Constrained Policy Optimization: A Tale of Regularization and Optimism

Dongsheng Ding

dongshed@utk.edu

THE UNIVERSITY OF TENNESSEE KNOXVILLE

IFAC Workshop on Contextual Control; Aug. 23, 2026

Motivating application

- ROBOTICS: ROBOT LOCOMOTION

Figure AI

$$\begin{array}{ll} \underset{\text{policy}}{\text{maximize}} & \text{task progress} \\ \text{subject to} & \text{safety} \geq \text{margin} \end{array}$$

CHALLENGE: constraint satisfaction

Context

- SUCCESS STORIES OF RL

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

- LESSONS LEARNED
  - ⋆ importance of policy optimization

    simple; scalable; model-free
  - ⋆ non-convex; optimality w/ one performance metric

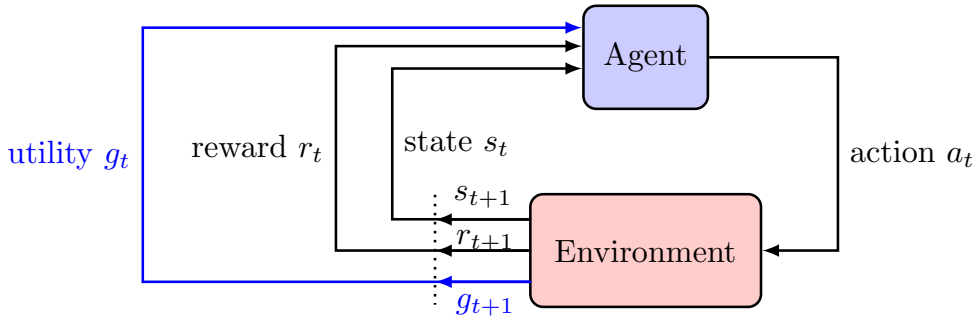
    reward
  - ⋆ difficult for multiple performance metrics
- WHAT NOW ?
  - ⋆ applications: robotics, healthcare, finance
  - ⋆ policy optimization: tremendous advances

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

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

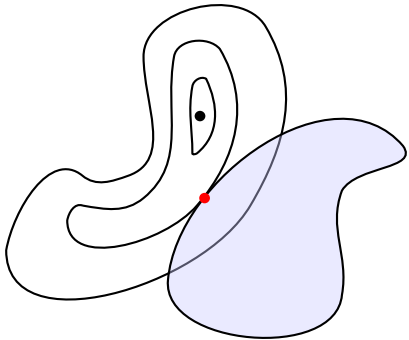
$$\downarrow$$

policy constraint

- ⋆ limit the policy space to an inequality constraint

2-D LEVEL NONCONVEX CURVES

optimal, but infeasible



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

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

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

- MIN-MAX POLICY OPTIMIZATION

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

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

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

SIMULTANEOUS ASCENT-DESCENT

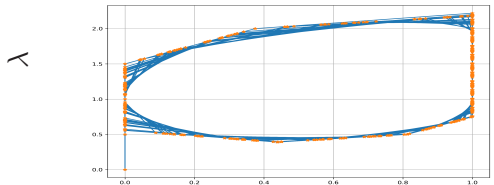
policy ascent direction: $\nabla_\pi V^\pi_{r+\lambda g}(\rho)$

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

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

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

single-time-scale w/ stepsize η



CHALLENGE

single-time-scale primal-dual methods in face of nonconvex Lagrangian

Glimpse of our efforts

policy convergence w/ (sub) linear error rate

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

DWZR, NeurIPS 2023
DHR, AISTATS 2024
R*__D__*MR, AAAI 2025

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

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

$$\downarrow$$

add stabilizing regularization

OBJECTIVE: find the regularized minimax point $(\pi^\star_\tau, \lambda^\star_\tau)$

- $\star$ **τ-near minimax point of** $V^\pi_{r+\lambda g}(\rho)$

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

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

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

Regularized policy gradient primal-dual method

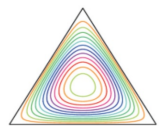
policy ascent direction: $Q^{\pi}_{L_\tau}(s,a) := Q^{\pi}_{r+\lambda g-\tau\log\pi}(s,a)$

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

- SINGLE-TIME-SCALE PRIMAL-DUAL UPDATE

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

⋆ 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** $(\pi_\tau^\star, \lambda_\tau^\star)$**:** $\mathbb{E}_{s \sim d_\rho^{\pi_\tau^\star}} [\, \mathsf{KL}_s(\pi_\tau^\star, \pi_t) \,] + \frac{1}{2}(\lambda_t - \lambda_\tau^\star)^2$

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 $\tau = \epsilon^2$

Distance of primal-dual iterates to $(\pi_\tau^\star, \lambda_\tau^\star)$

Step #1: performance difference & MWU

$$\boxed{L_\tau(\pi_\tau^\star, \lambda_t) - L_\tau(\pi_t, \lambda_t)}$$

$$\lesssim \mathbb{E}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\,\langle \pi_\tau^\star(\cdot\,|\,s) - \pi_t(\cdot\,|\,s), Q^{\pi_t}_{r+\lambda_t g - \tau\pi_t}(s,\cdot)\rangle\,\right] - \tau\mathbb{E}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\,\mathsf{KL}_t(s)\,\right]$$

$$\lesssim \frac{1}{\eta}\mathbb{E}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\,\mathsf{KL}_t(s) - \mathsf{KL}_{t+1}(s)\,\right] + \eta(C_\tau)^2 - \tau\mathbb{E}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\,\mathsf{KL}_t(s)\,\right]$$

$$= \frac{1}{\eta}\left[\,(1-\eta\tau)\mathsf{KL}_t(\rho) - \mathsf{KL}_{t+1}(\rho)\,\right] + \eta(C_\tau)^2$$

$$\mathsf{KL}_t(s) = \mathsf{KL}(\pi_\tau^\star(\cdot\,|\,s), \pi_t(\cdot\,|\,s))$$

$$\mathsf{KL}_t(\rho) = \mathbb{E}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\,\mathsf{KL}_t(s)\,\right]$$

$$\boxed{L_\tau(\pi_t, \lambda_t) - L_\tau(\pi_t, \lambda_\tau^\star)}$$

$$\lesssim \frac{1}{\eta}\left[\,(1-\eta\tau)(\lambda_\tau^\star - \lambda_t) - (\lambda_\tau^\star - \lambda_{t+1})\,\right] + \eta(C_\tau')^2$$

Step #2: saddle point property & geometric series

$$\cancel{\eta\left(\,L_\tau(\pi_\tau^\star, \lambda_t) - L_\tau(\pi_t, \lambda_\tau^\star)\,\right)} + \boxed{\mathsf{KL}_{t+1}(\rho) + \tfrac{1}{2}(\lambda_\tau^\star - \lambda_{t+1})^2}$$

$$\lesssim \boxed{(1-\eta\tau)\left(\mathsf{KL}_t(\rho) + \tfrac{1}{2}(\lambda_\tau^\star - \lambda_t)^2\right)} + \eta^2 \max\left((C_\tau)^2, (C_\tau')^2\right)$$

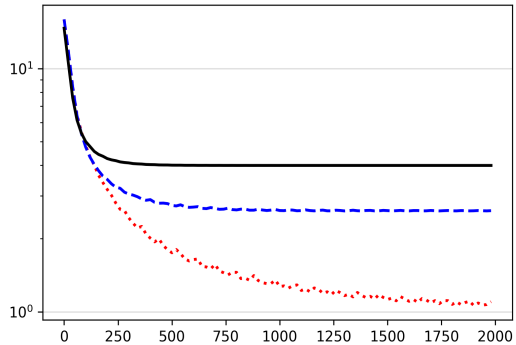
$$\vdots$$

$$\lesssim (1-\eta\tau)^t\left(\mathsf{KL}_1(\rho) + \tfrac{1}{2}(\lambda_\tau^\star - \lambda_1)^2\right)$$

$$+ \left(\eta^2\left(1 + (1-\eta\tau) + (1-\eta\tau)^2 + \cdots\right)\right)\max\left((C_\tau)^2, (C_\tau')^2\right)$$

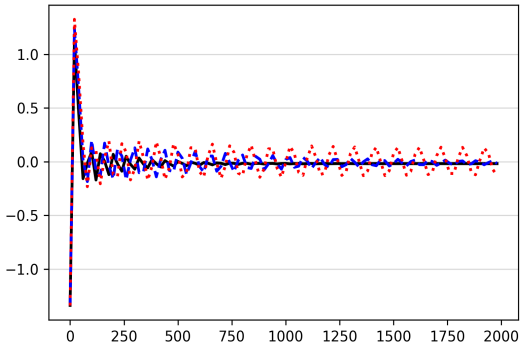
Sublinear convergence

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



smaller regularization τ $\longrightarrow$ better accuracy

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



(—) $\tau = 0.1$

(- - -) $\tau = 0.05$

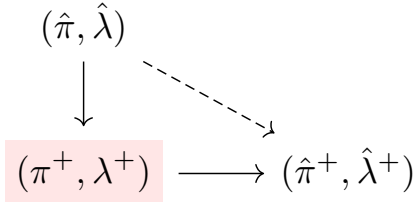
(···) $\tau = 0.01$

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

- ACCURACY VS OSCILLATION TRADE-OFF

Optimistic method

Optimistic policy gradient primal-dual method



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

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

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

$$\pi^+(\cdot \,|\, s) = \underset{\pi' \in \Delta}{\operatorname{argmax}} \; \langle \pi'(\cdot \,|\, s),\, Q^{\pi}_{r+\lambda g}(s,\cdot) \rangle - \frac{1}{2\eta} \|\pi' - \hat{\pi}\|_s^2$$

$$\lambda^+ = \underset{\lambda' \in \Lambda}{\operatorname{argmin}} \; \lambda' V_g^{\pi}(\rho) + \frac{1}{2\eta}(\lambda' - \hat{\lambda})^2$$

real step

$$\hat{\pi}^+(\cdot \,|\, s) = \underset{\pi' \in \Delta}{\operatorname{argmax}} \; \langle \pi'(\cdot \,|\, s),\, Q^{\pi^+}_{r+\lambda^+ g}(s,\cdot) \rangle - \frac{1}{2\eta} \|\pi' - \hat{\pi}\|_s^2$$

$$\hat{\lambda}^+ = \underset{\lambda' \in \Lambda}{\operatorname{argmin}} \; \lambda' V_g^{\pi^+}(\rho) + \frac{1}{2\eta}(\lambda' - \hat{\lambda})^2$$

Non-asymptotic last-iterate performance

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

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

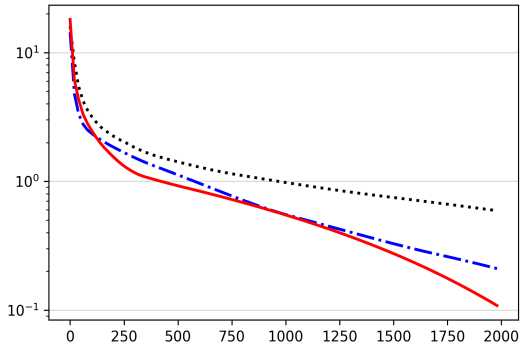
$$\log^2 \frac{1}{\epsilon} \quad \text{iterations}$$

linear rate

problem-dependent

Linear convergence

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



$(\cdots)$ $\eta = 0.05$

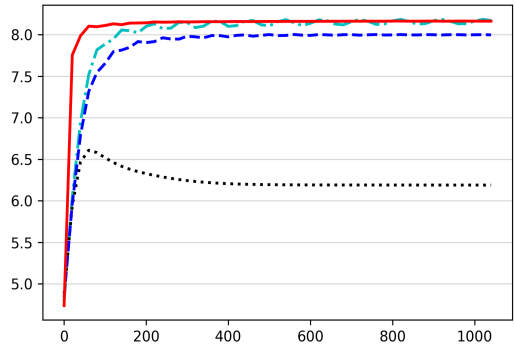
$(-\cdot)$ $\eta = 0.1$

$(—)$ $\eta = 0.2$

linear rate holds for a range of stepsizes

Comparison of primal-dual methods

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



(—) optimistic method

(- -) regularized method

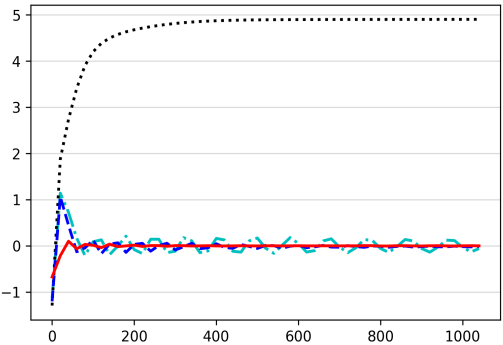
(–·) NPG-PD

(··) PID-Lagrangian

optimistic method $\approx$ **NPG-PD** $\geq$ **regularized method** $\geq$ **PID-Lagrangian**

Comparison of primal-dual methods

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



(··) PID-Lagrangian

(—) optimistic method

(- -) regularized method

(–·) NPG-PD

optimistic method $\approx$ **regularized method** $\leq$ **NPG-PD** $\leq$ **PID-Lagrangian**

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/o exploration
  - more practical considerations

Thank you for your attention.

Intuition of exponential decay

Distance of primal-dual iterates to $\Pi^\star \times \Lambda^\star$

Step #1: monotonic decreasing distance

$$\Theta_t = \frac{1}{2}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\|\hat{\pi}_t(\cdot\,|\,s) - \mathcal{P}_{\Pi^\star}(\hat{\pi}_t(\cdot\,|\,s))\|^2 + \frac{1}{2}\left(\hat{\lambda}_t - \mathcal{P}_{\Lambda^\star}(\hat{\lambda}_t)\right)^2$$

$$+ \frac{1}{16}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\|\hat{\pi}_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2 + \frac{1}{2}\left(\hat{\lambda}_t - \lambda_{t-1}\right)^2$$

$$\geq \Theta_{t+1} + \frac{7}{16}\underbrace{\zeta_t}$$

$$= \frac{1}{2}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\|\hat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\|^2 + \frac{1}{2}\left(\hat{\lambda}_{t+1} - \lambda_t\right)^2$$

$$+ \frac{1}{2}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\|\hat{\pi}_t(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\|^2 + \frac{1}{2}\left(\hat{\lambda}_t - \lambda_t\right)^2$$

Step #2: lower bound of ζ_t

$$\zeta_t \;\gtrsim\; \frac{1}{4}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\|\hat{\pi}_{t+1}(\cdot\,|\,s)-\pi_t(\cdot\,|\,s)\|^2 \;+\; \frac{1}{4}\left(\hat{\lambda}_{t+1}-\lambda_t\right)^2$$

$$+\,\eta^2 \underbrace{\sup_{\pi,\lambda}\frac{\Big[L(\pi,\hat{\lambda}_{t+1})-L(\hat{\pi}_{t+1},\lambda)\Big]_+^2}{\max_s\|\pi(\cdot\,|\,s)-\hat{\pi}_{t+1}(\cdot\,|\,s)\|+|\lambda-\hat{\lambda}_{t+1}|\Big)^2}}$$

$$\gtrsim\; \mathbb{E}_{s\sim d_\rho^{\pi^\star}}\|\hat{\pi}_{t+1}(\cdot\,|\,s)-\mathcal{P}_{\Pi^\star}\left(\hat{\pi}_{t+1}(\cdot\,|\,s)\right)\|^2 \;+\; \left(\hat{\lambda}_{t+1}-\mathcal{P}_{\Lambda^\star}(\hat{\lambda}_{t+1})\right)^2$$

$$\gtrsim\; \eta^2\Theta_{t+1}$$

problem-dependent constant

$$\Longrightarrow \quad \boxed{\Theta_{t+1} \;\lesssim\; \Theta_t \;-\; \eta^2\Theta_{t+1}}$$