Scalable Safe Policy Improvement for Factored Multi-Agent MDPs

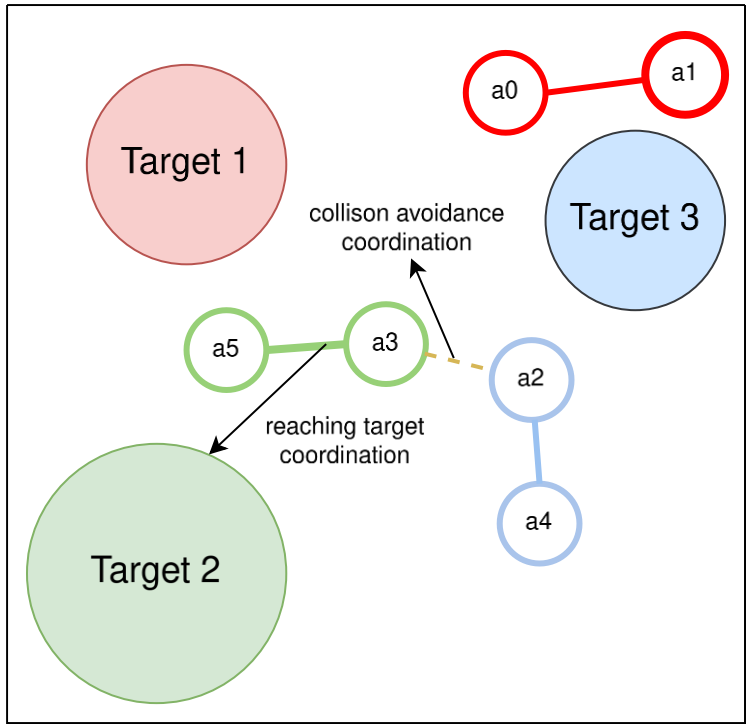
Federico Bianchi, Edoardo Zorzi, Alberto Castellini,
Thiago D. Simao, Matthijs T.J. Spaan, Alessandro Farinelli

federico.bianchi@univr.it

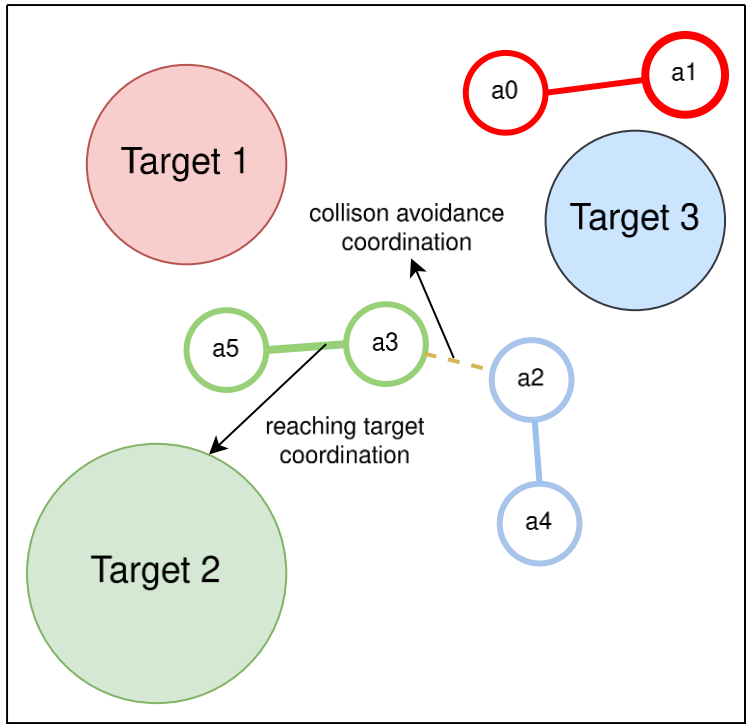
International Conference on Machine Learning (ICML 2024)
23/07/2024
Vienna, Austria

Problem Definition: Scaling SPI to multi-agents systems

Example on Multi-UAV Delivery

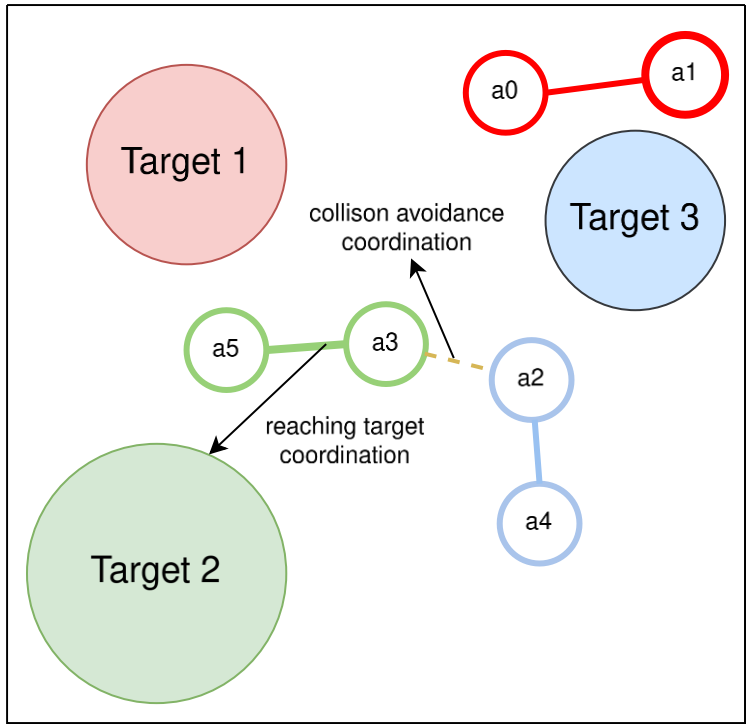


Example on Multi-UAV Delivery



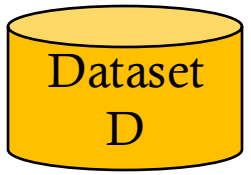
Behavior policy π_0: $\bar{s} \rightarrow \bar{a}$

Example on Multi-UAV Delivery

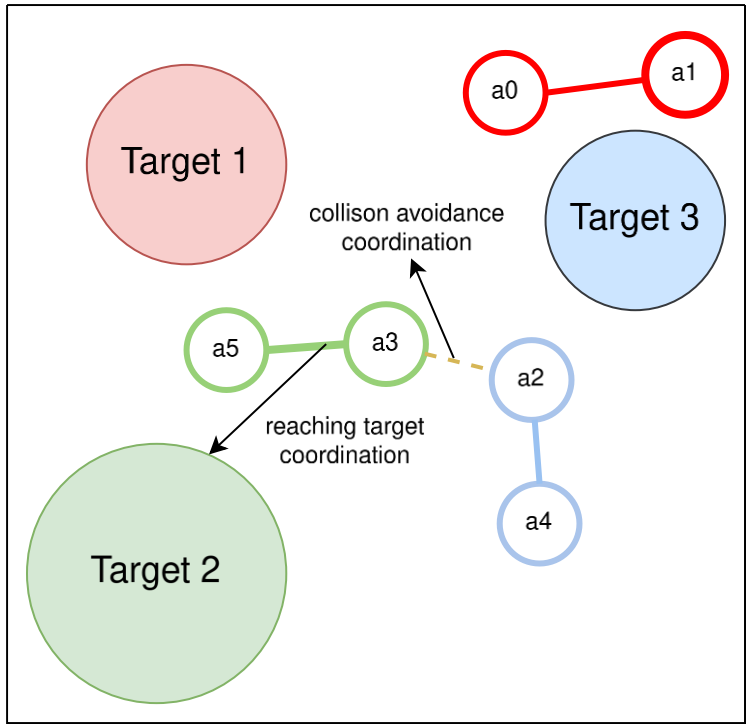


Behavior policy π_0: $\bar{s} \rightarrow \bar{a}$

Interaction with the environment



Example on Multi-UAV Delivery



Behavior policy $\pi_0: \bar{s} \rightarrow \bar{a}$

Interaction with the environment

Dataset D

MLE transition model T^D

$B_m = \{(\bar{s}, \bar{a}) \mid n_D(\bar{s}, \bar{a}) < m\}$

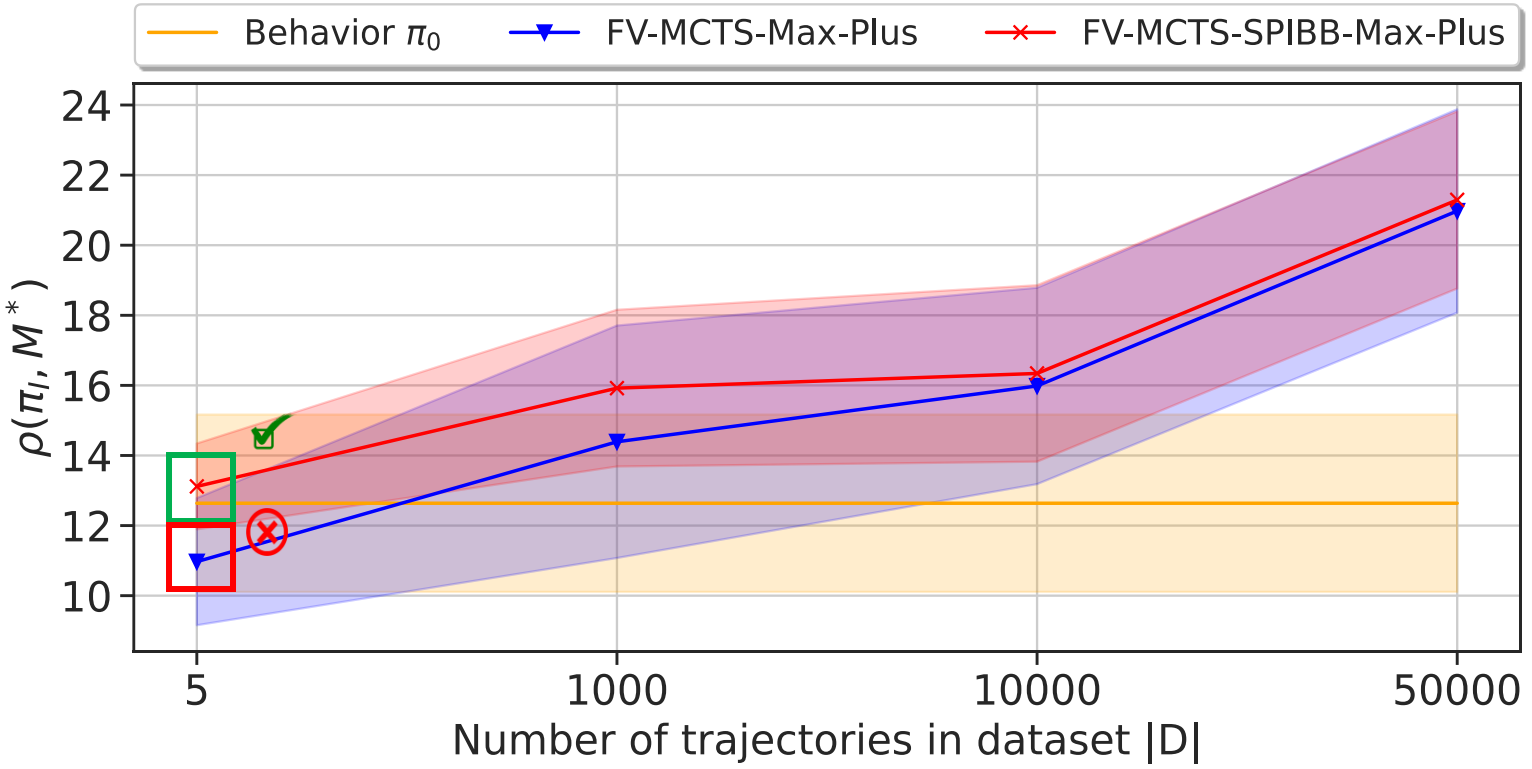
$\bar{B}_m = \{(\bar{s}, \bar{a}) \mid n_D(\bar{s}, \bar{a}) \geq m\}$

Safe Policy Improvement

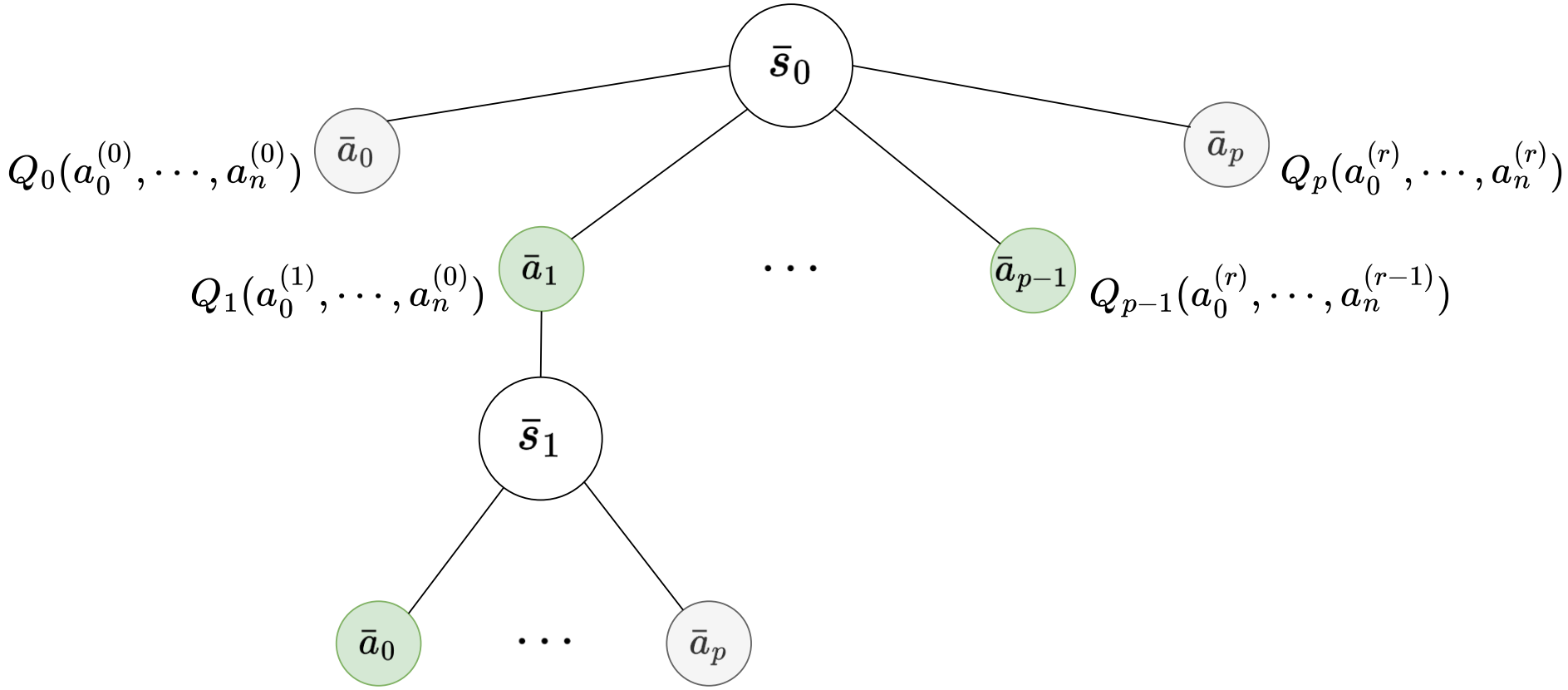
$$\mathbb{P}(\rho(\pi_I) \geq \rho(\pi_0) - \zeta) \geq 1 - \delta$$

Safe Policy Improvement: Goal

$$\mathbb{P}(\rho(\pi_I) \geq \rho(\pi_0) - \zeta) \geq 1 - \delta$$

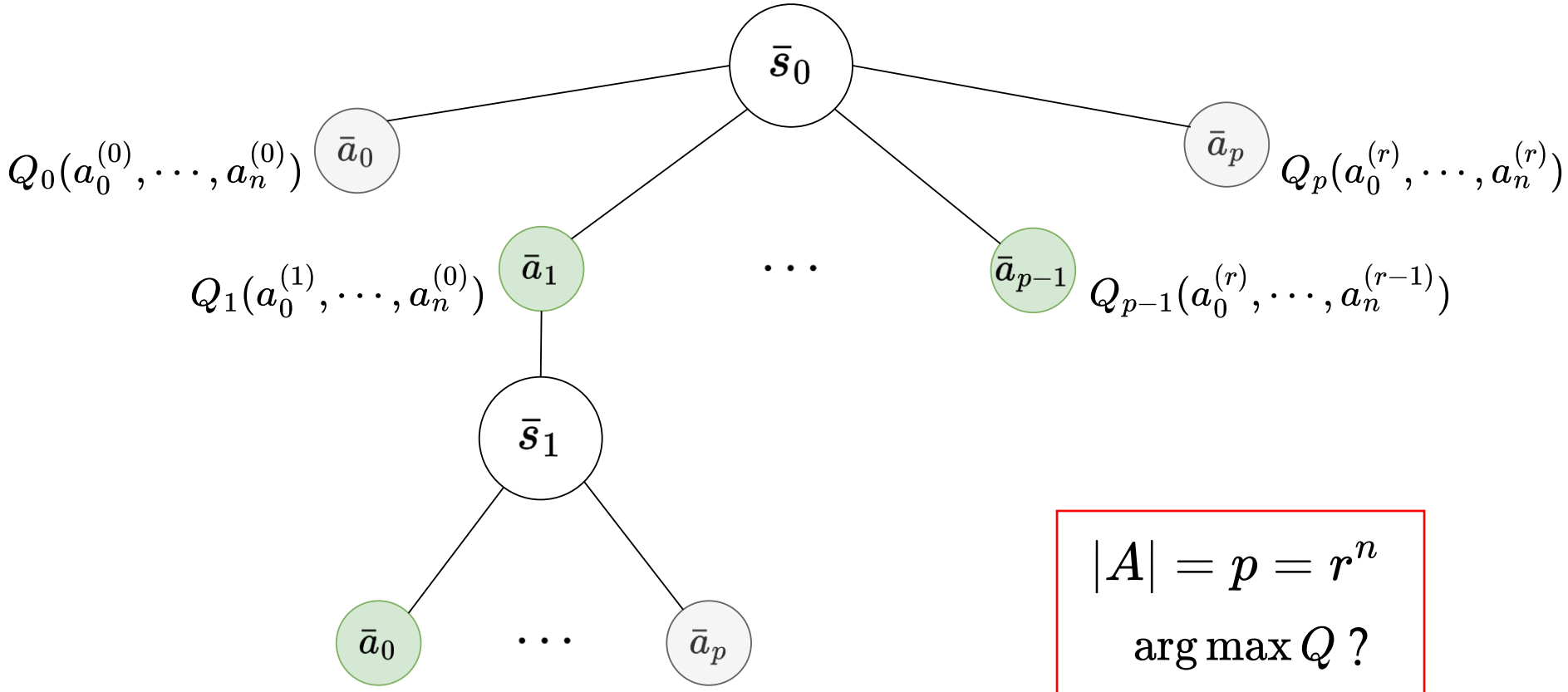


MCTS-SPIBB: Problem



A.Castellini, F.Bianchi, E.Zorzi, T.Simao, A.Farinelli, M.Spaan. Scalable Safe Policy Improvement via Monte Carlo Tree Search (ICML 2023)

MCTS-SPIBB: Problem



$$|A| = p = r^n$$

$$\arg\max Q \; ?$$

A.Castellini, F.Bianchi, E.Zorzi, T.Simao, A.Farinelli, M.Spaan. Scalable Safe Policy Improvement via Monte Carlo Tree Search (ICML 2023)

Method: Factored Value MCTS-SPIBB (FV-MCTS-SPIBB)

FV-MCTS-SPIBB

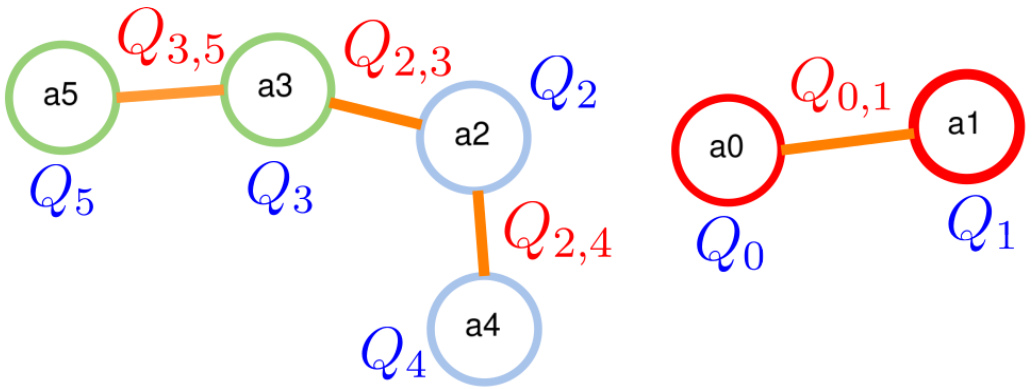
Scalable MCTS-based multi-agent SPI approach

FV-MCTS-SPIBB: Scalability

Scalable MCTS-based multi-agent SPI approach

- Factorization of the value function induced by Coordination graphs

Coordination Graphs

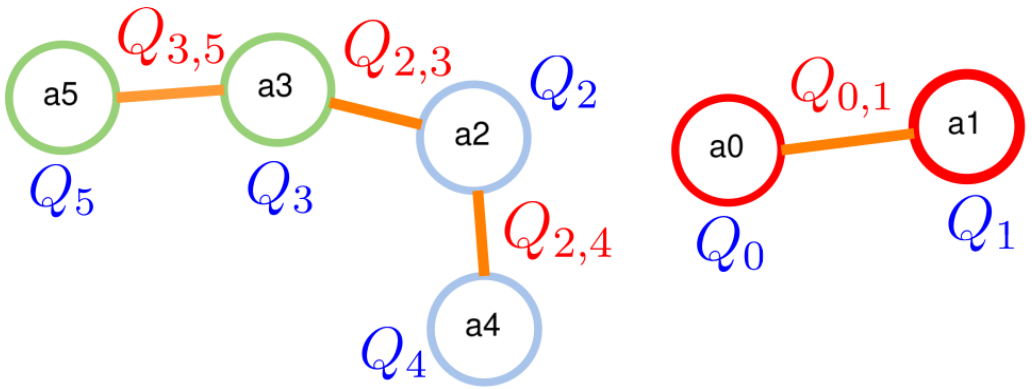


FV-MCTS-SPIBB: Scalability

Scalable MCTS-based multi-agent SPI approach

- Factorization of the value function induced by Coordination graphs

Coordination Graphs



Factorization of the value function

$$Q(\bar{a}) = \sum_{i \in \mathcal{V}} Q_i(a_i) + \sum_{i,j \in \mathcal{E}} Q_{ij}(a_i, a_j)$$

FV-MCTS-SPIBB: Safety

Two novel action selection strategies that guarantee safety

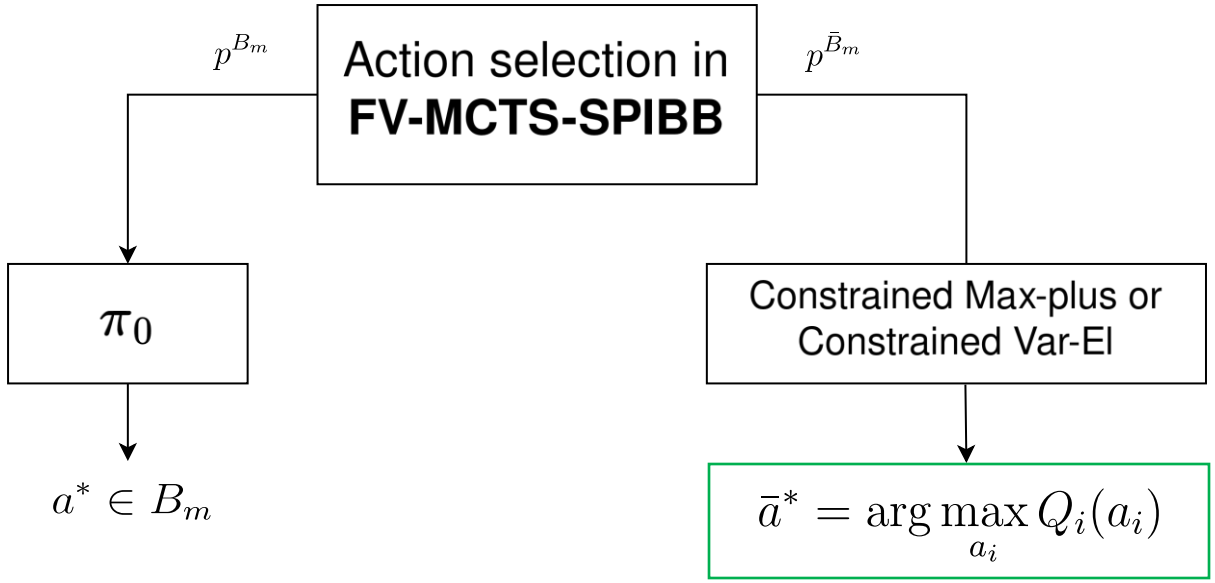
- Constrained Max-Plus
- Constrained Variable Elimination (Var-El)

FV-MCTS-SPIBB: Safety

Two novel action selection strategies that guarantee safety

- Constrained Max-Plus
- Constrained Variable Elimination (Var-El)

$$p^{B_m} = \sum p(\bar{s}, \bar{a})$$
$$p^{\bar{B}_m} = 1 - p^{B_m}$$



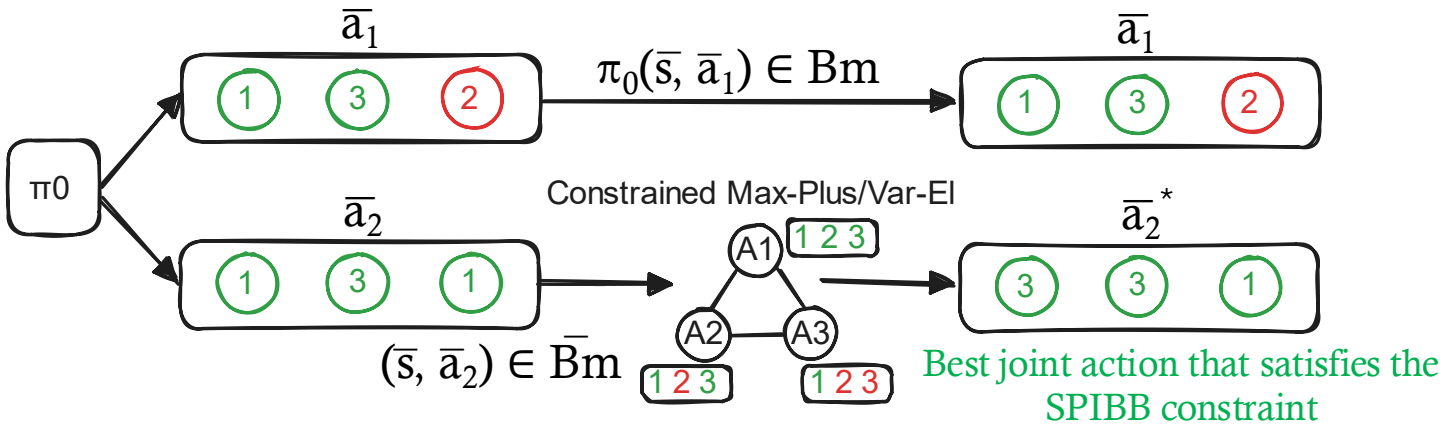
– No value function update
– No policy improvement
+ Safety is guaranteed

+ Value function update
+ Policy improvement
+ Safety is guaranteed

FV-MCTS-SPIBB: Sample scalability

Sample scalability

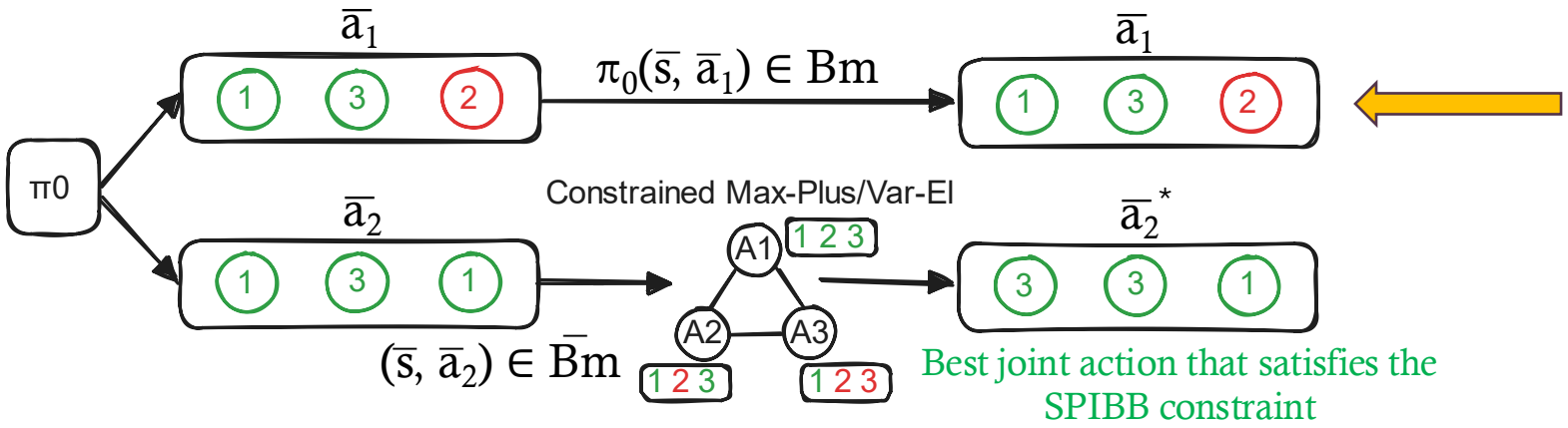
- Factorization of the transition model
- $\bar{B}_m = \{(\bar{s}, \bar{a}) \in S \times A \,|\, \exists S_k : n_D(\bar{s}, \bar{a}) < m_k\}$



FV-MCTS-SPIBB: Sample scalability

Sample scalability

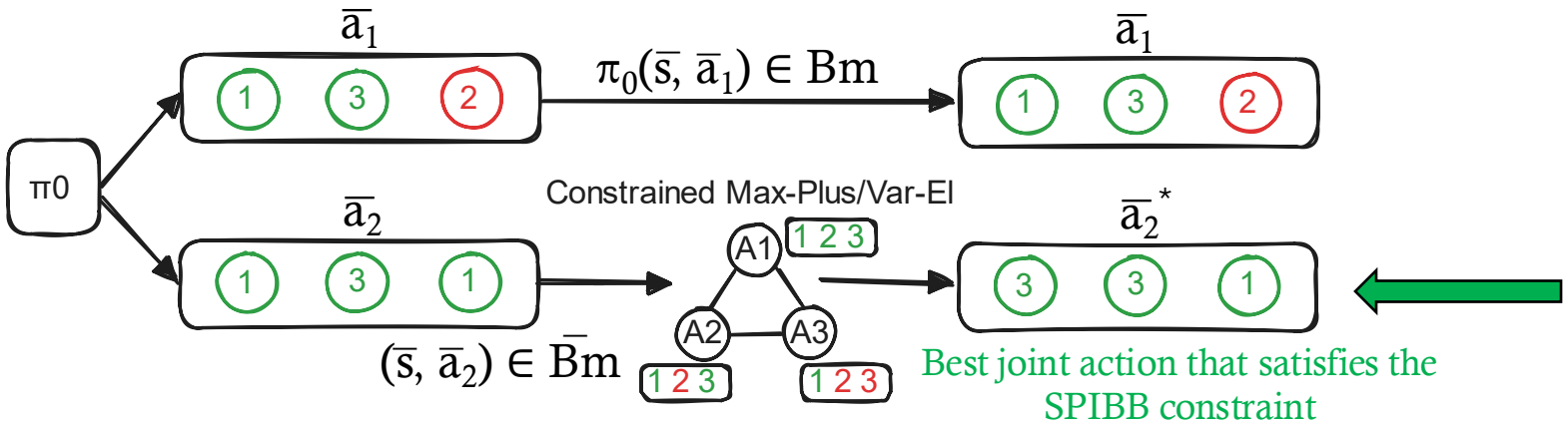
- Factorization of the transition model
- $\bar{B}_m = \{(\bar{s}, \bar{a}) \in S \times A \,|\, \exists S_k : n_D(\bar{s}, \bar{a}) < m_k\}$



FV-MCTS-SPIBB: Sample scalability

Sample scalability

- Factorization of the transition model
- $\bar{B}_m = \{(\bar{s}, \bar{a}) \in S \times A \,|\, \exists S_k : n_D(\bar{s}, \bar{a}) < m_k\}$



FV-MCTS-SPIBB: Theoretical analysis

Theorem 5.1 Safety for FMMDPs

Assuming:

- UCB be component wise in FV-MCTS-SPIBB
- A suitable factorization of the Q-value function

The improved policy π_I is a ζ-approximate safe policy improvement over π_0 with high probability 1-δ

Results: Domains, Scalability, Safety

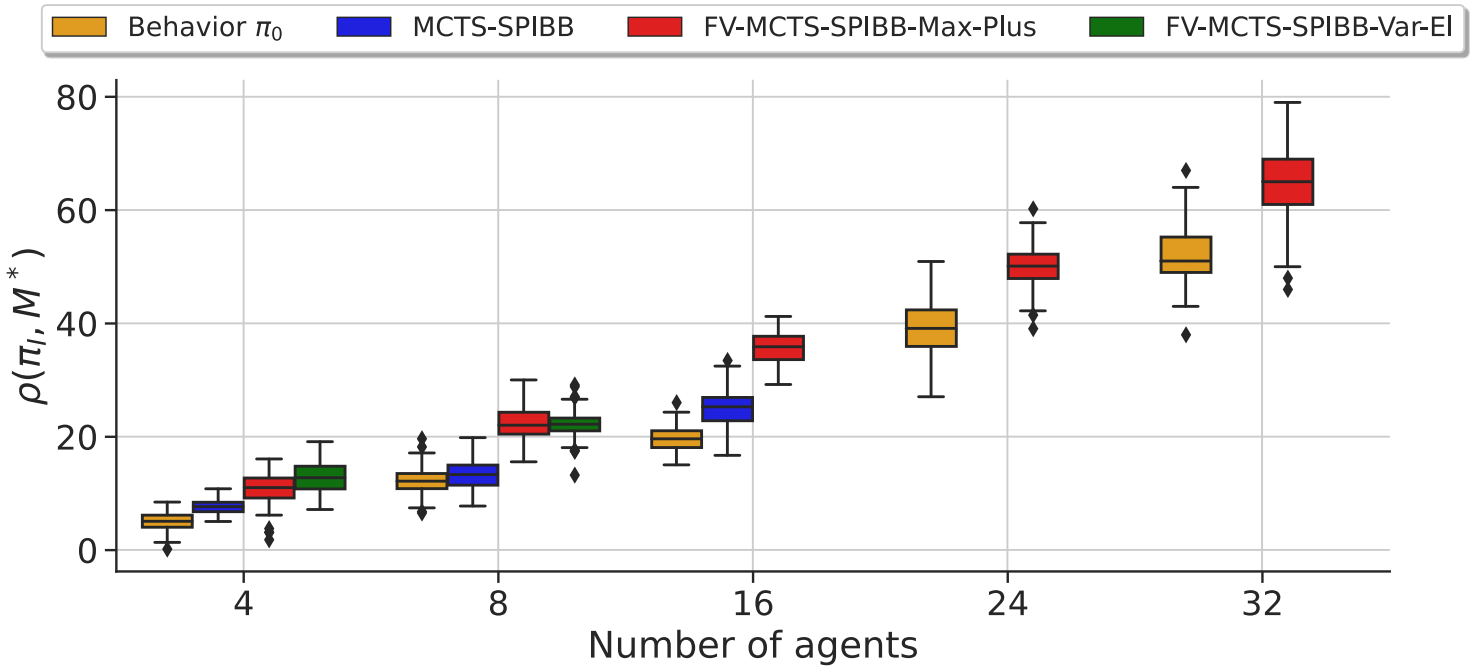
Results: Domains

Multi-agents SysAdmin

Multi-UAV Delivery

Significantly larger domains:
- Previous SPI works up to 25 states, 10 actions
- **Our settings span 10^{30} to 10^{41} states, 10^{9} to 10^{16} actions**

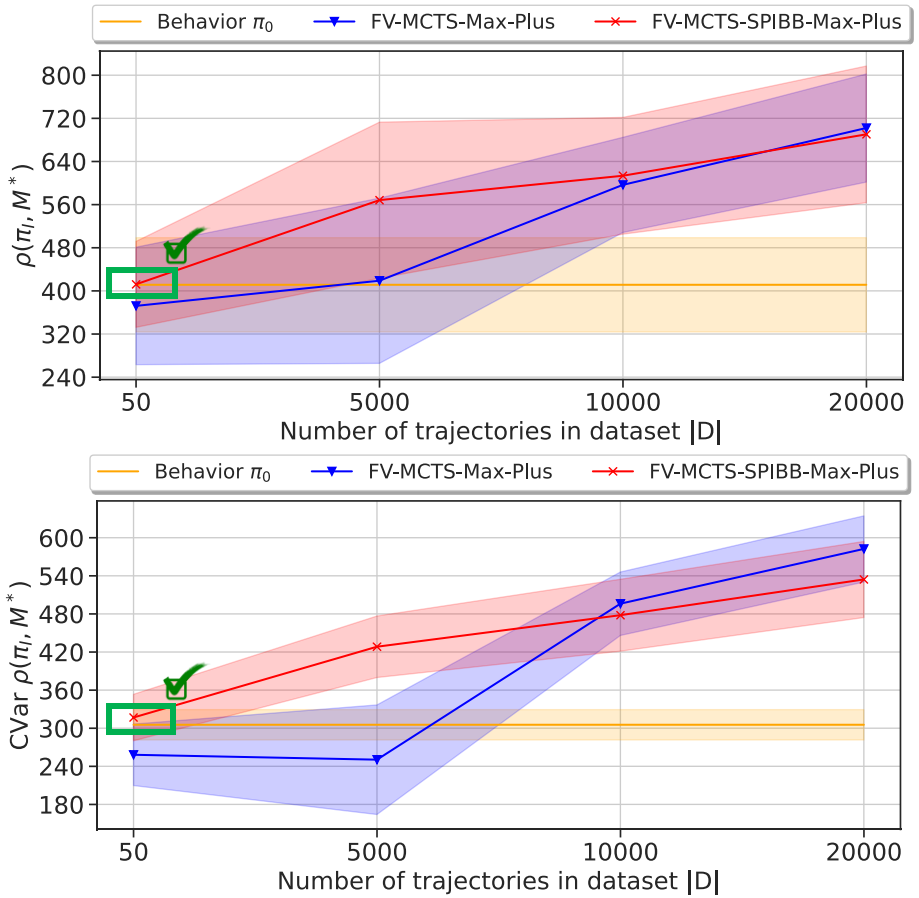
Results: Scalability



FV-MCTS-SPIBB can **scale** and **compute** the **improved policy** in **multi-agent domains** where **other SPI algorithms cannot work**

[MCTS-SPIBB, Castellini et al., ICML 2023]

Results: Safety



FV-MCTS-SPIBB preserves the safety guarantees of SPIBB. It achieves the behavior policy performance when the number of trajectories D is not large enough to improve the policy

[FV-MCTS-Max-Plus, Choudhury et al., AAMAS 2021]

Take Home Message:

- **FV-MCTS-SPIBB** is the **first multi-agent SPI method**
- **Key results towards applying SPI to real-world**

Thank you for your attention