Nested Actor-Critic algorithms for Constrained Cooperative Stochastic Games Supplementary Material

1 Convergence Analysis

In this section, we present the convergence analysis of decentralized multi-agent reinforcement algorithms for constrained cooperative stochastic games. We show that the algorithm converges to locally optimal policy which satisfies the constraints. The following are the assumptions required to show the convergence:

Assumption 1. *The underlying Markov chain for any agent i determining the dynamics $\{X_n^{\theta_i}\}$, under stationary random policy π^{θ_i}, is irreducible and aperiodic.*

Assumption 2. $\pi(\boldsymbol{a}|s) = \pi_1(a_1|s).\pi_2(a_2|s)\ldots\pi_n(a_n|s)$.

Assumption 3. *The algorithm uses the unbiased estimate of the value function V^π for parameter update of the policy.*

Assumption 4. *The policy $\pi^{\theta_i}(s,a)$ is continuously differentiable in θ_i.*

Assumption 5. *The step sizes a^n, b^n, c^n for all $n \geq 0$ satisfy the standard stochastic approximation algorithm conditions:*

$$\begin{aligned} &\Sigma_n a_n = \Sigma_n b_n = \Sigma_n c_n = \infty \\ &\Sigma_n (a_n)^2, \Sigma_n (b_n)^2, \Sigma_n (c_n)^2 < \infty \\ &a_n = o(b_n), b = o(c_n). \end{aligned} \tag{1}$$

Remark 1 (Unbiased estimate of V^π)**.** *The unbaised estimate of V^π can be obtained for full state representation and for linear function approximation in single agent RL. However, the "centralized critic" architecture allows the estimate of V^π to be unbiased in multi-agent RL as well.*

We have:

Theorem 1. *Under the Assumptions $1-5$, the sequence of policy parameter updates θ_i^n for an agent i converges to θ_i^* almost surely. Here, θ_i^* is a locally optimal policy parameter and $\pi^{\theta_i^*}$ satisfies the respective constraints.*

Proof. We use the ordinary differential equations (ODE) approach for stochastic approximation (SA) algorithms to prove the asymptotic convergence. The high level overview of the proof technique is as follows:

Step 1 (Two-time scale SA convergence [Borkar, 1997])**.** *We show that the updates of parameters (θ_i, λ_i) converges at different rates to the stationary point $(\theta_i^*, \lambda_i^*)$ almost surely.*

Step 2 (Lyapunov stability)**.** *We show that the Lagrangian $L(\theta_i, \lambda_i)$ is indeed the Lyapunov function in order to prove that the stationary point $(\theta_i^*, \lambda_i^*)$ is locally asymptotically stable.*

Step 3 (Saddle point theorem)**.** *We use the saddle point theorem [Bertsekas, 1999] to conclude that the θ_i^* of the stationary point $(\theta_i^*, \lambda_i^*)$ is the local optima for constraint cooperative stochastic game.*

Remark 2 (Step 1)**.** *We use the Two time scale SA scheme, i.e, θ_i converges faster compared to λ_i due to its faster time scale.*

The parameters which are in slower time scale are invariant to the parameter updates on faster time scale. While analyzing θ_i update the parameter λ_i is fixed. One can construct the ODE corresponding to the respective parameter updates to show that they converges to the stationary point $(\theta_i^*, \lambda_i^*)$.

Remark 3 (Step2)**.** *The Lyapunov functions to show that the iterates are asymptotically stable are as follows:*

$$\mathbb{L}_{\lambda_i}(\theta_i) = L(\theta_i, \lambda_i) - L(\theta_i^*, \lambda_i) \text{ for } \theta_i$$

and

$$\mathbb{L}(\lambda_i) = -L(\theta_i^*, \lambda_i) + L(\theta_i^*, \lambda_i^*) \text{ for } \lambda_i$$

Here, θ_i^* is local minimum where as λ_i^* is local maximum point.

We show that the sequence $\{\theta_i^n\}$ converges to the local minimum of $L(\theta_i, \lambda_i)$ for a fixed λ_i. Also, we show that λ_i sequence converges to local maximum point.

Remark 4 (Step3)**.** *We show that $(\theta_i^*, \lambda_i^*)$ is a local saddle point of the Lagrangian $L(\theta_i, \lambda_i)$. By saddle point theorem θ_i^* is a local optimal solution for constrained cooperative stochastic game.*

□

References

[Bertsekas, 1999] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[Borkar, 1997] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.