Decentralized Policy Gradient Method for Mean-Field Linear Quadratic Regulator with Global Convergence

Lewis Liu[1] Zhuoran Yang[2] Yuchen Lu[1] Zhaoran Wang[3]

1 Mila & University of Montreal
2 Princeton University
3 Northwestern University

July, 2020

Background and Contribution

Background and Challenges

- The complicated correlations in multi-agent systems.
- Large multi-agent systems result in an exponential growth of the capacity of the joint action space with the number of agents.
- The central controller is usually costly to install in practice.

Our Solution

- Mean field approximation: each agent has the same reward function and state transition function, which depends on the rest of the agents only through their aggregated effect.
- A novel decentralized algorithm (MF-DPGM) to effectively learn the optimal policy for mean-field MARL.
- A global convergence guarantee for MF-DPGM under mild assumptions and initial simulation results.

Problem Formulation and Algorithm

Problem Formulation

$$\text{minimize } C(\Theta) = \mathbb{E}_{x_0, w}\left[\sum_{t=0}^{\infty} \gamma^t c_t\right] \tag{1.1}$$

$$\begin{aligned} \text{s.t. } c_t &= \sum_{i=1}^{n} {x_t^{(i)}}^\top Q x_t^{(i)} + {u_t^{(i)}}^\top R u_t^{(i)} + \bar{x}_t^\top \bar{Q} \bar{x}_t, \ \bar{x}_t = 1/n \sum_{i=1}^{n} x_t^{(i)}, \\ x_{t+1}^{(i)} &= A x_t^{(i)} + B u_t^{(i)} + \bar{A}\bar{x}_t + w_t^{(i)}, \ x_0^{(i)} \sim \mathcal{D} \text{ for each } i \in [n]. \end{aligned} \tag{1.2}$$

Reparameterization

$$\begin{aligned} u_t^{(i)} = K x_t^{(i)} + L \bar{x}_t &= M(x_t^{(i)} - \bar{x}_t) + N \bar{x}_t \\ &\triangleq M y_t^{(i)} + N \bar{y}_t. \end{aligned} \tag{1.3}$$

for *path* p = *1 to* n_p **do**
  for t = *1 to* T **do**
    $u_t^{(i)} = M_k^{(i)} y_t^{(i)} + N_k^{(i)} \bar{y}_t$;
    $x_{t+1}^{(i)} \leftarrow A x_t^{(i)} + B u_t^{(i)} + \bar{A}\bar{x}_t + w_t^{(i)}$, for all $i \in [n]$ in paralell;
    $\widehat{Q}_{i,p}^{\pi}(x_t^{(i)}, u_t^{(i)}) \leftarrow \sum_{s=t}^{T} \gamma^{s-t} c_s^{(i)}$;
  end
end
Compute $\widehat{\nabla} C(\widetilde{\Theta}_k^{(i)}) \leftarrow 1/n_p \sum_{p=1}^{n_p} \sum_{t=1}^{T} \widehat{Q}_{i,p}^{\pi}(x_t^{(i)}, u_t^{(i)}) \cdot \nabla \log \pi_{\widehat{\Theta}^{(i)}}(u_t^{(i)} | x_t^{(i)})$, for all $i \in [n]$

(a) Policy running for estimating gradients of costs.

Communication and update: For all $i \in [n]$,

$$\widetilde{\Theta}_{k+1}^{(i)} \leftarrow \widetilde{\Theta}_k^{(i)} - \frac{1}{(\eta_i^{\theta})^2} \left(\frac{1}{n} \left(\nabla\widehat{C}(\widetilde{\Theta}_k^{(i)}) - \nabla\widehat{C}(\widetilde{\Theta}_{k-1}^{(i)}) \right) \right.$$
$$\left. - 2 \sum_{j:j\sim i} (\sigma_{ij}^{\theta})^2 \widetilde{\Theta}_k^{(j)} + (\gamma_i^{\theta})^2 \left(\widetilde{\Theta}_{k-1}^{(i)} - \widetilde{\Theta}_k^{(i)} \right) + \sum_{j:j\sim i} (\sigma_{ij}^{\theta})^2 \left(\widetilde{\Theta}_{k-1}^{(j)} + \widetilde{\Theta}_{k-1}^{(i)} \right) \right)$$

(b) Updating policies with neighborhood information.

Theoretical Results

Assumptions

The parameters of MF-DPGM are chosen to satisfy for any $k \geq 1$:

$$\frac{1}{2}\left(\Omega_\theta + \Gamma_\theta^2\right) \succeq \frac{(2c+1)\Phi_\theta}{n} + \frac{4\kappa}{n^2}\Phi_\theta \Gamma_\theta^{-2} \Phi_\theta, \tag{1.4}$$

$$c = \max\{1, 6\kappa\}, \quad \Gamma_\theta^2 \succeq \Phi_\theta \Gamma_\theta^{-2} \Phi_\theta / n^2, \tag{1.5}$$

where $\kappa = 1/\lambda_{\min}\left(\Omega F H^{-1} F^T \Omega\right)$ is a constant of the network.

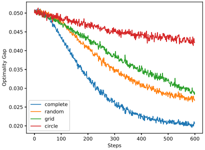
Main theorem

Given assumptions above, for time-step t MF-DPGM gives

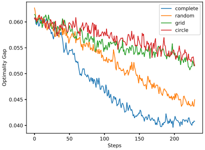
$$\min_{s\in[t]} \left| \frac{1}{n}\sum_{i=1}^{n} C(\widetilde{\Theta}_s^{(i)}) - \widetilde{C}(\widetilde{\Theta}^*) \right| + \|\Omega \widetilde{\mathcal{L}}\{\widetilde{\Theta}_s\}_{(1)}\|^2$$

$$\leq \underbrace{\frac{8\alpha_g \mathcal{C}\mathcal{C}'}{t}}_{\text{cost error bound}} + \underbrace{\frac{20\mathcal{C}'}{t}}_{\text{consensus error bound}} = \frac{4\mathcal{C}'}{t}(5 + 2\alpha_g \mathcal{C}), \tag{1.6}$$

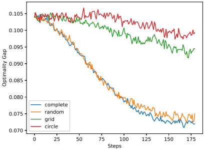
Some Empirical Results



(c) 3-dimension

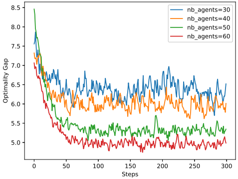


(d) 4-dimension

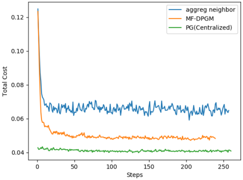


(e) 5-dimension

Figure: Simulation results (convergence curves) on complete (blue), random (orange), grid (green) and circle (red) networks, with different $d = 3, 4, 5$.



(a) Curves for different population sizes.



(b) Comparison on a circle.