of the previous theorem in $K \leq \|\lambda_0 - \lambda_\theta^\star\|^2/(2\eta\epsilon)$ iterations the dual solution is such that

$$P^\star + O(\eta,\beta,\varepsilon) \geq d_\theta(\lambda_K) \geq P^\star - \left(B_{r_0} + \|\lambda_\epsilon^\star\|_1 B_r\right)\frac{\epsilon}{1-\gamma}.$$

Example: Duality Gap



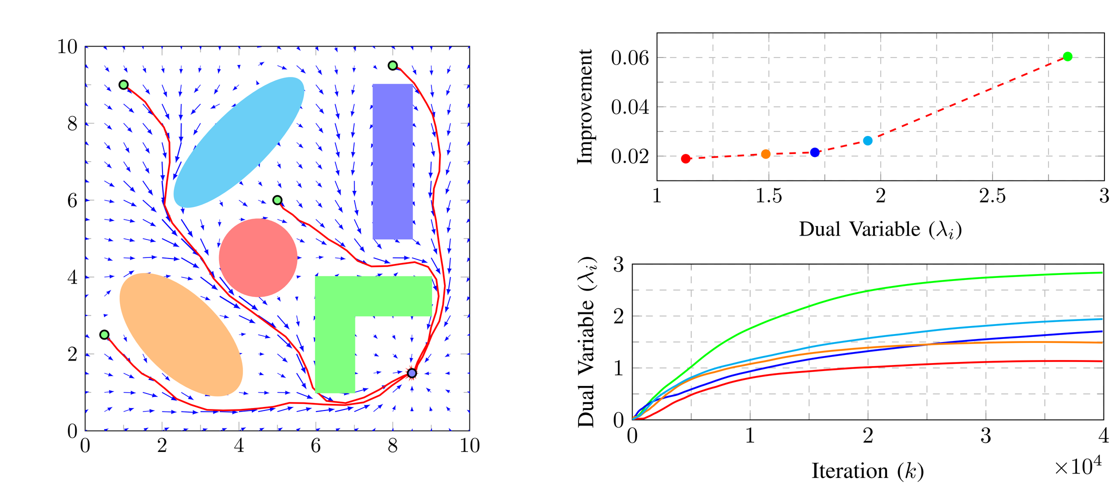
- We consider a gridworld navigation scenario
  - ⇒ Agent must navigate from left to right
  - ⇒ Red bridge is unsafe while blue bridge is safe
  - ⇒ Constrain the agent to not cross the unsafe bridge with 99%
- In this problem we can compute the global primal minimizer
  - ⇒ E.g., via Dijkstra's algorithm for a given value of the dual variables
  - ⇒ 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.



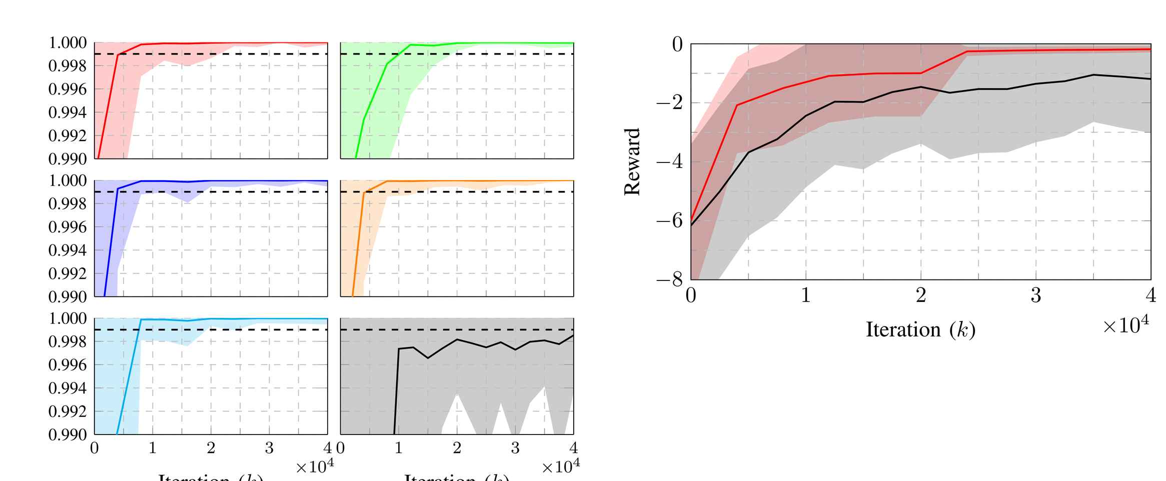
- Duality gap increases with parametrization coarseness

Example Application: Safe Navigation on Continuous Spaces

- Consider now safe navigation in an obstacle-ridden environment



- Constrained Reinforcement Learning learns to avoid obstacles
  - ⇒ 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

Conclusions

- Constrained RL problems have almost zero duality gap
  - ⇒ The gap depends of the how rich the parametrization is
  - ⇒ In some cases we can achieve zero duality gap
- Solving constrained RL problems is easy
  - ⇒ As easy as solving unconstrained RL problems
- Primal-Dual converges to the optimal solution
  - ⇒ If the computation of the primal is accurate