Modularity in Reinforcement Learning via Algorithmic Independence in Credit Assignment

Michael Chang*, Sidhant Kaushik*, Sergey Levine, Thomas L. Griffiths

Berkeley UNIVERSITY OF CALIFORNIA · PRINCETON UNIVERSITY

MOTIVATION

Want: Reinforcement learning agents that can re-use previously optimal decision for transferring to new tasks
Need: Learning algorithms that can modify the mechanisms for choosing certain actions independently of those for choosing others
Challenge: Currently we have no theory for how to achieve this kind of modular credit assignment, nor formalism within which to express this theory

CONTRIBUTIONS

Problem Formulation

Dynamic modularity: Our definition of dynamic modularity extends the traditional notion of static modularity to apply to learning systems that change with feedback.
Modularity constraint: Our definition of modular credit assignment is a constraint on the algorithmic mutual information among the gradients into different modules.
Theorem: We show that static modularity + modular credit assignment implies dynamic modularity and vice versa under certain conditions.
Challenge: Algorithmic mutual information is generally incomputable

Solution

Insight: Formally treat the learning algorithm as itself a causal graph[1]
Benefit: Reduces measuring algorithmic mutual information to inspecting the graph for d-separation
Theorem: We show how to evaluate, before any training, whether a learning algorithm exhibits modular credit assignment by simply inspecting its computational graph for d-separation.

Theoretical Results

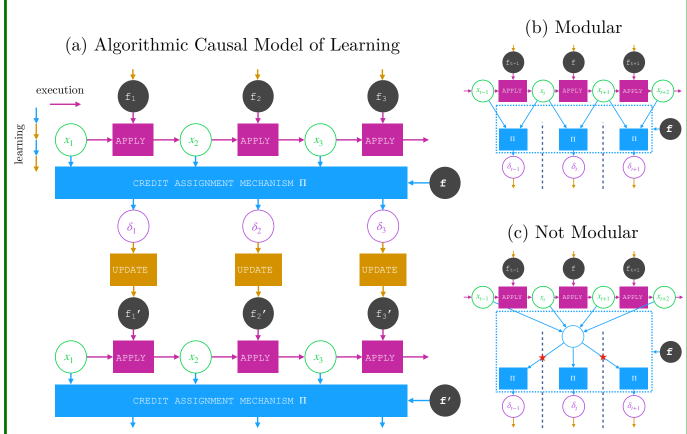
Which reinforcement learning algorithms satisfy the modularity constraint?
- Policy gradient methods: No
- N-step temporal difference methods ($n > 0$): No
- Single-step temporal difference methods: Yes, for acyclic trajectories

Which reinforcement learning algorithms enforce dynamic modularity?
- Tabular: Q-learning, SARSA, cloned Vickrey society[3]
- General function approximation: cloned Vickrey society[3]

Empirical Results

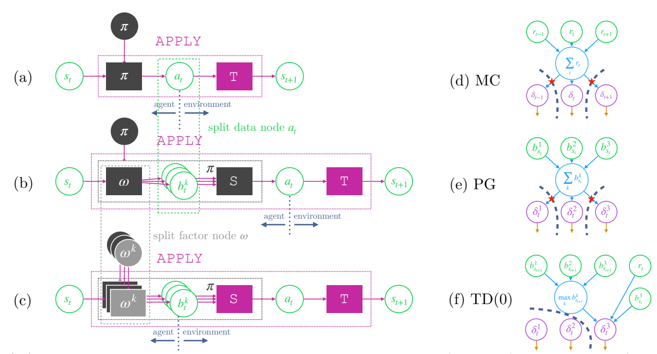
- RL algorithms that are statically modular are correlated with higher sample efficiency in transfer than non-modular RL algorithms
- RL algorithms that are dynamically modular are generally more sample efficient than RL algorithms that are statically modular

ALGORITHMIC CAUSAL MODEL OF LEARNING



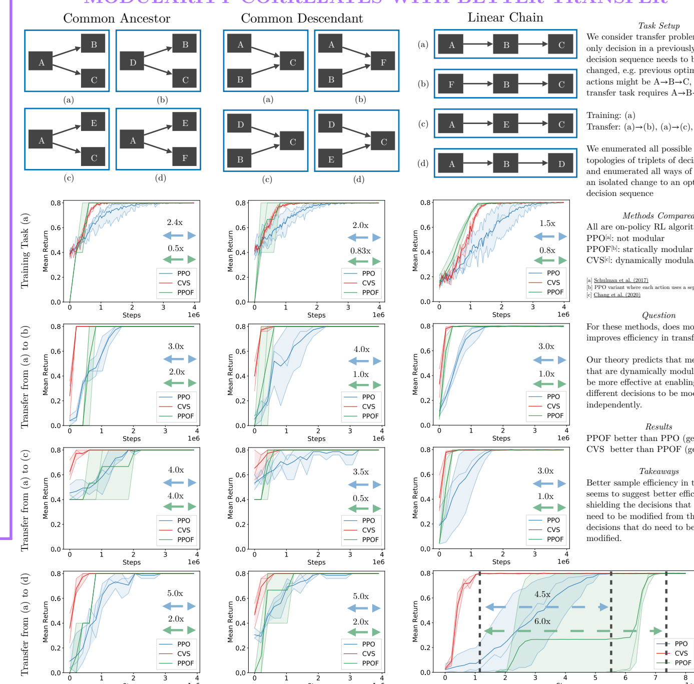
(a) The algorithmic causal model of learning treats the execution of the learning algorithm as a causal graph
(b) For reinforcement learning algorithms that are dynamically modular, the gradients δ of the modules f are d-separated from each other by the previous module weights and the current execution trace $\dots x_{t-1}, x_t, x_{t+1} \dots$
(c) For reinforcement learning algorithms that are *not* dynamically modular, the gradients δ of the modules f are coupled together by a hidden variable.

MODULARITY IN REINFORCEMENT LEARNING



(a-c) We can equivalently re-interpret a learnable discrete-action policy as composed of a society of learnable action-specific functions and a non-learnable selection mechanism[3].
(d) The hidden confounder for algorithms that use Monte Carlo returns is the sum of rewards.
(e) The hidden confounder for algorithms that use policy gradient is the normalization constant of the policy distribution.
(f) Single-step temporal difference algorithms have no confounders (for acyclic trajectories) and thus satisfy dynamic modularity.

MODULARITY CORRELATES WITH BETTER TRANSFER



Task Setup

We consider transfer problems where only decision in a previously optimal decision sequence needs to be changed, e.g. previous optimal actions might be A→B→C, but the transfer task requires A→B→D.

Training: (a)
Transfer: (a)→(b), (a)→(c), (a)→(d)

We enumerated all possible topologies of triplets of decisions, and enumerated all ways of making an isolated change to an optimal decision sequence

Methods Compared

All are on-policy RL algorithms:
PPO[a]: not modular
PPOF[b]: statically modular
CVS[c]: dynamically modular

[a] Schulman et al. (2017)
[b] PPO variant where each action uses a separate network
[c] Chang et al. (2020)

Question

For these methods, does modularity improves efficiency in transfer?

Our theory predicts that methods that are dynamically modular would be more effective at enabling different decisions to be modified independently.

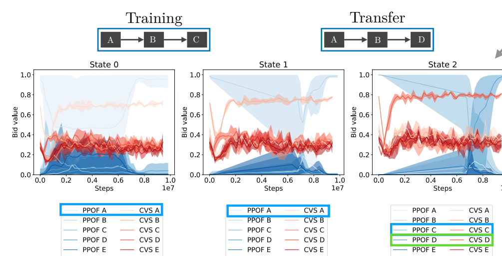
Results

PPOF better than PPO (generally)
CVS better than PPOF (generally)

Takeaways

Better sample efficiency in transfer seems to suggest better efficacy in shielding the decisions that do not need to be modified from the decisions that do need to be modified.

MODULARITY ALLOWS INDEPENDENT MODIFICATIONS



Analysis on Linear Chain

Previous optimal actions: A→B→C
Optimal actions for transfer task: A→B→D
Actions that should not be affected: A, B
Actions that should change: change C to D

Gradients into different decisions independent?
- PPOF: no
- CVS: yes

Takeaways

Independent gradients enables different modules to be modified independently, which empirically seems to improve sample efficiency in transfer.

References

[1] Janzing and Schökopf (2010)
[2] Schulman, et al. (2017)
[3] Chang, et al. (2020)