Policy gradient primal-dual method for constrained MDPs

Dongsheng Ding

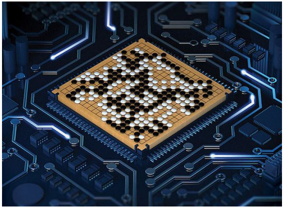
a joint work with

Kaiqing Zhang, Tamer Başar, Mihailo R. Jovanović

2022 American Control Conference, Atlanta

Success stories of RL

Go



AlphaZero, Silver et al., '17

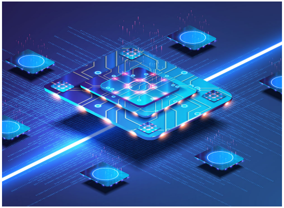
Video game

OpenAI Five, '18

Automated trading

Crypto trading bots, '20

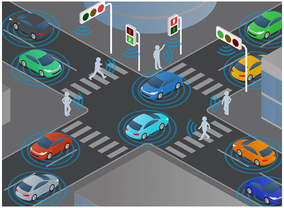
Chip design



AI chip, Azalia et al., Google, '21

Constrained RL

Automated vehicles



Keysight

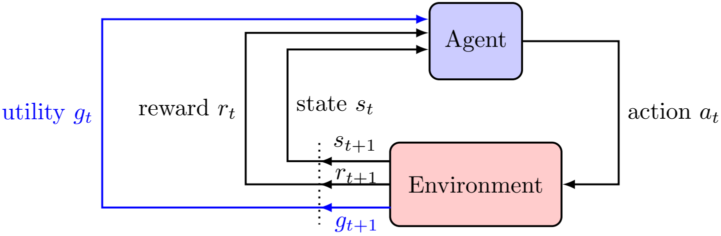
Industrial robot

Siemens

Applications	**Goal**	**Constraints**
Automated vehicles	Reach a destination	Fuel/Traffic rules
Industrial robot	Manufacture products	No damages
...	...	...

Framework for constrained RL

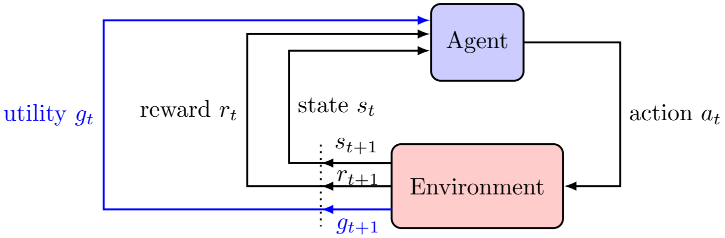
■ CONSTRAINED MDPS



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

Framework for constrained RL

CONSTRAINED MDPS



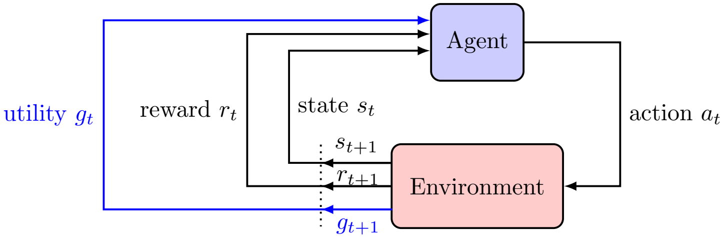
$\pi : S$ (states) $\to A$ (actions) – a policy

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

$\gamma \in [0, 1)$ – discounted factor

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

$\gamma \in [0, 1)$ – discounted factor

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

Constrained policy optimization

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

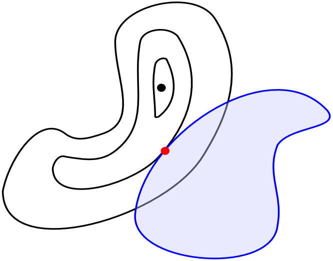
Altman, CRC Press '99

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 $\quad V_r^\pi(\rho)$

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

Ding, Zhang, Başar, Jovanović, NeurIPS '20

Constrained tabular policy optimization

- DIRECT TABULAR POLICY

$$\pi_\theta(a \mid s) \;=\; \theta_{s,a}, \;\; \theta \in \Theta$$

$$|S|, |A| < \infty$$

$$\Theta \;=\; \{\theta \in \mathbb{R}^{|S||A|} \mid \textstyle\sum_{a'} \theta_{s,a'} = 1,\; \theta_{s,a'} \geq 0, \forall s \in S\}$$

Constrained tabular policy optimization

- DIRECT TABULAR POLICY

$$\pi_\theta(a \mid s) \;=\; \theta_{s,a}, \;\; \theta \in \Theta$$

$$|S|, |A| < \infty$$

$$\Theta \;=\; \{\theta \in \mathbb{R}^{|S||A|} \mid \textstyle\sum_{a'} \theta_{s,a'} = 1, \; \theta_{s,a'} \geq 0, \forall s \in S\}$$

- PARAMETER OPTIMIZATION

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

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)

Question: convergence of first-order methods?

Policy gradient primal-dual method

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

$\mathcal{P}_\Theta$, $\mathcal{P}_\Lambda$ – projections

- ⋆ $\nabla_\theta L(\theta, \lambda)$ – policy gradient (PG)

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

- ⋆ $\nabla_\lambda L(\theta, \lambda) := V_g^\theta(\rho) - b$

Related work

ASYMPTOTIC CONVERGENCE

- ⋆ Average-reward case: policy in spherical coordinates

  Abad, Krishnamurthy, Martin, Baltcheva, CDC '02

- ⋆ Average-reward case: direct policy, actor-critic

  Borkar, SCL '05

- ⋆ Discounted-reward case: general policy, actor-critic

  Tessler, Mankowitz, Mannor, ICLR '19

Related work

■ ASYMPTOTIC CONVERGENCE

- ⋆ Average-reward case: policy in spherical coordinates

  Abad, Krishnamurthy, Martin, Baltcheva, CDC ’02

- ⋆ Average-reward case: direct policy, actor-critic

  Borkar, SCL ’05

- ⋆ Discounted-reward case: general policy, actor-critic

  Tessler, Mankowitz, Mannor, ICLR ’19

Question: non asymptotic convergence ?

Finite-time performance guarantee

- OPTIMALITY GAP

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \;\lesssim\; \frac{|A|\,|S|\,d_\infty^2}{T^{1/4}}$$

- CONSTRAINT VIOLATION

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \;\lesssim\; \frac{|A|\,|S|\,d_\infty^2}{T^{1/4}}$$

$d_\infty := \|d_\rho^{\pi^\star}/\rho\|_\infty$ – distribution mismatch

Finite-time performance guarantee

OPTIMALITY GAP

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \;\lesssim\; \frac{|A|\,|S|\,d_\infty^2}{T^{1/4}}$$

CONSTRAINT VIOLATION

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \;\lesssim\; \frac{|A|\,|S|\,d_\infty^2}{T^{1/4}}$$

$d_\infty := \|d_\rho^{\pi^\star}/\rho\|_\infty$ – distribution mismatch

- ⋆ $\eta_1 \simeq 1/|A|$, $\eta_2 \simeq |A||S|d_\infty^2/\sqrt{T}$ – stepsizes
- ⋆ $V_r^{\theta^{(0)}} \geq V_r^\star$ – initialization

Finite-time performance guarantee

OPTIMALITY GAP

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \;\lesssim\; \frac{|A|\,|S|\,d_\infty^2}{T^{1/4}}$$

CONSTRAINT VIOLATION

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \;\lesssim\; \frac{|A|\,|S|\,d_\infty^2}{T^{1/4}}$$

$d_\infty := \|d_\rho^{\pi^\star}/\rho\|_\infty$ – distribution mismatch

- ⋆ $\eta_1 \simeq 1/|A|$, $\eta_2 \simeq |A||S|d_\infty^2/\sqrt{T}$ – stepsizes
- ⋆ $V_r^{\theta^{(0)}} \geq V_r^\star$ – initialization
- ⋆ $O(1/\epsilon^4)$ – iteration complexity for ϵ-optimality

Two pillars

- OCCUPANCY MEASURE

$$q^{\pi}_{s,a} \;=\; \sum_{t=0}^{\infty} \gamma^t P^{\pi}(s_t = s, a_t = a \,|\, s_0 \sim \rho)$$

$$\mathcal{Q} \;:=\; \{q^{\pi} \in \mathbb{R}^{|S||A|} \,|\, \textstyle\sum_{a \in A}(I - \gamma P_a^{\top})q_a^{\pi} = \rho, q^{\pi} \geq 0\}$$

Two pillars

■ OCCUPANCY MEASURE

$$q^{\pi}_{s,a} \;=\; \sum_{t=0}^{\infty} \gamma^t P^{\pi}(s_t = s, a_t = a \,|\, s_0 \sim \rho)$$

$$\mathcal{Q} \;:=\; \{q^{\pi} \in \mathbb{R}^{|S||A|} \,|\, \textstyle\sum_{a \in A}(I - \gamma P_a^{\top})q_a^{\pi} = \rho, q^{\pi} \geq 0\}$$

⋆ $V_r^{\pi}(\rho) = \langle q^{\pi}, r\rangle$, $V_g^{\pi}(\rho) = \langle q^{\pi}, g\rangle$ – linear functions in q^{π}

Two pillars

OCCUPANCY MEASURE

$$q^{\pi}_{s,a} \;=\; \sum_{t=0}^{\infty} \gamma^t P^{\pi}(s_t = s, a_t = a \,|\, s_0 \sim \rho)$$

$$\mathcal{Q} \;:=\; \{q^{\pi} \in \mathbb{R}^{|S||A|} \,|\, \textstyle\sum_{a \in A}(I - \gamma P_a^{\top}) q_a^{\pi} = \rho, q^{\pi} \geq 0\}$$

⋆ $V_r^{\pi}(\rho) = \langle q^{\pi}, r \rangle$, $V_g^{\pi}(\rho) = \langle q^{\pi}, g \rangle$ – linear functions in q^{π}

STATE VISITATION DISTRIBUTION

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

Convergence in constrained optimality measure

Key property: linearity in occupancy measure

Borkar, PTRF ’88 & Altman, CRC Press ’99

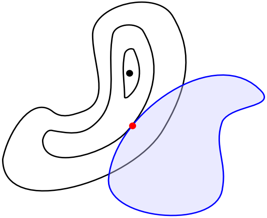
Convergence in constrained optimality measure

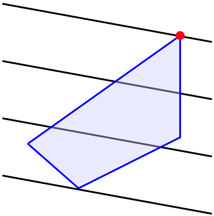
Key property: linearity in occupancy measure

Borkar, PTRF '88 & Altman, CRC Press '99

$$\underset{\theta \in \Theta}{\text{maximize}} \quad V_r^{\pi_\theta}(\rho) \qquad\qquad \underset{q^\pi \in \mathcal{Q}}{\text{maximize}} \quad \langle q^\pi, r \rangle$$

$$\text{subject to} \quad V_g^{\pi_\theta}(\rho) \;\geq\; b \qquad\qquad \text{subject to} \quad \langle q^\pi, g \rangle \;\geq\; b$$





$\theta \in \Theta \qquad \Longleftrightarrow \qquad q^\pi \in \mathcal{Q}$

one-to-one correspondence $\pi_\theta \;\leftrightarrow\; q^\pi$

Step #1: linearity in occupancy measure & smoothness

Step #1: linearity in occupancy measure & smoothness

$$
\begin{aligned}
& V_r^{(t+1)} + \lambda^{(t)} V_g^{(t+1)} \\
= \; & \boxed{\langle q^{(t+1)}, r + \lambda^{(t)} g \rangle} \\
\geq \; & \underset{\alpha \in [0,1]}{\text{maximize}} \;\; \alpha \boxed{\langle q^\star, r + \lambda^{(t)} g \rangle} + (1-\alpha) \langle q^{(t)}, r + \lambda^{(t)} g \rangle \\
& \qquad - \alpha^2 L d_\infty^2
\end{aligned}
$$

quadratic objective

AVERAGE PERFORMANCE

$$\boxed{V_r^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_r^{(t)}(\rho) + \lambda\left(V_g^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_g^{(t)}(\rho)\right) \;\lesssim\; \frac{1}{T^{1/4}}}$$

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

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

AVERAGE PERFORMANCE

$$V_r^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_r^{(t)}(\rho) + \lambda\left(V_g^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_g^{(t)}(\rho)\right) \;\lesssim\; \frac{1}{T^{1/4}}$$

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

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

Step #2: linear programming & strong duality

AVERAGE PERFORMANCE

$$V_r^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_r^{(t)}(\rho) + \lambda\left(V_g^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V_g^{(t)}(\rho)\right) \;\lesssim\; \frac{1}{T^{1/4}}$$

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

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

Step #2: linear programming & strong duality

CONSTRAINED OPTIMALITY MEASURE

$$\exists\, \pi',\ \underbrace{V_r^\star(\rho) - V_r^{\pi'}(\rho)}_{\text{optimality gap}} + C \times \underbrace{\left[\, b - V_g^{\pi'}(\rho) \,\right]_+}_{\text{constraint violation}} \;\lesssim\; \frac{1}{T^{1/4}}$$

Summary

- POLICY GRADIENT PRIMAL-DUAL METHOD
  - ⋆ finite-time performance guarantee in tabular case
  - ⋆ model-free algorithm & sample complexity

Summary

- POLICY GRADIENT PRIMAL-DUAL METHOD
  - ⋆ finite-time performance guarantee in tabular case
  - ⋆ model-free algorithm & sample complexity

- FUTURE DIRECTIONS
  - ⋆ better rate and dependence on problem parameters
  - ⋆ general policy

Thank you for your attention.