URJC
Penn University of Pennsylvania

Deterministic Policy Gradient Primal-Dual Methods for Continuous-Space Constrained MDPs

S. Rozada † D. Ding ‡ A. G. Marques † A. Ribeiro ‡

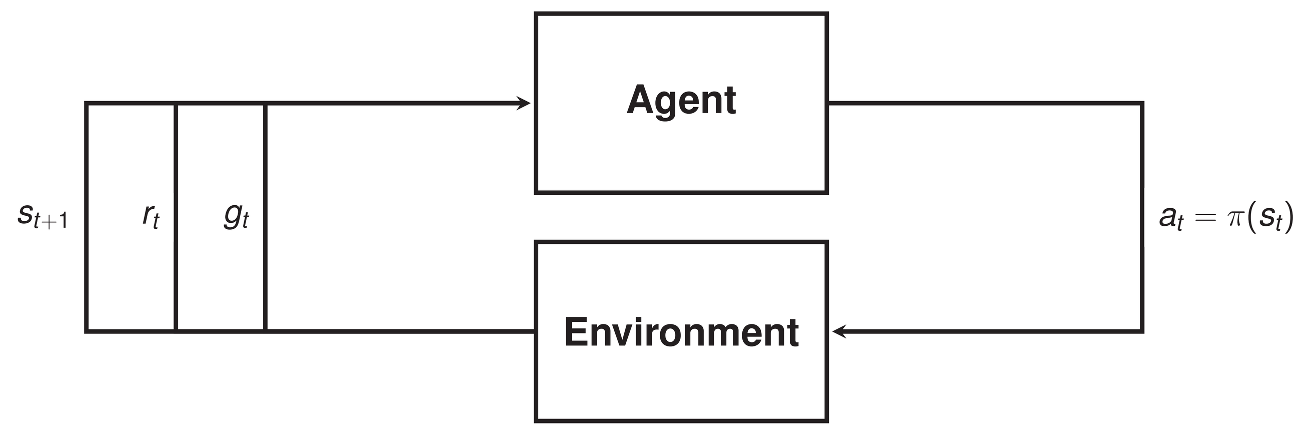
† King Juan Carlos University, Madrid, Spain ‡ University of Pennsylvania, Philadelphia, PA, US

Key Contributions

- This paper proposes **deterministic-policy search** for solving **constrained MDPs**
  C1) Zero-duality gap $\Rightarrow$ Despite deterministic policies
  C2) Deterministic PG primal-dual method $\Rightarrow$ Sub-linear convergence rate
  C3) Sample-based approximation $\Rightarrow$ Sub-linear convergence rate

Continuous-Space Constrained MDPs

- We solve **continuous-space constrained MDPs**
  - ▷ Continuous-state space $S \subseteq \mathbb{R}^{d_S}$ and continuous-action space $A \subseteq \mathbb{R}^{d_A}$
  - ▷ Probability transition function $p(s' \mid s, a)$ and initial-state distribution ρ
  - ▷ Reward function $r(s, a)$ and utility function $g(s, a)$
- We consider **deterministic policies** $\Rightarrow a = \pi(s)$
  - ▷ More practical for real-world applications
  - ▷ Crucial for safety-critical domains



- Goal $\Rightarrow$ Maximize $V_r(\pi) := \mathbb{E}_\rho[V_r^\pi(s)]$ ensuring $V_g(\pi) := \mathbb{E}_\rho[V_g^\pi(s)]$ is sufficiently good

$$V_r^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t) \mid s_0 = s\right] \quad \text{and} \quad V_g^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t g(s_t, a_t) \mid s_0 = s\right]$$

Problem Formulation

- Continuous-space constrained MDP optimizing over the class of deterministic policies Π

$$V_P^{\pi^\star} := \max_{\pi \in \Pi} V_r(\pi) \quad \text{s.t.} \quad V_g(\pi) \geq 0 \qquad \text{(P-CRL)}$$

- Workhorse of constrained RL $\Rightarrow$ **Lagrangian method**

$$L(\pi, \lambda) := V_r(\pi) + \lambda V_g(\pi) \quad \Leftrightarrow \quad L(\pi, \lambda) := V_\lambda(\pi) \quad \text{with} \quad r_\lambda(s, a) := r(s, a) + \lambda g(s, a)$$

- Minimize the dual function $\Rightarrow$ Upper bound of (P-CRL)

$$V_D^{\lambda^\star} := \min_{\lambda \in \mathbb{R}^+} D(\lambda) \quad \text{with} \quad D(\lambda) := \max_{\pi \in \Pi} V_\lambda(\pi) \qquad \text{(D-CRL)}$$

- The primal problem (P-CRL) and the dual problem (D-CRL) are
  - ▷ Tractable for stochastic policies $\Rightarrow$ Rich literature of methods
  - ▷ Considered to be challenging for deterministic policies

P1) Deterministic policies sub-optimal in discrete constrained MDPs [Altman, Rout.2021]
P2) Searching for deterministic policies is an NP-complete problem [Dolgov, IJCAI2005]

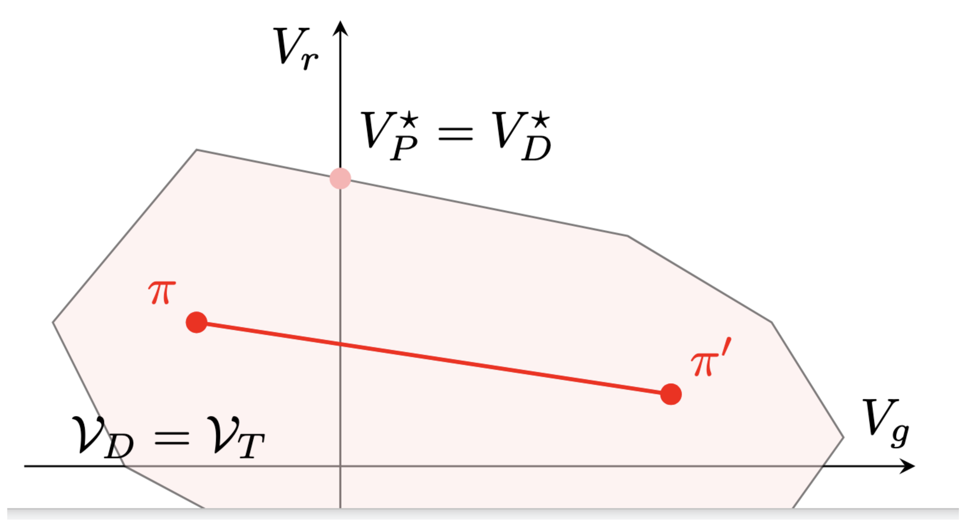
Addressing P1: Sufficiency of Deterministic Policies

- Deterministic policies are sufficient under **non-atomicity**
  - ▷ Vector of value functions $V(\pi) = [V_r(\pi), V_g(\pi)]^\top$
  - ▷ Value images $\mathcal{V}_T = \{V(\pi)$ for all policies$\}$ and $\mathcal{V}_D = \{V(\pi) \mid \pi \in \Pi\}$

Lemma: Sufficiency of deterministic policies [Feinberg, SICON2019]
For a non-atomic discounted MDP with continuous spaces, the deterministic value image $\mathcal{V}_D$ is convex, and equals the value image $\mathcal{V}_T$, i.e., $\mathcal{V}_D = \mathcal{V}_T$

- Continuous-space constrained RL with deterministic policies has **zero duality gap**

Theorem: Zero duality gap for deterministic policies
Under non-atomicity, problem (P-CRL) has zero duality gap, i.e., $V_P^{\pi^\star} = V_D^{\lambda^\star}$



Addressing P2: Regularized Lagrangian

- **Regularized Lagrangian** $\Rightarrow$ Smooth optimization landscape limiting optimality loss
  - ▷ Primal regularization $\Rightarrow H(\pi) := \mathbb{E}_\rho[H^\pi(s)]$ with $H^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=0}^\infty -\gamma^t \|\pi(s_t)\|^2 \mid s\right]$
  - ▷ Dual regularization $\Rightarrow h(\lambda) := \lambda^2$

$$L_\tau(\pi, \lambda) := V_\lambda(\pi) + \frac{\tau}{2}H(\pi) + \frac{\tau}{2}h(\lambda) \quad \Leftrightarrow \quad L_\tau(\pi, \lambda) := V_{\lambda,\tau}(\pi) + \frac{\tau}{2}h(\lambda)$$

- Solve the **saddle-point** problem

$$\min_{\lambda \in \Lambda} \max_{\pi \in \Pi} V_{\lambda,\tau}(\pi) + \frac{\tau}{2}h(\lambda) \qquad \text{(R-CRL)}$$

Deterministic Policy Gradient Primal-Dual Method (D-PGPD)

- **D-PGPD** $\Rightarrow$ Maximizes regularized advantage $A_{\lambda_t,\tau}^{\pi_t}$ associated with $V_{\lambda_t,\tau}^{\pi_t}$

$$\pi_{t+1}(s) = \operatorname{argmax}_{a \in A} \; A_{\lambda_t,\tau}^{\pi_t}(s, a) - \frac{1}{2\eta}\|a - \pi_t(s)\|^2 \qquad \text{(D-PGPD-P)}$$

$$\lambda_{t+1} = \operatorname{argmin}_{\lambda \in \Lambda} \; \lambda\left(V_g(\pi_t) + \tau\lambda_t\right) + \frac{1}{2\eta}\|\lambda - \lambda_t\|^2 \qquad \text{(D-PGPD-D)}$$

- Analysis requires mild technical conditions
  - ▷ Function $Q_{\lambda,\tau}^\pi(s, a) - \tau_0\|a - \pi_0(s)\|^2$ concave in a
- Convergence assessed via $\Rightarrow \Phi_t \approx \mathbb{E}\left[\|\mathcal{P}_{\Pi_\tau^\star}(\pi_t(s)) - \pi_t(s)\|^2\right] + \|\mathcal{P}_{\Lambda_\tau^\star}(\lambda_t) - \lambda_t\|^2$

Theorem: Sub-linear convergence of D-PGPD
For $\tau > \tau_0$, the primal-dual iterates of D-PGPD satisfy $\Phi_{t+1} \leq e^{-\beta_0 t}\Phi_1 + \beta_1 C_0^2$

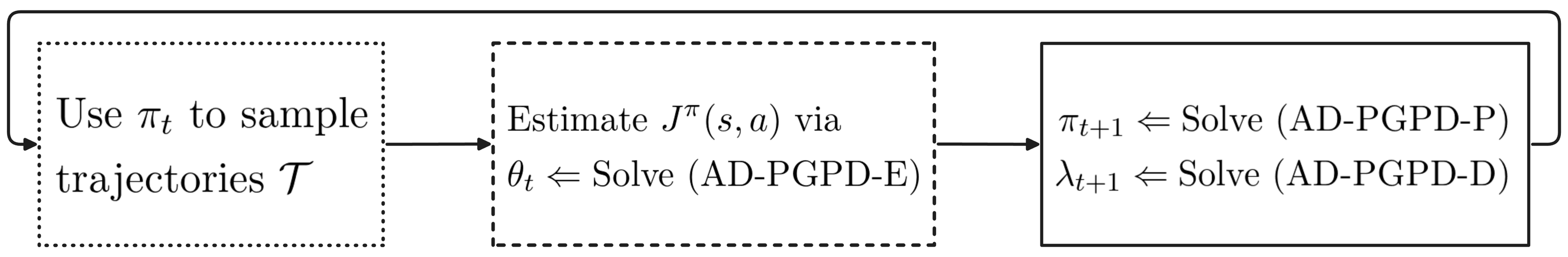
- Convergence to a neighborhood at sub-linear rate
  - ▷ C_0 depends on MDP parameters
  - ▷ β_0, β_1 depend on $\eta \Rightarrow \epsilon$-convergence in $O(\epsilon^{-1})$ iterations with $\eta = O(\epsilon)$

Corollary: Close-optimality of RD-CRL
If $\eta = O(\epsilon^4)$, $\tau = O(\epsilon^2) + \tau_0$, $t = \Omega(\epsilon^{-6}\log^2\epsilon^{-1})$, close-optimality and near-feasibility follow

$$V_r(\pi^\star) - V_r(\pi_t) \leq \epsilon - \tau_0 H(\pi^\star)$$
$$V_g(\pi_t) \geq -\epsilon + \tau_0 H(\pi^\star)(\lambda_{\max} - \lambda^\star)^{-1}$$

- D-PGPD requires computing value functions in closed form

Approximate Deterministic-Policy Search method



- **AD-PGPD** $\Rightarrow$ Approximate D-PGPD to avoid closed-form computations
  - ▷ Approximate augmented action-value function $J^\pi(s, a) := Q_{\lambda,\tau}^\pi(s, a) + \frac{1}{\eta}\pi(s)^\top a$

$$\theta_t = \operatorname{argmin}_\theta \; \mathbb{E}_{(s,a)\sim\nu}\left[\|\phi(s, a)^\top\theta - J^{\pi_t}(s, a)\|^2\right] \qquad \text{(AD-PGPD-E)}$$

$$\pi_{t+1}(s) = \operatorname{argmax}_{a \in A} \; \hat{J}_{\theta_t}(s, a) - \left(\frac{\tau}{2} + \frac{1}{2\eta}\right)\|a\|^2 \qquad \text{(AD-PGPD-P)}$$

$$\lambda_{t+1} = \operatorname{argmin}_{\lambda \in \Lambda} \; \lambda\left(V_g(\pi_t) + \tau\lambda_t\right) + \frac{1}{2\eta}\|\lambda - \lambda_t\|^2 \qquad \text{(AD-PGPD-D)}$$

- Extending convergence analysis requires boundedness of approximation error
  - ▷ Function $\hat{J}_\theta(s, a) - \tau_0\|a - \pi_0(s)\|^2 \Rightarrow$ Concave in a

Theorem: Sub-linear convergence of AD-PGPD
For $\tau > \tau_0$, the primal-dual iterates of AD-PGPD satisfy $\Phi_{t+1} \leq e^{-\beta_0 t}\Phi_1 + \beta_1 C_0^2 + \beta_2\epsilon_{\text{approx}}$

- Convergence depends on approximation error $\Rightarrow \beta_2$ depends on $1/(\tau - \tau_0)$
- **Sample-based AD-PGPD** $\Rightarrow$ Learn approximator from trajectories $\mathcal{T}$ using SGD
  - ▷ Basis functions, bias and approximation errors are bounded
  - ▷ Non-zero probability of sampling optimal state-action pairs

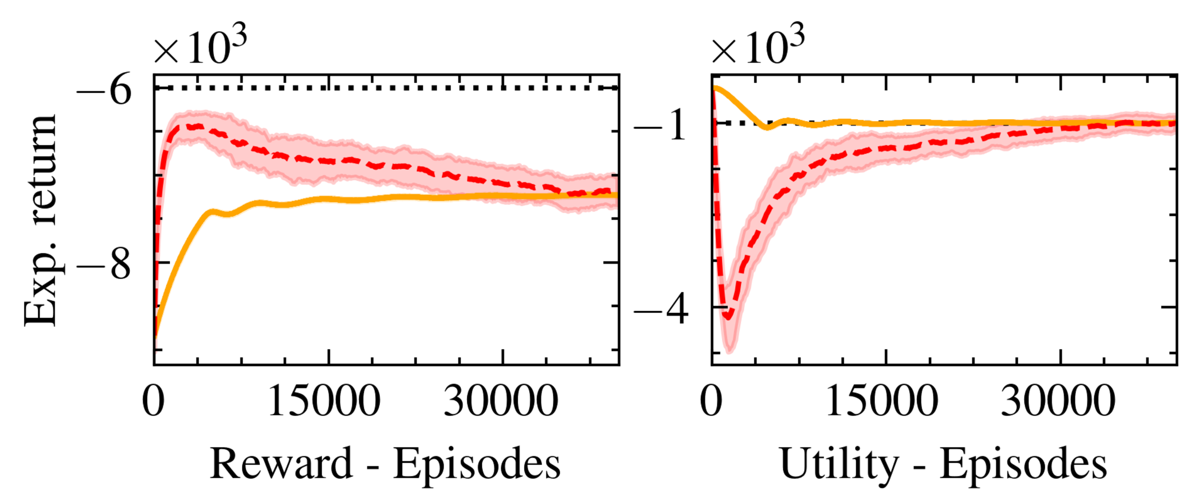
Corollary: Sub-linear convergence of sample-based AD-PGPD
For $\tau > \tau_0$, the iterates of the sample-based A-PGPD satisfy

$$\mathbb{E}[\Phi_{t+1}] \leq e^{-\beta_0 t}\mathbb{E}[\Phi_1] + \beta_1 C_0^2 + \beta_2\left(\frac{C_1^2}{\eta^2(N+1)} + \epsilon_{\text{bias}}\right)$$

- Convergence depends on number of samples and bias error
  - ▷ C_1 depends on MDP parameters
  - ▷ N is the number of samples for approximation

Numerical Experiments

- Continuous velocity-constrained robot navigation $\Rightarrow$ absolute-value rewards
  - ▷ Sample-based AD-PGPD (—) vs. dual-based baseline PGDual (- -)



- Continuous constrained fluid-velocity control $\Rightarrow$ quadratic dynamics