Policy gradient primal-dual mirror descent for constrained MDPs

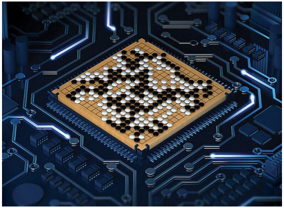
Dongsheng Ding
https://dongshed.github.io

a joint work with
Mihailo R. Jovanović

61st IEEE Conference on Decision and Control, Cancún, Mexico

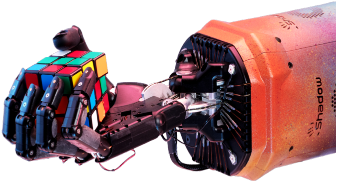
Success stories of RL

Go



AlphaZero, Silver et al., ’17

Robot hand



Rubik’s cube, ’19

Constrained RL

Automated vehicles

Waymo

Industrial robot

Siemens

Constrained RL

Automated vehicles

Waymo

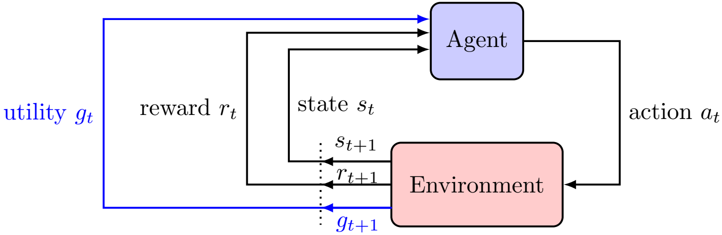
Industrial robot

Siemens

Applications	**Goal**	**Constraints**
Automated vehicles	Follow a path	Fuel efficiency
Industrial robot	Manufacture products	Risk-awareness
...	...	...

Framework for constrained RL

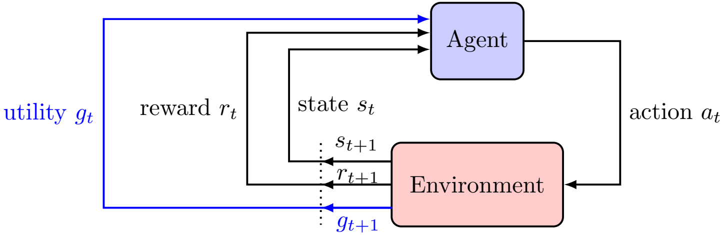
CONSTRAINED MDPS



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

Framework for constrained RL

CONSTRAINED MDPS



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

$V_r^\pi(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$ – reward value function

$V_g^\pi(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g_t\right]$ – utility value function

Constrained policy optimization

$$\begin{array}{ll} \underset{\pi}{\text{maximize}} & V_r^\pi(\rho) \;:=\; \mathbb{E}_{s_0 \sim \rho}\left[\, V_r^\pi(s_0)\,\right] \\ \text{subject to} & V_g^\pi(\rho) \;:=\; \mathbb{E}_{s_0 \sim \rho}\left[\, V_g^\pi(s_0)\,\right] \;\geq\; b \end{array}$$

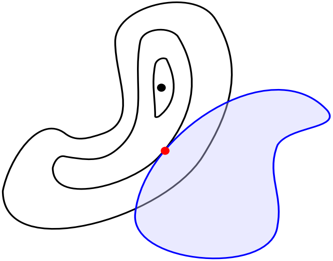
Altman, CRC Press ’99

Constrained policy optimization

$$\underset{\pi}{\text{maximize}} \quad V_r^\pi(\rho) \;:=\; \mathbb{E}_{s_0 \sim \rho}\left[\,V_r^\pi(s_0)\,\right]$$

$$\text{subject to} \quad V_g^\pi(\rho) \;:=\; \mathbb{E}_{s_0 \sim \rho}\left[\,V_g^\pi(s_0)\,\right] \;\geq\; b$$

Altman, CRC Press '99



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

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

Ding, Zhang, Başar, Jovanović, NeurIPS '20

Constrained parameter optimization

POLICY PARAMETRIZATION

- $\star$ $\pi_\theta(a \,|\, s) = \theta_{s,a}$ – direct policy
- $\star$ $\pi_\theta(a \,|\, s) = \frac{\mathrm{e}^{\theta_{s,a}}}{\sum_{a'} \mathrm{e}^{\theta_{s,a'}}}$ – softmax policy
- $\star$ $\pi_\theta(a \,|\, s) = \frac{\mathrm{e}^{f_\theta(s,a)}}{\sum_{a'} \mathrm{e}^{f_\theta(s,a')}}$ – general softmax policy

Agarwal, Kakade, Lee, Mahajan, JMLR ’21

Constrained parameter optimization

POLICY PARAMETRIZATION

- $\star$ $\pi_\theta(a\,|\,s) = \theta_{s,a}$ – direct policy
- $\star$ $\pi_\theta(a\,|\,s) = \dfrac{\mathrm{e}^{\theta_{s,a}}}{\sum_{a'} \mathrm{e}^{\theta_{s,a'}}}$ – softmax policy
- $\star$ $\pi_\theta(a\,|\,s) = \dfrac{\mathrm{e}^{f_\theta(s,a)}}{\sum_{a'} \mathrm{e}^{f_\theta(s,a')}}$ – general softmax policy

Agarwal, Kakade, Lee, Mahajan, JMLR '21

PARAMETER OPTIMIZATION

$$\boxed{\begin{array}{ll} \underset{\theta \,\in\, \Theta}{\text{minimize}} & V_r^{\pi_\theta}(\rho) \\[2ex] \text{subject to} & V_g^{\pi_\theta}(\rho) \;\geq\; b \end{array}}$$

Parameter domain Θ

Lagrangian method

SADDLE POINT PROBLEM

$$\underset{\theta \in \Theta}{\text{maximize}} \;\; \underset{\lambda \geq 0}{\text{minimize}} \;\; L(\theta, \lambda)$$

$L(\theta, \lambda) := V_r^{\pi_\theta}(\rho) + \lambda(V_g^{\pi_\theta}(\rho) - b)$ – Lagrangian

Existence of saddle points

Non concave (θ) and convex (λ)

Lagrangian method

- SADDLE POINT PROBLEM

$$\underset{\theta \in \Theta}{\text{maximize}} \ \underset{\lambda \geq 0}{\text{minimize}} \ L(\theta, \lambda)$$

$L(\theta, \lambda) := V_r^{\pi_\theta}(\rho) + \lambda(V_g^{\pi_\theta}(\rho) - b)$ – Lagrangian

Existence of saddle points

Non concave (θ) and convex (λ)

Popularity of primal-dual methods with different guarantees

Related work (incomplete)

- SMALL STATE/ACTION SPACES

  - ⋆ policy in spherical coordinates, policy gradient primal-dual

    Abad, Krishnamurthy, Martin, Baltcheva, CDC ’02

  - ⋆ direct policy, projected policy gradient primal-dual

    Borkar, SCL ’05; Bhatnagar, SCL ’10

    Ding, Zhang, Başar, Jovanović, ACC ’22

Related work (incomplete)

SMALL STATE/ACTION SPACES

- ⋆ policy in spherical coordinates, policy gradient primal-dual

  Abad, Krishnamurthy, Martin, Baltcheva, CDC ’02

- ⋆ direct policy, projected policy gradient primal-dual

  Borkar, SCL ’05; Bhatnagar, SCL ’10

  Ding, Zhang, Başar, Jovanović, ACC ’22

LARGE STATE/ACTION SPACES

- ⋆ general policy, projected policy gradient primal-dual

  Tessler, Mankowitz, Mannor, ICLR ’19

- ⋆ general softmax policy, natural policy gradient primal-dual

  Ding, Zhang, Başar, Jovanović, NeurIPS ’20

  Ding, Zhang, Duan, Başar, Jovanović, arXiv ’22 (arXiv: 2206.02346)

Related work (incomplete)

- SMALL STATE/ACTION SPACES
  - ⋆ policy in spherical coordinates, policy gradient primal-dual

    Abad, Krishnamurthy, Martin, Baltcheva, CDC '02

  - ⋆ direct policy, projected policy gradient primal-dual

    Borkar, SCL '05; Bhatnagar, SCL '10

    Ding, Zhang, Başar, Jovanović, ACC '22

- LARGE STATE/ACTION SPACES
  - ⋆ general policy, projected policy gradient primal-dual

    Tessler, Mankowitz, Mannor, ICLR '19

  - ⋆ general softmax policy, natural policy gradient primal-dual

    Ding, Zhang, Başar, Jovanović, NeurIPS '20

    Ding, Zhang, Duan, Başar, Jovanović, arXiv '22 (arXiv: 2206.02346)

Question: a unified famework ?

Two pillars

- Q-VALUE FUNCTION

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

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

Two pillars

- Q-VALUE FUNCTION

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

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

- BREGMAN DISTANCE

$$D(p,p') \;=\; h(p) \,-\, (\,h(p') \,+\, \langle \nabla h(p'), p - p' \rangle\,)$$

$h(\cdot)$ – strictly convex and continuously differentiable

Two pillars

Q-VALUE FUNCTION

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

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

BREGMAN DISTANCE

$$D(p,p') \;=\; h(p) \,-\, (\, h(p') \,+\, \langle \nabla h(p'), p - p' \rangle \,)$$

$h(\cdot)$ – strictly convex and continuously differentiable

- ⋆ $D(p,p') = \frac{1}{2}\|p - p'\|^2$ – squared Euclidean 2-norm ($h(p) = \frac{1}{2}\|p\|^2$)
- ⋆ $D(p,p') = \sum_a p_a \log \frac{p_a}{p'_a}$ – KL divergence ($h(p) = \sum_a p_a \log p_a$)

Policy gradient primal-dual mirror descent

$$\boxed{\begin{array}{rcl} \pi^+(\cdot\,|\,s) & \overset{\forall s}{=} & \underset{\pi'(\cdot\,|\,s)\,\in\,\Delta_A}{\operatorname{argmax}} \;\alpha\,\langle Q_r(s,\cdot)+\lambda Q_g(s,\cdot),\pi'(\cdot\,|\,s)\rangle \,-\, D_s(\pi',\pi) \\[2ex] \lambda^+ & = & \mathcal{P}_\Lambda\left(\,\lambda\,-\,\eta\,(V_g(\rho)-b)\,\right) \end{array}}$$

$D_s(\pi',\pi)$ – Bregman distance at s $\qquad$ $\mathcal{P}_\Lambda$ – projection

- ⋆ $Q_r(s,\cdot)+\lambda Q_g(s,\cdot)$ – direction of policy search
- ⋆ $b - V_\rho(\rho)$ – price of constraint violation
- ⋆ no dependence on policy parametrizations

Special case (I)

$$\boxed{\begin{aligned} \pi^+(\cdot\,|\,s) &\overset{\forall s}{=} \underset{\pi'(\cdot\,|\,s)\,\in\,\Delta_A}{\operatorname{argmax}} \; \alpha\,\langle Q_r(s,\cdot) + \lambda Q_g(s,\cdot), \pi'(\cdot\,|\,s)\rangle \,-\, D_s(\pi',\pi) \\ \lambda^+ &= \mathcal{P}_\Lambda\left(\lambda \,-\, \eta\,(V_g(\rho) - b)\right) \end{aligned}}$$

$D_s(\pi',\pi)$ – Bregman distance

$\mathcal{P}_\Lambda$ – projection

$\star$ $D_s(\pi',\pi) = \sum_a \pi'(a\,|\,s) \log \frac{\pi'(a\,|\,s)}{\pi(a\,|\,s)}$ – KL divergence

$$\boxed{\pi^+(\cdot\,|\,s) \;\propto\; \pi(\cdot\,|\,s)\,\mathrm{e}^{\alpha\,(\,Q_r(s,\cdot)\,+\,\lambda Q_g(s,\cdot)\,)}}$$

Natural policy gradient primal-dual method

Ding, Zhang, Başar, Jovanović, NeurIPS '20

Ding, Zhang, Duan, Başar, Jovanović, arXiv '22 (arXiv: 2206.02346)

Special case (II)

$$\boxed{\begin{aligned} \pi^+(\cdot\,|\,s) \;&\overset{\forall s}{=}\; \underset{\pi'(\cdot\,|\,s)\,\in\,\Delta_A}{\operatorname{argmax}}\;\; \alpha\,\langle Q_r(s,\cdot) + \lambda Q_g(s,\cdot), \pi'(\cdot\,|\,s)\rangle \,-\, D_s(\pi',\pi) \\ \lambda^+ \;&=\; \mathcal{P}_\Lambda\,(\,\lambda \,-\, \eta\,(V_g(\rho) - b)\,) \end{aligned}}$$

$D_s(\pi',\pi)$ – Bregman distance $\qquad$ $\mathcal{P}_\Lambda$ – projection

$\star$ $D_s(\pi',\pi) = \frac{1}{2}\|\pi'(\cdot\,|\,s) - \pi(\cdot\,|\,s)\|^2$ – squared Euclidean 2-norm

$$\boxed{\pi^+(\cdot\,|\,s) \;=\; \mathcal{P}_{\Delta_A}\,(\,\pi(\cdot\,|\,s) \,+\, \alpha\,(\,Q_r(s,\cdot) + \lambda Q_g(s,\cdot)\,)\,)}$$

Projected Q-policy gradient primal-dual method

Ding, Zhang, Başar, Jovanović, ACC ’22

Finite-time performance guarantee

OPTIMALITY GAP

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

CONSTRAINT VIOLATION

$$\boxed{\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \;\lesssim\; \frac{1}{(1-\gamma)^2\sqrt{T}}}$$

$\lesssim \;\equiv\; \leq$ up to an absolute constant

Finite-time performance guarantee

OPTIMALITY GAP

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

CONSTRAINT VIOLATION

$$\boxed{\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \lesssim \frac{1}{(1-\gamma)^2\sqrt{T}}}$$

$\lesssim \; \equiv \; \leq$ up to an absolute constant

- $\star$ no dependence on the size of state/action spaces
- $\star$ no dependence on the distribution mismatch coefficient κ
- $\star$ agnostic to distance metrics

Convergence in constrained optimality measure

Step #1: performance difference & telescope sum

Convergence in constrained optimality measure

Step #1: performance difference & telescope sum

$$\boxed{V_r^\star(s) - V_r^t(s) \,+\, \lambda^t\big(V_g^\star(s) - V_g^t(s)\big)}$$

Convergence in constrained optimality measure

Step #1: performance difference & telescope sum

$$\boxed{V_r^\star(s) - V_r^t(s) \,+\, \lambda^t\big(V_g^\star(s) - V_g^t(s)\big)}$$

$$= \tfrac{1}{1-\gamma}\mathbb{E}_{s' \sim d_s^\star}\big[\, \langle Q_L^t(s',\cdot), (\pi^\star - \pi^{t+1})(\cdot \,|\, s') + (\pi^{t+1} - \pi^t)(\cdot \,|\, s')\rangle \,\big]$$

Convergence in constrained optimality measure

Step #1: performance difference & telescope sum

$$\boxed{V_r^\star(s) - V_r^t(s) \,+\, \lambda^t\big(V_g^\star(s) - V_g^t(s)\big)}$$

$$= \; \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^\star}\big[\,\langle Q_L^t(s',\cdot), (\pi^\star - \pi^{t+1})(\cdot\,|\,s') + (\pi^{t+1} - \pi^t)(\cdot\,|\,s')\rangle\,\big]$$

$$\lesssim \; \tfrac{1}{\alpha}\mathbb{E}_{s'\sim d^\star}\,\Big[\,D_{s'}\big(\pi^\star,\,\pi^t\big) - D_{s'}\big(\pi^\star,\,\pi^{t+1}\big)\,\Big]$$

$$+\;\boxed{\big(V_r^{t+1}(d_\rho^\star) - V_r^t(d_\rho^\star)\big) + \lambda^t\big(V_g^{t+1}(d_\rho^\star) - V_g^t(d_\rho^\star)\big)}$$

Convergence in constrained optimality measure

Step #1: performance difference & telescope sum

$$\boxed{V_r^\star(s) - V_r^t(s) \,+\, \lambda^t\big(V_g^\star(s) - V_g^t(s)\big)}$$

$$= \tfrac{1}{1-\gamma}\mathbb{E}_{s' \sim d_s^\star}\big[\,\langle Q_L^t(s',\cdot), (\pi^\star - \pi^{t+1})(\cdot\,|\,s') + (\pi^{t+1} - \pi^t)(\cdot\,|\,s')\rangle\,\big]$$

$$\lesssim\; \tfrac{1}{\alpha}\mathbb{E}_{s' \sim d^\star}\,\big[\,D_{s'}\big(\pi^\star,\,\pi^t\big) - D_{s'}\big(\pi^\star,\,\pi^{t+1}\big)\,\big]$$

$$+\;\boxed{\big(V_r^{t+1}(d_\rho^\star) - V_r^t(d_\rho^\star)\big) + \lambda^t\big(V_g^{t+1}(d_\rho^\star) - V_g^t(d_\rho^\star)\big)}$$

$$\tfrac{1}{T}\textstyle\sum_{t=0}^{T-1}\lambda^t\big(V_g^{t+1}(d_\rho^\star) - V_g^t(d_\rho^\star)\big) \;\lesssim\; \tfrac{1}{\sqrt{T}}$$

AVERAGE PERFORMANCE

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

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

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

AVERAGE PERFORMANCE

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

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

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

Step #2: linear programming & strong duality

AVERAGE PERFORMANCE

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

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

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

Step #2: linear programming & strong duality

CONSTRAINED OPTIMALITY MEASURE

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

Linear function approximation

- LINEAR VALUE FUNCTION ASSUMPTION

$$Q_r^\pi(\cdot,\cdot) \;=\; \langle \phi_r(\cdot,\cdot), w_r^\pi \rangle \;\text{ and }\; Q_g^\pi(\cdot,\cdot) \;=\; \langle \phi_g(\cdot,\cdot), w_g^\pi \rangle$$

$\phi_r, \phi_g : S \times A \to \mathbb{R}^d$ – feature maps

Linear function approximation

■ LINEAR VALUE FUNCTION ASSUMPTION

$$Q_r^\pi(\cdot,\cdot) \;=\; \langle \phi_r(\cdot,\cdot), w_r^\pi \rangle \;\text{ and }\; Q_g^\pi(\cdot,\cdot) \;=\; \langle \phi_g(\cdot,\cdot), w_g^\pi \rangle$$

$\phi_r,\, \phi_g : S \times A \to \mathbb{R}^d$ – feature maps

⋆ linear MDPs – a special case

Jin, Yang, Wang, Jordan, COLT ’20

Linear function approximation

- LINEAR VALUE FUNCTION ASSUMPTION

$$Q_r^\pi(\cdot,\cdot) \;=\; \langle \phi_r(\cdot,\cdot), w_r^\pi \rangle \;\text{ and }\; Q_g^\pi(\cdot,\cdot) \;=\; \langle \phi_g(\cdot,\cdot), w_g^\pi \rangle$$

$\phi_r, \phi_g : S \times A \to \mathbb{R}^d$ – feature maps

- ⋆ linear MDPs – a special case

  Jin, Yang, Wang, Jordan, COLT '20

- ⋆ $V_g^\pi(\cdot) = \langle \varphi_g(\cdot), u_g^\pi \rangle$ – linear value function

Linear function approximation

LINEAR VALUE FUNCTION ASSUMPTION

$$Q_r^\pi(\cdot,\cdot) \;=\; \langle \phi_r(\cdot,\cdot), w_r^\pi \rangle \;\text{ and }\; Q_g^\pi(\cdot,\cdot) \;=\; \langle \phi_g(\cdot,\cdot), w_g^\pi \rangle$$

$\phi_r,\, \phi_g : S \times A \to \mathbb{R}^d$ – feature maps

- ⋆ linear MDPs – a special case

Jin, Yang, Wang, Jordan, COLT ’20

- ⋆ $V_g^\pi(\cdot) = \langle \varphi_g(\cdot), u_g^\pi \rangle$ – linear value function
- ⋆ $\hat{Q}_r^\pi(\cdot,\cdot)$, $\hat{Q}_g^\pi(\cdot,\cdot)$, and $\hat{V}_g^\pi(\rho)$ – unbiased estimates, e.g.,

$$\hat{Q}_r^\pi(\cdot,\cdot) \;=\; \langle \phi_g(\cdot,\cdot), \hat{w}_r^\pi \rangle$$

$$\hat{w}_r^\pi \;\simeq\; \mathrm{LR}(\{(\phi_r(s^k,a^k), R^k)\}_{k=1}^K)$$

Sample-based algorithm

$$\boxed{\begin{aligned} \pi^+(\cdot \,|\, s) \;\; &\overset{\forall s}{=} \;\; \underset{\pi'(\cdot \,|\, s) \,\in\, \Delta_A}{\operatorname{argmax}} \;\; \alpha \, \langle \hat{Q}_r(s,\cdot) + \lambda \hat{Q}_g(s,\cdot), \pi'(\cdot \,|\, s) \rangle \; - \; D_s(\pi', \pi) \\ \lambda^+ \;\; &= \;\; \mathcal{P}_\Lambda \Big(\lambda \; - \; \eta \, (\hat{V}_g(\rho) - b) \Big) \end{aligned}}$$

$D_s(\pi', \pi)$ – Bregman distance

$\mathcal{P}_\Lambda$ – projection

- $\star$ $\hat{Q}_r(s,\cdot) + \lambda \hat{Q}_g(s,\cdot)$ – estimated direction of policy search
- $\star$ $b - \hat{V}_\rho(\rho)$ – estimated price of constraint violation

Finite-time performance guarantee

- OPTIMALITY GAP

$$\boxed{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \;\lesssim\; \frac{1}{(1-\gamma)^2\sqrt{T}} + \sqrt{\kappa\,\epsilon_{\mathrm{stat}}}}$$

- CONSTRAINT VIOLATION

$$\boxed{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \;\lesssim\; \frac{1}{(1-\gamma)^2\sqrt{T}} + \sqrt{\kappa\,\epsilon_{\mathrm{stat}}}}$$

$\lesssim \; \equiv \; \leq$ up to an absolute constant

Finite-time performance guarantee

- **OPTIMALITY GAP**

$$\boxed{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \;\lesssim\; \frac{1}{(1-\gamma)^2\sqrt{T}} + \sqrt{\kappa\,\epsilon_{\text{stat}}}}$$

- **CONSTRAINT VIOLATION**

$$\boxed{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \;\lesssim\; \frac{1}{(1-\gamma)^2\sqrt{T}} + \sqrt{\kappa\,\epsilon_{\text{stat}}}}$$

$\lesssim \;\equiv\; \leq$ up to an absolute constant

⋆ ϵ_{stat} – estimation error, e.g., $O(1/K)$ for SGD

⋆ $O(1/\epsilon^4)$ – sample complexity

Summary

- POLICY GRADIENT PRIMAL-DUAL MIRROR DESCENT
  - ⋆ dimension-free finite-time performance bounds
  - ⋆ model-free algorithm & sample complexity

Summary

- POLICY GRADIENT PRIMAL-DUAL MIRROR DESCENT
  - ⋆ dimension-free finite-time performance bounds
  - ⋆ model-free algorithm & sample complexity

- FUTURE DIRECTIONS
  - ⋆ better sample complexity
  - ⋆ general function approximation
  - ⋆ policy-directed exploration

Thank you for your attention.