Reinforcement Learning from Demonstration through Shaping

Tim Brys and **Anna Harutyunyan**
Vrije Universiteit Brussel
{timbrys, aharutyu}@vub.ac.be

Halit Bener Suay and **Sonia Chernova**
Worcester Polytechnic Institute
{benersuay, soniac}@wpi.edu

Matthew E. Taylor
Washington State University
taylorm@eecs.wsu.edu

Ann Nowé
Vrije Universiteit Brussel
anowe@vub.ac.be

IJCAI 2015

Contents

Introduction

- A large number of environment samples are needed before the agent reaches a desirable level of performance.
- Learning from demonstrations (LfD) can directly derive behavior, but it can not guarantee the quality of demonstrations, which hurts the learning behavior.
- We propose to use demonstrations to shape rewards in the RL problems.

Preliminaries - Reinforcement learning

- Q-learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha\delta$$

- TD-error

$$\delta = R(s,a,s') + \gamma \max_{a'} Q(s',a') - Q(s,a)$$

Preliminaries - Reward shaping

Modifying the reward function may make the agent coverge to suboptimal policies.

- The extra reward **F** is added to the environment's reward R to create a new composite reward signal:

$$R_F(s, a, s') = R(s, a, s') + F(s, a, s')$$

- Define potential function $\Phi : S \to R$, and take **F** as follows, the total policies remain unchanged.

Ng et al, 1999

$$F(s, a, s') = \gamma\Phi(s') - \Phi(s)$$

Wiewiora et al, 2003

$$F(s, a, s', a') = \gamma\Phi(s', a') - \Phi(s, a)$$

Preliminaries - Reward shaping

(Ng et al, ICML, 1999)

$$\left(s_1 \to s_2 \to s_3 \cdots \to s_n \to s_1 \cdots\right)$$

"distracted" problem

$$F(s_1, a_1, s_2) + \cdots + F(s_{\text{n-1}}, a_{\text{n-1}}, s_{\text{n}}) + F(s_{\text{n}}, a_{\text{n}}, s_1) > 0$$

$$\sum R_F(s,a,s') = \sum R(s,a,s') + \underbrace{\sum F(s,a,s')}_{=0}$$

$$= \sum R(s,a,s') + \underbrace{\sum \gamma\Phi(s') - \Phi(s)}_{=0}$$

$$\max_{\pi} \sum R_F(s,a,s') \Leftrightarrow \max_{\pi} \sum R(s,a,s')$$

Shaping RL using Demonstrations

- Key idea:
  - We want the potential $\Phi^D(s,a)$ of a state-action pair **(s, a)** to be high when action **a** was demonstrated in a state s^d similar to **s** .
  - We want the potential to be low when the action was not demonstrated in the neighbourhood of **s**.
- Similarity:

$$g(s, s^d, \Sigma) = e^{\left(-\frac{1}{2}(s-s^d)^T \Sigma^{-1}(s-s^d)\right)}$$

  - where Σ is a covariance matrix. If two state-action pairs differ in the action, their similarity is 0, and the similarity is 1 when $s = s^d$.

Shaping RL using Demonstrations

- The potential function:

$$\Phi^D(s,a) = \max_{(s^d,a)} g(s, s^d, \Sigma)$$

- This potential function can then be integrated in two ways into the learning process

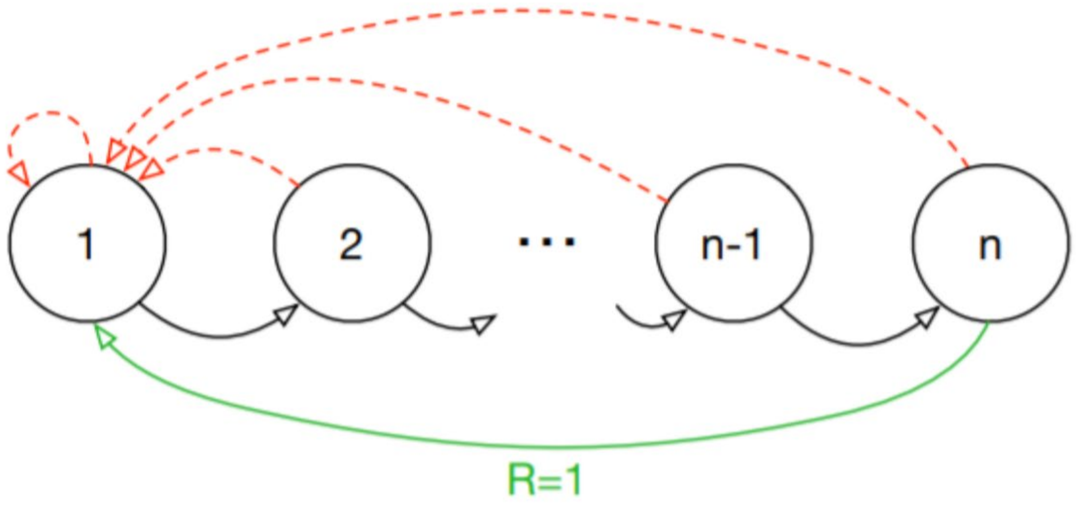
1. By creating a shaping function and adding it to the base reward.

$$F^D(s,a,s',a') = \gamma \Phi^D(s',a') - \Phi^D(s,a)$$

2. Initializing the Q function with potential function

$$Q_0(s,a) = \Phi^D(s,a)$$

Shaping RL using Demonstrations - Example: Blind Cliffwalk



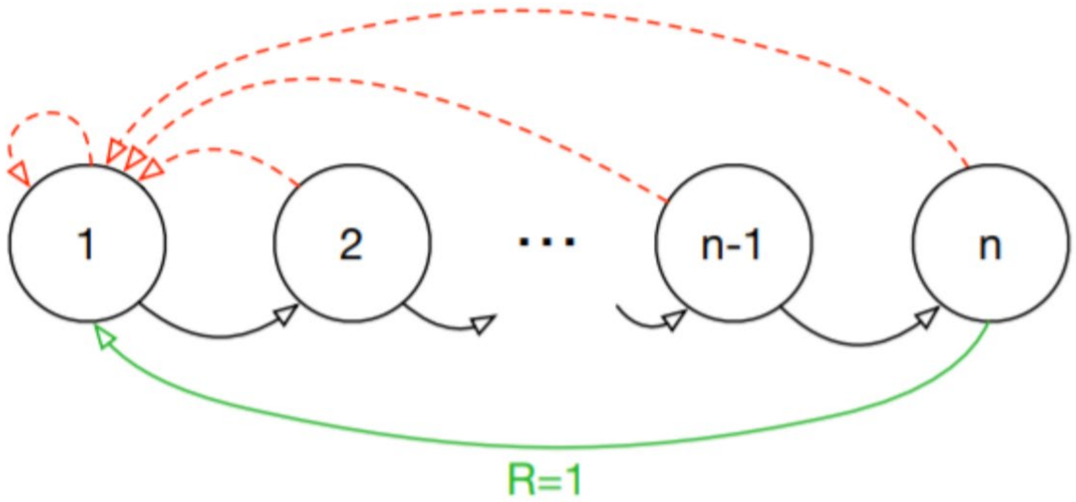
$$\{(1,R),(2,R),\cdots,(n-1,R),(n,L)\}$$

$$Q_0(s,a) = \Phi^D(s,a)$$

$$\{Q_0(1,R)=1, Q_0(2,R)=1,\cdots,Q_0(n-1,R)=1, Q_0(n,L)=1\}$$

This initialization allows the agent to immediately use the bias in action selection.

Shaping RL using Demonstrations - Example: Blind Cliffwalk



$$\{(1,R),(2,R),\cdots,(n-1,R),(n,L)\}$$

$$F^D(s,a,s',a') = \gamma\Phi^D(s',a') - \Phi^D(s,a)$$

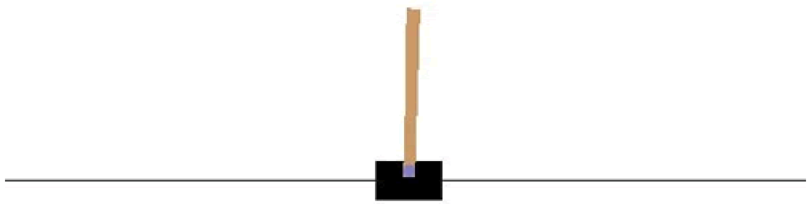
$$R(s,a,s') = \begin{cases} 1 & ,\mathrm{s = n\ \&\ s'=1} \\ 0 & ,\mathrm{otherwise} \end{cases}$$

$$R_F(s,a,s') = \begin{cases} 1 & ,\text{Go right for the first time} \\ 0 & ,\text{Go right} \\ \text{-}1 & ,\text{Go left} \end{cases}$$

$$R_F(s,a,s') = \begin{cases} <1 & ,\text{Go right for the first time} \\ <0 & ,\text{Go right} \\ \text{-}1 & ,\text{Go left} \end{cases}$$

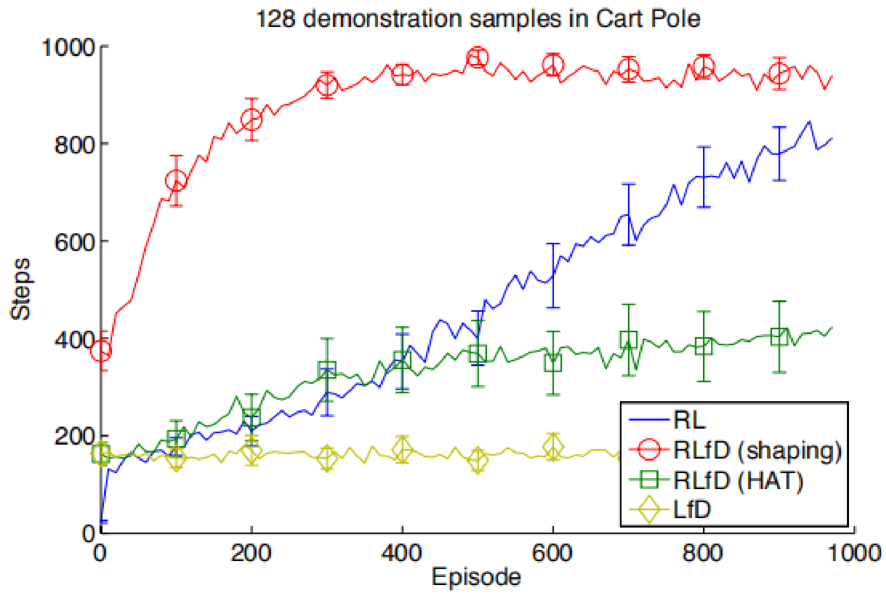
Experiments

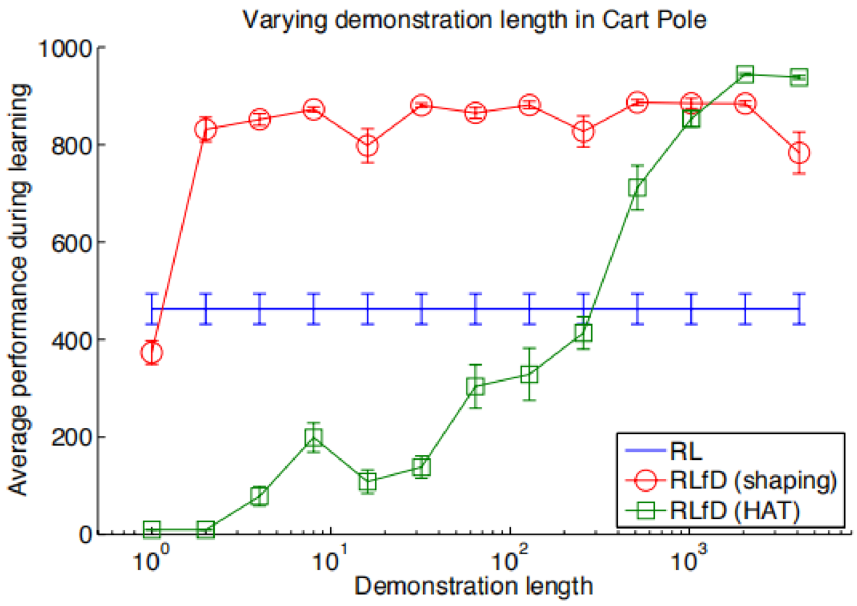
- Environments
  - Cart Pole
  - Super Mario Bros game



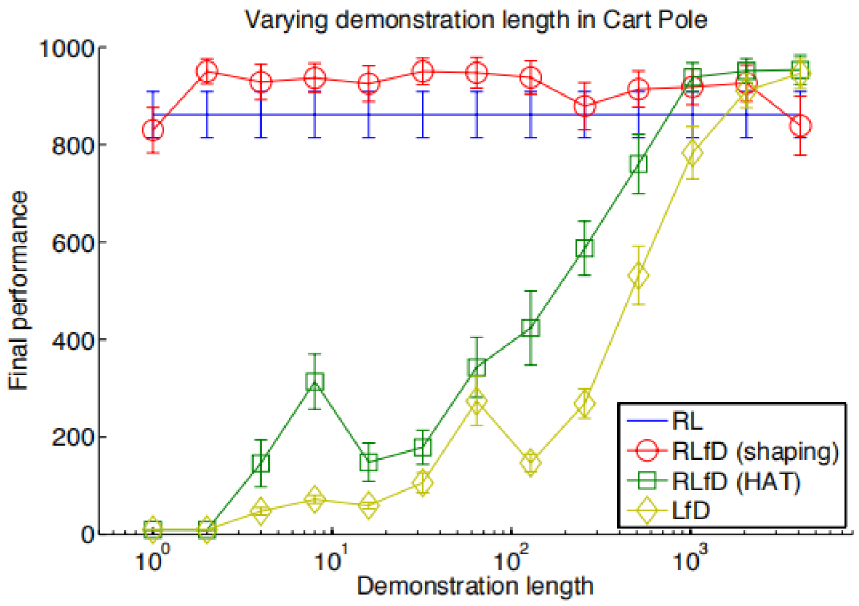
- Comparisons
  - RL
  - RLfD (shaping)
  - RLfD (HAT)
  - LfD

Experiments - CartPole



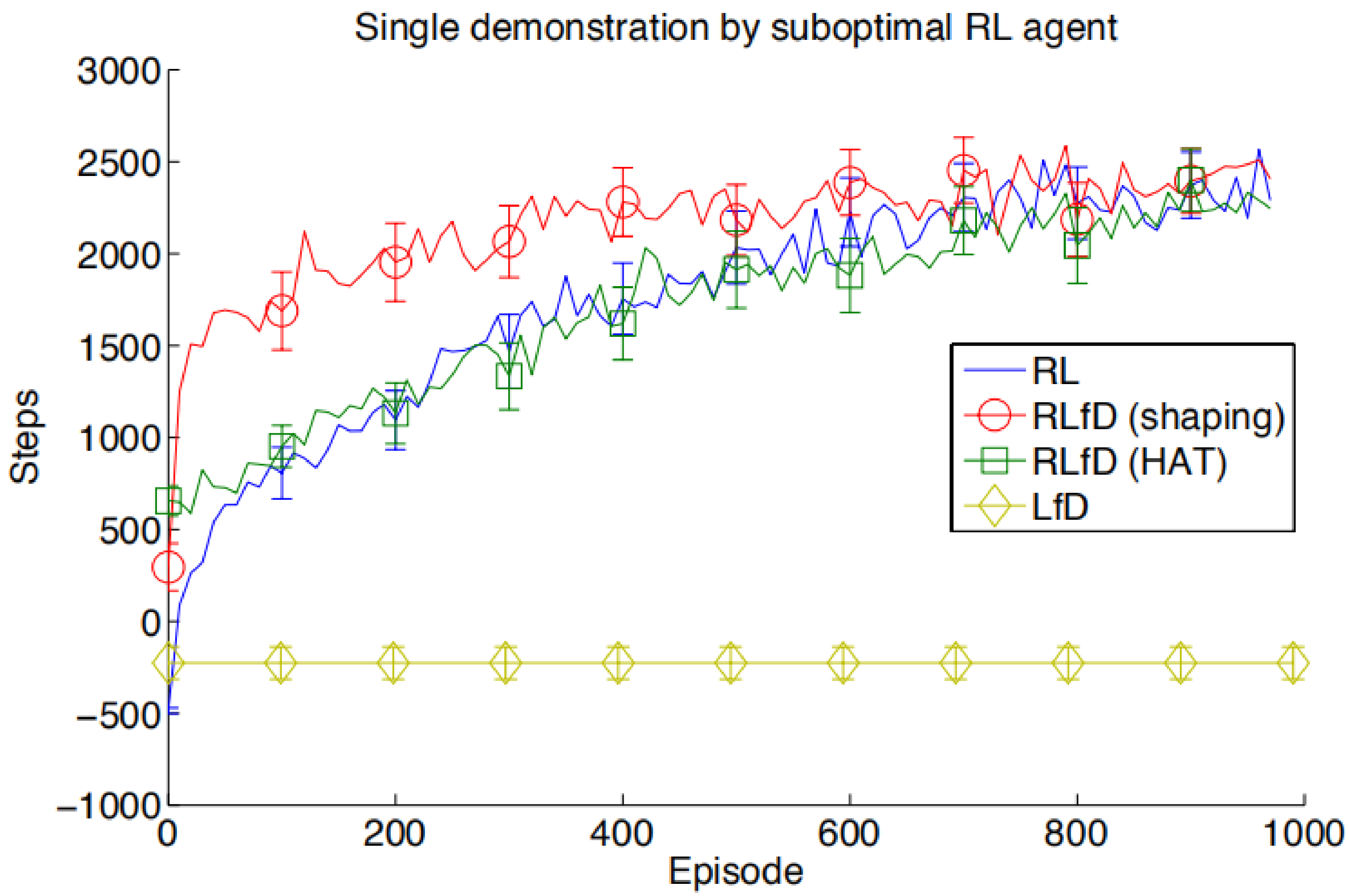


(a)



(b)

Experiments - Mario



Experiments - Mario

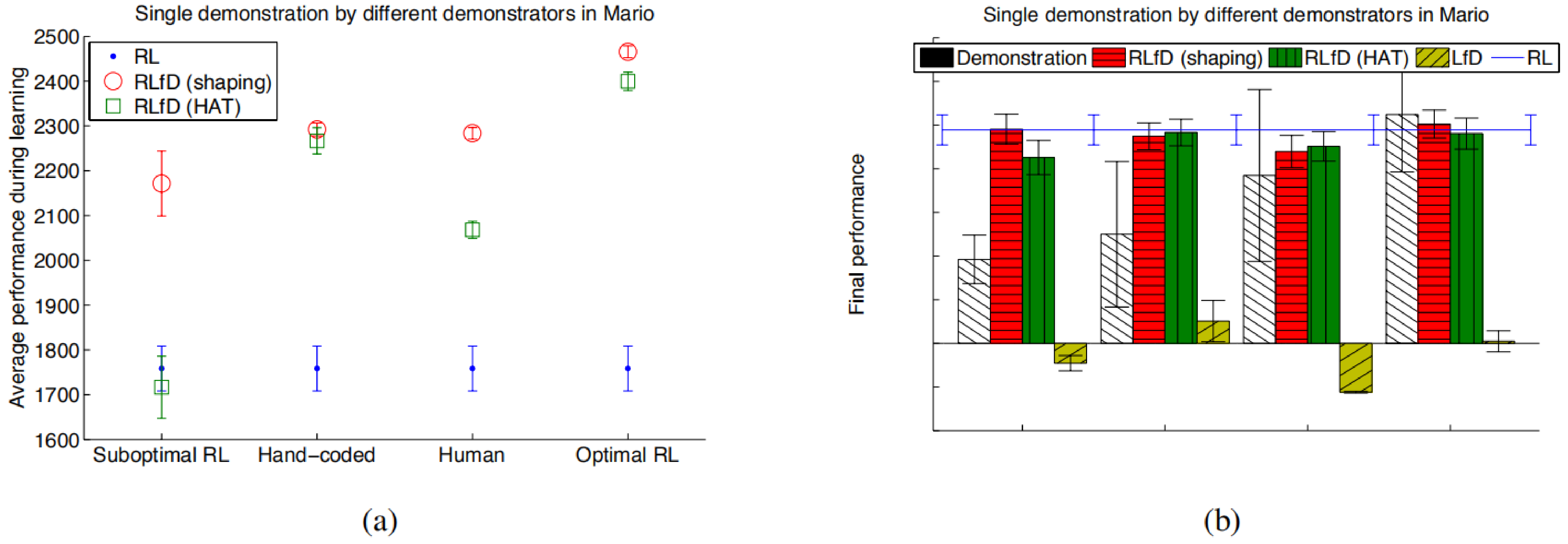


Figure 4: The effect the type of demonstrator has on RLfD and LfD in Mario (RL performance provided for comparison). Figure (a) shows average performance over 1000 learning episodes, an indication of the speed of learning (excluding LfD), (b) shows the final performance (after 1000 learning episodes) of the policies proposed by each technique. RLfD (shaping) always outperforms or matches the performance of other techniques.