Fast Adaptation of Behavior to Changing Goals with a γ-Ensemble

OIST Neural Computations Unit

Chris Reinke, Eiji Uchibe, Kenji Doya

1) Introduction & Problem Description

Adaptive agents need not only to adapt to changes in their environments, but also to changes in the goal they want to achieve in them. For example, during foraging the goal might change from obtaining the most desirable food to securing food as soon as possible. In reinforcement learning, the goal is defined by the reward function and how rewards are discounted over time. If the goal changes model-free methods need to adapt to the new reward function or discounting strategy. This relearning does not allow for a quick adaptation.

We formalize such problems by differentiating between external and internal reward. External reward $r_t = R(s_t)$ is given by the environment, e.g. the food quality. Internal reward defines the goal that the agent wants to fulfill in the environment. It depends on the external reward r_t and the time t to reach it: $\phi(r_t, t) \in \mathbb{R}$. For example, external reward might be reduced linearly over time: $\phi(r_t, t) = r_t - 10t$ (see Fig. 4 for more examples). The final goal is to maximize the internal reward while minimizing the required time, i.e. the number of steps:

$$\min_T \max_\pi E_\pi \left[\sum_{t=0}^{T} \phi(r_t, t) \right]$$

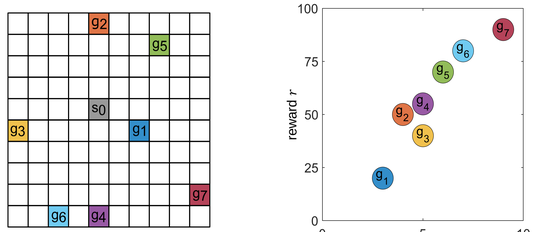
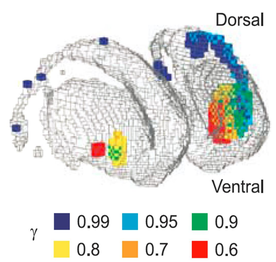
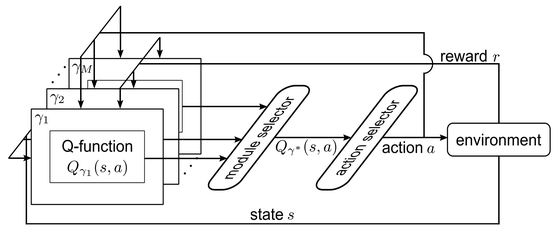


Fig. 1: Example grid world. The agent starts in s_0. It can reach 7 terminal states $(g_1, \dots, g_7)$ in a different number steps n_{s_0} and receive different external rewards r.

2) The Independent γ-Ensemble (IGE)

The IGE is inspired by neuroscientific findings [1] which suggest the brain computes expected reward sums with different discount factors. The IGE uses a similar architecture, consisting of several modules, each is a Q-function with a different γ:





$$Q_\gamma(s_t, a_t) = r_t + \gamma\, r_{t+1} + \gamma^2\, r_{t+2} + \cdots + \gamma^n\, r_{t+n}$$

$$Q_\gamma(s_t, a_t) = r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$$

a) Learning of Policy Sets

Because the γ-modules are independent and Q-learning is off-policy, they learn different policies in parallel dependent on their discount parameter γ:

- Small γ: steep discounting → prefers nearby, but small rewards
- Large γ: slow discounting → prefers large, but distant rewards

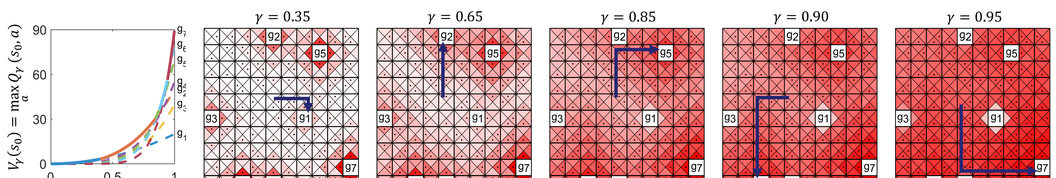


Fig. 2: The value function for state s_0 and the Q-functions for a subset of γ-modules of the example MDP (Fig 1). The agent will learn a set of policies to reach 5 out of 7 terminal states $(g_1, g_2, g_5, g_6, g_7)$.

b) Decoding Information about Policies

The values of modules that learn the same policy because their discount factors are similar can be used to decode their expected reward trajectory.

$$\text{Given: } V = \begin{bmatrix} V_{\gamma_0}(s) \\ \vdots \\ V_{\gamma_T}(s) \end{bmatrix}, \Gamma = \begin{bmatrix} 1 & \gamma_0 & \gamma_0^2 & \cdots & \gamma_0^T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \gamma_T & \gamma_T^2 & \cdots & \gamma_T^T \end{bmatrix} \text{ and } R_\tau = \begin{bmatrix} E[r_0] \\ \vdots \\ E[r_T] \end{bmatrix}$$

$$\mathrm{V} = \Gamma \times R_\tau \quad \Leftrightarrow \quad R_\tau = \Gamma^{-1} \times V$$

Unfortunately, this has practical problems for general MDPs because it is hard to identify modules with the same policy and the inverse Γ^{-1} can be ill-conditioned.

Nonetheless, this can be overcome in episodic, deterministic, and goal-only-reward (reward is only given if a terminal state is reached) MDPs. The value definition for them is simplified to $V_\gamma(s) = \gamma^{n_s} r$ with r as the reward by reaching a terminal state, and n_s as the number of steps required to reach it. Having two neighboring modules (γ_a, γ_b) that follow the same policy, r and n_s can be decoded:

$$V_{\gamma_a} = \gamma_a^{n_s} r \qquad n_s = \frac{\log V_{\gamma_a}(s) - \log V_{\gamma_b}(s)}{\log \gamma_a - \log \gamma_b}$$

$$V_{\gamma_b} = \gamma_b^{n_s} r \quad \Leftrightarrow \quad r = \frac{V_{\gamma_a}(s)}{\gamma_a^{n_s}}$$

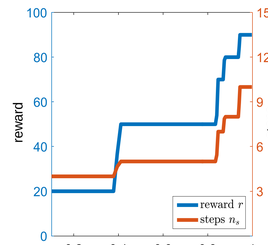


Fig. 3: The expected reward r and number of steps n_s of each γ-module pair based on their value functions (Fig. 2) for state s_0.

3) The Goal Adaptive Independent γ-Ensemble

For the changing goal formulation task, the IGE learns a set of policies based on the external reward. In the beginning of an episode it decodes the expected r and n_s of each module pair and selects one that maximizes the current internal reward function ϕ_k:

input: learning rate α, sorted list of discount factors $\Gamma = (\gamma_1, \gamma_2, \dots, \gamma_M)$
initialize: $Q_\gamma(s, a)$ to zero
repeat (for each episode)
  initialize: state s, goal formulation ϕ_k
  for all $(\gamma_i, \gamma_{i+1}) \in \Gamma$ **do** /** calculate internal reward for module pairs **/
    calculate external reward: $\mathrm{R(i)} \leftarrow r(\gamma_i, \gamma_{i+1})$
    calculate steps: $\mathrm{N(i)} \leftarrow n_s(\gamma_i, \gamma_{i+1})$
    calculate internal reward: $J(i) \leftarrow \phi_k(R(i),\ N(i))$
  for i **from** 2 **to** $M-1$ **do** /** remove invalid modules **/
    if $N(i-1) \neq N(i) \neq N(i+1)$ **do**
      remove invalid module pair: $J(i) \leftarrow \mathrm{NAN}$
  select module $\gamma^* \leftarrow \Gamma(\mathrm{argmax}_i J(i))$ /** select best module **/
  repeat (for each step in episode)
    choose action a for s derived from $Q_{\gamma^*}(s, a)$ (e.g. ϵ-greedy)
    take action a, observe r, s'
    for all $\gamma \in \Gamma$ **do**
      $\mathrm{Q}_\gamma(s, a) \leftarrow Q_\gamma(s, a) + \alpha(r + \gamma \max_{a'} Q_\gamma(s', a') - Q_\gamma(s, a))$
    $s \leftarrow s'$

4) Experimental Results

The IGE outperforms time-dependent Q-learning ($Q(t, s, a) = r + \gamma \max_{a'} Q(t+1, s', a')$) in the MDP of Fig 1. The agents had to adapt to 8 different goal formulations $(\phi_1, \dots, \phi_8)$. They had 3000 episodes to learn the task for ϕ_1 and then 1000 episodes for each successive goal. The IGE can adapt immediately after a goal change; whereas Q-learning needs many episode to adapt.

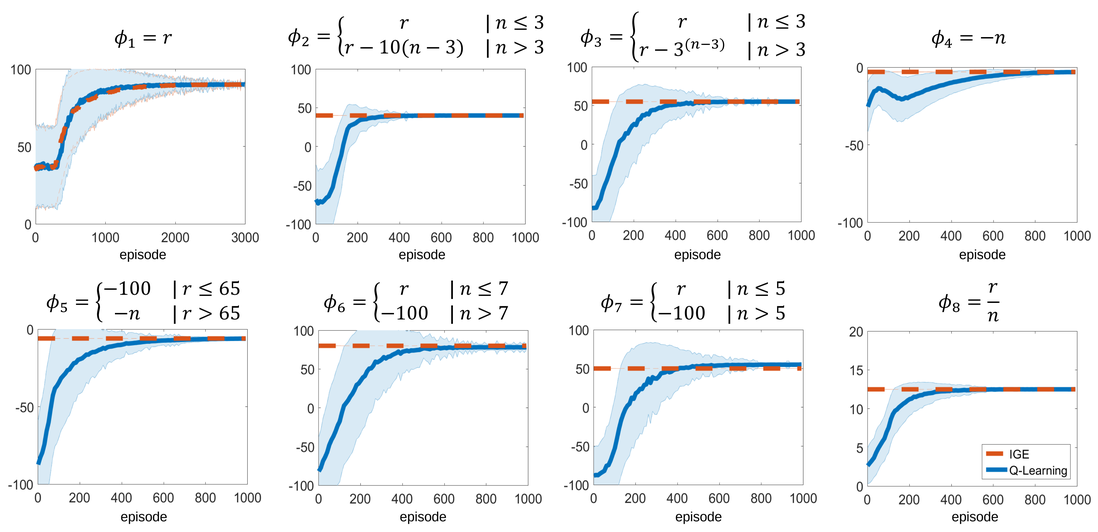


Fig. 4: The IGE outperforms Q-learning in the learning performance (internal reward sum per episode) for the 8 successive goal formulations. For convenience internal reward is only given if a terminal state is reached $(\phi_k(s, t) = 0 \mid s \notin G)$.

5) Discussion

Two key features of the IGE allow it to adapt quickly to goal formulation changes: 1) It learns a set of policies. 2) It can decode information about the policies to choose the most appropriate one for the current goal.

A drawback is that optimality cannot be guaranteed for each goal-formulation, as seen for goal ϕ_7 (Fig. 4), because the IGE does not learn the policies to reach all possible terminal states (Fig. 2). Nonetheless, it can guarantee optimality to maximize per episode the exponentially discounted reward sum for each γ, the maximum reward, the minimum number of steps, and most interestingly the average reward r/n and linearly discounted reward. For other goal formulations, the IGE can be viewed as a heuristic that produces often good results with the ability to adapt immediately.

The IGE provides also a potential model for the modular discounting structure found in the striatum. The model-free learning mechanism in the striatum is connected to habitual decision making [2]. Therefore, the striatum could learn different habits in parallel in its sub-regions. Serotonin has been shown to control which regions are active during decision-making [1]. Because serotonin is associated with stress, different habits could be used for different stress levels. Under low stress, habits are used that strive for large rewards, but that require a longer time to reach; whereas under high stress habits that go for short-term rewards can be activated to adapt quickly without any relearning.

References

[1] Daw, Niv & Dayan (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience, 8, 1704-1711

[2] Tanaka, Schweighofer, Asahi, Shishida, Okamoto, Yamawaki & Doya,. (2007). Serotonin differentially regulates short-and long-term prediction of rewards in the ventral and dorsal striatum. PLoS One, 2 (12), e1333.

Contact:
chris.reinke@oist.jp