The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

# Adversarial Goal Generation for Intrinsic Motivation

**Ishan Durugkar, Peter Stone**
University of Texas at Austin
2317 Speedway, Stop D9500, Austin, TX 78712
Phone: 512.471.7316
ishand, pstone @cs.utexas.edu

## Abstract

Generally in Reinforcement Learning the goal, or reward signal, is given by the environment and cannot be controlled by the agent. We propose to introduce an intrinsic motivation module that will select a reward function for the agent to learn to achieve. We will use a Universal Value Function Approximator (Schaul et al. 2015), that takes as input both the state and the parameters of this reward function as the goal to predict the value function (or action-value function) to generalize across these goals. This module will be trained to generate goals such that the agent's learning is maximized. Thus, this is also a method for automatic curriculum learning.

## Introduction

Reinforcement Learning (Sutton and Barto 1998) is widely studied as a method to train an agent by interaction with it's environment. There have been a wide variety of successes of this paradigm in robotics (Stone, Sutton, and Kuhlmann 2005), computer systems, board games (Tesauro 1995), video games (Mnih et al. 2015), online web services, etc.

An RL agent learns by trying to maximize a scalar reward function that the environment provides it. In most cases, a human has to hand-engineer a suitable reward signal for the agent to maximize, or the application itself needs to have an intuitive reward signal (the score in video games, for example). However, it is hard for an agent to learn a meaningful policy if this reward signal is sparse, or non-existent. This is evidenced by games that are difficult even for recent Deep Learning augmented systems to learn, like Montezuma's Revenge (Mnih et al. 2015).

Intrinsic Motivation (Barto and Simsek 2005) in Reinforcement Learning is a method to allow an agent to generate its own reward function. Such an intrinsic reward is a generated reward signal that is designed to facilitate learning a wide variety of problems, rather than learning to optimize a single one. If designed appropriately, such intrinsic rewards can act as a curriculum for an agent to bootstrap itself and learn faster. It can also be used to learn skills or options (Sutton, Precup, and Singh 1999) that can then be used to learn any task of interest in that environment faster (Chentanez, Barto, and Singh 2005).



In this project, we propose to learn a goal generator that will select or generate goals (reward signals) for the RL agent to learn to achieve. This generator will learn to generate goals that the agent will learn from most. Instead of using these intrinsic motivations to learn separate options, we propose to use Universal Value Function Approximators (Schaul et al. 2015), which are function approximators designed to generalize across states and goals.

There has been recent work in using UVFAs to generalize across goals (Held et al. 2017; Andrychowicz et al. 2017; Cabi et al. 2017) and for automatic curriculum learning (Graves et al. 2017; Matiisen et al. 2017; Sukhbaatar et al. 2017). We believe this area is a novel and interesting one with a lot more issues to explore.

## Background

A Markov Decision Process is a tuple $\langle S, A, T, R_0, \gamma \rangle$, where $S$ is the set of states, $A$ is the set of available actions, $T(s_{t+1}|s_t, a_t)$ is the state transition probability for $s_t, s_{t+1} \in S, a_t \in A$, $\gamma \in [0, 1)$ is the discounting factor and $R_0(s_t, a_t, s_{t+1})$ is a scalar reward function.

A policy $\pi(a|s)$ is defined as the probability of taking an action $a$ in state $s$. The value of a state $s$ given a reward function is defined as

$$V_\pi(s_t) \doteq \mathbb{E}\left[\sum_{i=t}^{\infty} \gamma^{i-t} R_0(s_i, a_i, s_{i+1})\right] \tag{1}$$

where the actions are taken according to policy $\pi$ and the next state is drawn from $T(s_{t+1}|s_t, a_t)$. The corresponding action value is defined as:

$$Q(s_t, a_t) \doteq \mathbb{E}[R_0(s_t, a_t, s_{t+1}) + \gamma V_\pi(s_{t+1})] \tag{2}$$

We redefine reward $R(s, a, s', g)$ as a function over state $s \sim S$, action $a \sim \pi(s)$, next state $s' \sim T(s'|s, a)$ and goal $g \sim P(G)$. Here, $G$ can be the set of states $S$ for single goal systems or a set of all possible goals, where a viable goal would be any subset of $S$. The value and action-value functions correspondingly change to reflect the dependence on the goal.

We will restrict ourselves right now to goals that are directly comparable to states. $R(s, a, s', g)$ can then be defined either as a sparse reward signal, which checks if $s' = g$, or

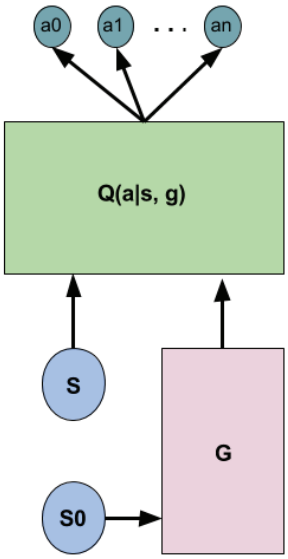


Figure 1: The goal generator generates a goal, and the Q function approximates action values for all actions conditioned on the current state and the given goal

as a distance or similarity measure to the goal state, for a more continuous signal.

## Adversarial Goal Generation

Consider an agent that has two components, value prediction and goal generation. The goal generator will draw goals from the distribution $P(G|\omega)$, and will be able influence this distribution using parameters $\omega$. The value prediction, $\hat{V}(s, g|\theta)$, will approximate the true value of a state for a given goal and will do so using parameters $\theta$.

$V^*(s, g)$ is the optimal value of the state $s$ for the goal $g$ on following the optimal policy $\pi(a|s, g)$ for that goal, defined by the Bellman equation.

$$V^*(s, g) = R(s, a, s', g) + \gamma V^*(s', g) \tag{3}$$

If we consider this goal generation as an adversarial game, then the generator can try to generate goals such that the agent's value prediction error is maximized. The objective that will be optimized is:

$$\max_{\omega} \min_{\theta} \sum_{s \sim S, g \sim P(G|\omega)} \|V^*(s, g) - \hat{V}(s, g|\theta)\|^2 \tag{4}$$

So in (4), we want the agent to learn to predict the optimal value function as defined in (3), or minimize its error with respect to this true value function.

However, knowing the optimal value function beforehand is unlikely. So instead we can also have the goal generator generate goals which maximize the change in the agent's value function (Şimşek and Barto 2006), or maximize the mean TD error over transitions for generated goals.

Overall, we feel an intrinsic motivation approach to automatic curriculum learning is very promising. We propose to use an adversarial framework to generate goals for the agent and using UVFAs to generalize this learning across diverse goals.

## References


Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Abbeel, P.; and Zaremba, W. 2017. Hindsight experience replay. *arXiv preprint arXiv:1707.01495*.

Barto, A. G., and Simsek, O. 2005. Intrinsic motivation for reinforcement learning systems. In *Proceedings of the Thirteenth Yale Workshop on Adaptive and Learning Systems*, 113–118.

Cabi, S.; Colmenarejo, S. G.; Hoffman, M. W.; Denil, M.; Wang, Z.; and de Freitas, N. 2017. The intentional unintentional agent: Learning to solve many continuous control tasks simultaneously. *arXiv preprint arXiv:1707.03300*.

Chentanez, N.; Barto, A. G.; and Singh, S. P. 2005. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, 1281–1288.

Graves, A.; Bellemare, M. G.; Menick, J.; Munos, R.; and Kavukcuoglu, K. 2017. Automated curriculum learning for neural networks. *arXiv preprint arXiv:1704.03003*.

Held, D.; Geng, X.; Florensa, C.; and Abbeel, P. 2017. Automatic goal generation for reinforcement learning agents. *arXiv preprint arXiv:1705.06366*.

Matiisen, T.; Oliver, A.; Cohen, T.; and Schulman, J. 2017. Teacher-student curriculum learning. *arXiv preprint arXiv:1707.00183*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Schaul, T.; Horgan, D.; Gregor, K.; and Silver, D. 2015. Universal value function approximators. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 1312–1320.

Şimşek, Ö., and Barto, A. G. 2006. An intrinsic reward mechanism for efficient exploration. In *Proceedings of the 23rd international conference on Machine learning*, 833–840. ACM.

Stone, P.; Sutton, R. S.; and Kuhlmann, G. 2005. Reinforcement learning for robocup soccer keepaway. *Adaptive Behavior* 13(3):165–188.

Sukhbaatar, S.; Kostrikov, I.; Szlam, A.; and Fergus, R. 2017. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407*.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112(1-2):181–211.

Tesauro, G. 1995. Temporal difference learning and td-gammon. *Communications of the ACM* 38(3):58–68.