Towards Painless Policy Optimization for CMDPs

Arushi Jain,
Sharan Vaswani, Reza Babanezhad, Doina Precup, Csaba Szepesvari

BACKGROUND: Safety-critical applications are subjected to constraints. **Constrained MDPs** models such framework by maximizing reward function, while satisfying constraints.

MOTIVATION

Current CMDPs algorithms are **highly sensitive to the choice of hyperparameters**.

Design robust algorithms that require minimal hyperparameter tuning.

CONTRIBUTION

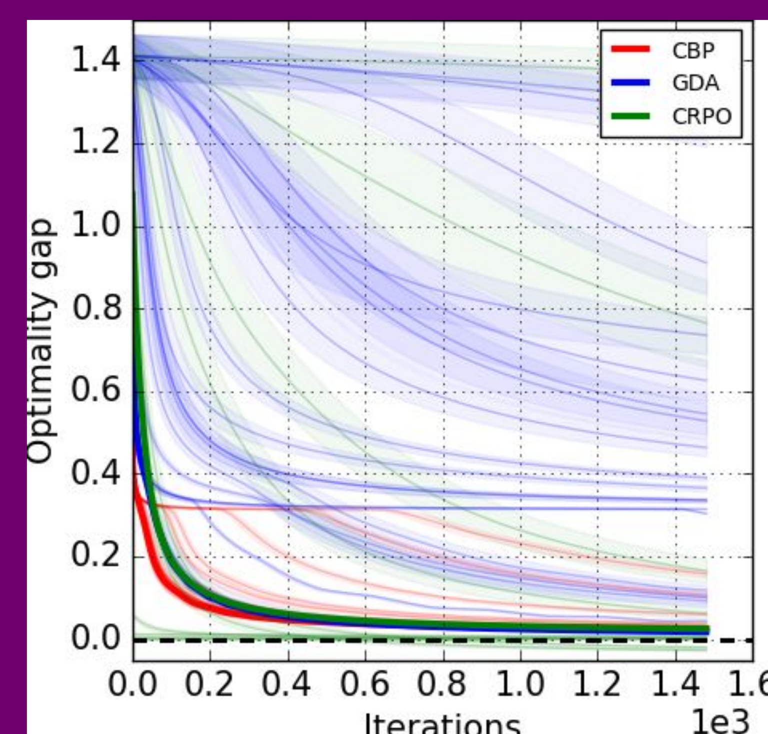
1. Propose **generic primal-dual** framework bounding optimality gap and constraint violation (performance metrics).
2. Use **parameter-free approach** called **coin-betting algorithm** from online linear optimization.
3. Proposed robust **Coin Betting Politex (CBP)** algorithm requiring minimal hyperparameter tuning.
4. Comparable bounds on performance metrics as other primal-dual and primal-only methods.

EMPIRICAL RESULTS

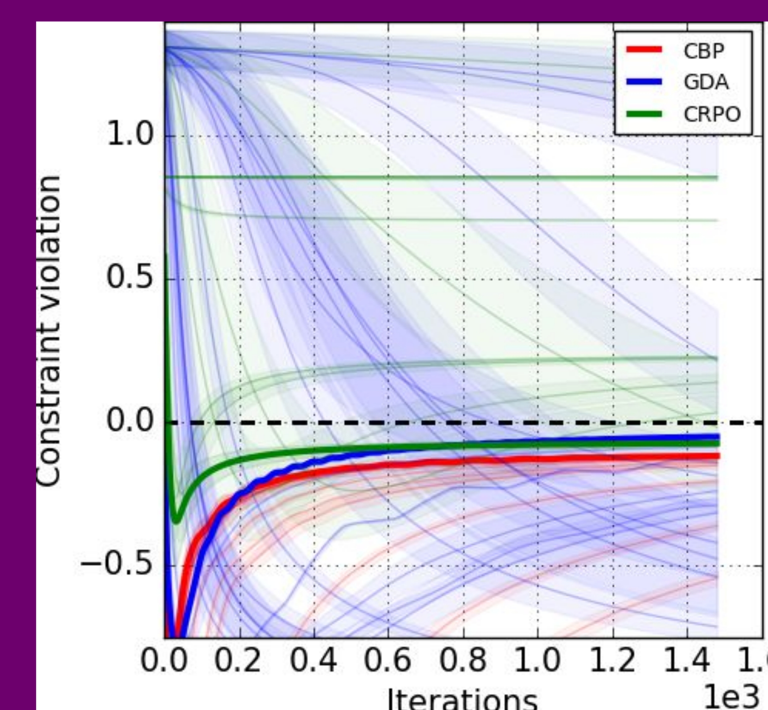
- Compare proposed **CBP** (primal-dual) with sota baselines **GDA** (primal-dual) and **CRPO** (primal-only) algorithms.
- Experiments in tabular and Cartpole (linear function approximation) environment demonstrate consistent effectiveness and **robustness of CBP**.

Reduce hyperparameter sensitivity for **policy optimization** using online linear optimization.

Policy Optimization $\xrightarrow[\text{algorithm}]{\text{Politex}}$ Online Linear Optimization $\xrightarrow[\text{Betting algorithm}]{\text{Coin}}$ Parameter-free algorithms



Ideal value = 0



Ideal value ≤ 0

Robustness of proposed Coin Betting Politex (CBP) algorithm to the choice of hyperparameters:

- CBP: proposed parameter-free algorithm for the policy optimization.
- Comparison of Optimality Gap and Constraint Violation in gridworld environment.
- Dark lines - best hyperparameter performance.
- Lighter lines - other hyperparameters performance.

Main Takeaway: CBP is robust to hyperparameter tuning.

Take a picture to **download** the **full paper**

Paper # 1.133

EQUATIONS

Objective function:

$$\max_\pi V_r^\pi(\rho) s.t. V_c^\pi(\rho) \geq b$$

Optimality Gap (OG):

$$\frac{1}{T}\sum_{t=0}^{T-1} V_r^*(\rho) - V_r^t(\rho)$$

Constraint Violation (CV):

$$\frac{1}{T}[\sum_{t=0}^{T-1} b - V_c^t(\rho)]_+$$

PERFORMANCE BOUNDS

1. OG is bounded by $O(\frac{1}{(1-\gamma)^3\sqrt{T}})$
2. CV is bounded by $O(\frac{1}{(1-\gamma)^2\sqrt{T}})$

REGRETS

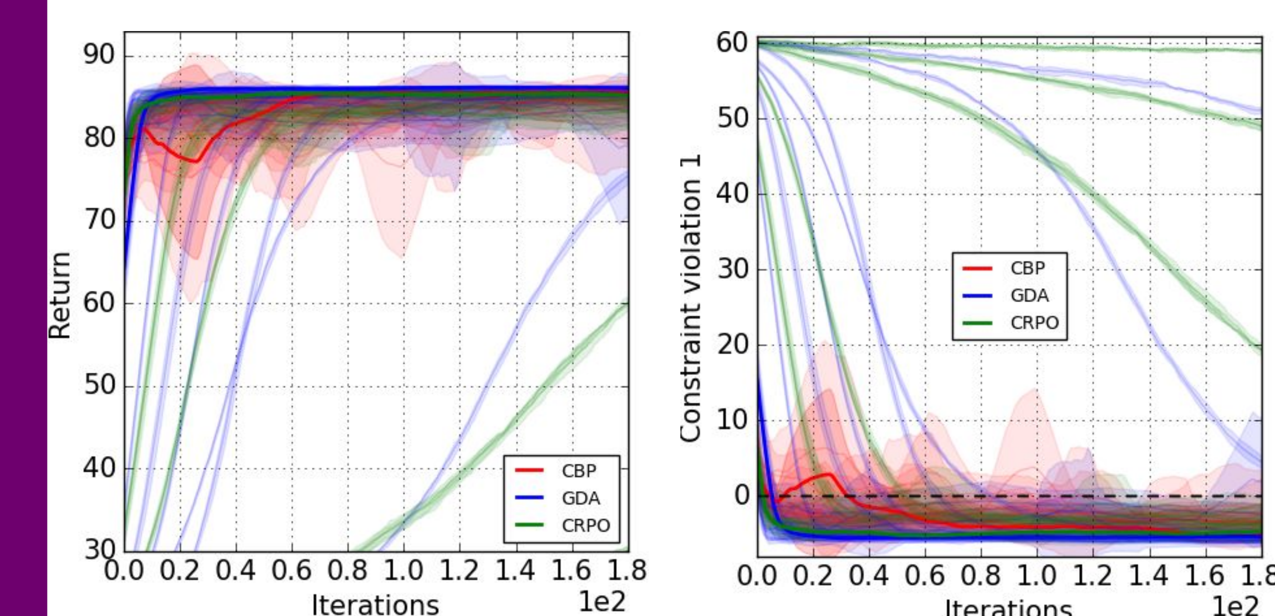
1. **Primal regret**

$$\sum_{t=0}^{T-1} < \pi^* - \pi_t, Q_r^t + \lambda_t Q_c^t >$$

2. **Dual regret**

$$\sum_{t=0}^{T-1} (\lambda_t - \lambda)(V_c^t(\rho) - b)$$

CARTPOLE ENVIRONMENT



Performance with reward and constraint functions. Dark line shows best performance. Light lines show other hyperparameters performance.