Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability

Thomy Phan[1,2], Fabian Ritz[2], Philipp Altmann[2], Maximilian Zorn[2], Jonas Nüßlein[2], Michael Kölle[2], Thomas Gabor[2], Claudia Linnhoff-Popien[2]

[1]University of Southern California,
[2]LMU Munich

Multi-Agent Reinforcement Learning in Dec-POMDPs

Most state-of-the-art MARL approaches assume <u>deterministic observability</u> and focus on

$$Q^*_{MPD}(s,a) = \mathcal{R}(s,a) + \gamma \max_{a'} Q^*_{MPD}(s',a')$$

However, the <u>true optimal value function</u> Q^* in Dec-POMDPs is actually defined by

$$Q^*(\tau,a) = b(s|\tau)\left(\mathcal{R}(s,a) + \gamma \sum_{s'}\sum_{z'} \mathcal{P}(s'|s,a)\boldsymbol{\Omega(z'|a,s')}Q^*(\tau',\pi^*(\tau'))\right)$$

Popular benchmarks do not exhibit <u>stochastic partial observability</u> unlike general Dec-POMDPs

https://github.com/oxwhirl/smac

https://ai.googleblog.com/2019/06/introducing-google-research-football.html

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Multi-Agent Recurrence

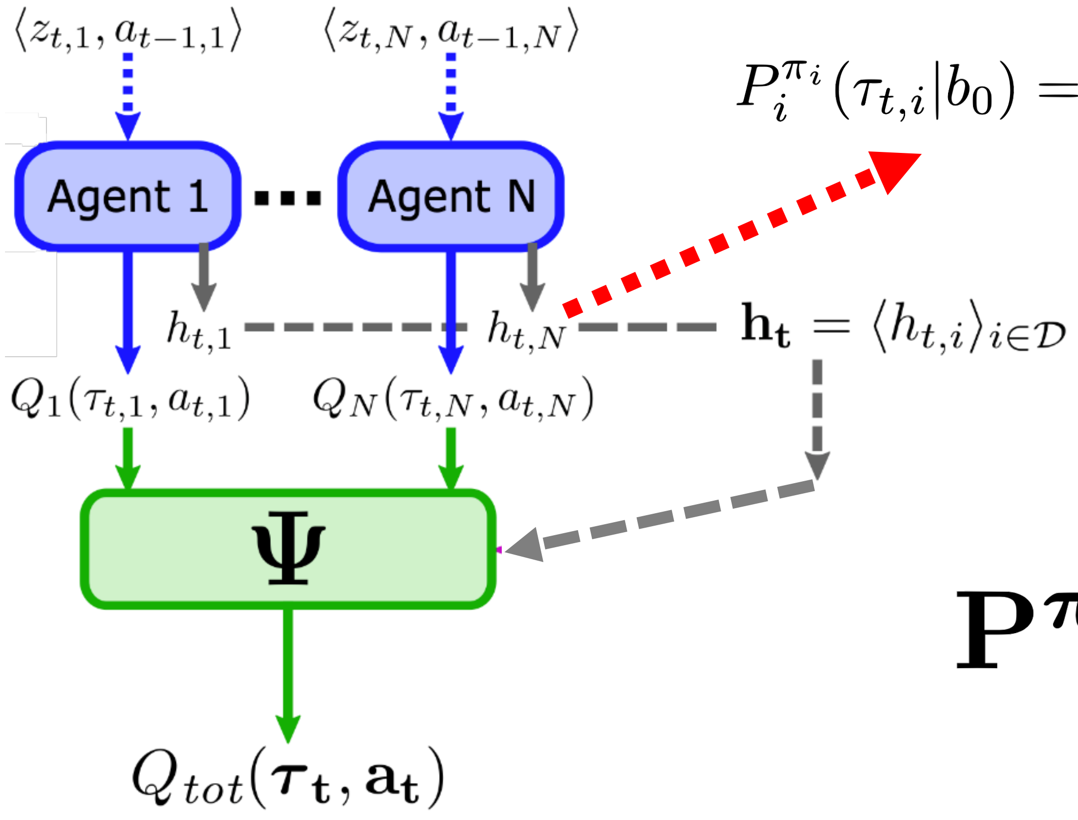
$$\boldsymbol{\pi}^* = argmax_{\boldsymbol{\pi}} \sum_{t=0}^{T-1} \sum_{\boldsymbol{\tau_t} \in (\mathcal{Z}^N \times \mathcal{A})^t} \mathcal{C}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}) \underbrace{\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0)}_{\text{multi-agent recurrence}} Q^*(\cdot)$$

$$\begin{aligned}\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0) &= \mathbf{P}(\mathbf{z_0}|b_0) \prod_{c=1}^{t} \mathbf{P}(\mathbf{z_c}|\boldsymbol{\tau_{c-1}}, \boldsymbol{\pi}) \\ &= \mathbf{P}(\mathbf{z_0}|b_0) \prod_{c=1}^{t} \sum_{s_c \in \mathcal{S}} \sum_{s_{c-1} \in \mathcal{S}} \mathcal{T}(\cdot)\Omega(\cdot)\end{aligned}$$

AERIAL: Attention-based Embeddings of Recurrence In multi-Agent Learning

Value factorization approach considering …

- memory representations of agents



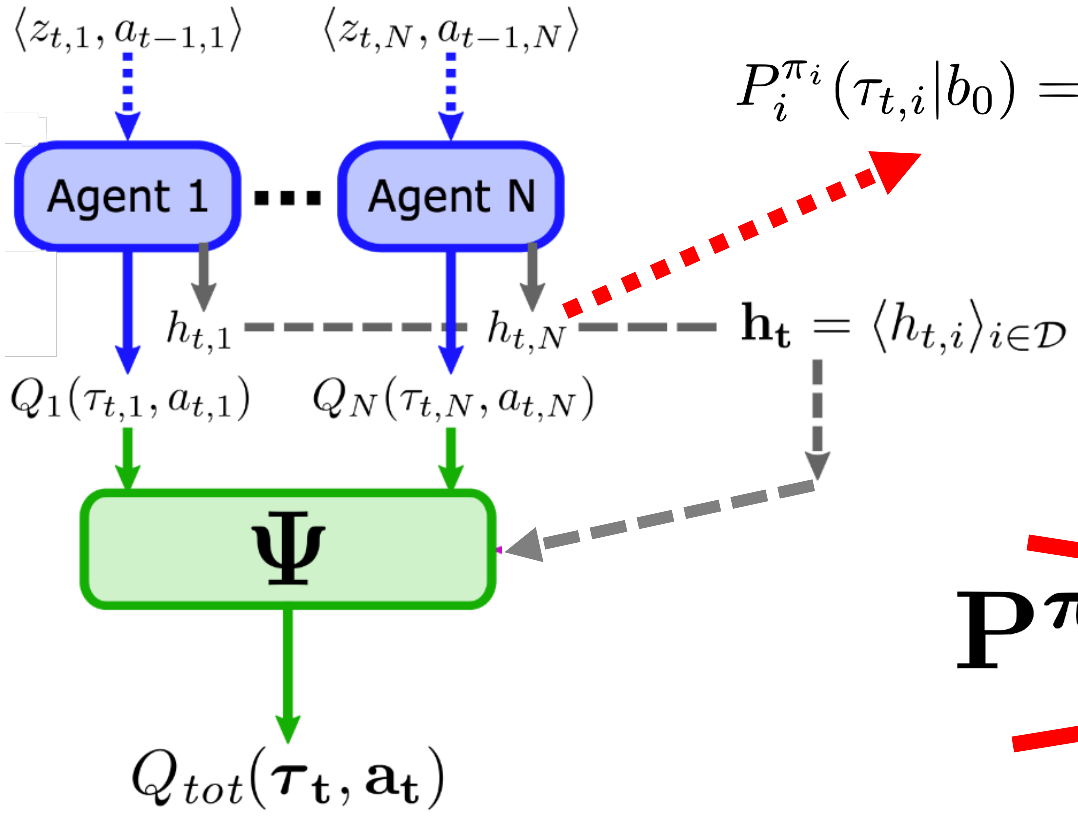
$$P_i^{\pi_i}(\tau_{t,i}|b_0) = P_i(z_{0,i}|b_0) \prod_{c=1}^{t} P_i(z_{c,i}|\tau_{c-1,i}, \pi_i)$$

$$\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau}_\mathbf{t}|b_0) = \prod_{i=1}^{N} P_i^{\pi_i}(\tau_{t,i}|b_0)$$

AERIAL: Attention-based Embeddings of Recurrence In multi-Agent Learning

Value factorization approach considering …

- memory representations of agents



$$P_i^{\pi_i}(\tau_{t,i}|b_0) = P_i(z_{0,i}|b_0) \prod_{c=1}^{t} P_i(z_{c,i}|\tau_{c-1,i}, \pi_i)$$

$$\cancel{\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau}_\mathbf{t}|b_0) = \prod_{i=1}^{N} P_i^{\pi_i}(\tau_{t,i}|b_0)}$$

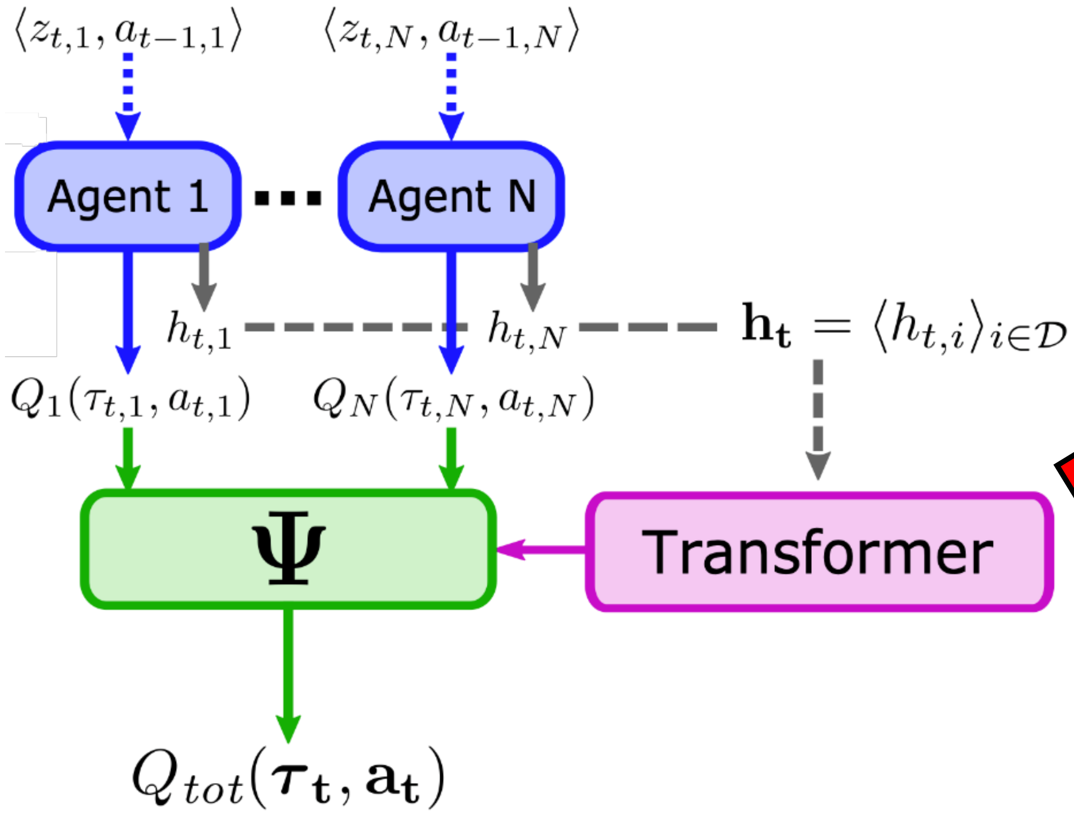
Individual recurrences not independent in general ☹

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

AERIAL: Attention-based Embeddings of Recurrence In multi-Agent Learning

Value factorization approach considering …

- memory representations of agents
- statistical dependence of these representations



$$all2all = \frac{1}{\sqrt{d_{att}}} \begin{bmatrix} W_q(h_{t,1}) \\ \dots \\ W_q(h_{t,N}) \end{bmatrix} \cdot [W_k(h_{t,1}) \; \dots W_k(h_{t,N})]$$

$$recurrence = MLP(softmax(all2all) \begin{bmatrix} W_v(h_{t,1}) \\ \dots \\ W_v(h_{t,N}) \end{bmatrix})$$

MessySMAC

Code

Modified variant of SMAC with …

- observation stochasticity
- initialization stochasticity