Learning General Parameterized Policies for Infinite Horizon Average Reward Constrained MDPs via Primal-Dual Policy Gradient Algorithm

Qinbo Bai[1], *Washim Uddin Mondal*[2], *Vaneet Aggarwal*[1]

1 Purdue University, 2 Indian Institute of Technology Kanpur

NEURAL INFORMATION PROCESSING SYSTEMS

Formulation: Gain Function

- CMDP(S, A, P, r, c, ρ), Gain function (long-term average reward)

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

- We consider a parametrized class of policies π_θ where $\theta \in \Theta \subset R^d$
- The original problem is defined as

$$\max_{\theta\in\Theta} J_r^{\pi_\theta} \text{ s.t. } J_c^{\pi_\theta} \geq 0$$

- Regret and violation are defined as

$$\mathrm{Reg}_T(\mathbb{A}, \mathcal{M}) \triangleq \sum_{t=0}^{T-1}\left(J_r^{\pi^*} - r(s_t, a_t)\right) \quad \mathrm{Vio}_T(\mathbb{A}, \mathcal{M}) \triangleq -\sum_{t=0}^{T-1} c(s_t, a_t)$$

Formulation: Ergodicity

- Assumption 1: The MDP is ergodic.
- Unique stationary distribution

$$d^{\pi_\theta}(s) = \lim_{T\to\infty} \frac{1}{T}\left[\sum_{t=0}^{T-1} \Pr(s_t = s | s_0 \sim \rho, \pi_\theta)\right]$$

- Define the mixing time

Definition 1. *The mixing time of an MDP $\mathcal{M}$ with respect to a policy parameter θ is defined as,*

$$t_{\text{mix}}^{\theta} \triangleq \min\left\{t \geq 1 \middle| \left\|(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}\right\| \leq \frac{1}{4}, \forall s \in \mathcal{S}\right\}$$

- One key property of mixing time under ergodic assumption is

$$\left\|(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}\right\| \leq 2 \cdot 2^{-\frac{t}{t_{mix}}}$$

Method

- Define the Lagrange function $J_L(\theta, \lambda) = J_r(\theta) + \lambda J_c(\theta)$
- The problem becomes a saddle-point problem

$$\max_{\theta \in \Theta} \min_{\lambda \geq 0} \; J_{\mathrm{L}}(\theta, \lambda)$$

- Use primal-dual method with policy gradient

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J_{\mathrm{L}}(\theta_k, \lambda_k), \;\; \lambda_{k+1} = \mathcal{P}_{[0, \frac{2}{\delta}]}[\lambda_k - \beta J_c(\theta_k)]$$

- It is well known that the gradient can be written as

$$\nabla_\theta J_{\mathrm{L}}(\theta, \lambda) = \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(s)} \left[A^{\pi_\theta}_{\mathrm{L}, \lambda}(s, a) \nabla_\theta \log \pi_\theta(a|s) \right]$$

Algorithm

Algorithm 1 Primal-Dual Parameterized Policy Gradient

Input: Episode length H, learning rates α, β, initial parameters θ_1, λ_1, initial state $s_0 \sim \rho(\cdot)$,

$K = T/H$

for $k \in \{1, \cdots, K\}$ **do**

  $\mathcal{T}_k \leftarrow \phi$

  for $t \in \{(k-1)H, \cdots, kH-1\}$ **do**

    Execute $a_t \sim \pi_{\theta_k}(\cdot|s_t)$

    Observe $r(s_t, a_t)$, $c(s_t, a_t)$ and s_{t+1}

    $\mathcal{T}_k \leftarrow \mathcal{T}_k \cup \{(s_t, a_t)\}$

  end for

  for $t \in \{(k-1)H, \cdots, kH-1\}$ **do**

    Obtain $\hat{A}^{\pi_{\theta_k}}_{\mathrm{L},\lambda_k}(s_t, a_t)$ via Algorithm $\boxed{2}$ and $\mathcal{T}_k$

  end for

  Compute ω_k using $(\boxed{15})$ $\omega_k \triangleq \hat{\nabla}_\theta J_{\mathrm{L}}(\theta_k, \lambda_k) = \frac{1}{H} \sum_{t=t_k}^{t_{k+1}-1} \hat{A}^{\pi_{\theta_k}}_{\mathrm{L},\lambda_k}(s_t, a_t) \nabla_\theta \log \pi_{\theta_k}(a_t|s_t)$

  Update the parameters:

$$\theta_{k+1} = \theta_k + \alpha \omega_k,$$

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

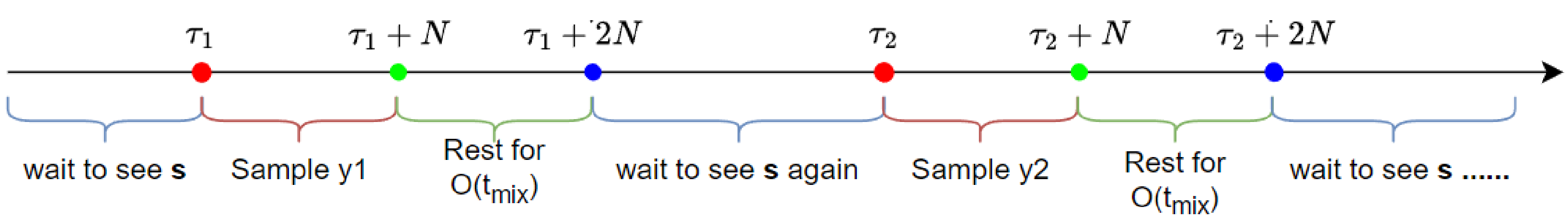
$$\text{where } \hat{J}_c(\theta_k) = \frac{1}{H-N} \sum_{t=(k-1)H+N}^{kH-1} c(s_t, a_t)$$

end for

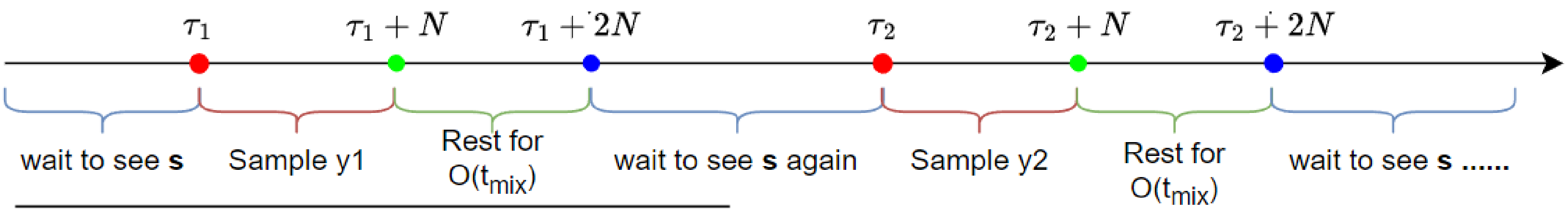
Challenge

$$\nabla_\theta J_{\mathrm{L}}(\theta,\lambda) = \mathbf{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(s)}\left[A^{\pi_\theta}_{\mathrm{L},\lambda}(s,a)\nabla_\theta \log \pi_\theta(a|s)\right]$$

- Challenge: How to get a good estimation of advantage function?
- Discounted setups typically assume access to simulator to obtain unbiased value estimator, while such estimate is needed from samples.
- Since it is single trajectory, unbiased estimator is not straightforward to obtain.
- Solution: Divide each episode into following sub-trajectories



Algorithm



Algorithm 2 Advantage Estimation

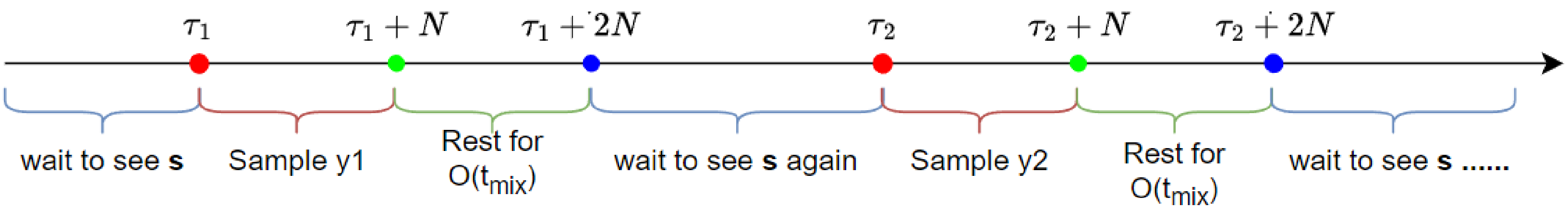
```
Input: Trajectory (s_{t_1}, a_{t_1}, ..., s_{t_2}, a_{t_2}), state s, action
    a, Lagrange multiplier λ, and parameter θ
Initialize: M ← 0, τ ← t_1
Define: N = 4 t_mix log_2 T.
while τ ≤ t_2 − N do
    if s_τ = s then
        M ← M + 1.
        τ_M ← τ
        g_M ← Σ_{t=τ}^{τ+N−1} g(s_t, a_t),  ∀g ∈ {r, c}
        τ ← τ + 2N.
    else
        τ ← τ + 1.
    end if
end while
if M > 0 then
    Q̂_g(s, a) = 1/π_θ(a|s) [ (1/M) Σ_{i=1}^{M} g_i 1(a_{τ_i} = a) ],
    V̂_g(s) = (1/M) Σ_{i=1}^{M} g_i,  ∀g ∈ {r, c}
else
    V̂_g(s) = 0, Q̂_g(s, a) = 0,  ∀g ∈ {r, c}
end if
return (Q̂_r(s, a) − V̂_r(s)) + λ(Q̂_c(s, a) − V̂_c(s))
```

$$\hat{V}^{\pi_{\theta_k}}(s) = \frac{1}{i}\sum_{j=1}^{i} y_j \qquad y_j = \sum_{t=\tau_i}^{\tau_i+N} r(s_t, a_t)$$

$$\hat{Q}^{\pi_\theta}(s,a) = \frac{1}{\pi_{\theta_k}(a|s)}\left[\frac{1}{i}\sum_{j=1}^{i} y_j 1(a_{\tau_j} = a)\right]$$

$$\hat{A}^{\pi_{\theta_k}}(s,a) = \hat{Q}^{\pi_{\theta_k}}(s,a) - \hat{V}^{\pi_{\theta_k}}(s)$$

Alogrithm: bound the variance



- Finally, define $H = O\left(T^{-\xi}\right)$, it can be proved that

$$\mathbf{E}\left[\left(\hat{A}^{\pi_{\theta_k}}_{\mathrm{L},\lambda_k}(s,a) - A^{\pi_{\theta_k}}_{\mathrm{L},\lambda_k}(s,a)\right)^2\right] \leq \mathcal{O}\left(\frac{t_{\mathrm{hit}} N^3 \log T}{\delta^2 H \pi_{\theta_k}(a|s)}\right) = \mathcal{O}\left(\frac{t^2_{\mathrm{mix}} (\log T)^2}{\delta^2 T^{\xi} \pi_{\theta_k}(a|s)}\right)$$

- Using Lemma 2 and Concentration inequality for Markov samples, the gradient estimator is close to the true gradient

$$\mathbf{E}\left[\|\omega_k - \nabla_\theta J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2\right] \leq \tilde{\mathcal{O}}\left(\delta^{-2} A G^2 t^2_{\mathrm{mix}} T^{-\xi}\right)$$

Challenge: Lack of strong duality

- The global convergence of Lagrange function can be bounded as

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(J_{\mathrm{L}}(\pi^*,\lambda_k)-J_{\mathrm{L}}(\theta_k,\lambda_k)\Big)\leq G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\sqrt{\beta}+\frac{\sqrt{A}Gt_{\mathrm{mix}}}{\delta T^{\xi/2}}+\frac{\sqrt{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}}{\delta T^{(1-\xi)/2}}\right)$$
$$+\frac{B}{L}\tilde{\mathcal{O}}\left(\frac{AG^2t_{\mathrm{mix}}^2}{\delta^2T^{\xi}}+\frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}{\delta^2T^{1-\xi}}+\beta\right)+\tilde{\mathcal{O}}\left(\frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}\mathbf{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))]}{T^{1-\xi}\delta}\right)+\sqrt{\epsilon_{\mathrm{bias}}}$$

- However, the strong duality property does not hold under the general parameterization

- The following Lemma serves as an important tool in disentangling the convergence rates of regret and constraint violation

Lemma *Let Assumption* $\boxed{2}$ *hold. For any constant* $C\geq 2\lambda^*$*, if there exists a* $\pi\in\Pi$ *and* $\zeta>0$ *such that* $J_r^{\pi^*}-J_r^{\pi}+C[-J_c^{\pi}]\leq\zeta$*, then*

$$-J_c^{\pi}\leq 2\zeta/C$$

Global convergence

- The sample complexity of reward and violation are

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left(J_r^{\pi^*}-J_r(\theta_k)\right)\leq\sqrt{\epsilon_{\text{bias}}}+\tilde{\mathcal{O}}\left(T^{-\eta/2}+T^{-\xi/2}+T^{-(1-\xi)/2}\right),$$

$$\mathbf{E}\left[\frac{1}{K}\sum_{k=1}^{K}-J_c(\theta_k)\right]\leq\delta\sqrt{\epsilon_{\text{bias}}}+\tilde{\mathcal{O}}\left(T^{-(1-\xi-\eta)}+T^{-\eta/2}+T^{-\xi/2}+T^{-(1-\xi)/2}\right)$$

- The best choice for ξ and η can be solved by

$$\max_{(\eta,\xi)\in(0,1)^2}\min\left\{1-\xi-\eta,\frac{\eta}{2},\frac{\xi}{2},\frac{1-\xi}{2}\right\}$$

- Finally, the regret and violation can be achieved as

$$\mathbf{E}\left[\mathrm{Reg}_T\right]\leq T\sqrt{\epsilon_{\text{bias}}}+\tilde{\mathcal{O}}(T^{4/5})+\mathcal{O}(t_{\text{mix}})$$

$$\mathbf{E}\left[\mathrm{Vio}_T\right]\leq T\delta\sqrt{\epsilon_{\text{bias}}}+\tilde{\mathcal{O}}(T^{4/5})+\mathcal{O}(t_{\text{mix}})$$