Planning in Sokoban Puzzle Environment

Łukasz Kuciński, Maciej Klimek, Piotr Miłoś

KU KDM 2019
Zakopane, 07.03.2019

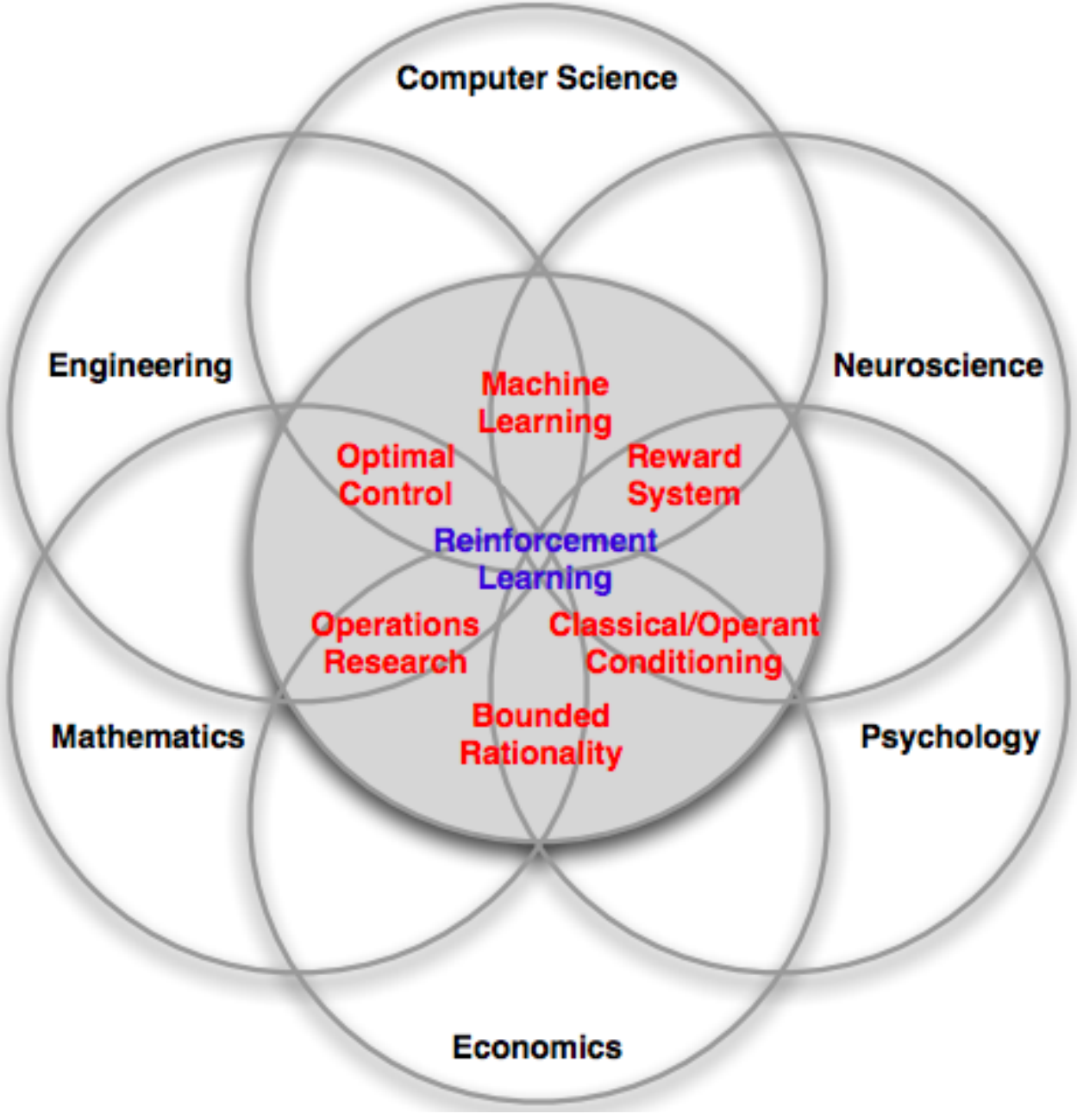
Agenda

- Reinforcement Learning
- Sokoban environment
- Model-free
- Model-based: known model
- Model-based: unknown model

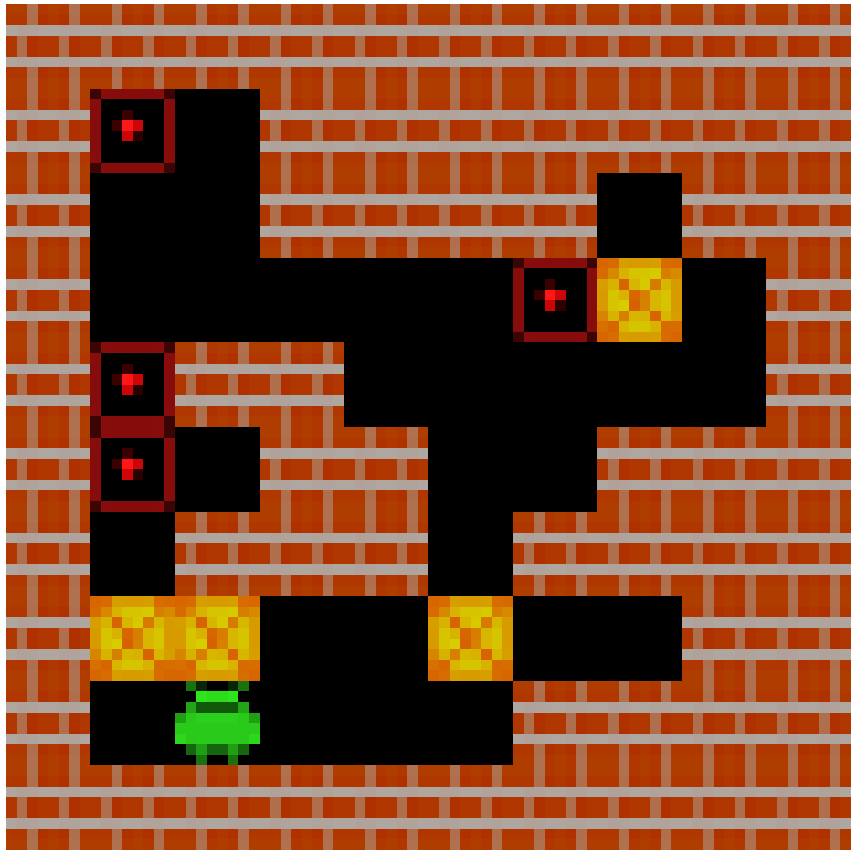
Reinforcement Learning

- Reinforcement Learning (RL): computational approach to solving sequential decision-making problems.
- Model-free RL: learning only through interactions with environment, the embodiment of trial-and-error learning.
- Model-based RL: learning with the use of a model of environment.
- Planning: any computational process that takes a model as input and produces or improves a policy for interacting with the modeled environment.

Reinforcement Learning roots



Sokoban environment



- Goal: Push boxes onto goal locations within step count limit.
- New challenging Reinforcement Learning environment.

Sokoban: problem setup

Solving Sokoban

- Rewards:
  - $+1$ for pushing a box onto a goal location.
  - $+10$ for solving the level.
  - -0.1 for each step.
  - -1 for pushing the box of a goal location.
- Levels are generated procedurally (with solvability guarantees).
- Each episode is run on a random level (both training and test)
- Metric: % of solved levels during test.

Some comments

- Current number of steps is a hidden variable.
- Maximum number of steps influences value function.
- Generating mechanism defines a subclass of Sokoban puszzles.

Sokoban: why is it hard?

- Deciding if Sokoban level is solvable is NP-hard.
- Sparse rewards.
- Graph search problem.
- Game graph has cycles.
- Irreversible states.
- No learning signal for 'dead states'.
- No obvious similarity measures for trajectories.
- Random agent has very low probability of success.

Model-free

- Weber, et. al. "Imagination-Augmented Agents for Deep Reinforcement Learning", 2018.
  - A3C: 60% winrate (10x10x4).
- Guez, et. al. "An investigation of model-free planning", 2019.
  - planning = generalization + sample efficiency + scaleability with compute.
  - Deep Repeated ConvLSTM (DRC): 99%.

Model-based: known model

- Silver, et. al. "Mastering the Game of Go without Human Knowledge", 2017.
  - AlphaZero - state-of-the-art Monte Carlo Tree Search (MCTS) algorithm.
  - MCTS: 87% (25k env steps) - 95% (100k env steps).

Model-based: unknown model

- Weber, et. al. "Imagination-Augmented Agents for Deep Reinforcement Learning", 2018.
  - I2A with learned model (poor or good): 87%.

Conjecture

Sokoban agent can be improved using some of the following:

- HER, Anrychowicz, et. al. "Hindsight Experience Replay", 2017.
- Ensemble, Lowrey, et. al., "Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control", 2018.
- NRPA, Rosin, "Nested Rollout Policy Adaptation for Monte Carlo Tree Search", 2010.
- Dense rewards

Further work: include model training in the loop.

Acknowledgments

- Number of experiments $\sim$ 2K (and counting).
- Experiments CPU intensive.
- Research supported also by the NCN grant UMO-2017/26/E/ST6/00622.

Thank you!