Bootstrapping Skills

Daniel J. Mankowitz $^1$, Timothy A. Mann $^{1,2}$, Shie Mannor $^1$

$^1$Department of Electrical Engineering
The Technion - Israel Institute of Technology
Haifa, Israel

$^2$Google Deepmind
London, UK
Outline

1. Motivation
2. Skills
3. Algorithm Learning Skills via Bootstrapping (LSB)
4. Convergence Guarantee and Analysis
5. Experiments
Monolithic Policy

- One policy
- Big and Complex
- No attempt to decompose
Example: Monolithic Policy

- Task: Leave the room
- Skill to Learn: Walk to door, grasp door knob, open the door and walk through door opening
Skills

- Accomplish a subgoal (decompose)
- Can be applied in different contexts (reusable)
- Special form of an option [1]
Example: Skills

- Task: Leave the room
- Skills to Learn:
  - Walk
  - Grasp door knob
  - Open the door
Learning Skills

- Given a partition of states
- Find the best ‘local’ policy
- Inspired by Skill Chaining [2]
First theoretical convergence guarantees for iteratively learning skills in a continuous state MDP
Model Iteration

Bootstrapping Skills
Daniel J. Mankowitz, Timothy A. Mann, Shie Mannor

Motivation
Skills
Algorithm Learning Skills via Bootstrapping (LSB)
Convergence Guarantee and Analysis
Experiments

Misspecified Model → Improved Model → Optimal Model

Iteration 1

Iteration 2

... Iteration 5

G

G

σ5
σ4
σ3
σ2
σ1

Optimal Model
Main Theorem

Theorem

Let $\varepsilon > 0$. If we run LSB with partition $\mathcal{P}$ for $K \geq \log_\gamma (\varepsilon(1 - \gamma))$ iterations, then the algorithm returns policy $\varphi = \langle \mu, \Sigma \rangle$ such that

$$
\| V^*_M - V^\varphi_M \|_\infty \leq \frac{m \eta_{\mathcal{P}}}{(1 - \gamma)^2} + \varepsilon ,
$$

where $m$ is the number of classes in $\mathcal{P}$.

- LSB learns a near-optimal policy
Experiment: Puddle World

Goal

Average Cost for Different Partitions

Optimal Policy

Average Cost

1x1 2x2 3x3 4x4

0 50 100 150 200
Experiments: Puddle World
Experiments: Pinball

- Maze-world
- More complex dynamics
- 4 dimensional state space

![Graph showing average reward over iterations](image)

![Map of the Pinball Maze-world](image)
Experiments: Pinball

- Pinball-world
- Sharp obstacles, non-linear dynamics at obstacle edges
- 4 dimensional state space
Conclusion

- Monolithic Approach is not feasible for many real-world problems
- Decomposing the task and iteratively learning skills allows us to scale
- **We provide the first theoretical convergence guarantees for skill learning in a continuous state environment**
  - Skills *work together*
  - Skill learning requires *iterative improvements*
The research leading to these results has received funding from the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.306638.
For Further Reading I

Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning.

G. Konidaris, A. Barto.
Skill Discovery in Continuous Reinforcement Learning Domains using Skill Chaining