Last-Iterate Convergent Policy Gradient Primal-Dual Methods for Constrained MDPs

Dongsheng Ding

a joint work with

Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro

NeurIPS 2023

Constrained policy optimization

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

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

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

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

FEATURES

- ⋆ non-convex functional constrained optimization
- ⋆ randomized optimal policy
- ⋆ no uniform optimal policy across all states

Lagrangian-based approaches

$$L(\pi, \lambda) \;:=\; V_r^{\pi}(\rho) + \lambda V_g^{\pi}(\rho)$$

ISSUES

- ⋆ scalarization fallacy — suboptimal

  e.g., Zahavy et al., NeurIPS 2021

- ⋆ dual methods — two-time-scale

  e.g., Ying, et al., AISTATS 2021; Gladin et al., AISTATS 2023

- ⋆ primal-dual methods — oscillation

  e.g., Stooke, et al., ICML 2020; Ding et al., NeurIPS 2020

Question

Can the policy iterates of a single-time-scale policy-based primal-dual algorithm converge to an optimal constrained policy with non-asymptotic rate?

Non-asymptotic last-iterate performance

REGULARIZED POLICY GRADIENT PRIMAL-DUAL METHOD

policy last-iterate convergence with sublinear error rate

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

OPTIMISTIC POLICY GRADIENT PRIMAL-DUAL METHOD

policy last-iterate convergence with linear error rate

- ⋆ tabular — problem-dependent

error rate – optimality gap & constraint violation

Constrained saddle-point problem

$$\underset{\pi \in \Pi}{\text{maximize}} \ \underset{\lambda \in \Lambda}{\text{minimize}} \ \ L(\pi, \lambda) \ \ = \ \ \underset{\lambda \in \Lambda}{\text{minimize}} \ \underset{\pi \in \Pi}{\text{maximize}} \ \ L(\pi, \lambda)$$

■ CHALLENGES

- ⋆ non-convex constrained saddle-point problem
- ⋆ randomized optimal policy
- ⋆ no uniform optimal policy across all states
- ⋆ asymmetric two-player game

Settlement I: Regularized method

REGULARIZED LAGRANGIAN

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

Li, et al., arXiv 2021

τ – regularization parameter

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

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

Regularized policy gradient primal-dual method

REGULARIZED POLICY PRIMAL-DUAL UPDATE

$$\boxed{\begin{aligned} \pi^+(\cdot\,|\,s) &\;\propto\; \pi(\cdot\,|\,s)\,\text{exp}\left(\tfrac{\eta}{1-\gamma} Q_{L_\tau}^\pi(s,\cdot)\right) \quad \text{(MWU)} \\ \lambda^+ &\;=\; \mathcal{P}\left((1-\eta\tau)\lambda \,-\, \eta\, V_g^\pi(\rho)\right) \end{aligned}}$$

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

- $\star$ $\tau = 0$ – NPG-PD (Ding et al., NeurIPS 2020)
- $\star$ $\eta > 0$ – single-time-scale

Non-asymptotic last-iterate performance

Theorem (informal)

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

$$\text{Dist}(\pi_t, \pi_\tau^\star) + \frac{1}{2}(\lambda_t - \lambda_\tau^\star)^2 \;\lesssim\; \mathrm{e}^{-\eta\tau t} + \frac{\eta}{\tau} \quad \text{for any } \; t \geq 0$$

Dist – visitation-weighted KL divergence

⋆ (π_t, λ_t) – exponential stability

Implication (informal)

★ Optimality gap & Constraint violation

$$V_r^\star(\rho) - V_r^{(\pi_T)}(\rho) \;\le\; \epsilon \quad \text{and} \quad -V_g^{(\pi_T)}(\rho) \;\le\; \epsilon$$

$$T = \Omega\left(\frac{1}{\epsilon^6}\right)$$

$$\eta \;=\; \Theta(\epsilon^4)$$

$$\tau \;=\; \Theta(\epsilon^2)$$

$\star$ optimality of instantaneous policy iterate

Settlement II: Optimistic method

OPTIMISTIC POLICY GRADIENT PRIMAL-DUAL UPDATE

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

prediction step

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

Popov, USSR 1980

- ⋆ $(\hat{\pi}, \hat{\lambda}) = (\pi^+, \lambda^+)$ – PG-PD MD (Ding et al., CDC 2022)
- ⋆ $\eta > 0$ – single-time-scale

Non-asymptotic last-iterate performance

Theorem (informal)

★ Distance of $(\hat{\pi}_t, \hat{\lambda}_t)$ to the set of saddle points $\Pi^\star \times \Lambda^\star$

$$\text{Dist}(\hat{\pi}_t, \mathcal{P}_{\Pi^\star}(\hat{\pi}_t)) + \frac{1}{2}(\hat{\lambda}_t - \mathcal{P}_{\Lambda^\star}(\hat{\lambda}_t))^2 \;\lesssim\; \left(\frac{1}{1+C}\right)^t \quad \text{for any } \; t \geq 0$$

Dist – visitation-weighted norm square distance

C – problem-dependent constant

⋆ $(\hat{\pi}_t, \hat{\lambda}_t)$ – exponential stability

Implication (informal)

★ Optimality gap & Constraint violation

$$V_r^{\star}(\rho) - V_r^{(\hat{\pi}_T)}(\rho) \;\le\; \epsilon \quad \text{and} \quad -V_g^{(\hat{\pi}_T)}(\rho) \;\le\; \epsilon$$

$$T = \Omega\left(\log^2 \tfrac{1}{\epsilon}\right)$$

η – problem-dependent constant

⋆ optimality of instantaneous policy iterate

Poster #1114: 6 pm EST, 13 Dec (Wed)

Thank you for your attention.