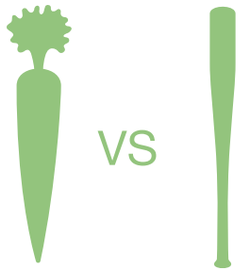
# Designing Rewards for Fast Learning

Henry Sowerby, Zhiyuan Zhou, Michael Littman henry_sowerby@brown.edu

paper on Arxiv



**What are principles of good reward design?**

We look for rewards that encourage a certain behavior while resulting in fast learning.

## Reward design principles

- **Penalizing step** > rewarding goal
- **Subgoal** rewards should **gradually increase towards the goal**
- Dense reward is only good when designed carefully

## Background

Visitations

$$D(s,i) = F(s,i) + \gamma \sum_{s'} T(s, \pi^{+}(s), s') \cdot D(s', i)$$

$$D_a(s,i) = F(s,i) + \gamma \sum_{s'} T(s, a, s') \cdot D(s', i)$$

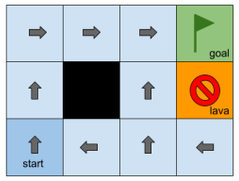
Expressing the value functions

$$V^{\pi^{+}}(s) = \sum_i D(s,i) \cdot R(i)$$

$$Q^{\pi^{+}}(s,a) = \sum_i D_a(s,i) \cdot R(i)$$

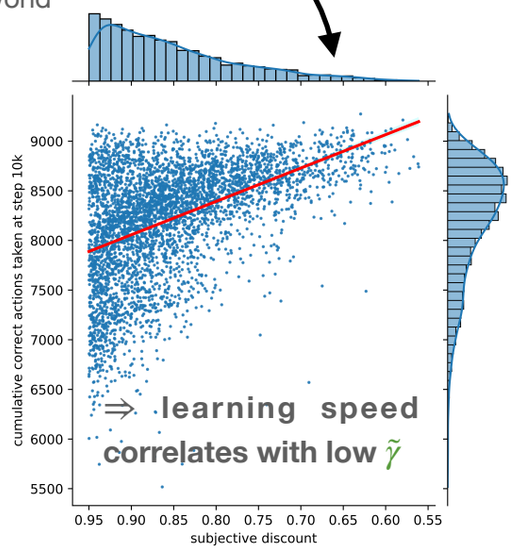
**Good rewards have big action gaps and small "subjective discounts"**

## Not all rewards are born equal…




Russell/Norvig grid world

Sample random reward functions that lead to correct behavior




## Algorithm to design rewards

*Find a reward that **maintains the correct behavior** while maximizing $\delta$ with linear programming and minimizing $\tilde{\gamma}$ through binary search:*

$$\sum_i D(s,i) \cdot R(i) \geq \sum_i D_a(s,i) \cdot R(i) + \delta$$

$$\sum_i \widetilde{D}(s,i) \cdot R(i) \geq \sum_i \widetilde{D}_a(s,i) \cdot R(i) + \delta$$

$$-1 \leq R(i) \leq 1$$

$$\forall\, i, s \in S, a \in A \backslash \{\pi^{+}(s)\}$$

## Action gap & subjective discount

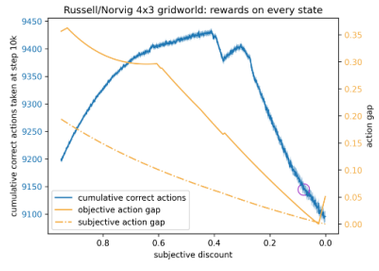
**Action gap** $\delta = Q(s, a_{optimal}) - Q(s, a_{second})$.
Large action gaps are beneficial for learning.

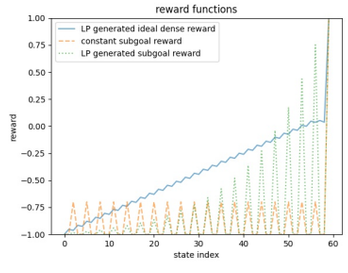
The **objective discount** is a constant of the environment, while the **subjective discount** is a property of the reward function:

$$\tilde{\gamma} = \min \{\tilde{\gamma} : \pi_R^{\gamma'} = \pi^{target}, \forall \gamma' \in [\tilde{\gamma}, \gamma]\}$$

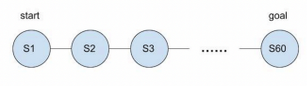
It is the **smallest discount factor that still encourages the target behavior**.

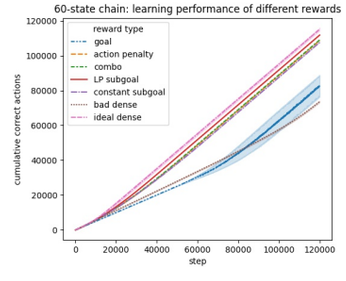
## LP-designed rewards and faster learning




The environment: chain




- $\delta$ and $\tilde{\gamma}$ tradeoff
- $\tilde{\gamma}$ as a regularization
- Dense reward shape
- Increasing subgoal reward