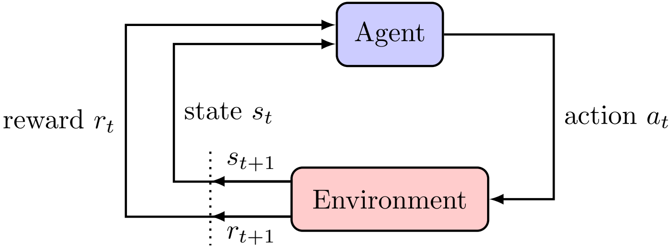
Provable constrained policy optimization in RL

Dongsheng Ding
https://dongshed.github.io

June 28, 2023 @ Safe Reinforcement Learning Online Seminar

Framework for RL

- MARKOV DECISION PROCESS (MDP)



$\pi : S$ (states) $\to A$ (actions) – policy

$P(s_{t+1} \mid s_t, a_t)$ – unknown transition kernel

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

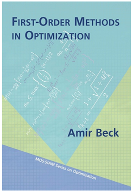
Policy optimization

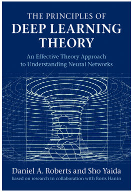
Objective		Direct policy search
$\underset{\pi}{\text{maximize}} \quad V_r^\pi(\rho)$	$\leftrightarrows$	$\pi^+ \;\leftarrow\; \pi \;+\; \nabla_\pi V_r^\pi$

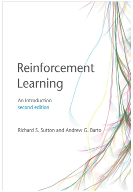
Increasingly use, e.g., ChatGPT

■ FEATURES

- ⋆ simple
- ⋆ scalable
- ⋆ model-free

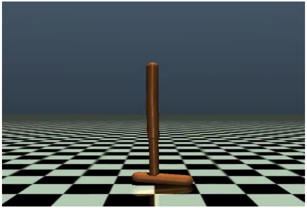






RL under constraints

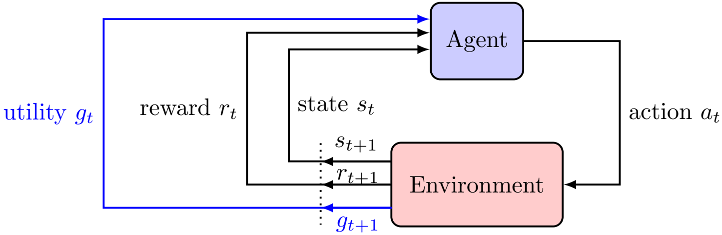
MuJoCo robotics



Goal	forward moving
Constraints	smoothness energy risk-awareness ⋮

Framework for RL under constraints

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

$$\underset{\pi}{\text{maximize}} \quad V_r^\pi(\rho)$$

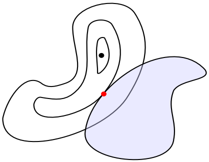
$$\text{subject to} \quad V_g^\pi(\rho) \;\geq\; b$$

$$L(\pi,\lambda) \;:=\; V_r^\pi(\rho) + \lambda\left(V_g^\pi(\rho) - b\right)$$ – Lagrangian

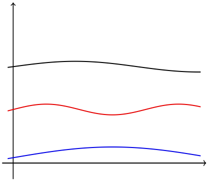
Altman, CRC Press '99

STRUCTURAL PROPERTIES

non-convexity



non-uniformity



Question

Can we identify constrained policy optimization methods with provable efficiency guarantees?

RL UNDER CONSTRAINTS

- ⋆ nearly or even exactly meeting specific constraints
- ⋆ establishing finite-time convergence guarantees

Part I: Finite-time average-value performance

NATURAL POLICY GRADIENT PRIMAL-DUAL METHOD

average-value convergence with subliner error rate

- ⋆ tabular — dimension-free
- ⋆ function approximation — up to approx. error

error rate – optimality gap & constraint violation

Ding, Zhang, Başar, Jovanović, NeurIPS '20

Ding, Zhang, Duan, Başar, Jovanović, arXiv:2206.02346 (under revision)

Part II: Finite-time last-iterate performance

- REGULARIZED POLICY GRADIENT PRIMAL-DUAL METHOD

last-iterate convergence with sublinear error rate

- ⋆ tabular — dimension-free
- ⋆ function approximation — up to approx. error

- OPTIMISTIC POLICY GRADIENT PRIMAL-DUAL METHOD

last-iterate convergence with linear error rate

- ⋆ tabular — problem-dependent

error rate – optimality gap & constraint violation

Ding, Wei, Zhang, Ribeiro, arXiv:2306.11700 (submitted)

Part I
Finite-time average-value performance

Tabular case

(exact gradient, small state space)

Constrained softmax policy optimization

SOFTMAX POLICY

$$\pi_\theta(a \,|\, s) \;=\; \frac{\mathrm{e}^{\theta_{s,a}}}{\sum_{a'} \mathrm{e}^{\theta_{s,a'}}}, \quad \text{parameter } \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

complete & differentiable

CONSTRAINED PARAMETER OPTIMIZATION

$$\begin{array}{ll} \underset{\theta \,\in\, \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}}{\text{minimize}} & V_r^{\pi_\theta}(\rho) \\ \text{subject to} & V_g^{\pi_\theta}(\rho) \;\geq\; b \end{array}$$

non-convex optimization

Q-value function & visitation measure

Q-VALUE FUNCTION

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

- ⋆ $A_r^\pi(s,a) = Q_r^\pi(s,a) - V_r^\pi(s)$ – advantage

$Q_g^\pi(s,a)$, $A_g^\pi(s,a)$ – use g to define them similarly

STATE VISITATION DISTRIBUTION

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

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

Lagrangian-based primal-dual method

$$\begin{aligned} \theta^+ &= \theta + \eta_1 \nabla_\theta L(\theta, \lambda) \\ \lambda^+ &= \mathcal{P}\left(\lambda - \eta_2 \left(V_g^\theta(\rho) - b \right) \right) \end{aligned}$$

$L(\theta, \lambda) := V_r^\theta(\rho) + \lambda \left(V_g^\theta(\rho) - b \right)$ – Lagrangian

λ – dual variable

Abad, Krishnamurthy, Martin, Baltcheva, CDC '02

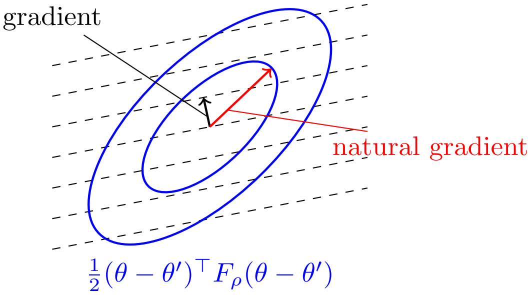
Borkar, SCL '05

Tessler, Mankowitz, Mannor, ICLR '18

Observation I: asymptotic convergence

Observation II: stationary point

Natural (policy) gradient



$$F_\rho(\theta) \;:=\; \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)} \left[\nabla_\theta \log \pi_\theta \, (\nabla_\theta \log \pi_\theta)^\top \right]$$

steepest descent in Fisher information distance

Amari, ’83

Natural policy gradient primal-dual method

$$\begin{aligned} \theta^+ &= \theta + \eta_1 F_\rho(\theta)^\dagger \nabla_\theta L(\theta, \lambda) \\ \lambda^+ &= \mathcal{P}\left(\lambda - \eta_2 \left(V_g^\theta(\rho) - b\right)\right) \end{aligned}$$

$L(\theta, \lambda) := V_r^\theta(\rho) + \lambda \left(V_g^\theta(\rho) - b\right)$ – Lagrangian

λ – dual variable

⋆ $F_\rho(\theta)^\dagger \nabla_\theta L(\theta, \lambda)$ – natural policy gradient (NPG)

$$F_\rho(\theta)^\dagger \nabla_\theta L(\theta, \lambda) = \underbrace{F_\rho(\theta)^\dagger \nabla_\theta V_r^\theta(\rho)}_{\text{NPG for reward}} + \lambda \underbrace{F_\rho(\theta)^\dagger \nabla_\theta V_g^\theta(\rho)}_{\text{NPG for utility}}$$

NPG as A-regression

$$\underset{w}{\text{minimize}} \;\; \mathbb{E}_{(s,a)\sim\nu}\Big[\,\big(A^{\pi_\theta} - w^\top \nabla_\theta \log \pi_\theta\big)^2\,\Big]$$

$$\nu \;=\; d_\rho^{\pi_\theta}(s)\pi_\theta(a\,|\,s)$$

$$A^{\pi_\theta} \;=\; A_r^{\pi_\theta} \text{ or } A_g^{\pi_\theta}$$

- ⋆ optimal solution

$$\begin{aligned} w^\star &= F_\rho(\theta)^\dagger \cdot \mathbb{E}_{(s,a)\sim\nu}\big[\,\nabla_\theta \log \pi_\theta(a\,|\,s) A^{\pi_\theta}(s,a)\,\big] \\ &= (1-\gamma)\, F_\rho(\theta)^\dagger \cdot \nabla_\theta V^{\pi_\theta}(\rho) \\ &\simeq A^{\pi_\theta} \end{aligned}$$

NPG $\simeq$ advantage function

Policy primal-dual update

PRIMAL UPDATE AS MULTIPLICATIVE WEIGHT UPDATE

$$\theta^+ \;=\; \theta + \frac{\eta_1}{1-\gamma} A_L^{\pi_\theta}$$

$$A_L^{\pi_\theta} \;:=\; A_r^{\pi_\theta} + \lambda A_g^{\pi_\theta}$$

$$\Downarrow$$

$$\boxed{\pi_\theta^+(\cdot \,|\, s) \;\propto\; \pi_\theta(\cdot \,|\, s) \exp\left(\frac{\eta_1}{1-\gamma} A_L^{\pi_\theta}(s, \cdot)\right)}$$

multiplicative weights update (MWU)

- ⋆ $A_L^{\pi_\theta} \leftarrow Q_L^{\pi_\theta}$ – invariant to action-independent terms
- ⋆ NPG as A-regression $\leftarrow$ NPG as Q-regerssion

Finite-time average-value performance

Theorem (informal)

★ Optimality gap & Constraint violation

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(V_r^\star(\rho)-V_r^{(t)}(\rho)\big),\ \frac{1}{T}\sum_{t=0}^{T-1}\big(b-V_g^{(t)}(\rho)\big)\ \leq\ \epsilon \text{ for } T=O\left(\frac{1}{\epsilon^2}\right)$$

T – number of iterations

⋆ $O(\cdot)$ – dimension-free: free of $|\mathcal{S}|$, $|\mathcal{A}|$, and ρ

⋆ t_{mix} – mixture policy

$$V_r^\star(\rho)-\mathbb{E}\big[V_r^{(t_{\text{mix}})}(\rho)\big]\ \leq\ \epsilon \text{ and } b-\mathbb{E}\big[V_g^{(t_{\text{mix}})}(\rho)\big]\ \leq\ \epsilon$$

Function approximation case

(inexact gradient, large state space)

General softmax policy

$$\pi_\theta(a \,|\, s) \;=\; \frac{e^{f_\theta(s,a)}}{\sum_{a'} e^{f_\theta(s,a')}}, \quad \text{parameter } \theta \in \mathbb{R}^d$$

$f_\theta(s,a)$ – neural network

$f_\theta(s,a) = \theta_{s,a}$ – softmax policy

LOG-LINEAR POLICY

$$\pi_\theta(a \,|\, s) \;=\; \frac{e^{\theta^\top \phi_{s,a}}}{\sum_{a'} e^{\theta^\top \phi_{s,a'}}}, \quad \text{parameter } \theta \in \mathbb{R}^d$$

$\phi_{s,a} \in \mathbb{R}^d$ – linear feature map

Log-linear policy primal-dual update

$$w \;\approx\; \underset{\|w\|\le W}{\text{argmin}}\; \mathbb{E}_{(s,a)\sim\nu}\left[\left(Q^{\pi_\theta}(s,a) - w^\top \phi_{s,a}\right)^2\right]$$

$$\nu \;=\; d_\rho^{\pi_\theta}(s)\pi_\theta(a\,|\,s)$$

$$Q^{\pi_\theta} \;=\; Q_r^{\pi_\theta} \text{ or } Q_g^{\pi_\theta}$$

- PRIMAL UPDATE VIA EMPIRICAL SOLUTION

$$\begin{aligned} \theta^+ &= \theta + \frac{\eta_1}{1-\gamma} w \\ \lambda^+ &= \mathcal{P}_\Lambda\left(\lambda - \eta_2\left(V_g^{\pi_\theta}(\rho) - b\right)\right) \end{aligned}$$

$w \;:=\; w_r + \lambda w_g$ – NPG

Approximation error

Exact solution

$$w_\star \in \operatorname*{argmin}_{\|w\| \leq W} \mathcal{E}^\nu(w; \pi_\theta)$$

- ESTIMATION ERROR

$$\mathcal{E}_{\text{est}} := \mathbb{E}\Big[\mathcal{E}^\nu(w; \pi_\theta) - \mathcal{E}^\nu(w_\star; \pi_\theta)\Big] \sim \frac{1}{K}$$

w can be different from $w_\star$

Lacoste-Julien, Schmidt, Bach, '12

- TRANSFER ERROR

$$\mathcal{E}_{\text{bias}} := \mathbb{E}\Big[\mathcal{E}^{\nu^\star}(w_\star; \pi_\theta)\Big]$$

e.g., 0 for tabular case

the best linear fit $w_\star$ may mismatch Q^{π_θ}

Finite-time average-value performance

Theorem (informal)

★ Optimality gap & Constraint violation

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho)-V_r^{(t)}(\rho)\right)\right],\ \ \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b-V_g^{(t)}(\rho)\right)\right]$$

$$\leq\ \ O\left(\epsilon+\sqrt{\epsilon_{\text{bias}}}+\sqrt{\kappa\,\epsilon_{\text{est}}}\right)\ \text{ for } T=O\left(\frac{1}{\epsilon^2}\right)$$

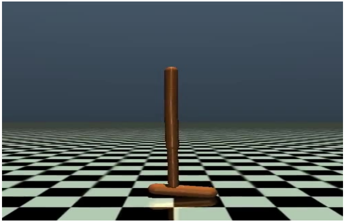
T – number of iterations

- ⋆ $\epsilon_{\text{bias}}=0$ for tabular case – transfer error
- ⋆ $\epsilon_{\text{est}}\simeq\frac{1}{K}$ for K SGD steps – estimation error
- ⋆ $\kappa<\infty$ – relative condition number

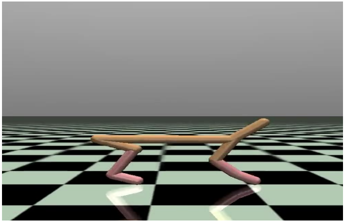
generalization to general smooth policy

MuJoCo robotics

Hopper



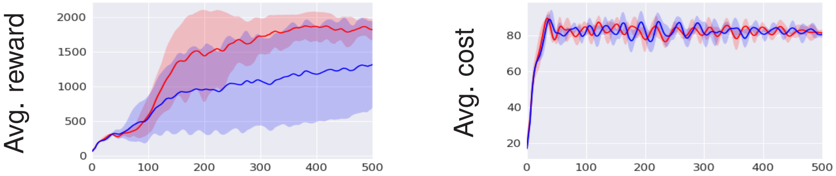
HalfCheetah



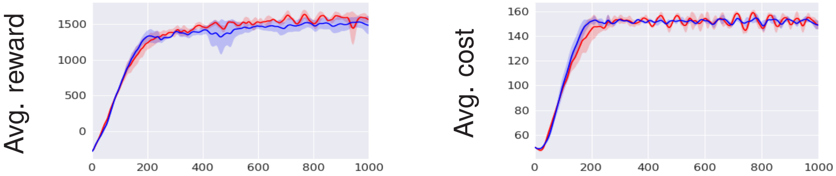
- ⋆ walk with energy efficiency – constrained objective
- ⋆ energy efficiency $= 50\%$ speed from unconstrained PPO:

  83 – Hopper   152 – Halfcheetah

Hopper



HalfCheetah



horizontal axis – # iterations

- ⋆ (red line) natural policy gradient primal-dual method
- ⋆ (blue line) FOCOPS (NeurIPS '20)

Summary of Part I

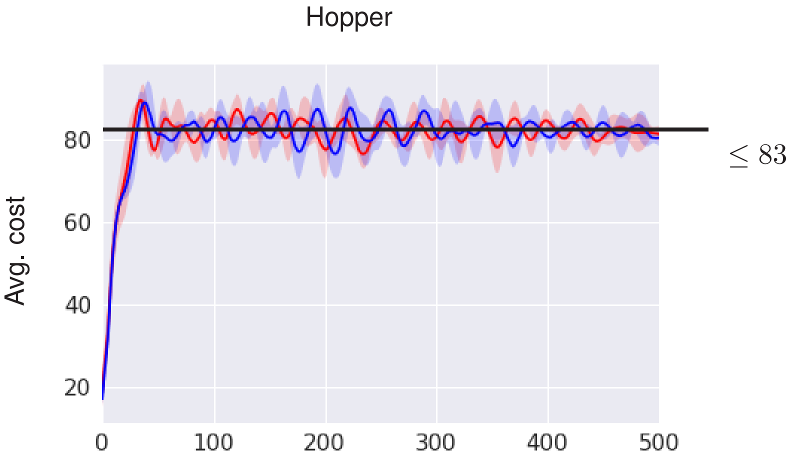
- FINITE-TIME AVERAGE-VALUE PERFORMANCE
  - ⋆ natural policy gradient primal-dual method
  - ⋆ tabular case
  - ⋆ function approximation case

Ding, Zhang, Başar, Jovanović, NeurIPS '20

Ding, Zhang, Duan, Başar, Jovanović, arXiv:2206.02346 (in revision)

Part II
Finite-time last-iterate performance

Oscillation is intrinsic



horizontal axis – # iterations

- ⋆ (—) natural policy gradient primal-dual method
- ⋆ (—) FOCOPS (NeurIPS '20)

Prior art

- ⋆ PID Lagrangian

  Stooke, Achiam, Abbeel, ICML ’20

- ⋆ state augmentation

  Calvo-Fullana, Paternain, Chamon, Ribeiro, arXiv:2102.11941

- ⋆ occupancy measure-based approaches

  Zheng, You, Mallada, arXiv:2212.01505

  Moskovitz, O’Donoghue, Veeriah, Flennerhag, Singh, Zahavy, arXiv:2302.01275

Observation: asymptotic convergence

Settlement I: Regularized method

REGULARIZED LAGRANGIAN

$$L_{\tau}(\pi, \lambda) \;\; = \;\; L(\pi, \lambda) + \tau \left(\mathcal{H}(\pi) + \tfrac{1}{2}\lambda^2 \right)$$

$L(\pi, \lambda) \; := \; V_r^{\pi}(\rho) + \lambda \left(V_g^{\pi}(\rho) - b \right)$ – Lagrangian

$\mathcal{H}(\pi) \; := \; (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t \,|\, s_t) \right]$ – entropy-like term

τ – regularization parameter

$\star$ $(\pi_\tau^\star, \lambda_\tau^\star)$ – τ-near saddle point of $L(\pi, \lambda)$

Regularized policy gradient primal-dual method

REGULARIZED POLICY PRIMAL-DUAL UPDATE

$$\pi^+(\cdot \mid s) \;\propto\; \pi(\cdot \mid s) \exp\left(\frac{\eta}{1-\gamma} Q^{\pi}_{L_\tau}(s,\cdot)\right) \qquad \text{(MWU)}$$

$$\lambda^+ \;=\; \mathcal{P}\left((1-\eta\tau)\lambda \,-\, \eta\left(V^{\pi}_{g}(\rho) - b\right) \right)$$

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

- ⋆ $\tau = 0$ – natural policy gradient primal-dual method
- ⋆ $\eta > 0$ – single-time-scale

Finite-time last-iterate performance

Theorem (informal)

★ Distance to $(\pi_\tau^\star, \lambda_\tau^\star)$

$$\Phi_{t+1} \;:=\; \mathsf{KL}(\pi_t, \pi_\tau^\star) + \frac{1}{2}(\lambda_t - \lambda_\tau^\star)^2 \;\;\lesssim\;\; \mathrm{e}^{-\eta\tau t} \;+\; \frac{\eta}{\tau}$$

KL – visitation-weighted KL divergence

- $\eta\tau$ – linear rate
- (π_t, λ_t) – exponential stability
- $\eta = \epsilon\tau$ – ϵ-near regularized saddle point

$$\Phi_t \;=\; O(\epsilon) \quad \text{for all } \; t \;\geq\; \frac{1}{\epsilon\tau^2}\log\left(\frac{1}{\epsilon}\right)$$

Implication (informal)

★ Optimality gap & Constraint violation

$$V_r^\star(\rho) - V_r^{(T)}(\rho) \;\leq\; \epsilon \quad \text{and} \quad b - V_g^{(T)}(\rho) \;\leq\; \epsilon \quad \text{for} \;\; T = \Omega\left(\frac{1}{\epsilon^6}\right)$$

$$\eta \;=\; \Theta(\epsilon^4) \qquad\qquad \tau \;=\; \Theta(\epsilon^2)$$

- ⋆ optimality of instantaneous policy iterate
- ⋆ $g' = g - \delta$ – zero constraint violation

$$V_r^\star(\rho) - V_r^{(T)}(\rho) \;\leq\; \epsilon \quad \text{and} \quad b - V_g^{(T)}(\rho) \;\leq\; 0$$

Settlement II: Optimistic method

OPTIMISTIC POLICY GRADIENT PRIMAL-DUAL UPDATE

$$\begin{aligned} \pi^{+}(a \,|\, s) &= \mathcal{P}_{\Delta(A)}\left(\hat{\pi}(\cdot \,|\, s) + \eta\, Q^{\pi}_{r+\lambda g}(s,\cdot) \right) \\ \lambda^{+} &= \mathcal{P}_{\Lambda}\left(\hat{\lambda} - \eta \left(V_g^{\pi}(\rho) - b \right) \right) \end{aligned}$$

prediction step

$$\begin{aligned} \hat{\pi}^{+}(a \,|\, s) &= \mathcal{P}_{\Delta(A)}\left(\hat{\pi}(\cdot \,|\, s) + \eta\, Q^{\pi^{+}}_{r+\lambda^{+} g}(s,\cdot) \right) \\ \hat{\lambda}^{+} &= \mathcal{P}_{\Lambda}\left(\hat{\lambda} - \eta \left(V_g^{\pi^{+}}(\rho) - b \right) \right) \end{aligned}$$

real update

- ⋆ $(\hat{\pi}, \hat{\lambda}) = (\pi^{+}, \lambda^{+})$ – natural policy gradient primal-dual method
- ⋆ $\eta > 0$ – single-time-scale

Finite-time last-iterate performance

Theorem (informal)

★ Distance to the set of saddle points $\Pi^\star \times \Lambda^\star$

$$\text{Dist}(\hat{\pi}_t, \mathcal{P}_{\Pi^\star}(\hat{\pi}_t)) + \frac{1}{2}(\hat{\lambda}_t - \mathcal{P}_{\Lambda^\star}(\hat{\lambda}_t))^2 \;\; \leq \;\; \left(\frac{1}{1+C}\right)^t$$

Dist – visitation-weighted norm square distance

η, C – problem-dependent constants

- $\frac{1}{1+C}$ – linear rate
- (π_t, λ_t) – exponential stability

Implication (informal)

★ Optimality gap & Constraint violation

$$V_r^\star(\rho) - V_r^{(T)}(\rho) \;\leq\; \epsilon \quad \text{and} \quad b - V_g^{(T)}(\rho) \;\leq\; \epsilon \quad \text{for} \;\; T = \Omega\left(\log^2 \frac{1}{\epsilon}\right)$$

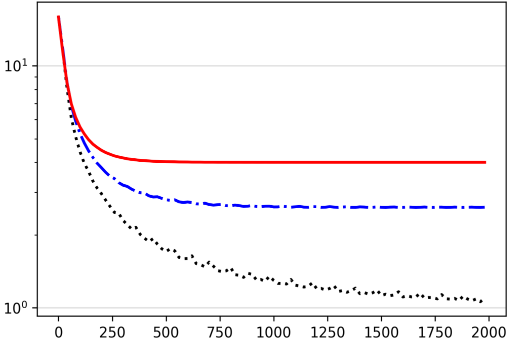
η – problem-dependent constant

- optimality of instantaneous policy iterate
- $g' = g - \delta$ – zero constraint violation

$$V_r^\star(\rho) - V_r^{(T)}(\rho) \;\leq\; \epsilon \quad \text{and} \quad b - V_g^{(T)}(\rho) \;\leq\; 0$$

Sublinear convergence of regularized method

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

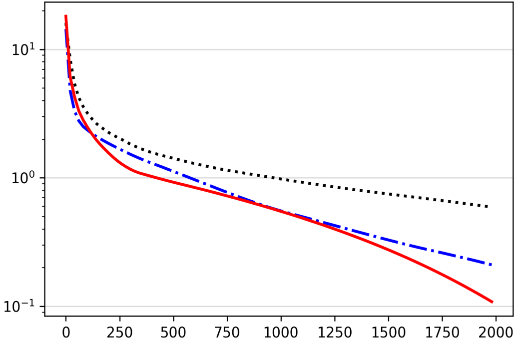


horizontal axis − # iterations $\eta = 0.1$ − stepsize

⋆ (—) $\tau = 0.1$ (-·) $\tau = 0.05$ $(\cdot\cdot)$ $\tau = 0.01$

Linear convergence of optimistic method

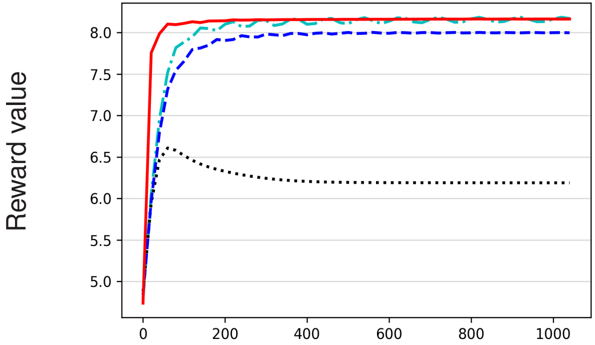
$$\sum_s \|\pi_t(\cdot \mid s) - \pi^\star(\cdot \mid s)\|^2$$



horizontal axis − # iterations

⋆ (—) $\eta = 0.2$ (−·) $\eta = 0.1$ (··) $\eta = 0.05$

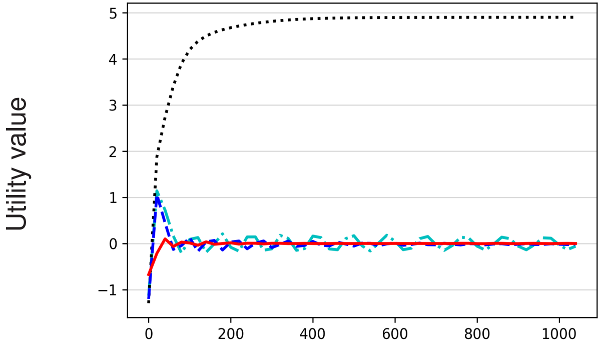
Comparison of primal-dual methods (I)



horizontal axis – # iterations $\eta = 0.1$ – stepsize

- (—) optimistic method (- -) regularized method ($\tau = 0.08$)
- (–·) NPG-PD (NeurIPS '20) (··) PID-Lagrangian (ICML '20)

Comparison of primal-dual methods (II)



horizontal axis – # iterations $\eta = 0.1$ – stepsize

- ⋆ (—) optimistic method (- -) regularized method ($\tau = 0.08$)
- ⋆ (–·) NPG-PD (NeurIPS '20) (··) PID-Lagrangian (ICML '20)

Summary of Part II

■ FINITE-TIME LAST-ITERATE PERFORMANCE

- ⋆ regularized policy gradient primal-dual method
- ⋆ optimistic policy gradient primal-dual method
- ⋆ single-time-scale

Ding, Wei, Zhang, Ribeiro, arXiv:2306.11700 (submitted)

Future directions

- ⋆ constrained policy optimization with exploration
- ⋆ finite-time last-iterate convergence in the online setting
- ⋆ other constrained MDP settings
- ⋆ real-life applications of constrained RL

Thank you for your attention.