Compositional Behavioral Semantics
for State Abstraction in Reinforcement Learning

Yivan Zhang[1, 2] Ziyan "Ray" Luo[3, 4] Manuel Baltieri[5, 6]
[1]The University of Tokyo [2]RIKEN AIP [3]Mila - Quebec Artificial Intelligence Institute [4]McGill University
[5]Araya Inc. [6]University of Sussex

arXiv:2606.25357

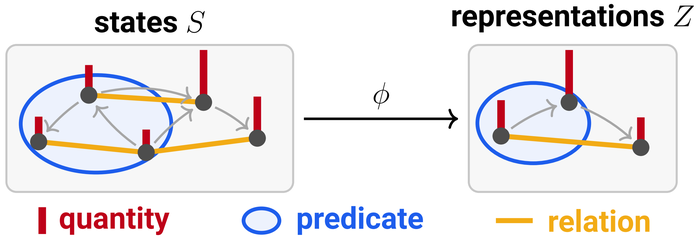
Behavioral structures can be *defined* compositionally and *transferred* through state abstraction.

Summary

State abstraction compresses a system while preserving the behaviors needed for reasoning.
We present a **compositional framework** for defining and transfering behavioral structures for state abstraction.

Background: structure preservation in state abstraction

State abstraction: a smaller representation preserves the information needed to reason about behavior, such as values, safety predicates, or behavioral metrics.



Problem: lack of reusable principles

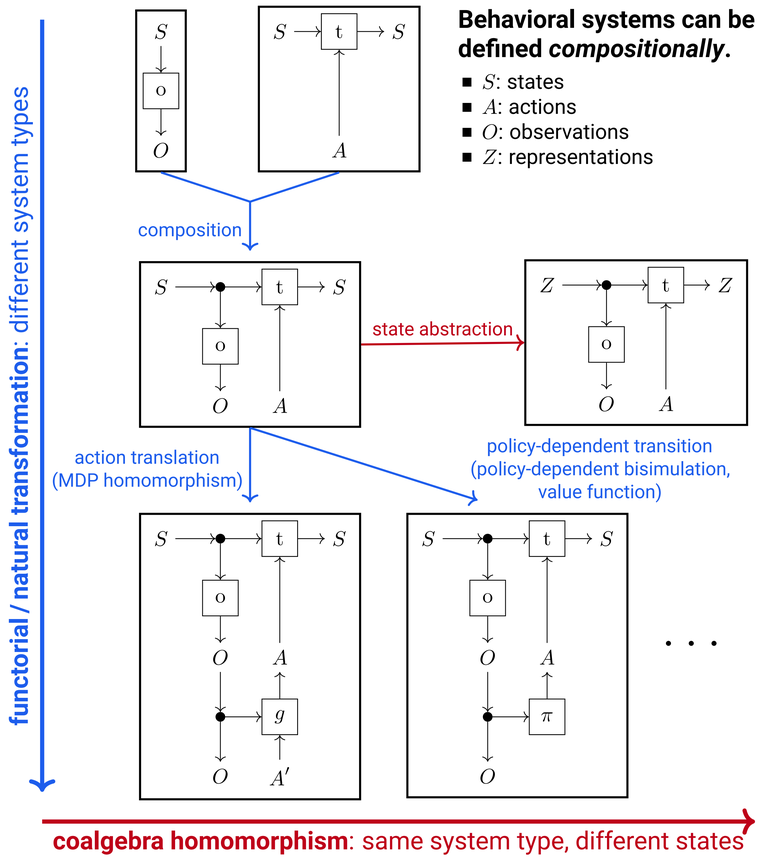
- Abstraction criteria preserve different things: rewards and one-step dynamics, values, safety requirements, bisimulation relations, or behavioral metrics.
- Existing theories analyze them separately, with limited reusable principles for definition, proof, and transfer.

Main Results

- **Definition principle.** A candidate structure becomes *behavioral* when updating it through one-step dynamics leaves it stable.
- **Transfer principle.** A dynamics-preserving abstraction makes two operations sound: read abstract structures back on concrete states, and compress concrete structures onto representations.
- **Metric construction.** Compatible logical and quantitative operators turn logical semantics into quantitative behavioral structures.

Behavioral Systems

How can we explicitly specify the system's input-output interface, and combine its components compositionally?



Behavioral systems can be defined *compositionally*.

- S: states
- A: actions
- O: observations
- Z: representations

Definition (Behavioral system). A *state-based transition system* as an F-coalgebra, whose *input-output interface* is specified by the functor F.

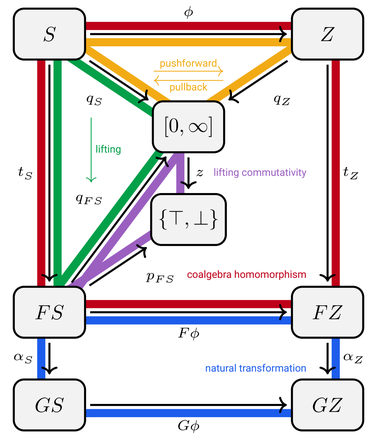
Definition (Homomorphic abstraction). A map from a concrete behavioral system to an abstract behavioral system, *preserving the coalgebraic structure*.

- Explicit interface and composition beyond (PO)MDPs
- **coalgebra homo.**: self-prediction + reward prediction
- **MDP homo.**: coalgebra homo. + natural transformation

Behavioral Semantics

Among all structures on the state space, which ones account for the system's behavior?

Definition (Bundle). A map $S^n \to V$, where V can be truth values $\{\top, \bot\}$, quantitative values $[0, \infty]$ or $\mathbb{R}$, etc.



Bundle operations:

- **Pullback**: from $Z \to V$ to $S \to V$ along ϕ
- **Pushforward**: from $S \to V$ to $Z \to V$ along ϕ
- **Lifting**: from $S \to V$ to $FS \to V$, compositionally
- **Closure**: lifting + pullback along a coalgebra t
- **Lifting commutativity**: how liftings of different bundles are related

Definition (Behavioral structure). A bundle that is a post-fixed point of a (lifting + pullback) closure operator.

Example (Value function). $v^\pi(s) = \gamma\, \mathbb{E}_{s' \sim \mathrm{t}^\pi(s)}\, v^\pi(s') + \mathrm{r}(s)$ is a fixed point of the Bellman closure operator: lift a quantity by expectation and discounted addition, then pull it back along the policy-dependent transition system.

Example (Bisimulation relation). A bisimulation relation is a post-fixed point of the closure operator that lifts a relation on states to a relation on action-conditioned state distributions and observations, and then pulls it back along the transition and observation functions.

With suitable liftings, other behavioral structures like safety predicates and successor features also share the same **definition and transfer principles**.

Future Work

- approximate abstraction via lax homomorphism
- weaker behavioral-structure-preserving abstraction