Global Convergence for Average Reward Constrained MDPs with Primal-Dual Actor Critic Algorithm

Yang Xu*[1], Swetha Ganesh*[1], Washim Uddin Mondal[2], Qinbo Bai[1], Vaneet Aggarwal[1]

[1]Purdue University; [2]Indian Institute of Technology Kanpur

Problem Setup

Motivation: Average-reward reinforcement learning with constraints has many crucial applications since it targets long-term performance in safety-critical systems.

Key question: Can we use general parameterized policy gradient methods in this setup and achieve a near-optimal global convergence rate?

Setting:

- Constrained average-reward MDP:

$$J_g^\pi \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi \left[\sum_{t=0}^{T-1} g(s_t,a_t)\middle| s_0\sim\rho\right], g\in\{r,c\}$$

- Goal $\max_{\theta\in\Theta} J_r^{\pi_\theta}$ s.t. $J_c^{\pi_\theta}\geq 0 \qquad c:\mathcal{S}\times\mathcal{A}\to[-1,1]$

Policy gradients: $\{\pi_\theta|\theta\in\Theta\}\quad \theta\in\mathbb{R}^{\mathrm{d}}\quad \mathrm{d}\ll|\mathcal{S}||\mathcal{A}|$

Primal-Dual Reformulation

- Saddle point optimization:

$$\max_{\theta\in\Theta}\min_{\lambda\geq 0} \mathcal{L}(\theta,\lambda)\triangleq J_r(\theta)+\lambda J_c(\theta)$$

- Primal-dual NPG method:

$$\theta_{k+1}=\theta_k+\alpha F(\theta_k)^{-1}\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)$$

$$\lambda_{k+1}=\mathcal{P}_{[0,\frac{2}{\delta}]}\left[\lambda_k-\beta J_c(\theta_k)\right]$$

Assumptions:

- The CMDP is ergodic
- Slater condition
  There exists a $\delta\in(0,1)$ and $\bar{\theta}\in\Theta$ such that $J_c(\bar{\theta})\geq\delta$
- Fisher non-degeneracy
- Lipschitz and smoothness of score function
- Function approximation error bounded

Algorithm

Algorithm 1 Primal-Dual Natural Actor-Critic

for $k=0,\cdots,K-1$ **do**
  $\omega_{g,0}^k\leftarrow\omega_0,\ \xi_{g,0}^k\leftarrow\xi_0\ \forall g\in\{r,c\}$;
  for $h=0,\cdots,H-1$ **do**
    $s_{kh}^0\leftarrow s_0, P_h^k\sim\mathrm{Geom}(1/2)$
    $l_{kh}\leftarrow(2^{P_h^k}-1)\mathbf{1}(2^{P_h^k}\leq T_{\max})+1$
    for $t=0,\ldots,l_{kh}-1$ **do**
      Take action $a_{kh}^t\sim\pi_{\theta_k}(\cdot|s_{kh}^t)$
      Observe $s_{kh}^{t+1}\sim P(\cdot|s_{kh}^t,a_{kh}^t)$
      Observe $g(s_{kh}^t,a_{kh}^t), g\in\{r,c\}$
    end for
    $s_0\leftarrow s_{kh}^{l_{kh}}$
    Update $\xi_{g,h}^k$ using (18) and (20).
  end for
  for $h=0,\cdots,H-1$ **do**
    $s_{kh}^0\leftarrow s_0, Q_h^k\sim\mathrm{Geom}(1/2)$
    $l_{kh}\leftarrow(2^{Q_h^k}-1)\mathbf{1}(2^{Q_h^k}\leq T_{\max})+1$
    for $t=0,\ldots,l_{kh}-1$ **do**
      Take action $a_{kh}^t\sim\pi_{\theta_k}(\cdot|s_{kh}^t)$
      Observe $s_{kh}^{t+1}\sim P(\cdot|s_{kh}^t,a_{kh}^t)$
      Observe $g(s_{kh}^t,a_{kh}^t), g\in\{r,c\}$
    end for
    $s_0\leftarrow s_{kh}^{l_{kh}}$
    Update $\omega_{g,h}^k$ using (21) and (23)
  end for
  $\xi_g^k\leftarrow\xi_{g,H}^k, \omega_g^k\leftarrow\omega_{g,H}^k, g\in\{r,c\}$
  $\omega_k\leftarrow\omega_r^k+\lambda_k\omega_c^k$
  $\theta_{k+1}=\theta_k+\alpha\omega_k$
  $\lambda_{k+1}=\mathcal{P}_{[0,\frac{2}{\delta}]}[\lambda_k-\beta\eta_c^k]$
end for

Multi-Level Monte Carlo with ergodic chain Z_t

$$g_{\mathrm{MLMC}}=g^0+\begin{cases}2^Q\left(g^Q-g^{Q-1}\right), & \text{if } 2^Q\leq T_{\max}\\ 0, & \text{otherwise}\end{cases}$$

$$\text{where } g^j=2^{-j}\sum_{t=1}^{2^j}\nabla F(x,Z_t)$$

Results

Algorithm	Global Convergence	Violation	Mixing Time	Model-free	Setting
Algorithm 1 in [4]	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	Unknown	No	Tabular
Algorithm 3 in [4]	$\tilde{\mathcal{O}}\left(1/T^{1/3}\right)$	$\tilde{\mathcal{O}}\left(1/T^{1/3}\right)$	Known	No	Tabular
UC-CURL and PS-CURL [1]	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	0	Known	No	Tabular
Algorithm 2 in [5]	$\tilde{\mathcal{O}}\left(1/T^{1/4}\right)$	$\tilde{\mathcal{O}}\left(1/T^{1/4}\right)$	Known	-	Linear MDP
Algorithm 3 in [5]	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	Unknown	-	Linear MDP
Triple-QA [6]	$\tilde{\mathcal{O}}\left(1/T^{1/6}\right)$	0	Known	Yes	Tabular
Algorithm 1 in [3]	$\tilde{\mathcal{O}}\left(1/T^{1/5}\right)$	$\tilde{\mathcal{O}}\left(1/T^{1/5}\right)$	Unknown	Yes	General
This paper (Theorem 4.9)	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	Unknown	Yes	General
This paper (Theorem 4.10)	$\tilde{\mathcal{O}}\left(1/T^{0.5-\epsilon}\right)$	$\tilde{\mathcal{O}}\left(1/T^{0.5-\epsilon}\right)$	Known	Yes	General
Lower bound [2]	$\Omega\left(1/\sqrt{T}\right)$	N/A	N/A	N/A	N/A

Table 1: This table summarizes the different model-based and model-free state-of-the-art algorithms available in the literature for average reward CMDPs and their results for global convergence rate and average constraint violation. The bounds of global convergence describe convergence to the best performance permitted by the chosen features and policy class; the residual approximation floor is fixed (independent of T), and the listed rates govern how fast the statistical gap decays toward that floor. General parameterization refers to parameterizations whose policy score $\nabla_\theta \log\pi_\theta(a|s)$ is uniformly bounded and Lipschitz in θ, together with a Fisher non-degeneracy condition.

Reference

[1] Concave utility reinforcement learning with zero-constraint violations, TMLR 2022

[2] Near-optimal regret bounds for reinforcement learning, NeurIPS 2008

[3] Learning general parameterized policies for infinite horizon average reward constrained MDPs via primal-dual policy gradient algorithm, NeurIPS 2024

[4] Learning infinite-horizon average-reward Markov decision process with constraint, ICML 2022

[5] Achieving sub-linear regret in infinite horizon average reward constrained MDP with linear function approximation, ICLR 2023

[6] A provably-efficient model-free algorithm for infinite-horizon average-reward constrained Markov decision processes, AAAI 2022