# Finite-time performance of policy optimization methods for constrained reinforcement learning


Dongsheng Ding

https://dongshed.github.io

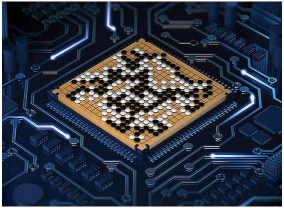
with Kaiqing Zhang, Jiali Duan, Tamer Başar, Mihailo R. Jovanović



2022 INFORMS Annual Meeting, Indianapolis, Indiana

# Policy optimization successes in RL

Go



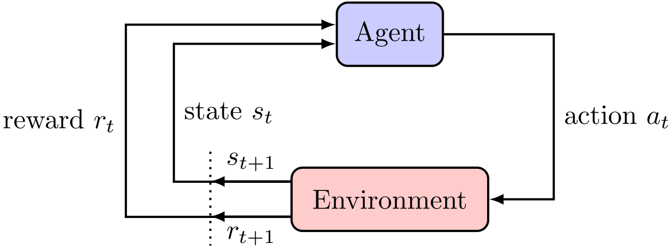
AlphaZero, Silver et al., '17

Video game

OpenAI Five, '18

# Framework for RL

## MARKOV DECISION PROCESSES (MDPS)

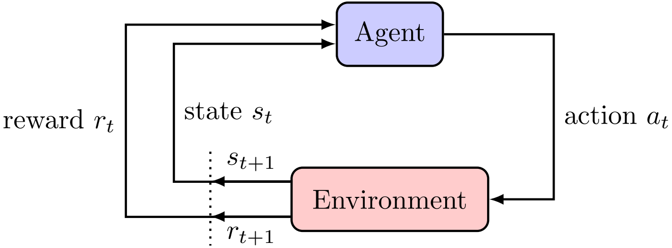



$\pi : S$ (states) $\to A$ (actions) – a policy

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

# Framework for RL

## ■ MARKOV DECISION PROCESSES ( MDPS )




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

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

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

# Real-world constraints

Automated vehicles


Waymo

Industrial robot

Siemens

# Real-world constraints

Automated vehicles


Waymo

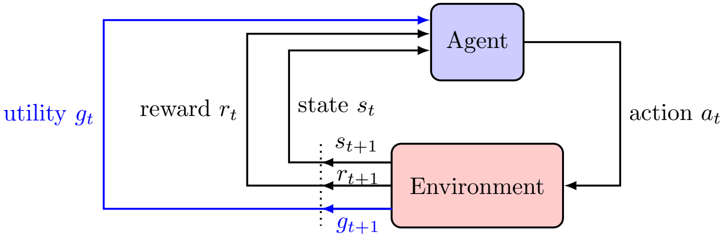
Industrial robot

Siemens

| Application | Goal | Constraints |
|---|---|---|
| Automated vehicles | Follow a path | Fuel efficiency |
| Industrial robot | Manufacture products | Risk-awareness |
| ⋮ | ⋮ | ⋮ |

# Framework for constrained RL

## CONSTRAINED MDPS




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

$$V_r^{\pi}(\rho) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,|\, s_0 \sim \rho]$$

$$V_g^{\pi}(\rho) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(s_t, a_t) \,|\, s_0 \sim \rho]$$

# Constrained policy optimization

$$\begin{aligned} \underset{\pi}{\text{maximize}} \quad & V_r^\pi(\rho) \\ \text{subject to} \quad & V_g^\pi(\rho) \;\geq\; b \end{aligned}$$

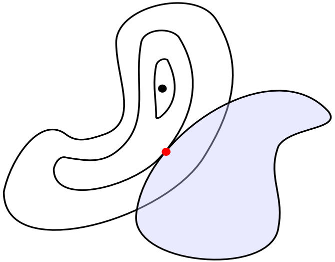
Altman, CRC Press '99

# Constrained policy optimization

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

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

Altman, CRC Press '99

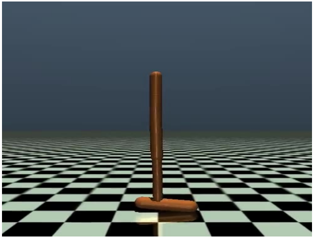


non-convex objective $V_r^\pi(\rho)$

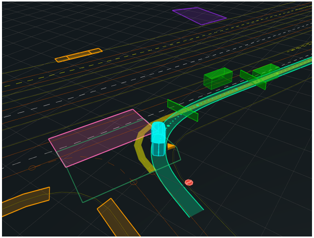
non-convex feasible set $\{\pi \mid V_g^\pi(\rho) \;\geq\; b\}$

# Model-free policy search

MuJoCo



Waymo Driver



**folklore:** asymptotic convergence (to a stationary point)

Achiam, Held, Tamar, Abbeel, ICML ’17

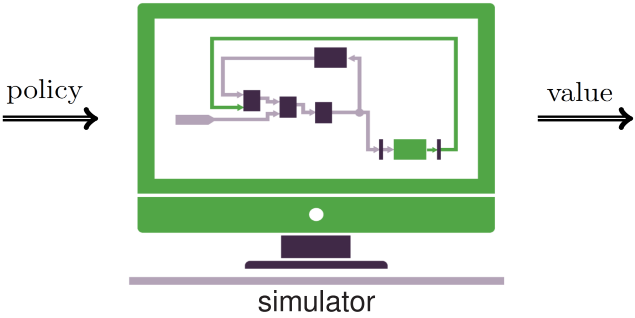
Tessler, Mankowitz, Mannor, ICLR ’18

Le, Voloshin, Yue, ICML, ’19

**Limitation I:** lack of finite-time performance guarantee

**Limitation II:** lack of optimality guarantee

# Simulation setting



# Contribution

■ EFFECTIVE CONSTRAINED POLICY SEARCH METHODS

Finite-time performance

error bound $O\left(\frac{1}{\sqrt{T}}\right)$

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

$T$ – number of iterations

error bound – optimality gap & constraint violation

Ding, Zhang, Başar, Jovanović, NeurIPS '20

Ding, Zhang, Duan, Başar, Jovanović, arXiv:2206.02346 (submitted)

## softmax policy class

( exact gradient, tabular case )

# 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 softmax policy optimization

- SOFTMAX POLICY

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

complete & differentiable

- CONSTRAINED PARAMETER OPTIMIZATION

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & V_r^{\pi_\theta}(\rho) \\ \text{subject to} & V_g^{\pi_\theta}(\rho) \;\geq\; b \end{array}$$

# Constrained softmax policy optimization

## SOFTMAX POLICY

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

complete & differentiable

## CONSTRAINED PARAMETER OPTIMIZATION

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & V_r^{\pi_\theta}(\rho) \\ \text{subject to} \quad & V_g^{\pi_\theta}(\rho) \;\geq\; b \end{aligned}$$

Non-convex objective & feasible set

# $Q$-value function & visitation measure

## $Q$-VALUE FUNCTION

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

# $Q$-value function & visitation measure

- $Q$-VALUE FUNCTION

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

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

# $Q$-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

# $Q$-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

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

# $Q$-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

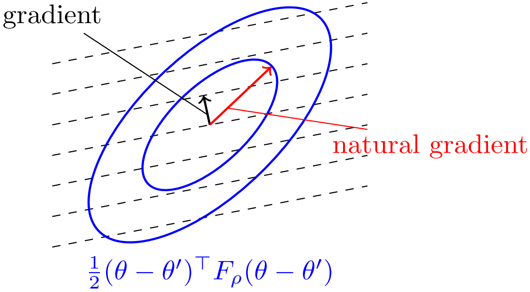
$Q_g^\pi(s,a)$, $A_g^\pi(s,a)$ – use $g$ to define them similarly

## 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 $s_0 \sim \rho$

# Natural ( policy ) gradient




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

steepest descent in Fisher information distance

Amari, '83

# Natural policy gradient primal-dual method

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

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

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

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

$F_\rho(\theta)$ – Fisher information

$\lambda$ – price of constraint violation

# NPG as $A$-regression

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

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

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

# NPG as $A$-regression

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

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

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

- ⋆ 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 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 $=$ stretched advantage function

# Policy primal-dual update

- PRIMAL UPDATE AS MULTIPLICATIVE WEIGHT UPDATE

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

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

# Policy primal-dual update

## PRIMAL UPDATE AS MULTIPLICATIVE WEIGHT UPDATE

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

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

$$\Downarrow$$

$$\pi_\theta^+(a\,|\,s) \;=\; \pi_\theta(a\,|\,s)\,\frac{\exp\left(\frac{\eta_1}{1-\gamma} A_L^{\pi_\theta}(s,a)\right)}{Z(s)} \qquad \text{(MWU)}$$

$$\lambda^+ \;=\; \mathcal{P}_\Lambda\left(\lambda - \eta_2\left(V_g^{\pi_\theta}(\rho) - b\right)\right)$$

$$Z(s) := \sum_a \pi(a\,|\,s)\exp\left(\frac{\eta_1}{1-\gamma} A_L^{\pi_\theta}(s,a)\right)$$

# Policy primal-dual update

## PRIMAL UPDATE AS MULTIPLICATIVE WEIGHT UPDATE

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

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

$$\Downarrow$$

$$\pi_\theta^+(a \,|\, s) = \pi_\theta(a \,|\, s) \frac{\exp\left(\frac{\eta_1}{1-\gamma} A_L^{\pi_\theta}(s,a)\right)}{Z(s)} \quad \text{(MWU)}$$

$$\lambda^+ = \mathcal{P}_\Lambda \left( \lambda - \eta_2 \left( V_g^{\pi_\theta}(\rho) - b \right) \right)$$

$$Z(s) := \sum_a \pi(a \,|\, s) \exp\left( \frac{\eta_1}{1-\gamma} A_L^{\pi_\theta}(s,a) \right)$$

- ⋆ $A_L^{\pi_\theta} \leftarrow Q_L^{\pi_\theta}$ – the same policy update
- ⋆ NPG as $A$-regression $\leftarrow$ NPG as $Q$-regerssion

# Finite-time performance

## Theorem (informal)

- ★ Optimality gap

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \;\leq\; O\left(\frac{1}{(1-\gamma)^2}\frac{1}{\sqrt{T}}\right)$$

- ★ Constraint violation

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \;\leq\; O\left(\frac{1}{(1-\gamma)^2}\frac{1}{\sqrt{T}}\right)$$

$T$ – number of iterations

$\star$ $O(\cdot)$ – dimension-free: no $|\mathcal{S}|$, $|\mathcal{A}|$, and $\rho$

## general policy class

( inexact gradient, function approximation case )

# General softmax policy

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

$f_\theta(s,a)$ – neural network

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

# 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

$$\boxed{\pi_\theta(a \,|\, s) \;=\; \frac{e^{\theta^\top \phi_{s,a}}}{\sum_{a'} e^{\theta^\top \phi_{s,a'}}}}$$

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

# Log-linear policy primal-dual update

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

$\nu = d_\rho(s)\pi_\theta(a \,|\, s)$ – 'on-policy' distribution

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

# Log-linear policy primal-dual update

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

$\nu = d_\rho(s)\pi_\theta(a\,|\,s)$ – 'on-policy' distribution

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

■ PRIMAL UPDATE VIA EMPIRICAL SOLUTION

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

$w := w_r + \lambda w_g$ – approximate NPG direction

# Finite-time performance

## Theorem (informal)

★ Optimality gap & Constraint violation

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

$$\leq \;\; O\left(\frac{1}{\sqrt{T}} + \sqrt{\epsilon_{\text{bias}}} + \sqrt{\kappa\, \epsilon_{\text{est}}}\right)$$

$T$ – number of iterations

- $\kappa := \sup_{w \in \mathbb{R}^d} \frac{w^\top \Sigma_{\nu^\star} w}{w^\top \Sigma_{\nu_0} w} < \infty$ – relative condition number
- $\epsilon_{\text{est}} / \epsilon_{\text{bias}}$ – estimation / transfer errors

# Finite-time performance

## Theorem (informal)

★ Optimality gap & Constraint violation

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

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

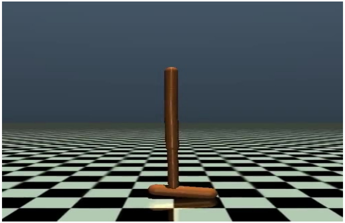
$T$ – number of iterations

- $\kappa := \sup_{w\in\mathbb{R}^d}\frac{w^\top\Sigma_{\nu^\star}w}{w^\top\Sigma_{\nu_0}w}<\infty$ – relative condition number
- $\epsilon_{\text{est}}/\epsilon_{\text{bias}}$ – estimation / transfer errors

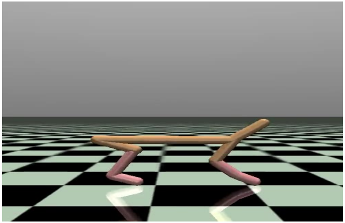
This holds for general smooth policy.

# MuJoCo robotics

Hopper-v3



HalfCheetah-v3

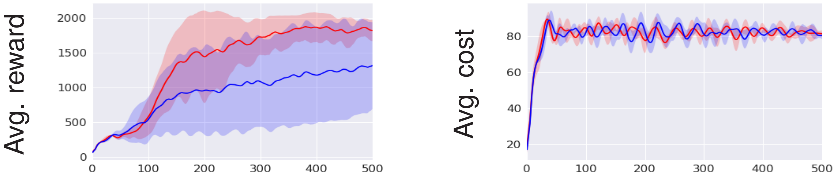


- energy efficiency $= 50\%$ speed from unconstrained PPO:

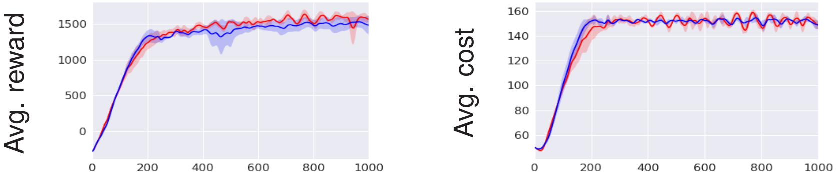
  83 – Hopper-v3 152 – Halfcheetah-v3

- constraint – reward tradeoff: walk with energy efficiency

## Hopper-v3




## HalfCheetah-v3




horizontal axis – # iterations

- (—) – our method
- (—) – FOCOPS, NeurIPS '20

# Summary

■ THEORY OF NPG PRIMAL-DUAL METHOD

- ⋆ softmax tabular case
- ⋆ function approximation case
- ⋆ sample-based algorithms & sample complexity

Ding, Zhang, Başar, Jovanović, NeurIPS ’20

Ding, Zhang, Duan, Başar, Jovanović, arXiv:2206.02346 (submitted)

■ FUTURE DIRECTIONS

- ⋆ better performance
- ⋆ policy-directed exploration
- ⋆ other types of constraints

# Backup slides

# Proof sketch
# Softmax policy class

# Convergence in constrained optimality measure

**Step #1:** performance difference & telescope MWU

$$\boxed{V_r^\star(\rho) - V_r^{(t)}(\rho)}$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\,\sim\, d^\star}\big[\sum_{a\,\in\, A}\pi^\star(a\,|\,s)A_r^{(t)}(s,a)\big]$$

$$\leq \frac{1}{\eta_1}\mathbb{E}_{s\,\sim\, d^\star}\Big[D_{\mathsf{KL}}\big(\pi^\star(\cdot\,|\,s),\,\pi^{(t)}(\cdot\,|\,s)\big) - D_{\mathsf{KL}}\big(\pi^\star(\cdot\,|\,s),\,\pi^{(t+1)}(\cdot\,|\,s)\big)\Big]$$

$$- \boxed{\lambda^{(t)}\big(V_g^\star(\rho) - V_g^{(t)}(\rho)\big)}$$

$$+ \frac{1}{\eta_1}\mathbb{E}_{s\,\sim\, d^\star}\log Z^{(t)}(s)$$

$$\frac{1}{\eta_1 T}\sum_{t\,=\,0}^{T-1}\mathbb{E}_{s\,\sim\, d^\star}\log Z^{(t)}(s) \;\lesssim\; \frac{1}{\sqrt{T}}$$

## AVERAGE PERFORMANCE

$$V_r^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_r^{(t)}(\rho) + \lambda\left( V_g^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_g^{(t)}(\rho)\right) \;\lesssim\; \frac{1}{\sqrt{T}}$$

any $\lambda \in [0, C]$, $C > 0$

$$V_g^\star(\rho) \;\geq\; b$$

**Step #2:** linear programming & strong duality

## CONSTRAINED OPTIMALITY MEASURE

$$\exists\,\pi',\; \underbrace{V_r^\star(\rho) - V_r^{\pi'}(\rho)}_{\text{optimality gap}} + C \times \underbrace{\left[\, b - V_g^{\pi'}(\rho)\,\right]_+}_{\text{constraint violation}} \;\lesssim\; \frac{1}{\sqrt{T}}$$