Multitask Model-free Reinforcement Learning

Andrew M. Saxe

Stanford University

Overview

Conventional model-free reinforcement learning algorithms are limited to performing only one task, such as navigating to a single goal location in a maze, or reaching one goal state in the Tower of Hanoi block manipulation problem. Yet in ecological settings, our tasks change often and we respond flexibly—we may wish to navigate to some other point in the maze, or reach some state other than the typical end goal in the Tower of Hanoi problem. It has been thought that in most cases, only model-based algorithms provide such flexibility.

We present a novel model-free algorithm, multitask Z-learning, capable of flexible adaptation to new tasks without a forward search through future states. The algorithm learns about many different tasks simultaneously, and mixes these together to perform novel, never-before-seen tasks. Crucially, it performs any linear blend of previously learned tasks optimally.

The algorithm learns a distributed representation of tasks, thereby avoiding the curse of dimensionality afflicting other hierarchical reinforcement learning approaches like the options framework that cannot blend subtasks. That is, while it is easy to learn a fixed set of alternative tasks using an off-policy model-free algorithm, it has not previously been possible to perform an infinite set of tasks by blending a fixed set. We present applications to sequential choice, spatial navigation, and the Tower of Hanoi problem. The model has a number of limitations, which make novel predictions about problem manipulations that should show rapid adaptation, versus those that should require slow practice for improvement.

Beyond model-free and model-based

Model-free learning reinforces actions that lead to reward in the past

Model-based learning uses a model of the world to perform a forward search, selecting actions that lead to reward

- Model-free, thus, cannot rapidly adapt to changing reward structure
- Model-based appears to require an explicit, sequential search through the environment

Yet intuitively,

- If you learn how to navigate to one spot in a maze, you should know something about how to get to other spots
- If you learn how to reach your arm to one point in space, you should know something about reaching other points
- If you learn how to solve the Tower of Hanoi task, you should learn how to get to states other than just the goal

Can we go beyond the model-free, model-based distinction? (Doll et al., 2012)

Needed: Compositionality

To perform novel tasks using knowledge from previously learned tasks, we must be able to **compose** them, blending them together to perform an infinite variety of tasks

- Composition is a key intuition underlying theories of perceptual systems
- How can we transfer this intuition to control systems?

Compositionality does not come naturally in control because the **Bellman equation is nonlinear**:

- If we have an optimal cost-to-go function for reward structure A and an optimal cost-to-go function for reward structure B,
- Does **not** mean that the optimal cost-to-go function for reward structure A + B is the sum of these

We use a careful problem formulation to restore this compositionality, and exploit it to permit flexible execution of a variety of tasks.

Multitask Z-learning

Linearly-solvable Markov Decision Processes

We use the first-exit LMDP formulation of Todorov, 2006.

States: *x*, partitioned into interior states and absorbing boundary states

Passive dynamics: $p(y|x)$ = Prob(state(t+1)=y | state(t)=x)

Action: *u*, choose new transition probabilities $p(y|x,u) = u(y|x)$

Cost: $l(x,u) = q(x) + \mathrm{KL}(u(\cdot|x)||p(\cdot|x)) = \text{state cost} + \text{action cost}$

- *q(x)* specifies instantaneous rewards
- KL term penalizes deviation from passive dynamics

Goal: Minimize total cost

Multitask Z-learning

For LMDPs, optimal action directly computable from cost-to-go function *v(x)*
Define exponentiated cost-to-go (desireability) function: $z(x) = \exp(-v(x))$
Bellman equation *linear* in *z*:

$$z(x) = \exp(-q(x))\mathbf{E}_{y\sim p(\cdot|x)}[z(y)]$$

Or $z_i = Mz_i + n_b$ where z_i encodes desireability of interior states

Crucial property: Solutions for two different boundary reward structures linearly compose (Todorov, 2009)

$$\tilde{q}_b^{1+2} = a\tilde{q}_b^1 + b\tilde{q}_b^2 \Rightarrow z_i^{1+2} = az_i^1 + bz_i^2$$

Multitask Z-learning: Learn about a set of boundary reward structures $\tilde{q}_b^c, c = 1,\ldots,m$

- represent any new task as a linear combination of these
- optimal *z(x)* is linear combination of component tasks' $z^c(x)$

Off-policy update: For each component task at each time step, update:

$$\hat{z}(y_t) \leftarrow (1-\eta)\hat{z}(y_t) + \eta\exp(-q_t)\hat{z}(y_{t+1})$$

Limitations

Multitask Z-learning is capable of instantaneous optimal adaptation to any novel task whose (exponentiated) boundary cost structure is a linear combination of previously learned component tasks

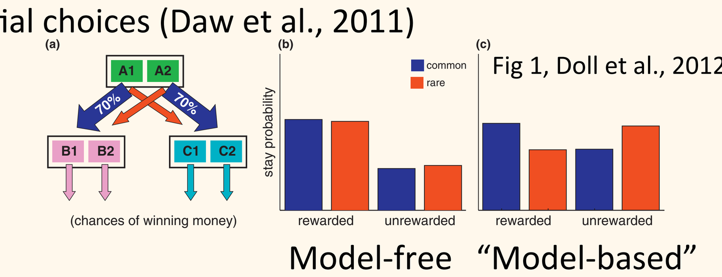
Yet it has serious limitations:

- It can only adapt rapidly to changes in *boundary* rewards; changing internal reward structure requires slower z-learning updates
- It cannot adapt quickly to changes in the *passive dynamics*. The optimal action computation uses the passive dynamics, so changing this will immediately change behavior but not in an optimal way until z-learning has time to act
- The LMDP formulation differs from the standard discrete action choice MDP formulation (though a traditional MDP can be embedded), and is more natural for quasi continuous motor control style problems than discrete action selection problems

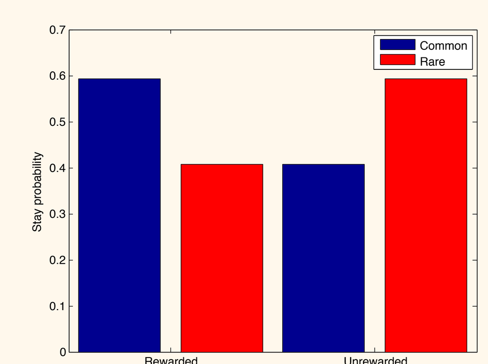
Outcome revaluation, latent learning and specific satiety

Outcome revaluation in sequential choice

Humans and animals can rapidly adapt to changing rewards in sequential choices (Daw et al., 2011)



- Two stage choice task
- Multitask Z-learning behaves like model-based methods
- Does not rely on forward model search

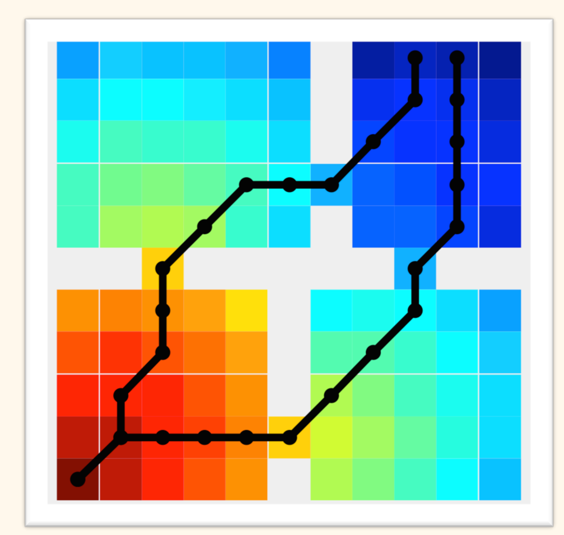


Multitask Z-learning

Latent learning in spatial navigation

After random exploration of a maze environment, introduction of a reward at one location leads to instant goal-directed behavior towards that point (Tolman, 1948)

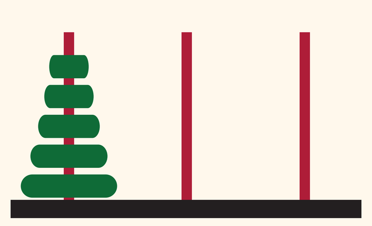
- Covert multitask z-learning during exploration enables immediate navigation to rewarded locations when reward structure becomes known
- Covert tasks need not include navigation to every point, but must only provide a linear basis in which each point can be represented



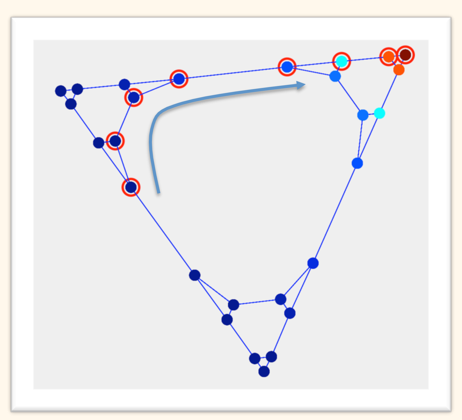
Tower of Hanoi

Applicable to goal-directed action in more complex domains (Diuk et al., 2013)

- Move blocks to peg 3; smaller blocks must always be stacked on larger blocks



- After exploration, multitask Z-learning is capable of navigating to arbitrary configurations

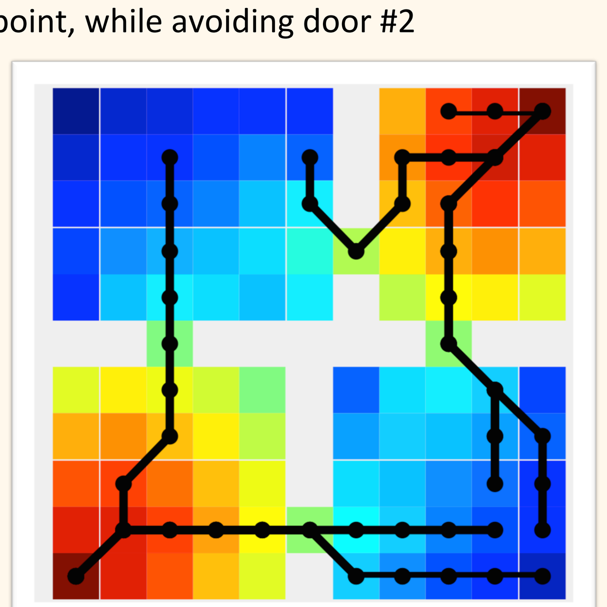

State graph with cost-to-go and optimal trajectory

Exploiting compositionality: Complex task blends

"Navigate to room A or B"

Can respond flexibly to a variety of navigation tasks

- Find food or water (specific satiety experiments)
- Go to a point, while avoiding door #2

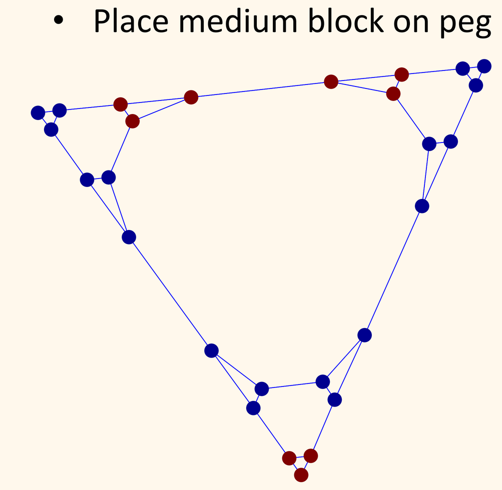


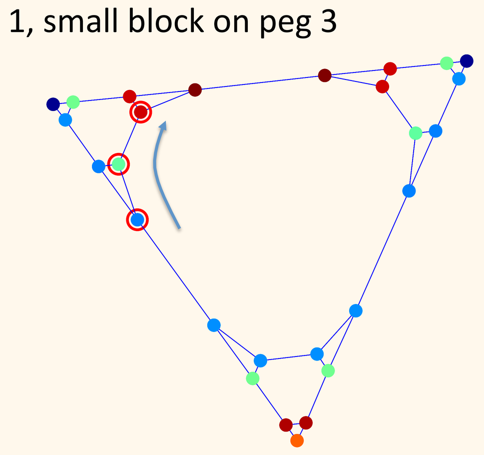
- Important note: Not the same as planning through arbitrary cost map because of boundary state formulation.

"Place medium size block on middle peg"

Compositionality enables rapid response to complex queries

- Stack small block on large block
- Place medium block on peg 1, small block on peg 3


Instantaneous rewards


Cost-to-go/trajectory

- Models highly practiced expert quite familiar with domain
- Can be combined with model-based search

Conclusions

Multitask Z-learning is a new reinforcement learning algorithm with interesting properties:

- **Instantaneous optimal** adaption to new absorbing boundary rewards
- Relies on careful problem formulation to permit compositionality
- Off-policy algorithm over states (not state/action pairs)
- Compatible with function approximation

It suggests that, with enough experience in a domain, complex new tasks can be optimally implemented without an explicit forward search process

Similar in spirit to the successor representation (Dayan, 1993), but generalizes this to off-policy, states-based, multitask rewards: a more powerful representation than successors is one that already accounts for possibly complicated internal reward structure

Compatible with model-based & model-free accounts, which are tractable in the LMDP

Multitask z-learning introduces new potentially relevant distinctions:

- Absorbing boundary reward change => instant adaptation
- Internal reward change => slow adaptation
- Transition change => suboptimal instant adaptation, slow optimal adaptation

A.S. was supported by a NDSEG Fellowship and MBC Traineeship.