Constrained Policy Optimization: A Tale of Regularization and Optimism

Dongsheng Ding

Penn
UNIVERSITY *of* PENNSYLVANIA

joint work with:

Chen-Yu Wei

Kaiqing Zhang

Alejandro Ribeiro

2024 IOS Conference, Houston, TX; Mar. 23, 2024

Motivating application

- ROBOTICS: ROBOT RESCUE

PennToday

$$\underset{\text{navigation policy}}{\text{maximize}} \quad \text{distance from home}$$

$$\text{subject to} \quad \text{obstacle distance} \geq \text{margin}$$

CHALLENGE: constraint satisfaction

Context

- SUCCESS STORIES OF RL

  Go/Atari game, drone/car racing, etc.

- LESSONS LEARNED

  ⋆ importance of policy optimization

  simple; scalable; model-free

  ⋆ non-convex; optimality w/ one performance metric

  reward

  ⋆ difficult for multiple performance metrics

- WHAT NOW ?

  ⋆ applications: robotics, healthcare, finance

  ⋆ policy optimization: tremendous advances

OBJECTIVE

find a policy that

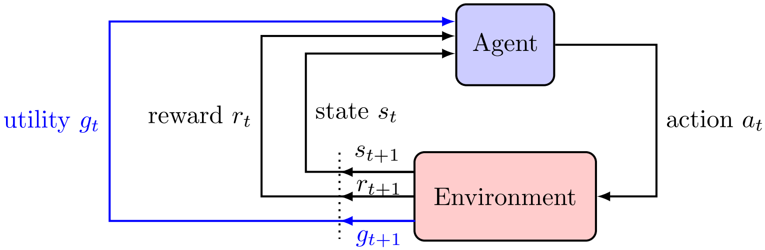
maximizes a performance metric

subject to a constraint on

another performance metric

Framework of RL

■ CONSTRAINED MDP



$\pi : \mathcal{S}$ (states) $\to \mathcal{A}$ (actions) – a policy

$V_r^\pi(\rho) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \mid s_0 \sim \rho]$ – reward value function

$V_g^\pi(\rho) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t g(s_t, a_t) \mid s_0 \sim \rho]$ – utility value function

Constrained policy optimization

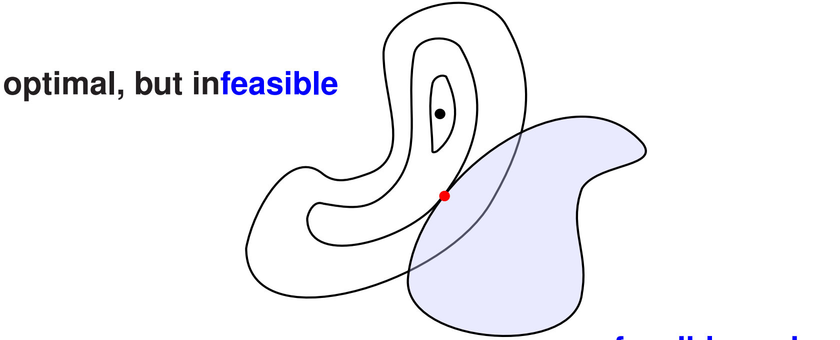
$$
\begin{aligned}
&\underset{\pi}{\text{maximize}} && V_r^\pi(\rho) && \longrightarrow && \textbf{standard objective} \\
&\text{subject to} && V_g^\pi(\rho) \;\geq\; 0
\end{aligned}
$$

$$\downarrow$$

policy constraint

- ⋆ limit the policy space to an inequality constraint

2-D LEVEL NONCONVEX CURVES



optimality w/ **feasibility**

$\Longrightarrow$ **init. dependent** & **stochastic**

Lagrangian approach

Lagrangian: $L(\pi, \lambda) \;=\; V_r^{\pi}(\rho) \;+\; \lambda\, V_g^{\pi}(\rho)$

Lagrange multiplier: $\lambda\ (\geq 0)$

composite reward: $r \;+\; \lambda\, g$

- POLICY OPTIMIZATION

$$\underset{\pi}{\text{maximize}} \qquad V_{r\,+\,\lambda\,g}^{\pi}(\rho)$$

multiple solutions, but infeasible

scalarization fallacy

Primal-dual method

dual problem: $\underset{\lambda}{\text{minimize}} \ \underset{\pi}{\text{maximize}} \ \ V^{\pi}_{r+\lambda g}(\rho)$

primal problem: $\underset{\pi}{\text{maximize}} \ \underset{\lambda}{\text{minimize}} \ \ V^{\pi}_{r+\lambda g}(\rho)$

■ MIN-MAX POLICY OPTIMIZATION

strong duality $\Longrightarrow$ **exchange of min and max**

$$\underset{\lambda}{\text{minimize}} \ \underset{\pi}{\text{maximize}} \ \ V^{\pi}_{r+\lambda g}(\rho) \ = \ \underset{\pi}{\text{maximize}} \ \underset{\lambda}{\text{minimize}} \ \ V^{\pi}_{r+\lambda g}(\rho)$$

OBJECTIVE: find a minimax point $(\pi^\star, \lambda^\star)$

PRIMAL-DUAL UPDATE

policy ascent direction: $\nabla_\pi V^\pi_{r+\lambda g}(\rho)$

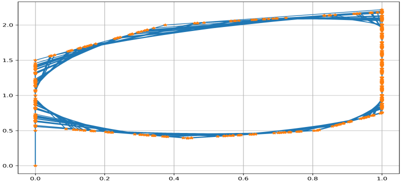
dual descent direction: $\nabla_\lambda V^\pi_{r+\lambda g}(\rho)$

$$\pi^+ \;\leftarrow\; \pi \;+\; \eta\, \nabla_\pi V^\pi_{r+\lambda g}(\rho)$$

$$\lambda^+ \;\leftarrow\; \lambda \;-\; \eta\, \nabla_\lambda V^\pi_{r+\lambda g}(\rho)$$

single-time-scale w/ stepsize η

λ



π ($\pi^\star = 0.66$)

CHALLENGE

single-time-scale primal-dual methods

in face of nonconvex Lagrangian

Glimpse of our results

policy convergence w/ (sub) linear error rate

error rate – optimality gap & feasibility gap

- REGULARIZED POLICY GRADIENT PRIMAL-DUAL METHOD
  - ⋆ **tabular** — dimension-free
  - ⋆ function approximation — up to approx. error
- OPTIMISTIC POLICY GRADIENT PRIMAL-DUAL METHOD
  - ⋆ **tabular** — problem-dependent

Ding, Wei, Zhang, Ribeiro, NeurIPS 2023

Regularized method

Regularized Lagrangian approach

entropy-like term: $$\mathcal{H}(\pi) \;=\; \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t \,|\, s_t)\right]$$

■ REGULARIZED LAGRANGIAN

$$L_\tau(\pi,\lambda) \quad = \quad V_{r+\lambda g}^{\pi}(\rho) \,+\, \tau\left(\mathcal{H}(\pi) + \tfrac{1}{2}\lambda^2\right)$$

$$\downarrow$$

"convexify" Lagrangian $V_{r+\lambda g}^{\pi}(\rho)$

OBJECTIVE: find the regularized minimax point $(\pi_\tau^\star, \lambda_\tau^\star)$

⋆ **τ-near minimax point of** $V_{r+\lambda g}^{\pi}(\rho)$

Two pillars

Q-VALUE FUNCTION

$$Q_r^\pi(s,a) \;=\; \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, s_0=s, a_0=a\right]$$

STATE VISITATION DISTRIBUTION

$$d_{s_0}^\pi(s) \;=\; (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P^\pi(s_t=s\,|\,s_0)$$

- $\star$ expectation $d_\rho^\pi(s) = \mathbb{E}_{s_0\sim\rho}\left[d_{s_0}^\pi(s)\right]$

Regularized policy gradient primal-dual method

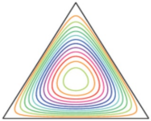
policy ascent direction: $Q^{\pi}_{L_\tau}(s,a) \;:=\; Q^{\pi}_{r+\lambda g-\tau \log \pi}(s,a)$

dual descent direction: $V^{\pi}_{g}(\rho) \;+\; \tau\,\lambda$

■ SINGLE-TIME-SCALE PRIMAL-DUAL UPDATE

$$\boxed{\begin{aligned} \pi^{+}(\cdot\,|\,s) \;&=\; \underset{\pi' \in \Delta_{\epsilon_0}}{\operatorname{argmax}} \;\langle \pi'(\cdot\,|\,s),\, Q^{\pi}_{L_\tau}(s,\cdot)\rangle \;-\; \frac{1}{\eta}\,\mathsf{KL}_s(\pi',\pi) \\ \lambda^{+} \;&=\; \underset{\lambda \in \Lambda}{\operatorname{argmin}} \;\lambda'\big(V^{\pi}_{g}(\rho)+\tau\lambda\big) \;+\; \frac{1}{2\eta}(\lambda'-\lambda)^2 \end{aligned}}$$

⋆ restricted policy set $\Delta_{\epsilon_0} = \{p_a \geq \epsilon_0, \forall a\}$



Non-asymptotic last-iterate performance

distance of (π_t, λ_t) to $(\pi_\tau^\star, \lambda_\tau^\star)$: $\mathbb{E}_{s \sim d_\rho^{\pi_\tau^\star}} \left[\mathsf{KL}_s(\pi_\tau^\star, \pi_t) \right] + \frac{1}{2}(\lambda_t - \lambda_\tau^\star)^2$

Theorem (informal)

★ **distance of (π_t, λ_t) to $(\pi_\tau^\star, \lambda_\tau^\star)$** is bounded by

$$\mathrm{e}^{-\eta \tau t} + \frac{\eta}{\tau}$$

exponential decay up to a ratio

⋆ ϵ-near regularized minimax point requires

$$\eta = \epsilon \tau \quad \text{and} \quad t = \frac{1}{\epsilon \tau^2}$$

sublinear rate

optimality gap: $V_r^{\star}(\rho) - V_r^{\pi_t}(\rho)$

feasibility gap: $0 - V_g^{\pi_t}(\rho)$

Implication (informal)

★ **ϵ-optimality gap & ϵ-feasibility gap** require

$$\frac{1}{\epsilon^6} \quad \text{iterations}$$

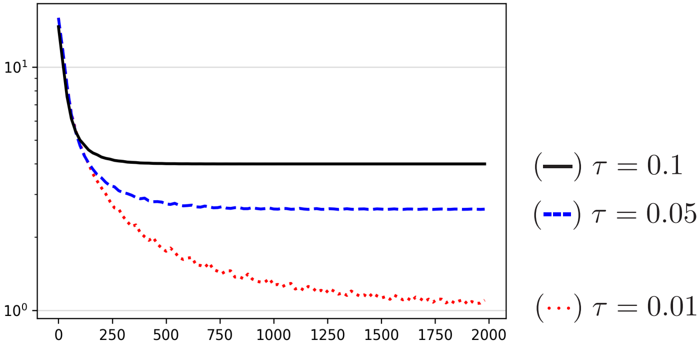
sublinear rate

stepsize $\eta = \epsilon^4$

regularization $\tau = \epsilon^2$

Sublinear convergence

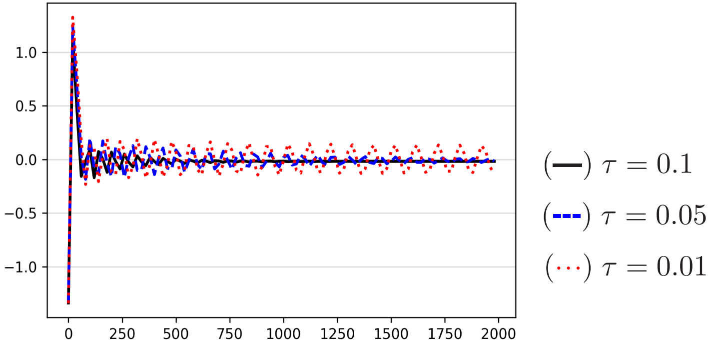
$\sum_s \|\pi_t(\cdot \mid s) - \pi^\star(\cdot \mid s)\|^2$



\# iterations

smaller regularization $\tau \longrightarrow$ **better accuracy**

$V_g^{\pi_t}(\rho)$



iterations

smaller regularization τ $\longrightarrow$ **larger oscillation**

■ ACCURACY VS OSCILLATION TRADE-OFF

Optimistic method

Optimistic policy gradient primal-dual method

$$
\begin{tikzcd}
(\hat{\pi}, \hat{\lambda}) \arrow[d] \arrow[dr, dashed] & \\
(\pi^+, \lambda^+) \arrow[r] & (\hat{\pi}^+, \hat{\lambda}^+)
\end{tikzcd}
$$

policy ascent direction: $Q^{\pi}_{r+\lambda g}(s,a)$

dual descent direction: $V^{\pi}_{g}(\rho)$

policy ascent direction: $Q^{\pi^+}_{r+\lambda^+ g}(s,a)$

dual descent direction: $V^{\pi^+}_{g}(\rho)$

SINGLE-TIME-SCALE PRIMAL-DUAL UPDATE

prediction step

$$\boxed{\begin{aligned} \pi^+(\cdot \,|\, s) &= \operatorname*{argmax}_{\pi' \in \Delta} \; \langle \pi'(\cdot \,|\, s), \, Q^{\pi}_{r+\lambda g}(s,\cdot) \rangle \; - \; \frac{1}{2\eta} \, \|\pi' - \hat{\pi}\|_s^2 \\ \lambda^+ &= \operatorname*{argmin}_{\lambda \in \Lambda} \; \lambda' V_g^{\pi}(\rho) \; + \; \frac{1}{2\eta}(\lambda' - \hat{\lambda})^2 \end{aligned}}$$

real step

$$\boxed{\begin{aligned} \hat{\pi}^+(\cdot \,|\, s) &= \operatorname*{argmax}_{\pi' \in \Delta} \; \langle \pi'(\cdot \,|\, s), \, Q^{\pi^+}_{r+\lambda^+ g}(s,\cdot) \rangle \; - \; \frac{1}{2\eta} \, \|\pi' - \hat{\pi}\|_s^2 \\ \hat{\lambda}^+ &= \operatorname*{argmin}_{\lambda \in \Lambda} \; \lambda' V_g^{\pi^+}(\rho) \; + \; \frac{1}{2\eta}(\lambda' - \hat{\lambda})^2 \end{aligned}}$$

Non-asymptotic last-iterate performance

distance of $(\hat{\pi}_t, \hat{\lambda}_t)$ to $\Pi^\star \times \Lambda^\star$: $\mathbb{E}_{s \sim d_\rho^{\pi^\star}} \left[\mathrm{Dist}_s(\hat{\pi}_t, \Pi^\star) \right] + \mathrm{Dist}(\hat{\lambda}_t, \Lambda^\star)$

Theorem (informal)

★ **distance of $(\hat{\pi}_t, \hat{\lambda}_t)$ to $\Pi^\star \times \Lambda^\star$** is bounded by

$$\left(\frac{1}{1 + C} \right)^t$$

exponential decay to zero

problem-dependent constants C, η

optimality gap: $V_r^\star(\rho) - V_r^{\pi_t}(\rho)$

feasibility gap: $0 - V_g^{\pi_t}(\rho)$

Implication (informal)

★ **ϵ-optimality gap & ϵ-feasibility gap** require

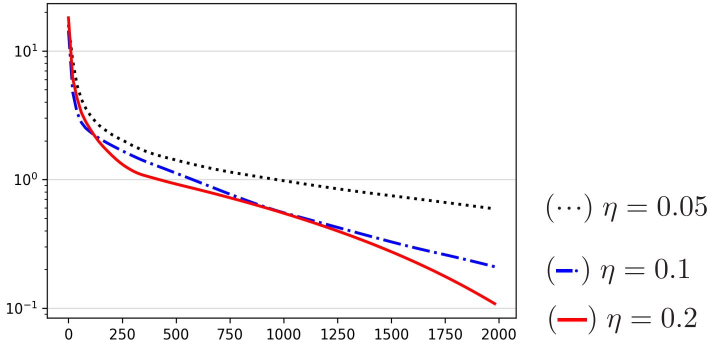
$$\log^2 \frac{1}{\epsilon} \quad \text{iterations}$$

linear rate

problem-dependent

Linear convergence

$$\sum_s \|\hat{\pi}_t(\cdot \mid s) - \pi^\star(\cdot \mid s)\|^2$$



iterations

linear rate holds for a range of stepsizes

Summary

- SINGLE-TIME-SCALE PRIMAL-DUAL METHODS
  - ⋆ regularized policy gradient primal-dual method
  - ⋆ optimistic policy gradient primal-dual method
  - ⋆ non-asymptotic last-iterate policy convergence

- ON-GOING EFFORTS
  - ⋆ constrained policy optimization w/ exploration
  - ⋆ more practical considerations

Thank you for your attention.