ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Lê Bằng Giang

ÁP DỤNG MẠNG NƠ-RON ĐỒ THỊ ĐỂ TÌM KIẾM TỐI ƯU PARETO MẠNH TRONG HỌC TĂNG CƯỜNG ĐA TÁC NHÂN

Ngành: Khoa học máy tính
Chuyên ngành: Khoa học máy tính
Mã số: 8480101.01

TÓM TẮT LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Hà Nội - 2024

Chương 1

Giới thiệu

1.1. Động lực nghiên cứu

Học tăng cường đa tác nhân (MARL) là một nhánh học máy trong đó nhiều tác nhân học cách ra quyết định tối ưu thông qua học tăng cường. Mặc dù MARL đã đạt được những kết quả đáng chú ý trong các trò chơi có tổng bằng không và môi trường hợp tác hoàn toàn, nhiều vấn đề trong thế giới thực liên quan đến các tác nhân tự lợi với mục tiêu có thể mâu thuẫn. Những tình huống này yêu cầu cân bằng giữa hợp tác và cạnh tranh để đạt được tính tối ưu Pareto, nơi không có chính sách nào là tốt hơn hoàn toàn so với chính sách khác. Các phương pháp MARL hiện tại thường hội tụ về các điểm cân bằng Nash, điều này có thể không phải là tối ưu đối với những vấn đề như vậy. Để giải quyết vấn đề này, luận văn giới thiệu học tập vị tha, trong đó các tác nhân tối ưu hóa không chỉ phần thưởng của chính mình mà còn phần thưởng của những tác nhân khác, kết nối MARL với tối ưu hóa đa mục tiêu (MOO). Thuật toán Xuống Dốc Gradient Đa (MGDA), một phương pháp MOO phổ biến, gặp khó khăn trong việc tìm ra các giải pháp Pareto yếu trong MARL, dẫn đến sự phát triển của MGDA++, một thuật toán cải tiến được chứng minh là đạt được tính tối ưu Pareto mạnh mẽ. Công trình này cũng giới thiệu một điều kiện đủ mới để đạt được tối ưu Pareto mạnh mẽ với các mục tiêu lồi. Hơn nữa, những thách thức như quan sát không đầy đủ, không ổn định và sự thiếu phối hợp được giải quyết bằng cách kết hợp mạng nơ-ron đồ thị (GNNs), giúp tăng cường việc tập hợp thông tin giữa các tác nhân, dẫn đến hiệu suất tốt hơn trong các thiết lập hợp tác.

1.2. Đóng góp của luận văn

Các đóng góp chính của luận văn là:

- **Kết nối MOO với MARL**.
- **Thuật toán MOO mới để tìm các giải pháp Pareto mạnh**.
- **Áp dụng GNN để xử lý sự tương tác của các tác nhân**.
- **Thử nghiệm thực nghiệm trên hai bộ dữ liệu chuẩn**.

1.3. Cấu trúc luận văn

Bố cục của luận văn bao gồm các phần: ChatGPT

- Chương 1: Giới thiệu.
- Chương 2: Cơ sở lý thuyết và các công trình liên quan.
- Chương 3: Phương pháp mới để tìm các chính sách Pareto mạnh.
- Chương 4: Thử nghiệm.
- Chương 5: Kết luận.

Chương 2

Cơ sở lý thuyết và các công trình liên quan

2.1. Học Tăng Cường Đa Tác Nhân

Trò chơi Markov. Học Tăng Cường Đa Tác Nhân (MARL) được mô hình hóa dưới dạng một trò chơi Markov, được định nghĩa bởi $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, r, \gamma, \rho \rangle$, trong đó $\mathcal{N}$ là tập hợp các tác nhân N, $\mathcal{S}$ là không gian trạng thái, và $\mathcal{A} = (\mathcal{A}_1 \times \cdots \times \mathcal{A}_N)$ là không gian hành động chung. ρ là phân phối trạng thái ban đầu, và $\gamma \in [0,1)$ là hệ số giảm giá. Mỗi tác nhân i có một hàm phần thưởng $r^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Lợi ích kỳ vọng được định nghĩa là $J_i(\pi) = \mathbb{E}_\pi \sum_t \gamma^t r_t^i$. Khác với các môi trường hợp tác, công trình này tập trung vào việc tìm các giải pháp tối ưu Pareto, được định nghĩa sau.

Khái niệm tối ưu trong MARL. Các khái niệm chính bao gồm tính tối ưu Pareto (PO), trong đó không có tác nhân nào cải thiện mà không làm tổn hại đến những tác nhân khác, và Cân bằng Nash (NE), trong đó không có tác nhân nào có động cơ thay đổi chiến lược của mình.

Môi trường hợp tác hoàn toàn vs. môi trường hợp tác. Trong môi trường hợp tác hoàn toàn, các tác nhân sử dụng tổng hợp phần thưởng, chia sẻ một phần thưởng nhóm duy nhất, điều này dẫn đến các thách thức trong việc phân bổ tín dụng. Trong môi trường hợp tác, các tác nhân tối ưu hóa phần thưởng của riêng mình, thường thông qua học độc lập (ví dụ, IQL, IPPO, MADDPG).

Cả hai môi trường đều đối mặt với các vấn đề như hội tụ về NE không tối ưu.

MAPPO. MAPPO điều chỉnh PPO cho môi trường hợp tác hoàn toàn, sử dụng chính sách tập trung π và hàm giá trị V. Nó sử dụng cơ chế cắt vùng tin cậy:

$$L^{\text{MAPPO}}_{\pi_{\theta_{\text{old}}}}(\pi_\theta) = \sum_i^n \mathbb{E}_{s,\mathbf{a}\sim\pi_{\theta_{\text{old}}}}\left[\min\left(\frac{\pi_\theta(\mathbf{a}^i|s)}{\pi_{\theta_{\text{old}}}(\mathbf{a}^i|s)}A_{\pi_{\theta_{\text{old}}}}(s,\mathbf{a}), \text{clip}\left(\frac{\pi_\theta(\mathbf{a}^i|s)}{\pi_{\theta_{\text{old}}}(\mathbf{a}^i|s)}, 1\pm\varepsilon\right)A_{\pi_{\theta_{\text{old}}}}(s,\mathbf{a})\right)\right] \tag{2.1}$$

trong đó $A_{\pi_{\theta_{\text{old}}}}$ là hàm lợi thế.

IQL. Q-Learning Độc lập mở rộng DQN cho các miền đa tác nhân, tối ưu hóa hàm Q của mỗi tác nhân:

$$L_{\text{TD}} = \mathbb{E}\left[\left(r+\gamma\text{sg}\left(\max_{a'}Q_{\theta'}(s',a')\right)-Q_\theta(s,a)\right)^2\right], \tag{2.2}$$

trong đó $\text{sg}(\cdot)$ dừng đạo hàm, và θ' là tham số của mạng mục tiêu.

IPPO. IPPO phân tán PPO cho các quan sát cục bộ, đạt được huấn luyện và thực thi phân tán. Mặc dù có những lo ngại lý thuyết, các vùng tin cậy tập trung giúp giảm sự bất ổn.

2.2. Tối Ưu Hóa Đa Mục Tiêu

Kí hiệu. Định nghĩa $[n] = 1,\ldots,n$, chuẩn $|\cdot|$ là chuẩn L2, và $\langle\cdot,\cdot\rangle$ là tích vô hướng. Đối với các vectơ x, y, x^i là phần tử thứ i, với $x \preceq y$ và $x \prec y$ chỉ sự không đều về mặt phần tử.

Thiết lập bài toán và định nghĩa. Xem xét việc tối thiểu hóa $F : \mathbb{R}^m \to \mathbb{R}^n$:

$$\min_{x\in\mathbb{R}^m} F(x) = [F_1(x), F_2(x), \ldots, F_n(x)]. \tag{2.3}$$

Xuống dốc gradient nhanh nhất. Phương pháp này tìm vectơ cập nhật d sao cho giảm thiểu:

$$\min_d \max_{i=1,\ldots,n} \langle\nabla F_i(x), d\rangle + \frac{1}{2}\|d\|^2. \tag{2.4}$$

Biểu thức này có thể được viết lại như sau:

$$\begin{aligned} &\text{minimize} && \alpha + \frac{1}{2}\|d\|^2 \\ &\text{s.t.} && \langle\nabla F_i(x), d\rangle - \alpha \le 0, \quad \forall i = 1,\ldots,n. \end{aligned} \tag{2.5}$$

MGDA. Phương trình kép của MGDA là:

$$\begin{array}{ll} \text{minimize} & \left\|\sum_i^n \lambda_i \nabla F_i(x)\right\|^2 \\ \text{s.t.} & \sum_i \lambda_i = 1 \\ & \lambda_i \geq 0 \qquad \forall i = 1, \ldots, n. \end{array} \tag{2.6}$$

2.3. Mạng Nơ-ron Đồ Thị cho MARL

Các tác nhân trong MARL có thể được mô hình hóa như các nút trong một đồ thị động với các láng giềng B_i.

DGN. DGN sử dụng các mạng nơ-ron đồ thị (GNN) tích chập với các kernel dựa trên chú ý:

$$\alpha_{ij}^m = \frac{\exp(\tau\langle W_Q^m h_i, W_K^m h_j\rangle)}{\sum_k^{B_i} \exp(\tau\langle W_Q^m h_i, W_K^m h_k\rangle)}, \tag{2.7}$$

cập nhật các vectơ ẩn:

$$h^{i'} = \text{MLP}\left(\text{concat}m\left[\sum j \in B_i \alpha_{ij}^m W_V^m h_j\right]\right). \tag{2.8}$$

2.4. Công Trình Liên Quan

Học Tăng Cường Đa Tác Nhân (MARL) cho Chính Sách Tối Ưu Pareto. Trong MARL hợp tác hoàn toàn, tính tối ưu Pareto đơn giản hóa thành một hàm giá trị toàn cục duy nhất, khớp với các giải pháp Pareto yếu và mạnh. Các phương pháp chính bao gồm phân tách giá trị và các phương pháp gradient chính sách. Các nghiên cứu ban đầu đã khám phá học độc lập cho các tác nhân, điều này có thể hội tụ về Cân bằng Nash. Các khung huấn luyện tập trung như của Lowe et al. đã cải thiện các nhiệm vụ hợp tác-cạnh tranh. Các nghiên cứu gần đây (ví dụ, Zhao et al.) chỉ ra tính không tối ưu trong các phương pháp MARL và thúc đẩy việc khám phá chính sách tối ưu Pareto. Christianos et al. đã giới thiệu Pareto Actor-Critic, nhưng chi phí tính toán và sự phụ thuộc vào môi trường hợp tác hoàn toàn giới hạn tính khả dụng của nó.

Tối Ưu Hóa Đa Mục Tiêu (MOO)/Học Nhiệm Vụ Đa Dạng. Các phương pháp MOO dựa trên gradient, như MGDA, đã được điều chỉnh để tìm các điểm dừng Pareto, với những cải tiến cho gradient bậc cao. Hầu hết các công trình coi tối ưu Pareto yếu và mạnh là tương đương, mặc dù một số công trình (ví dụ, Roy et

al.) khám phá các điều kiện mà chúng đồng nhất. Các phương pháp tuyến tính hóa và các phương pháp giảm xung đột (ví dụ, Yu et al.) cân bằng các mục tiêu nhưng không khám phá đầy đủ các giải pháp Pareto khác nhau. Thuật toán đề xuất trong luận văn này xây dựng trên MGDA để nâng cao sự hội tụ về Pareto mạnh mẽ trong khi vẫn giữ nguyên việc khám phá mặt cắt Pareto.

Mô Hình Hóa Đồ Thị trong MARL. Mạng Nơ-ron Đồ Thị (GNN) được tận dụng để mô hình hóa sự giao tiếp và tương tác giữa các tác nhân, giải quyết các thách thức như không ổn định và quan sát không đầy đủ. Các công trình gần đây, như CommFormer và GNN phân cấp, đã tiến xa trong việc chia sẻ thông tin và phối hợp giữa các tác nhân. Mặc dù có tiềm năng, các phương pháp này thường được thiết kế riêng cho các nhiệm vụ cụ thể, yêu cầu có kiến thức trước về môi trường. Để giảm độ phức tạp, luận văn này tập trung vào Mạng Nơ-ron Tích chập (CNN) thay vì GNN cho MARL hợp tác, tránh sự phụ thuộc quá sớm vào các phương pháp dựa trên đồ thị.

Chương 3

Phương pháp mới để tìm các chính sách Pareto mạnh

3.1. Những Thách Thức Trong Việc Tìm Tính Tối Ưu Pareto Trong Các Bài Toán MARL

Một thách thức kéo dài trong học tăng cường hợp tác đa tác nhân (MARL) là các thuật toán PG thường hội tụ vào các chính sách bị chi phối bởi Pareto. Mặc dù các phương pháp PG có thể tìm ra các Điểm Cân Bằng Nash (NE) trong các Trò chơi Tiềm năng Markov, chúng không đạt được tính tối ưu Pareto ngay cả khi có sự giao tiếp tùy ý giữa các tác nhân. Ví dụ, trong trò chơi ma trận trong Hình 3.1, bất kỳ cấu hình chính sách nào cũng là một NE, nhưng chỉ có (A,A) là tối ưu Pareto. Ở đây, PG thất bại vì các đạo hàm là bằng không khi phần thưởng của tác nhân không phụ thuộc vào hành động của nó. Hạn chế này cũng áp dụng cho các phương pháp dựa trên giá trị như IQL và MADDPG, trong đó việc tối ưu hóa ích lợi cá nhân dẫn đến các kết quả bị chi phối bởi Pareto. Để giải quyết vấn đề này, các tác nhân phải chuyển từ việc tối ưu

		Player 2	
		A	B
Player 1	A	$1,2$	$0,2$
	B	$1,0$	$0,0$

Hình 3.1: Ví dụ về trò chơi ma trận, bộ giá trị trong mỗi ô bảng chứa phần thưởng của tác nhân 1 và 2, tương ứng.

hóa phần thưởng cá nhân sang việc xem xét mục tiêu của những tác nhân khác trong quá trình huấn luyện.

3.2. MGDA với Tối ưu Hóa Chính Sách Khu Vực Tin Cậy (Trust Region Policy Optimization)

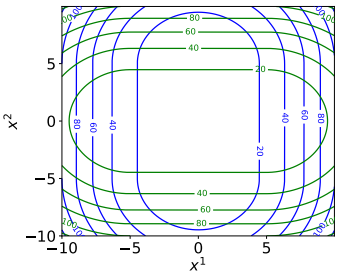
Để đạt được tính tối ưu Pareto trong MARL, tôi kết hợp MAPPO với MGDA. Mục tiêu chính sách cho tác nhân i liên quan đến phần thưởng j được sửa đổi như sau:

$$L_{\text{PPO}}^{i,j} = \min\left(\frac{\pi_{\text{new}}^{i}}{\pi_{\text{old}}^{i}}\hat{A}_{j}^{\pi_{\text{old}}^{i}}, \text{clip}\left(\frac{\pi_{\text{new}}^{i}}{\pi_{\text{old}}^{i}}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{j}^{\pi_{\text{old}}^{i}}\right), \tag{3.1}$$

trong đó $\hat{A}\,j^{\pi^{i}\text{old}}$ là lợi thế ước lượng cho phần thưởng j, với các cơ sở từ một hàm giá trị đa đầu. Các đạo hàm từ phần thưởng của tất cả các tác nhân được truyền vào MGDA (hoặc MGDA++) để tìm hướng giảm, đảm bảo sự cải thiện trên tất cả các mục tiêu.

3.3. Vấn Đề của MGDA

MGDA thường hội tụ vào các điểm tối ưu Pareto yếu, có thể là không tối ưu so với các giải pháp Pareto mạnh. Trong các trường hợp lồi, bất kỳ điểm dừng Pareto nào cũng đều là tối ưu yếu. Hạn chế này càng trở nên rõ rệt bởi ảnh hưởng của các đạo hàm giảm dần từ các mục tiêu đã hội tụ, như được minh họa trong ví dụ hai mục tiêu trong Hình 3.3. Mặc dù việc chuẩn hóa đạo hàm có thể giảm bớt sự hội tụ chậm trong thực nghiệm, nhưng nó thiếu cơ sở lý thuyết. Để giải quyết vấn đề này, tôi đề xuất MGDA++, khắc phục những điểm yếu của MGDA bằng cách đảm bảo sự hội tụ Pareto mạnh, ngay cả trong các môi trường MARL phức tạp. Mã giả (pseudocode) được trình bày trong 1.



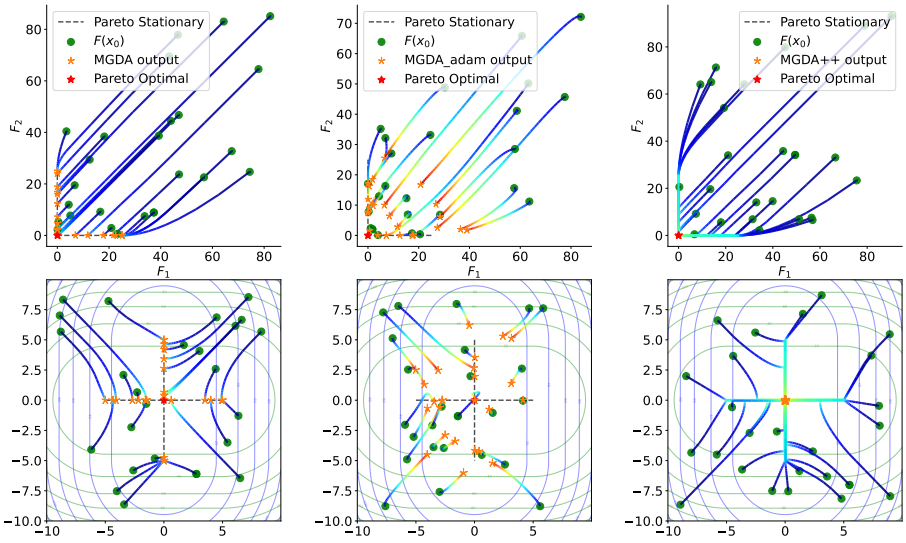
Hình 3.2: Objective landscape in 2D space.

Chúng tôi đưa ra các kết quả lý thuyết của thuật toán được đề xuất.

Algorithm 1 Thuật toán MGDA++

Input: $\varepsilon > 0$, nghiệm ban đầu x_0

for $k = 0, 1, \ldots$ **do**
  $S_k \leftarrow \emptyset$
  for $i = 1, \ldots, n$ **do**
    Tính toán $\nabla F_i(x_k)$
    if $\|\nabla F_i(x_k)\| > \varepsilon$ **then**
      $S_k = S_k \cup \{i\}$
    end if
  end for
  if $S_k = \emptyset$ **then**
    Dừng
  end if
  Tìm $\{\lambda_i\}_{i \in S_k}$ bằng cách giải phương trình (2.6) trên tập con các gradient $\{\nabla F_i(x_k)\}_{i \in S_k}$
  $d_k \leftarrow \sum_{i \in S_k} \lambda_i \nabla F_i(x_k)$
  Chọn bước nhảy t_k
  $x_{k+1} \leftarrow x_k - t_k d_k$
end for



(a) MGDA (b) MGDA w/ Adam (c) MGDA++

Hình 3.3: So sánh MGDA, MGDA với Adam và MGDA++. Bên trái: MGDA bị mắc kẹt ở các điểm Pareto Stationary, nơi quá trình học hoàn toàn dừng lại. Ở giữa: Việc thay đổi bộ tối ưu hóa không giúp tránh được sự hội tụ không tối ưu. Bên phải: MGDA++ có khả năng hội tụ vào các giải pháp Pareto Tối Ưu mạnh trong khi tránh bị mắc kẹt ở các điểm Pareto Stationary.

Giả thuyết 1. *Tất cả các hàm mục tiêu F_i đều là lồi và L-mượt.*

Giả thuyết 2. *Với một điểm cho trước x_0, tổng của các hàm mục tiêu F có một tập mức giới hạn bị chặn $\Gamma = \{x | \sum_i F_i(x) \leq \sum_i F_i(x_0)\}$. Hơn nữa, tập tối ưu của mỗi hàm mục tiêu F_i là không rỗng và được ký hiệu là X_i^*, giá trị tối ưu được ký hiệu tương tự là F_i^*.*

Mệnh đề 1. *Dưới giả thuyết 1, nếu tồn tại một tổ hợp lồi của tập con các vector gradient không bằng không*

$$\sum_{i \in S} \lambda_i \nabla F_i(x) = 0; \quad \lambda_i > 0, \quad |\nabla F_i(x)| > 0 \quad \forall i \in S \tag{3.2}$$

với $S \subseteq [n], S \neq \emptyset$, thì x là tối ưu Pareto.

Mệnh đề 2. *Dưới các giả thuyết 1 và 2, với mọi $\varepsilon > 0$, có thể chọn một ε sao cho $0 < \varepsilon \leq \sqrt{2L\varepsilon}$ sao cho nếu x thỏa mãn $|\nabla F_i(x)| < \varepsilon$, thì $F_i(x) \leq F_i^* + \varepsilon, \forall i \in [n]$. Những x như vậy nhất thiết là các giải pháp ε-Tối Ưu Pareto.*

Định lý 3. *Dưới các giả thuyết 1 và 2, với mọi $\varepsilon > 0$, khi $n = 2$ và chọn $\varepsilon < \sqrt{2L\varepsilon}$ và với các lựa chọn thích hợp cho các bước cập nhật t_k như sau:*

$$t_k = \begin{cases} \max\left(\dfrac{|S_k|\|d_k\|^2 + \left\langle \sum_{i\in\overline{S}_k} \nabla F_i(x_k), d_k \right\rangle}{nL\|d_k\|^2}, 0\right) & \text{if } \|d_k\| > 0 \\ 0 & \text{if } \|d_k\| = 0 \end{cases}, \tag{3.3}$$

với $\overline{S}_k$ là bổ sung của S_k, thì mỗi chuỗi con hội tụ của MGDA++ sẽ hội tụ vào các giải pháp Pareto tối ưu hoặc ε-Pareto tối ưu.

3.4. Áp Dụng Mạng Nơ-ron Đồ Thị (Graph Neural Networks) Để Xử Lý Tương Tác Giữa Các Tác Nhân

Tiếp theo, tôi sẽ xem xét việc áp dụng Mạng Lưới Nơ-ron Đồ Thị (GCN) vào thuật toán của mình trong Mục 3.2.. Cụ thể, tôi xây dựng việc triển khai GCN cho MARL dựa trên DGN (Jiang et al., 2018). Các quan sát của mỗi tác nhân đầu tiên được mã hóa bởi một bộ mã hóa đặc trưng f_e và sau đó được đưa vào lớp tích chập, là một mạng multi-head attention.

Với mỗi đầu attention, tôi chiếu mã hóa đặc trưng của lớp trước vào các vector truy vấn (query), khóa (key), và giá trị (value). Vector giá trị sau đó là tổng có trọng số của các vector giá trị, trong đó trọng số được tính bằng phân phối softmax của tích vô hướng giữa các vector khóa và truy vấn tương ứng giữa các tác nhân láng giềng trong đồ thị kề. Đầu ra của lớp tích chập là sự kết hợp của tất cả các vector giá trị đã được chú ý, được chiếu vào một không gian khác với chiều thấp hơn bằng một MLP một lớp.

Một số lựa chọn thiết kế, chẳng hạn như kiến trúc mạng, sẽ được trình bày trong phần Hyperparameters. Mã giả của Mạng Lưới Nơ-ron Đồ Thị trong phương pháp của chúng tôi được trình bày chi tiết trong Thuật toán 2.

Ở đây, tôi nhấn mạnh sự khác biệt chính giữa phương pháp Mạng Lưới Nơ-ron Đồ Thị của tôi và DGN: 1) DGN theo cách tiếp cận tối ưu hóa tác nhân đơn lẻ như đã thảo luận trong Mục 3.1., trong khi phương pháp của chúng tôi theo khuôn khổ tối ưu hóa MOO với học tập vị tha, và 2) DGN sử dụng một thuật toán RL off-policy dựa trên giá trị (DQN), trong khi tôi dựa vào phương pháp on-policy, trust region MAPPO (Yu et al., Mục 3.2.).

Algorithm 2 Chính sách Đa tác nhân Dựa trên Mạng Đồ thị

Input: Các quan sát $\{o_i\}_i^N$ của tất cả các tác nhân, số lượng tác nhân N, mạng mã hóa f_e, ma trận trọng số của truy vấn, khóa và giá trị $(W_{Qk}^m, W_{Kk}^m, W_{Vk}^m)$, số lượng lớp Mạng Lưới Đồ thị Convolutional K (tức là số bước nhảy), số lượng đầu chú ý đa M, ma trận Kề B

Output: Các hành động $\{a_i\}_i^N$

Mã hóa các quan sát bằng mạng mã hóa để có được biểu diễn ẩn $z_i = f_e(o_i), \forall i \leq N$
Đặt $h_0^i := z_i, \forall i \leq N$
for $k = 0, \ldots, K-1$ **do**
    for $i = 1, \ldots, N$ **do**
        Tính toán trọng số chú ý $\{\alpha_{ij}^m\}_j^{B_i}$ theo công thức Eq (2.7) $\forall m \leq M$
        Tính toán các vector giá trị $V_i^m = \sum_j^{B_i} a_{ij}^m W_{Vk}^m h_k^j, \quad \forall m \leq M$
        Tính toán $h_{k+1}^i = MLP(\text{concat}[V_i^m, \forall m \leq M])$ (Eq 2.8).
    end for
end for
for $i = 1, \ldots, N$ **do**
    // Tìm logit (rời rạc) hoặc trung bình và độ lệch chuẩn (Gaussian)
    $h^i = MLP(\text{concat}[h_k^i, k \leq K])$
    // Lấy mẫu từ phân phối (ví dụ: Categorical, Gaussian, hoặc Squashed Gaussian)
    $a_i \sim \text{dist}(h_i)$
end for

Chương 4

Thực nghiệm

Trong chương này, tôi trình bày thiết lập thí nghiệm để đánh giá phương pháp đề xuất so với các phương pháp cơ sở, giải quyết ba câu hỏi: 1) Học tập vị tha có giúp tìm ra các giải pháp Pareto không? 2) MGDA++ có cải thiện tính tối ưu mạnh Pareto so với các phương pháp cơ sở không? 3) Việc kết hợp các mạng nơ-ron đồ thị tác động thế nào đến hiệu suất tổng thể?

4.1. Thiết lập thí nghiệm

4.1.1. Các Môi trường đánh giá

Các phương pháp được đánh giá trên hai chuẩn mực MARL: Gridworld và MPE.

Gridworld. Môi trường Gridworld được lấy cảm hứng từ các thiết lập học tập vị tha (Franzmeyer et al., 2022), nơi các tác nhân di chuyển trong lưới, kiếm được phần thưởng (+10) khi đạt được mục tiêu có màu phù hợp và bị phạt (-0.1) khi va chạm. Các cánh cửa, được mở bởi các tác nhân đứng ở vị trí chìa khóa có màu tương ứng, thêm yếu tố hợp tác vào môi trường. Bốn biến thể của môi trường được nghiên cứu (Hình 4.1):

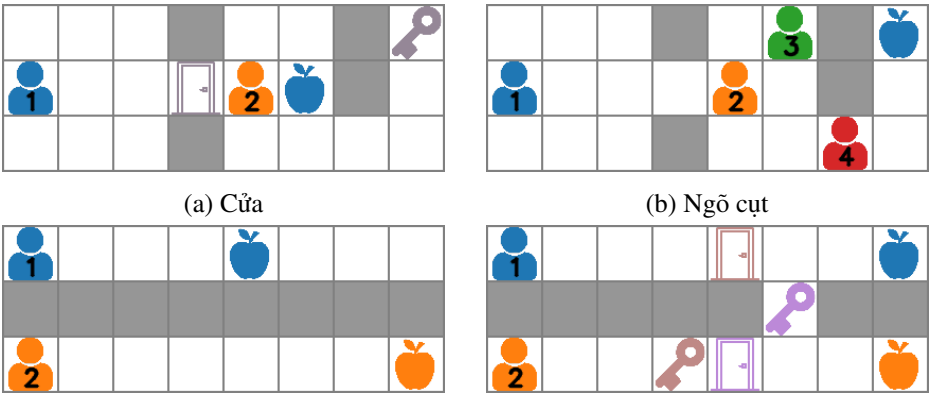
a) **Cửa (Door)** (Franzmeyer et al., 2022): Tác nhân thứ hai mở cửa cho tác nhân đầu tiên, cho phép các hành vi vị tha thành công, trong khi các tác nhân ích kỷ thất bại.

b) **Ngõ cụt (Dead End)** (Franzmeyer et al., 2022): Các động lực MARL phức

tạp xuất hiện khi các tác nhân chặn các con đường, minh họa sự can thiệp trong các bài toán tối ưu hóa.

c) **Hai hành lang (Two Corridors)**: Các tác nhân độc lập đối mặt với độ khó nhiệm vụ bất đối xứng, dẫn đến các giải pháp Pareto yếu do tốc độ hội tụ khác nhau.

d) **Hai phòng (Two Rooms)**: Hợp tác ngầm xảy ra khi một tác nhân có thể mở cửa cho tác nhân kia, cân bằng giữa vị tha và lợi ích cá nhân mà không cần các động cơ rõ ràng.



(a) Cửa (b) Ngõ cụt

(c) Hai hành lang (d) Hai phòng

Hình 4.1: Bốn kịch bản của môi trường Gridworld.

Môi trường nhiều hạt tác tử (MPE) (Lowe et al., 2017). MPE bao gồm các kịch bản với các cấu trúc phần thưởng khác nhau, từ cạnh tranh đến hợp tác. Các tác nhân được biểu diễn dưới dạng các hạt trong mặt phẳng 2D có thể di chuyển, giao tiếp và tương tác. Để đánh giá, tôi tập trung vào các môi trường hợp tác (Yu et al., 2022; Zhong et al., 2024), cụ thể là các kịch bản Spread và Reference, loại trừ kịch bản Speaker vì nó tập trung vào các tác nhân không đồng nhất (Zhong et al., 2024).

1. **Reference**. Hai tác nhân và ba cột mốc màu sắc. Mỗi tác nhân phải đạt được cột mốc mục tiêu của mình, chỉ được biết đến bởi tác nhân kia, do đó cần phải có giao tiếp. Cả hai tác nhân có thể nói và lắng nghe cùng một lúc.

2. **Spread**. Ba tác nhân và ba cột mốc. Các tác nhân cố gắng bao phủ tất cả các cột mốc trong khi tránh va chạm, với phần thưởng dựa trên khoảng cách đến tác nhân gần nhất.

4.1.2. Các phương pháp cơ sở

Để đánh giá phương pháp của tôi, tôi thực hiện hai thí nghiệm: (1) đánh giá MGDA++ trong MARL với MAPPO (Yu et al., 2022), IPPO (de Witt et al., 2020), và IQL (Tampuu et al., 2017) là các phương pháp cơ sở, và (2) tích hợp mạng nơ-ron đồ thị vào trong khuôn khổ, so sánh với DGN (Jiang et al., 2018) và các phương pháp cơ sở đã bổ sung đồ thị.

1. **Multi-head MAPPO**. Một mở rộng của MAPPO (Yu et al., 2022) với các mạng giá trị đa đầu, dự đoán phần thưởng cho từng tác nhân trong khi tuân theo CTDE.

2. **IPPO** (de Witt et al., 2020). PPO phân tán, trong đó các tác nhân học độc lập, bỏ qua tính không ổn định do các tác nhân khác gây ra, nhưng vẫn cạnh tranh hiệu quả với các phương pháp tập trung (Sun et al., 2022).

3. **IQL** (Tampuu et al., 2017). Một mở rộng off-policy, dựa trên giá trị của DQN cho môi trường nhiều tác nhân, sử dụng các mạng Q độc lập và bộ nhớ hồi phục cho mỗi tác nhân.

4. **MGPO**. Sử dụng kiến trúc đa đầu của MAPPO nhưng áp dụng MGDA để có một phương hướng tối ưu hóa chung cho tất cả các tác nhân.

5. **MGPO++**. Tương tự MGPO nhưng sử dụng phương pháp MGDA++ cho tối ưu hóa.

Đối với các phương pháp dựa trên đồ thị, tôi mở rộng MAPPO, MGPO, và MGPO++ với Mạng nơ-ron Chập Đồ thị (GCN) dựa trên DGN (Jiang et al., 2018), cho phép các tác nhân tổng hợp các đặc trưng từ các tác nhân láng giềng qua sự chú ý đa đầu.

1. **DGN** (Jiang et al., 2018). Kết hợp DQN (Mnih et al., 2013) với mạng nơ-ron đồ thị, trong đó các nút đại diện cho các tác nhân và các cạnh đại diện cho các kết nối cục bộ. Sự chú ý đa đầu tổng hợp các đặc trưng của các tác nhân láng giềng.

2. **GCN-MAPPO**. Mở rộng MAPPO với GCN để tổng hợp đặc trưng nhận thức về láng giềng, tối ưu hóa phần thưởng cá nhân.
3. **GCN-MGPO**. Tương tự GCN-MAPPO nhưng tối ưu hóa các mục tiêu chung thông qua MGDA.
4. **GCN-MGPO++**. Tương tự GCN-MGPO nhưng sử dụng MGDA++ cho tối ưu hóa.

4.1.3. Cài đặt siêu tham số

Các siêu tham số mặc định từ các bài báo gốc được sử dụng cho tất cả các phương pháp cơ sở. Các MLP ba lớp với kích thước ẩn là 64 được sử dụng cho tất cả các phương pháp. MAPPO, IPPO và IQL theo mặc định của kho lưu trữ; MGPO và MGPO++ nhận các siêu tham số của MAPPO. Trong MGDA++, ε được đặt là 0.05, ngoại trừ trong kịch bản Ngõ cụt nơi nó là 0.1. Các phương pháp được huấn luyện trong 500k bước đối với Gridworld và 5M bước đối với MPE, không chia sẻ tham số trong các mạng tác nhân hay mạng giá trị để so sánh công bằng.

Các phương pháp dựa trên đồ thị sử dụng GCN với kiến trúc từ Jiang et al. (2018). Giả định các đồ thị giao tiếp dày đặc do số lượng tác nhân nhỏ. Để giảm thiểu tính toán, các phương pháp đồ thị sử dụng một lớp chập và một đầu chú ý. Bộ mã hóa là một MLP hai lớp đối với MPE và một CNN ba lớp đối với Gridworld, trong khi các MLP khác chỉ có một lớp với các hàm kích hoạt ReLU.

4.2. Kết quả

4.2.1. Không sử dụng Mạng Nơ-ron Đồ thị (Graph Neural Network)

Trong môi trường Gridworld, phương pháp của chúng tôi vượt trội so với tất cả các phương pháp cơ bản trong các kịch bản khác nhau.

Trong kịch bản Cửa (Door), MAPPO với tối ưu hóa đa mục tiêu là phương pháp duy nhất hội tụ đến giải pháp tối ưu cho tác nhân đầu tiên, trong khi tất cả các thuật toán khác với mục tiêu đơn đều thất bại trong việc học hợp tác.

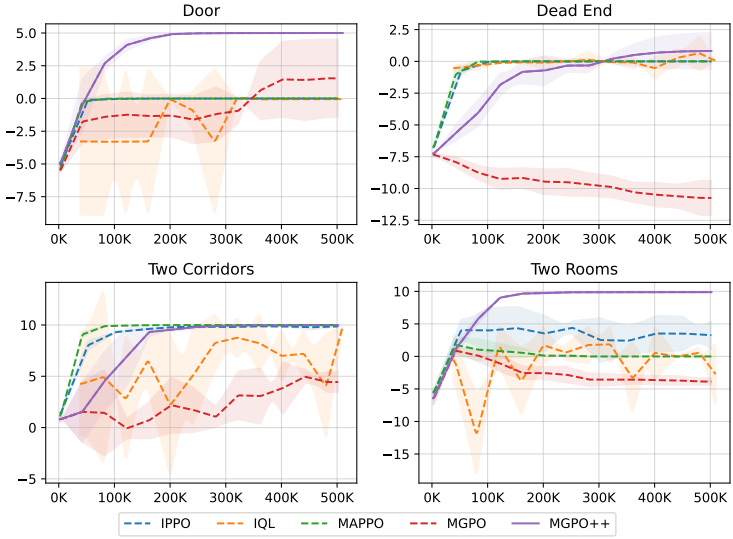
Trong kịch bản Ngõ cụt (Dead End), MGPO không thể học được trừ tác nhân đầu tiên, điều này cho thấy MGDA chỉ có thể học được chính sách tối ưu

cho một tác nhân duy nhất. Trong MGPO++, tất cả các tác nhân có thể học được giải pháp tốt mà không bị cản trở sau khi một số tác nhân hội tụ.

Trong kịch bản Hai hành lang (Two Corridors), tất cả các phương pháp mục tiêu đơn đều hội tụ, nhưng MGPO không hội tụ được đối với tác nhân thứ hai, người có nhiệm vụ khó khăn hơn.

Trong kịch bản Hai phòng (Two Rooms), chỉ MGDA++ đạt được các chính sách tối ưu cho cả hai tác nhân. Mặc dù MGDA có thể tìm được phần thưởng cao cho tác nhân đầu tiên, nhưng tác nhân thứ hai bị kẹt lại sau khi tác nhân đầu tiên hội tụ. MAPPO với hệ số entropy bằng 0 cho phép tác nhân đầu tiên tìm được phần thưởng cao ban đầu, nhưng đây chỉ là kết quả do sự khám phá của tác nhân thứ hai. Khi mức entropy của cả hai tác nhân giảm dần theo thời gian, hiệu suất của tác nhân thứ hai giảm xuống. Hiệu suất của IQL không ổn định, có thể là do sự không ổn định mà tác nhân đầu tiên nhận thấy do sự khám phá của tác nhân thứ hai.

Những phát hiện này cho thấy hiệu quả của phương pháp của chúng tôi trong việc đạt được các chính sách tối ưu trong môi trường đa tác nhân.

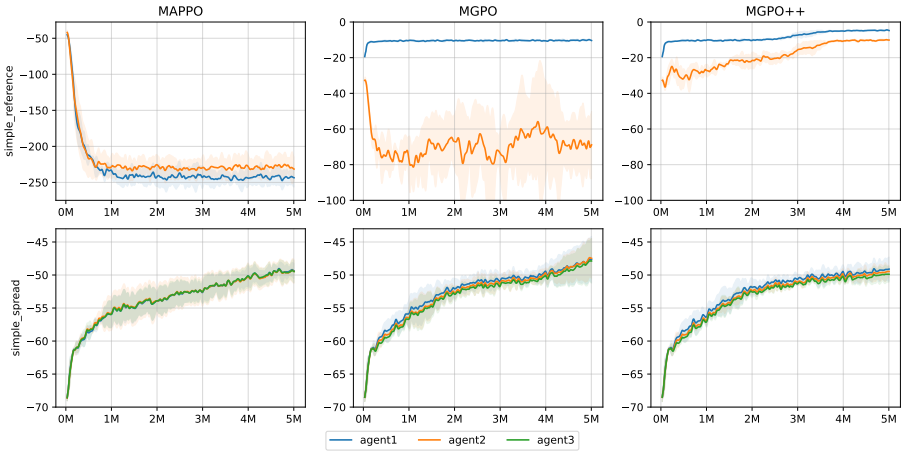


Hình 4.2: Phần thưởng trung bình của tất cả các tác nhân trong bốn kịch bản. Lưu ý rằng, toán tử trung bình có thể được xem là một cách tuyến tính hóa đặc biệt của vector phần thưởng.

MPE. Trong môi trường MPE (Hình 4.3), kết quả cho thấy các điểm sau:

Trong kịch bản Tham chiếu (Reference), MAPPO gặp khó khăn vì các tác nhân cần phải giao tiếp để đạt được mục tiêu, nhưng không được thưởng cho việc giao tiếp. Ban đầu, các tác nhân tìm thấy các điểm mốc mục tiêu một cách tình cờ, nhưng hiệu suất của họ giảm khi chính sách trở nên quyết đoán hơn. MGPO cho thấy tác nhân thứ hai bị "đóng băng"sau khi tác nhân đầu tiên hội tụ, điều này ảnh hưởng xấu đến việc học của tác nhân đầu tiên. Ngược lại, MGPO++ thấy một sự gia tăng nhẹ về phần thưởng cho cả hai tác nhân vào khoảng 3.5 triệu bước thời gian, chỉ ra sự hợp tác cải thiện giữa các tác nhân. Sự gia tăng phần thưởng này không có trong MGPO.

Trong kịch bản Phân tán (Spread), sự khác biệt giữa các phương pháp là rất nhỏ, vì cảnh quan phần thưởng rất hợp tác. Trong môi trường này, cả các giải pháp Pareto mạnh và yếu đều phù hợp, và tất cả các phương pháp dựa trên Gradient Chính sách đều hội tụ với tốc độ giống nhau.



Hình 4.3: Kết quả trong bài kiểm tra MPE, cho thấy hiệu suất của từng tác nhân trong suốt quá trình huấn luyện.

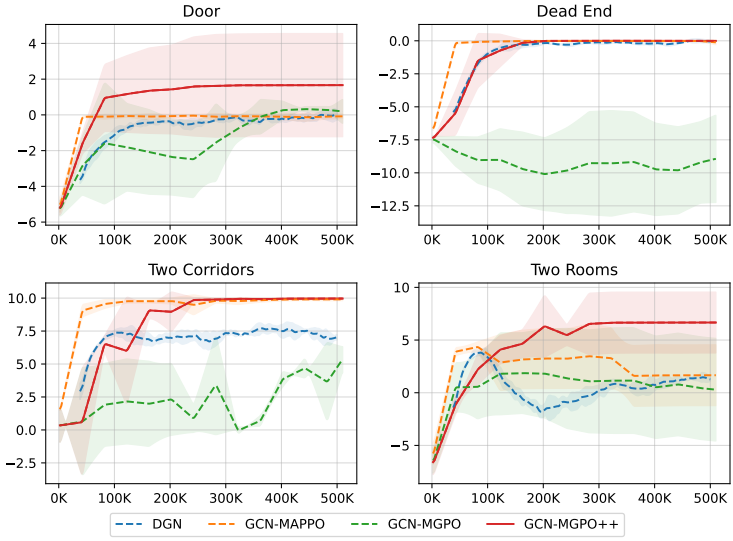
4.2.2. Với Mạng Nơ-ron Đồ thị (Graph Neural Network)

Trong phần này, chúng tôi trình bày kết quả thí nghiệm trên MPE và Gridworld với các phương pháp dựa trên đồ thị.

Gridworld: Trong kịch bản Cửa, các phương pháp học tập vị tha (altruistic learning) thành công trong việc thúc đẩy hợp tác. GCN-MGPO hoạt động tương

tự như trong môi trường không sử dụng đồ thị, chỉ có tác nhân đầu tiên hội tụ, trong khi GCN-MGPO++ hội tụ sau đó. Kịch bản Ngõ cụt là một bài toán khám phá khó khăn, trong đó hầu hết các phương pháp đều thất bại trong việc học chính sách tốt. GCN-MGPO cho thấy một chút hội tụ cho tác nhân đầu tiên, nhưng không cho các tác nhân khác.

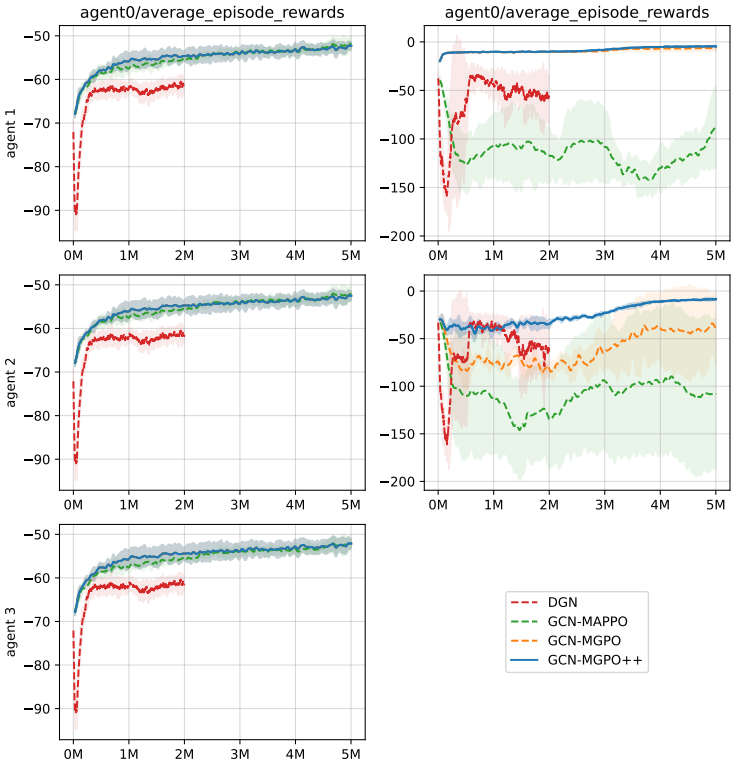
Trong kịch bản Hai hành lang, tất cả các phương pháp đều hội tụ ngoại trừ tác nhân thứ hai trong GCN-MGPO, do bị hội tụ Pareto yếu. DGN hoạt động kém trong môi trường đơn giản và không hợp tác này, cho thấy nó gặp khó khăn trong việc giao tiếp dựa trên đồ thị. GCN-MGPO++ vẫn vượt trội trong kịch bản Hai phòng, với cả hai tác nhân học cách hợp tác. GCN-MGPO hoạt động tốt đối với tác nhân đầu tiên, nhưng gây cản trở cho tác nhân thứ hai, trong khi GCN-MAPPO và DGN thất bại trong kịch bản này.



Hình 4.4: Phần thưởng trung bình của tất cả các tác nhân trong bốn kịch bản với các phương pháp dựa trên đồ thị.

MPE: Trong MPE, kết quả tương tự như phần trước. Trong kịch bản Phân tán, các phương pháp Gradient Chính sách hoạt động tương tự nhau, trong khi DGN tụt lại phía sau. Trong kịch bản Tham chiếu, các mạng nơ-ron đồ thị cải thiện các phương pháp không vị tha như GCN-MAPPO, giúp nó vượt trội so với MAPPO không sử dụng đồ thị. Tuy nhiên, GCN-MGPO và MGPO++ không cải thiện đáng kể, vì chúng giải quyết việc giao tiếp qua tối ưu hóa thay vì chia sẻ thông

tin.



Hình 4.5: Kết quả trên MPE với các phương pháp dựa trên đồ thị.

4.2.3. Tóm tắt Các Phát Hiện Thực Nghiệm

MGDA++ vượt trội so với tất cả các phương pháp cơ bản trong cả Gridworld và MPE, tránh được các chính sách không tối ưu và thúc đẩy hợp tác. Mạng nơ-ron đồ thị mang lại lợi ích hạn chế, với GCN-MAPPO thể hiện sự cải thiện đáng kể trong kịch bản Tham chiếu, nhưng GCN-MGPO++ vẫn là phương pháp vượt trội tổng thể. Hội tụ Pareto yếu vẫn tồn tại trong MGPO. Các công trình tương lai có thể nghiên cứu việc cắt tỉa giao tiếp và các chiến lược khác để giảm thiểu các vấn đề giao tiếp trong MARL.

Kết luận

Trong luận văn này, tôi khám phá việc huấn luyện các tác nhân đa năng trong các tình huống hợp tác bằng cách coi phần thưởng dựa trên vectơ như một nhiệm vụ đa mục tiêu, áp dụng thuật toán MGDA vào MARL. Tuy nhiên, MGDA chuẩn chỉ hội tụ tới tối ưu Pareto yếu. Tôi giới thiệu MGDA++ để cải thiện sự hội tụ bằng cách lọc các gradient mục tiêu có độ norm nhỏ và cung cấp phân tích lý thuyết cho các vấn đề với hai mục tiêu. Tôi cũng áp dụng mạng nơ-ron đồ thị (GNN) để mô hình hóa sự tương tác giữa các tác nhân. Sự kết hợp giữa MGDA++ và GCN được đánh giá trong Gridworld và MPE, cho thấy hiệu suất cải thiện hướng tới tối ưu Pareto mạnh.

Công việc tương lai:

- *Mở rộng cho $n > 2$ nhiệm vụ*: Mở rộng phân tích lý thuyết cho nhiều mục tiêu và đánh giá MGDA++ trên các nhiệm vụ khác như Học nhiều nhiệm vụ và Học tăng cường nhiều nhiệm vụ.
- *Bảng kiểm thử phức tạp hơn*: Kiểm tra MGPO++ trên các nhiệm vụ MARL lớn hơn và phức tạp hơn ngoài Gridworld và MPE.
- *Cải thiện thành phần GNN*: Nâng cao kiến trúc GNN cho các tình huống có sự liên quan giao tiếp hạn chế và kiểm thử các nhiệm vụ MARL tập trung vào giao tiếp.

Danh sách các công bố liên quan

1. Le, Bang Giang, and Viet Cuong Ta. "Toward Finding Strong Pareto Optimal Policies in Multi-Agent Reinforcement Learning."arXiv preprint arXiv:2410.193 (2024). (To be presented at ACML 2024 Journal Track)
2. Le, Bang-Giang, Thi-Linh Hoang, Hai-Dang Kieu, and Viet-Cuong Ta. "Structural and Compact Latent Representation Learning on Sparse Reward Environments."In Asian Conference on Intelligent Information and Database Systems, pp. 40-51. Singapore: Springer Nature Singapore, 2023.
3. Le, Bang Giang, and Viet Cuong Ta. "Distill Knowledge in Multi-task Reinforcement Learning with Optimal-Transport Regularization."In 2022 14th International Conference on Knowledge and Systems Engineering (KSE), pp. 1-6. IEEE, 2022.
4. Le, Bang Giang, and Viet Cuong Ta. "On the Effectiveness of Regularization Methods for Soft Actor-Critic in Discrete-Action Domains"(Accepted with minor revision to IEEE Transactions on Systems, Man and Cybernetics: Systems).
5. Le, Bang Giang, and Viet Cuong Ta. "Low Variance Trust Region Optimization with Independent Actors and Sequential Updates in Cooperative Multi-agent Reinforcement Learning"(Major revision, submited to Journal of Autonomous Agents and Multi-Agent Systems).