Different Approaches for Reward Shaping

Paniz Behboudian

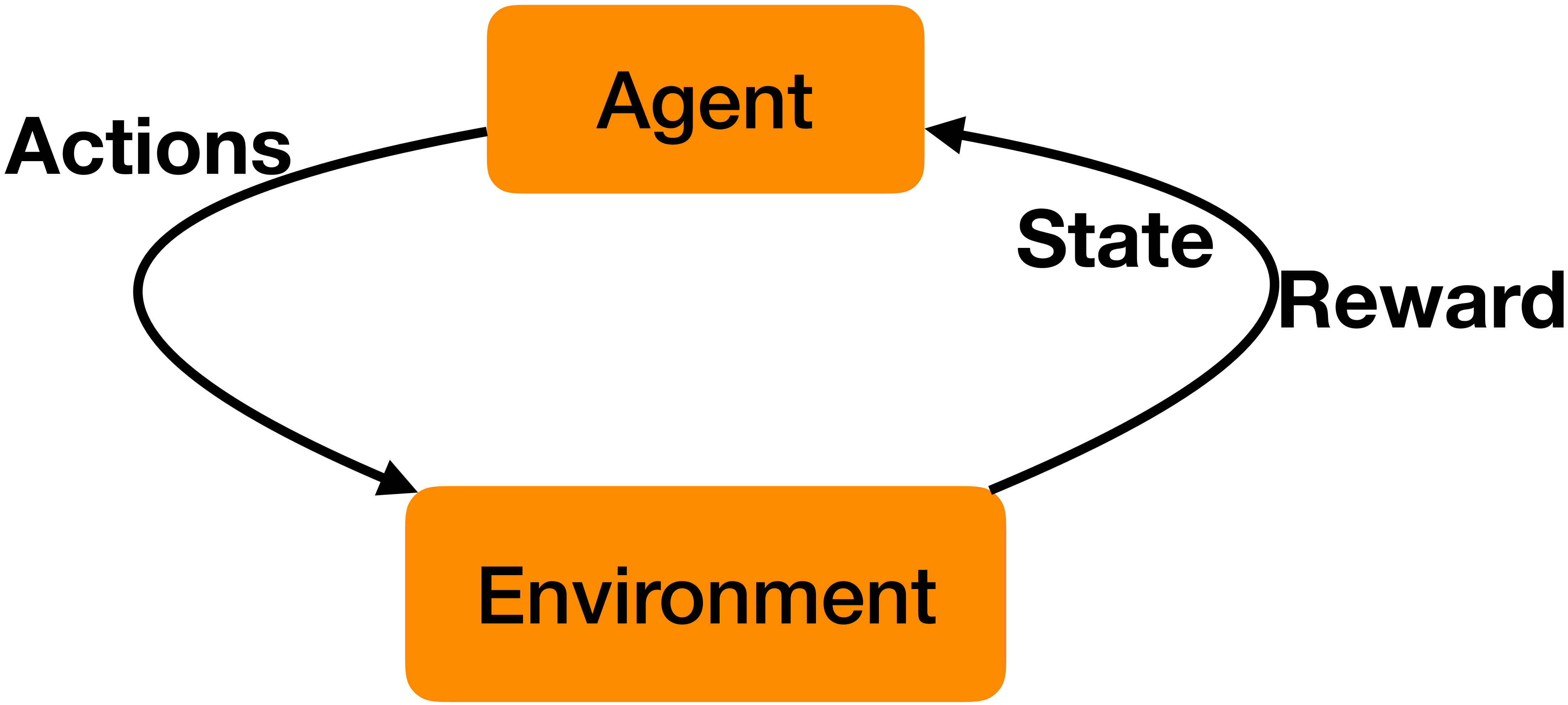
Joint work with: Yash Satsangi, Matthew E. Taylor, Michael Bowling

Summer 2019

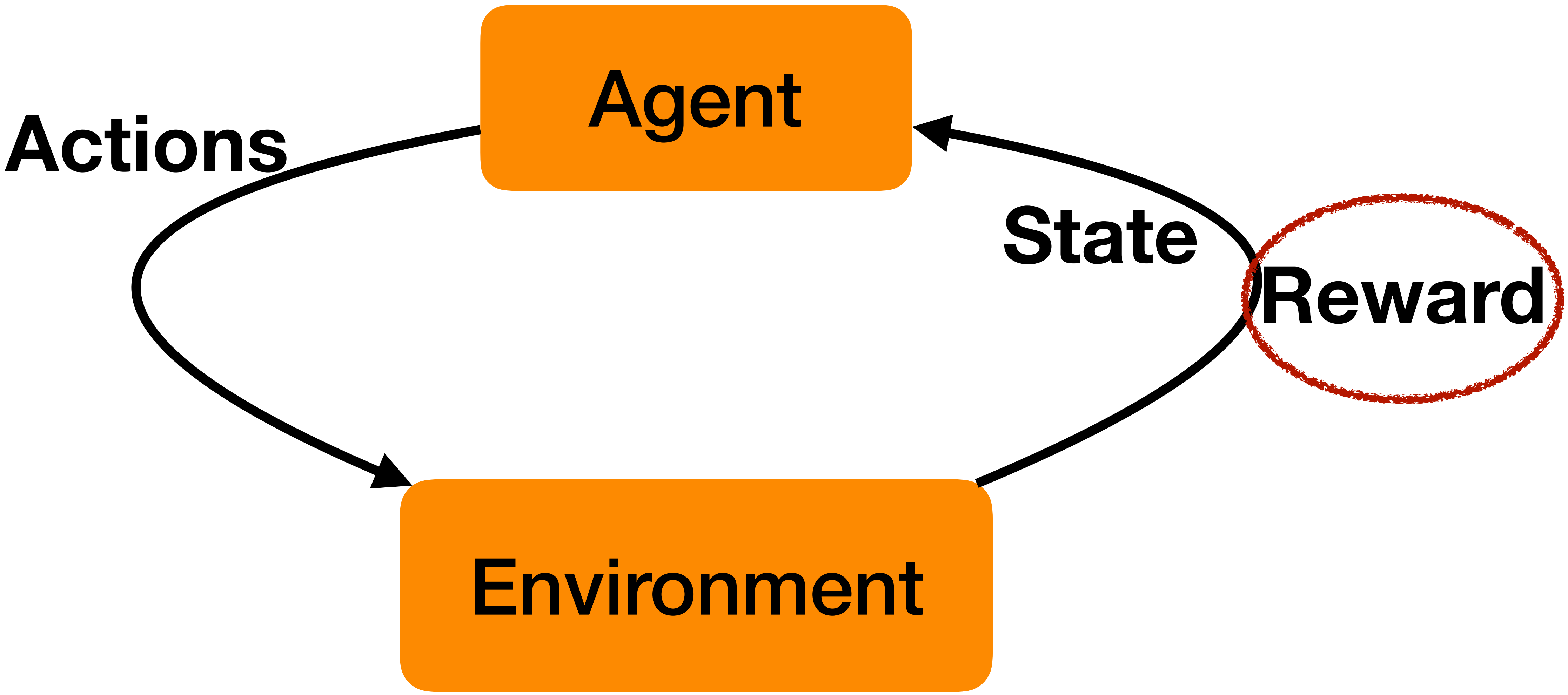
Outline

- Motivation
- Reward Shaping without Constraints
- Potential-Based Reward Shaping (PBRS)
    - State Potentials
    - State-Action Potentials
- Dynamic Potential-Based Reward Shaping
    - Transforming any Signal into PBRS
- The Problem with Transforming any Signal into PBRS
- One Possible Solution

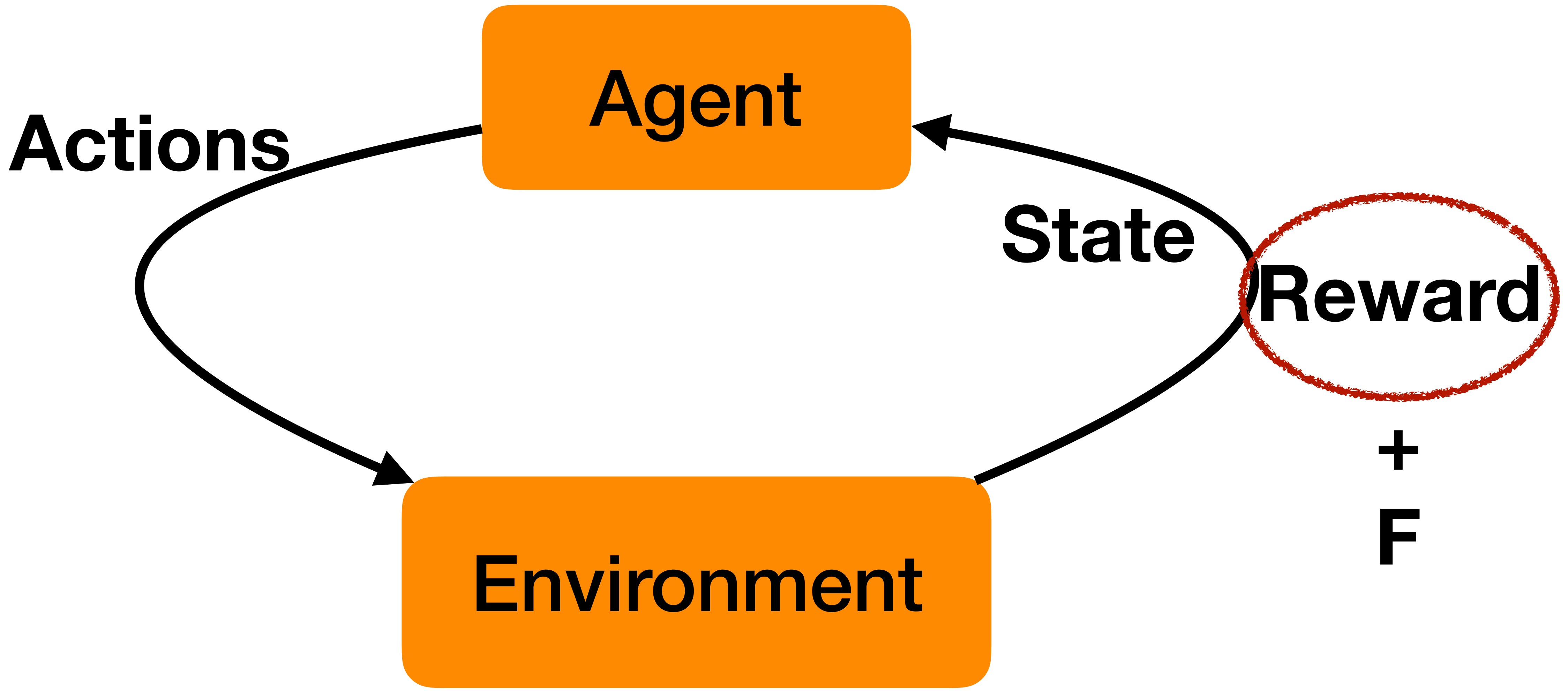
Motivation



Motivation



Motivation



Markov Decision Process (MDP)

$$M = \langle S, A, P, \gamma, R \rangle, \qquad \pi(s) : S \rightarrow A$$

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^{k} R(s_{t+k}, a_{t+k}) \,|\, s_t = s, a_t = a, \pi\right]$$

Markov Decision Process (MDP)

$$M = \langle S, A, P, \gamma, R \rangle, \qquad \pi(s) : S \to A$$

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}) \,|\, s_t = s, a_t = a, \pi\right]$$

$$Q^*(s, a) = \max_{\pi \in \prod} Q^{\pi}(s, a), \qquad \pi^*(s) = \arg\max_{a \in A} Q^*(s, a)$$

Shaping in Reinforcement Learning (RL)

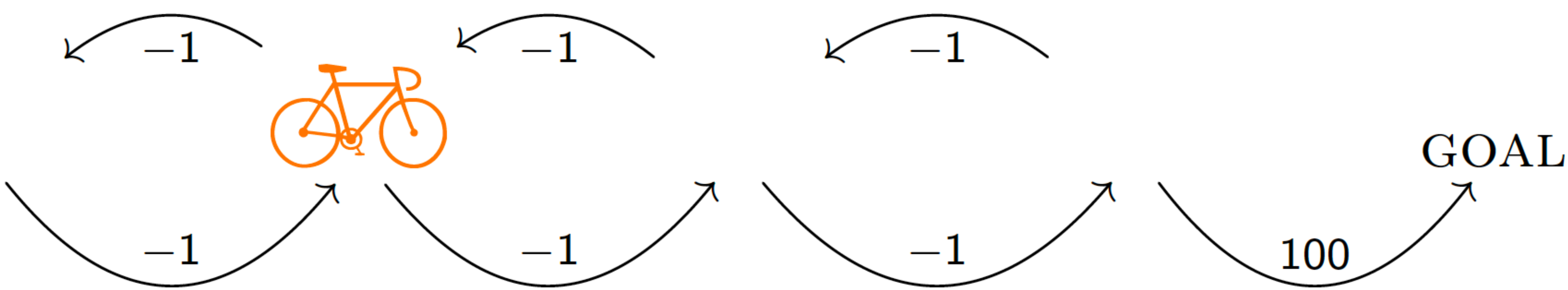
$$R' := R + F$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

Shaping in Reinforcement Learning (RL)

Adding a reward in without constraint [1]:

$$R' := R + F$$

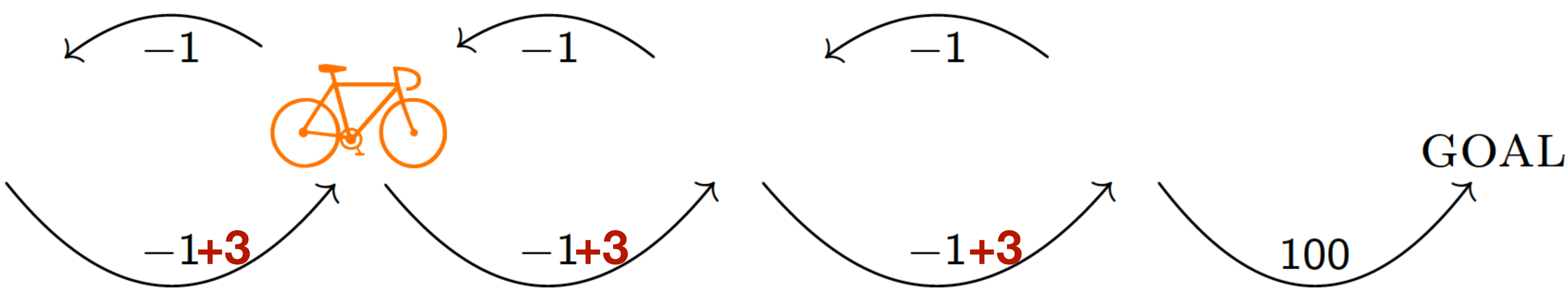


1. Randløv, J., and Alstrøm, P. 1998. Learning to drive a bicycle using reinforcement learning and shaping

Shaping in Reinforcement Learning (RL)

Adding a reward in without constraint [1]:

$$R' := R + F$$

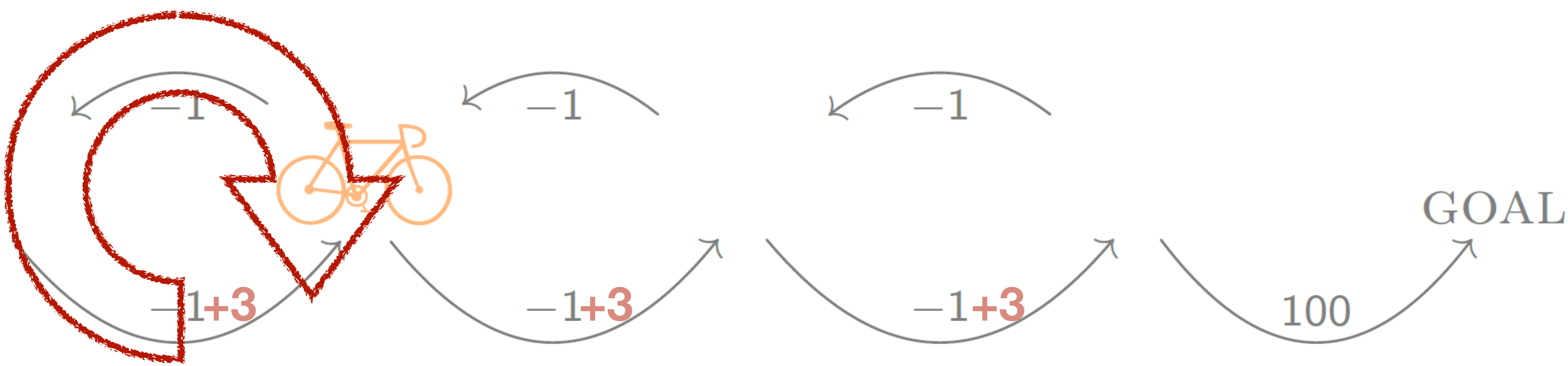


1. Randløv, J., and Alstrøm, P. 1998. Learning to drive a bicycle using reinforcement learning and shaping

Shaping in Reinforcement Learning (RL)

Adding a reward in without constraint [1]:

$$R' := R + F$$



1. Randløv, J., and Alstrøm, P. 1998. Learning to drive a bicycle using reinforcement learning and shaping

Potential-Based Reward Shaping (PBRS)

Constrain with PBRS [2]:

$$R' := R + \overbrace{F}^{\gamma\Phi(s') - \Phi(s)}$$

2. Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping.

Potential-Based Reward Shaping (PBRS)

Constrain with PBRS [2]:

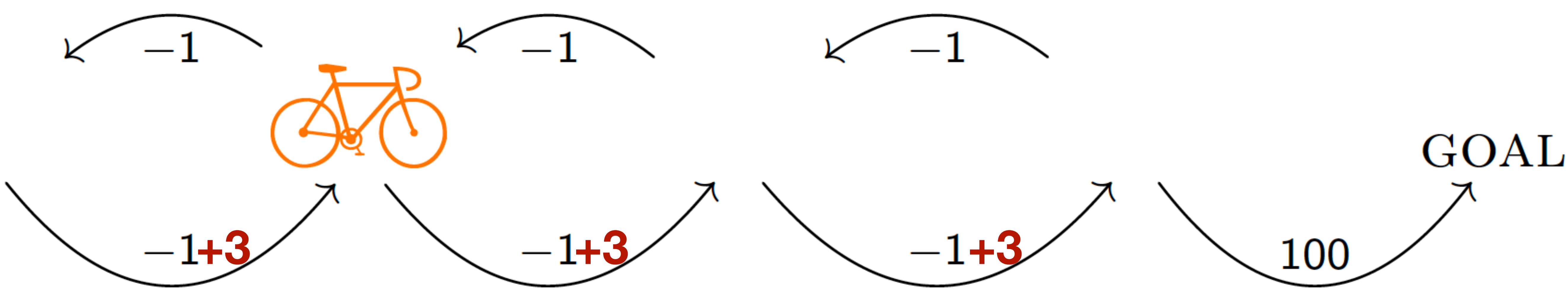
$$R' := R + \overbrace{F}^{\gamma\Phi(s') - \Phi(s)}$$

2. Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping.

Potential-Based Reward Shaping (PBRS)

Constrain with PBRS [2]:

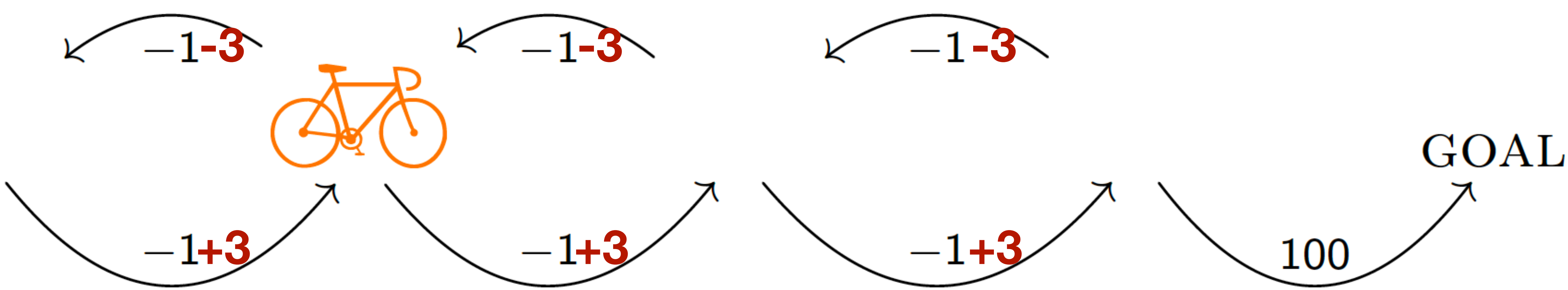
$$R' := R + \overbrace{F}^{\gamma\Phi(s') - \Phi(s)}$$



Potential-Based Reward Shaping (PBRS)

Constrain with PBRS [2]:

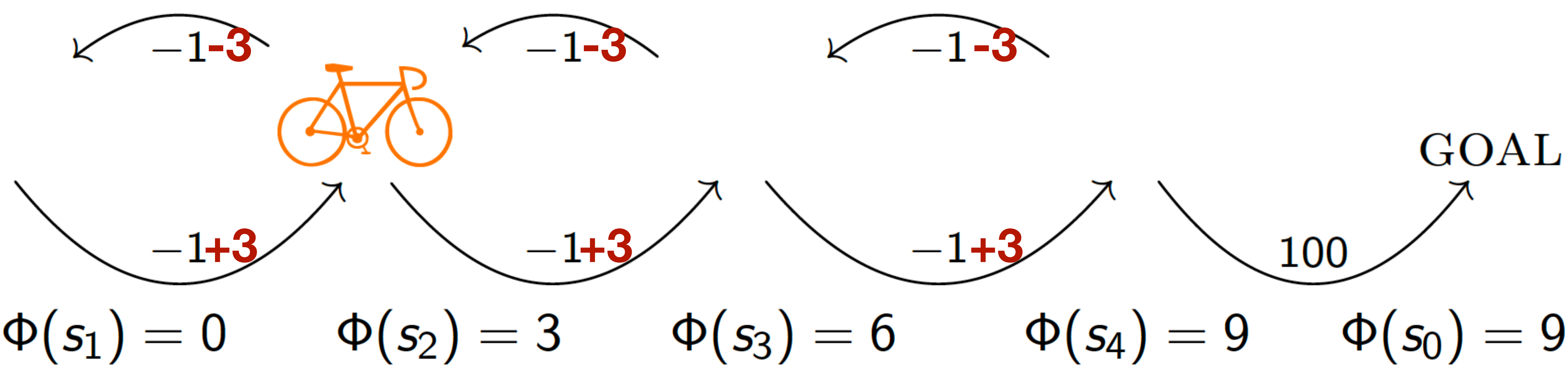
$$R' := R + \overbrace{F}^{\gamma\Phi(s') - \Phi(s)}$$



Potential-Based Reward Shaping (PBRS)

Constrain with PBRS [2]:

$$R' := R + \overbrace{F}^{\gamma\Phi(s') - \Phi(s)}$$



Potential-Based Reward Shaping (PBRS)

Constrain with PBRS:

$$R' := R + \overbrace{F}^{\gamma\Phi(s') - \Phi(s)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \Phi(s)$$

Potential-Based Reward Shaping (PBRS)

Constrain with PBRS:

$$R' := R + \overbrace{F}^{\gamma\Phi(s')-\Phi(s)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \boxed{\Phi(s)}$$

Bias Term

Potential-Based Reward Shaping (PBRS)

Constrain with PBRS:

$$R' := R + \overbrace{F}^{\gamma\Phi(s') - \Phi(s)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \Phi(s)$$

$$\pi^*_M = \pi^*_{M'}$$

State-Action PBRS

Constrain with state-action PBRS [3]:

$$R' := R + \overbrace{F}^{\gamma\Phi(s',a') - \Phi(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi(s,a)$$

$$\pi^*_M \neq \pi^*_{M'}$$

3. Wiewiora, E.; Cottrell, G. W.; and Elkan, C. 2003. Principled methods for advising reinforcement learning agents.

State-Action PBRS

Constrain with state-action PBRS [3]:

$$R' := R + \overbrace{F}^{\gamma\Phi(s',a') - \Phi(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \boxed{\Phi(s,a)}$$

$$\pi^*_M \neq \pi^*_{M'}$$

3. Wiewiora, E.; Cottrell, G. W.; and Elkan, C. 2003. Principled methods for advising reinforcement learning agents.

State-Action PBRS

Constrain with state-action PBRS [3]:

$$R' := R + \overbrace{F}^{\gamma\Phi(s',a') - \Phi(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \Phi(s, a)$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s, a) + \Phi(s, a))$$

3. Wiewiora, E.; Cottrell, G. W.; and Elkan, C. 2003. Principled methods for advising reinforcement learning agents.

State-Action PBRS

Constrain with state-action PBRS [3]:

$$R' := R + \overbrace{F}^{\gamma\Phi(s',a')-\Phi(s,a)}$$

Equivalent to state-action value initialization

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi(s,a)$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s,a) + \Phi(s,a))$$

Dynamic PBRS

- Dynamic PBRS [4]:
  - Used state-based dynamic PBRS in single and multi-agent RL
  - Proved the policy invariance
  - Even before Φ stabilize

4. Devlin, S. M., & Kudenko, D. (2012). Dynamic potential-based reward shaping.

Still need to define Φ

Still need to define Φ

- Expressing any arbitrary rewards as potential-based advice [5]:
  - Dynamic state-action shaping
  - Learning Φ as a value function

5. Harutyunyan, A., Devlin, S., Vrancx, P., & Nowé, A. (2015, February). Expressing arbitrary reward functions as potential-based advice.

Expressing any arbitrary rewards as potential-based advice

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a') - \Phi_t(s,a)}$$

$$\Phi_{t+1}(s, a) := \Phi_t(s, a) + \beta\delta_t^{\Phi}$$

$$\delta_t^{\Phi} := R^{\Phi}(s, a) + \gamma\Phi_{t+1}(s', a') - \Phi_t(s, a)$$

Expressing any arbitrary rewards as potential-based advice

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a')-\Phi_t(s,a)}$$

$$\Phi_{t+1}(s,a) := \Phi_t(s,a) + \beta\delta_t^{\Phi}$$

$$\delta_t^{\Phi} := R^{\Phi}(s,a) + \gamma\Phi_{t+1}(s',a') - \Phi_t(s,a)$$

$$:= -R^{expert}$$

Expressing any arbitrary rewards as potential-based advice

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a') - \Phi_t(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \Phi_0(s, a)$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s, a) + \Phi_0(s, a))$$

Expressing any arbitrary rewards as potential-based advice

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a') - \Phi_t(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \boxed{\Phi_0(s, a)}$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s, a) + \Phi_0(s, a))$$

Expressing any arbitrary rewards as potential-based advice

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a')-\Phi_t(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

0

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \cancel{\Phi_0(s, a)}$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s, a) + \Phi_0(s, a))$$

Expressing any arbitrary rewards as potential-based advice

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a') - \Phi_t(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

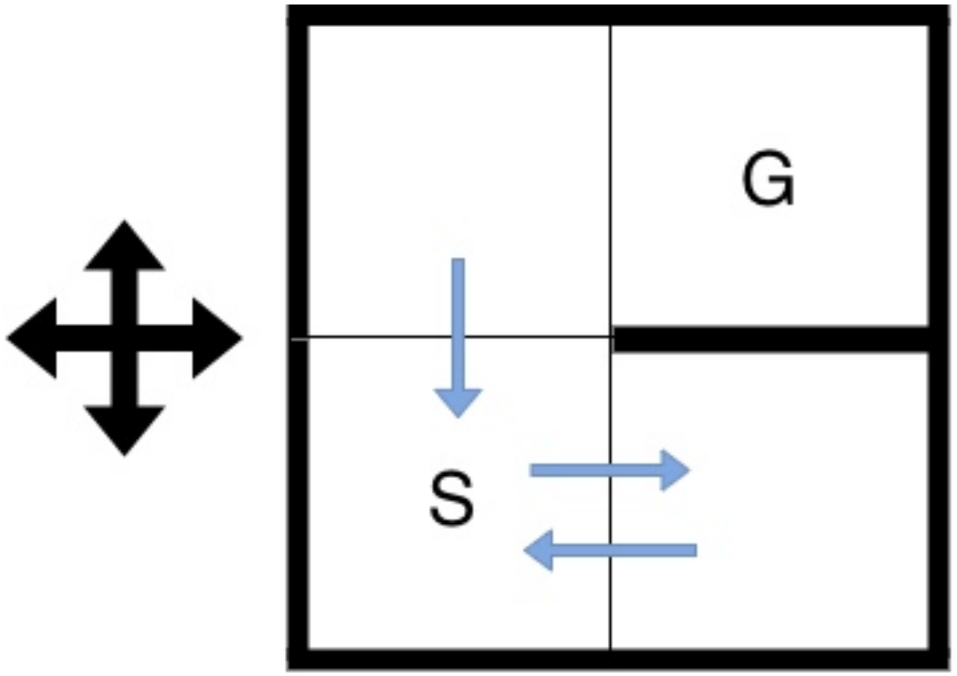
0

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \cancel{\Phi_0}(s,a)$$

$$\pi^*_M = \pi^*_{M'}$$

Experiments

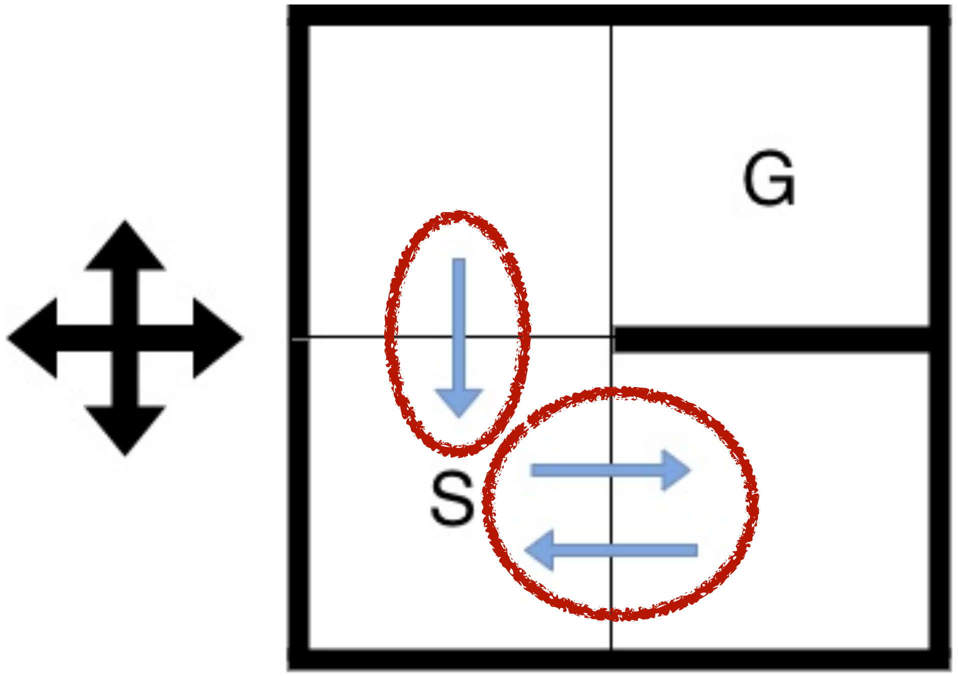
$$R(s, s') := \begin{cases} 1 & \text{if } s' = G \\ 0 & \text{o.w.} \end{cases}$$



$$R_{expert}(s, s') := -||next\_state(s) - s'||$$

Experiments

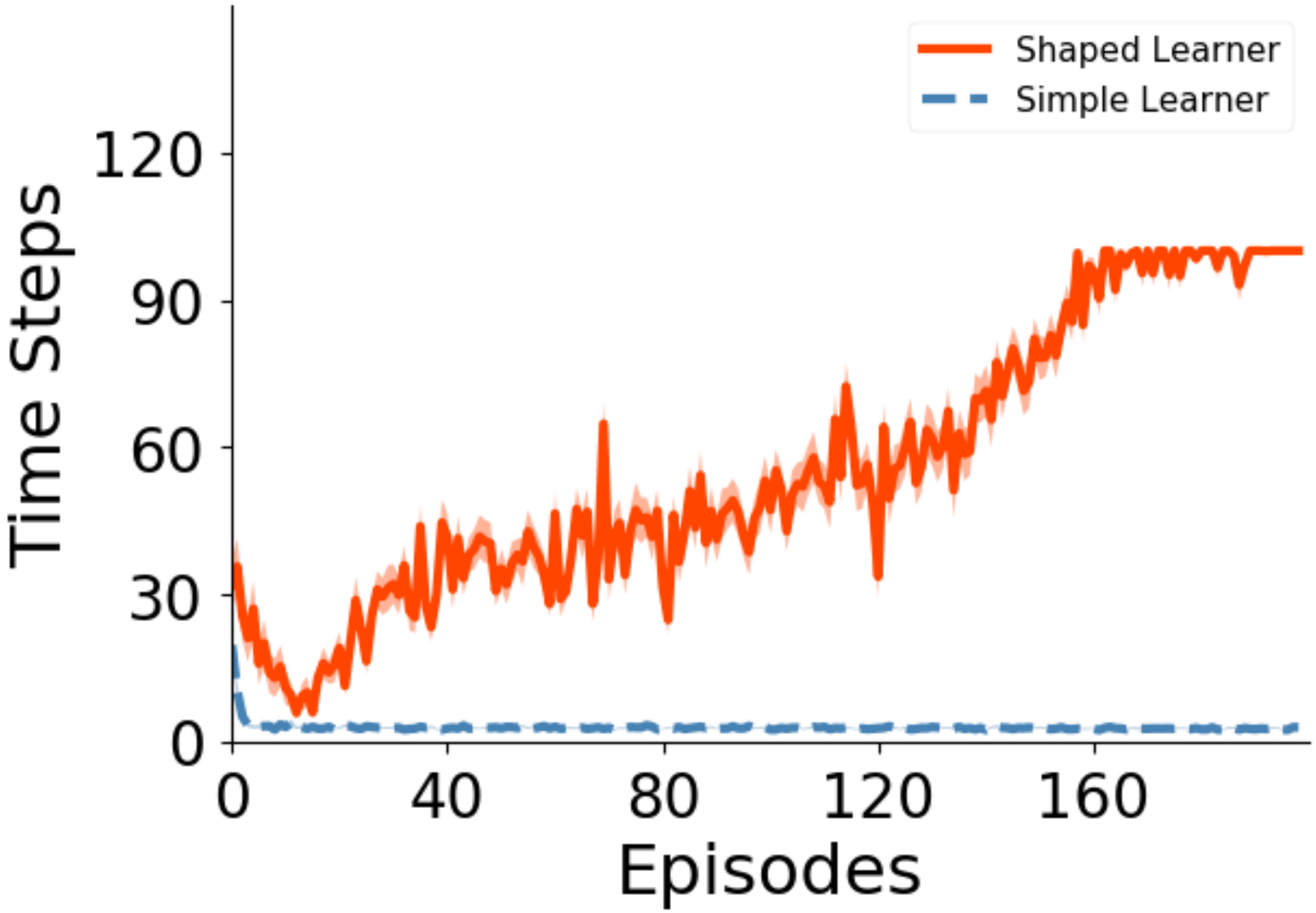
$$R(s, s') := \begin{cases} 1 & \text{if } s' = G \\ 0 & \text{o.w.} \end{cases}$$

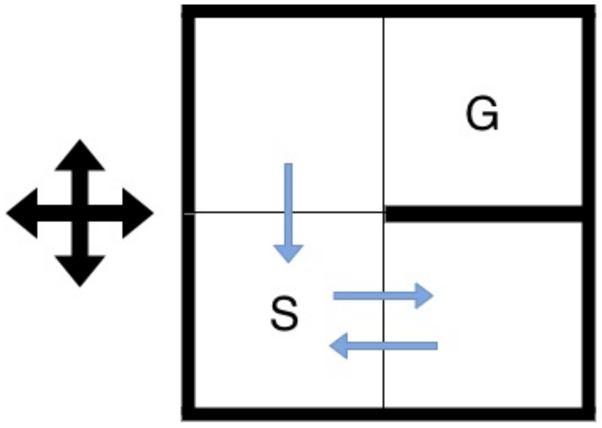


$$R_{expert}(s, s') := -||next\_state(s) - s'||$$

Experiments: Dynamic PBRS

Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy





Expressing any arbitrary rewards as potential-based advice

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a') - \Phi_t(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi_0(s,a)$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s,a) + \Phi_0(s,a))$$

The wrong bias

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a') - \Phi_t(s,a)}$$

$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \boxed{\Phi_0(s, a)}$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s, a) + \boxed{\Phi_0(s, a)})$$

The corrected bias

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a')-\Phi_t(s,a)}$$

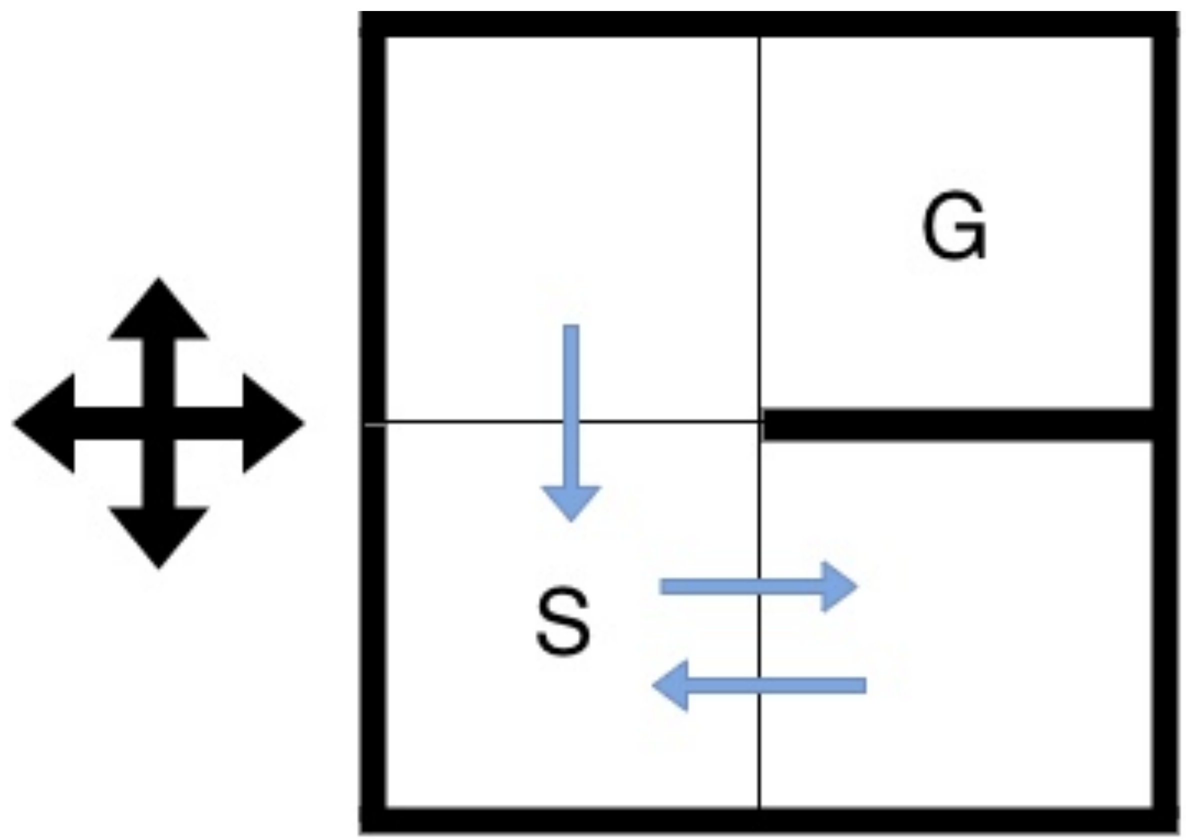
$$M = \langle S, A, P, \gamma, R \rangle \longrightarrow M' = \langle S, A, P, \gamma, R' \rangle$$

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \Phi_t(s, a)$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s, a) + \Phi_t(s, a))$$

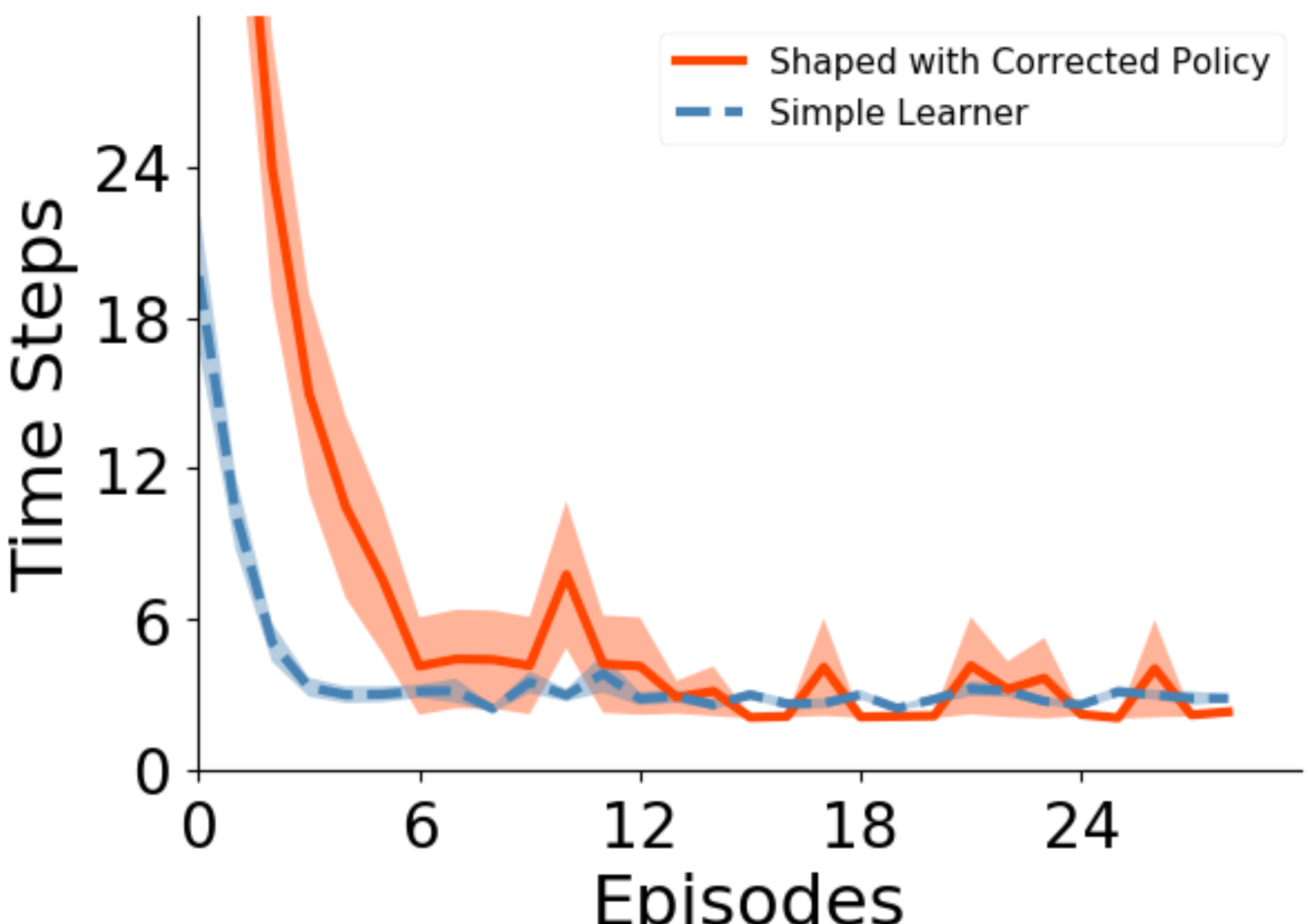
Experiments

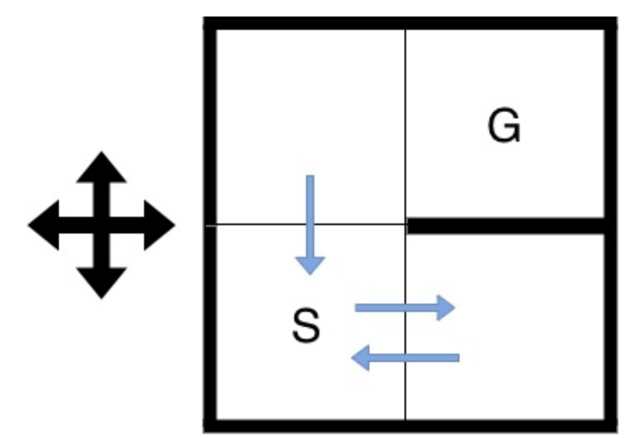
Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy



Experiments: Corrected Dynamic PBRS

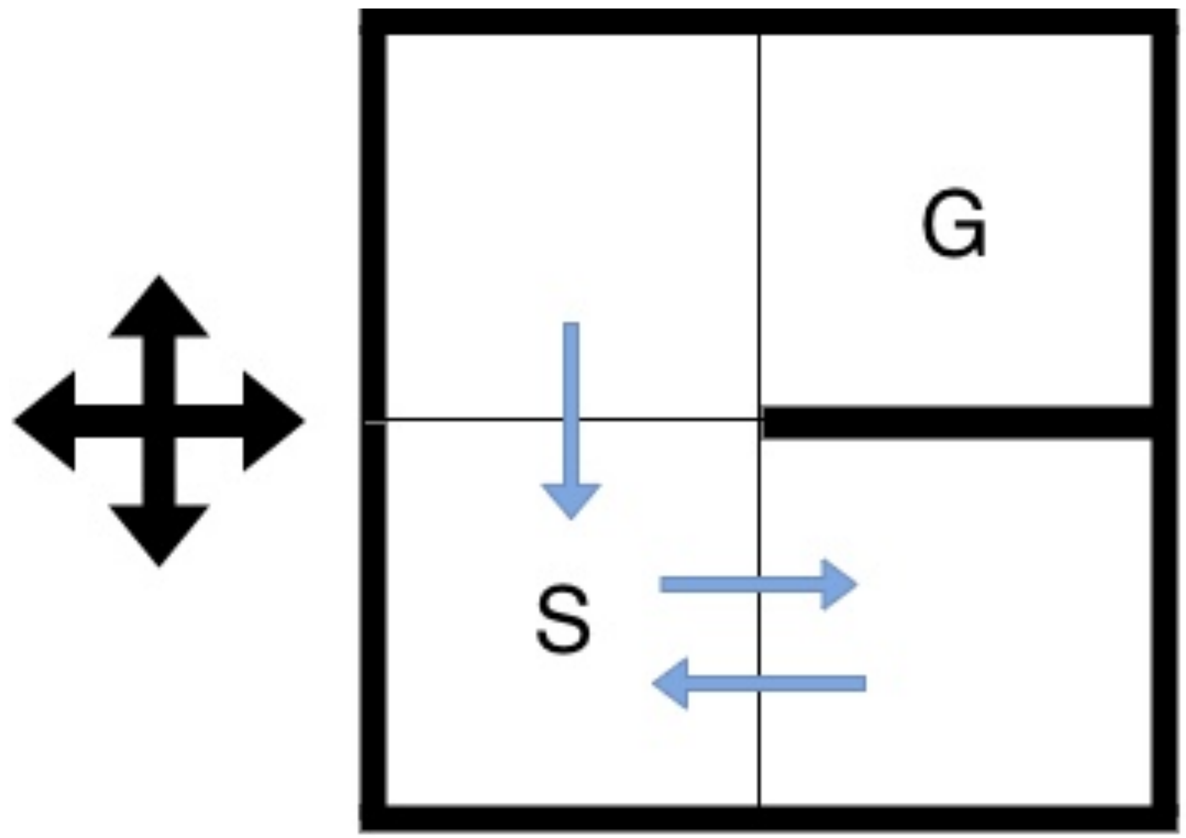
Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy





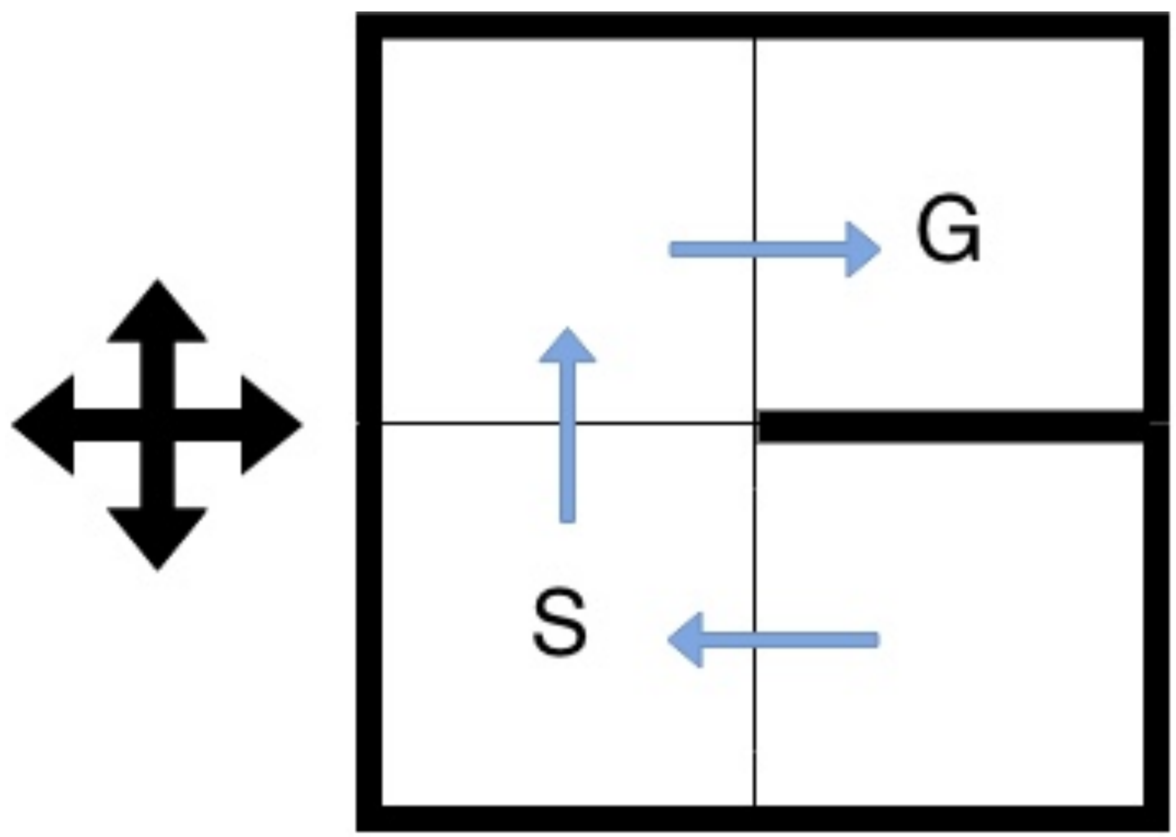
Experiments

Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy



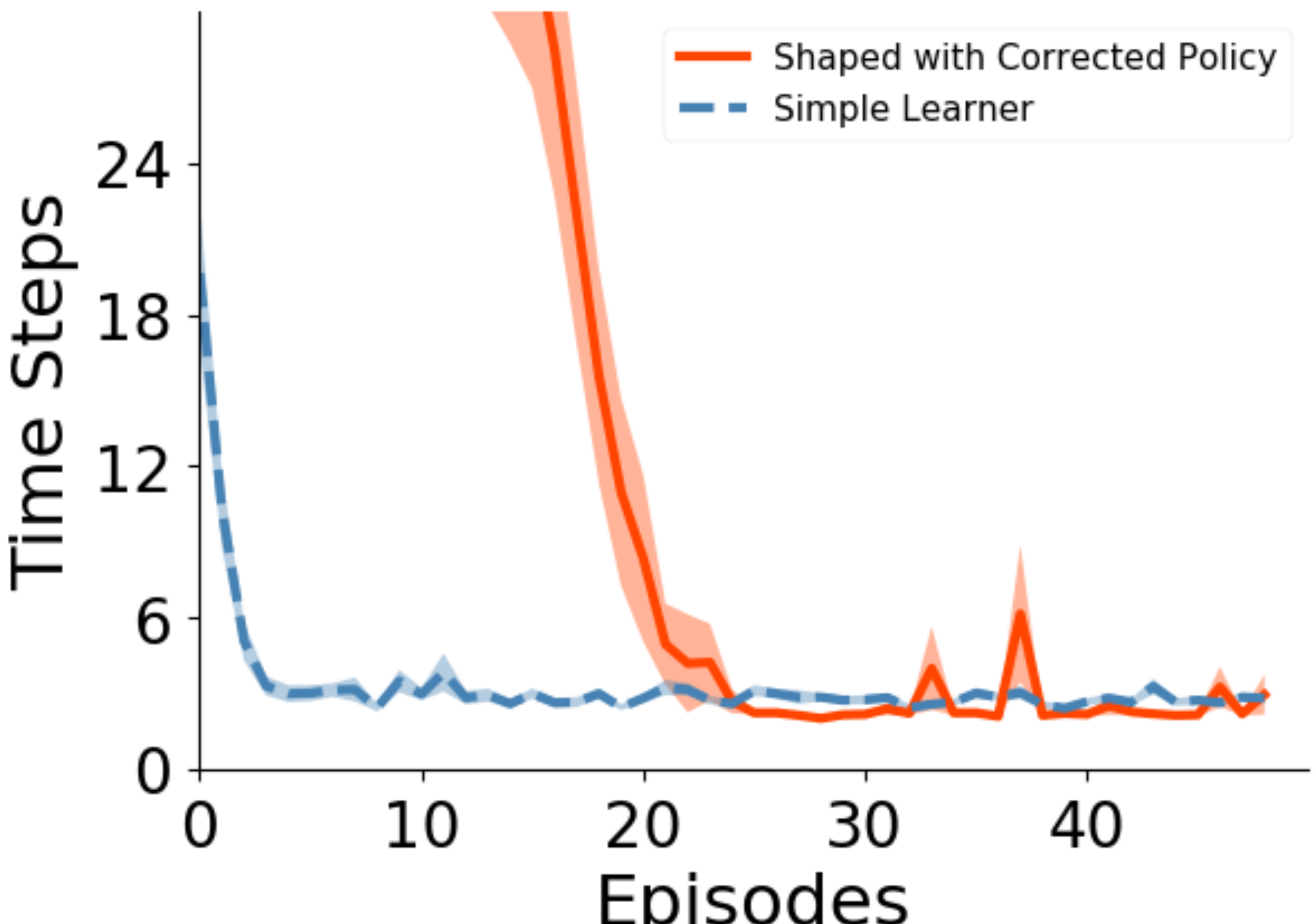
Experiments

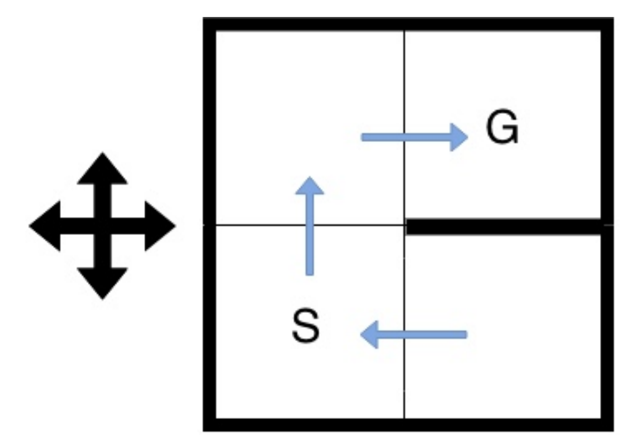
Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy



Experiments: Corrected Dynamic PBRS

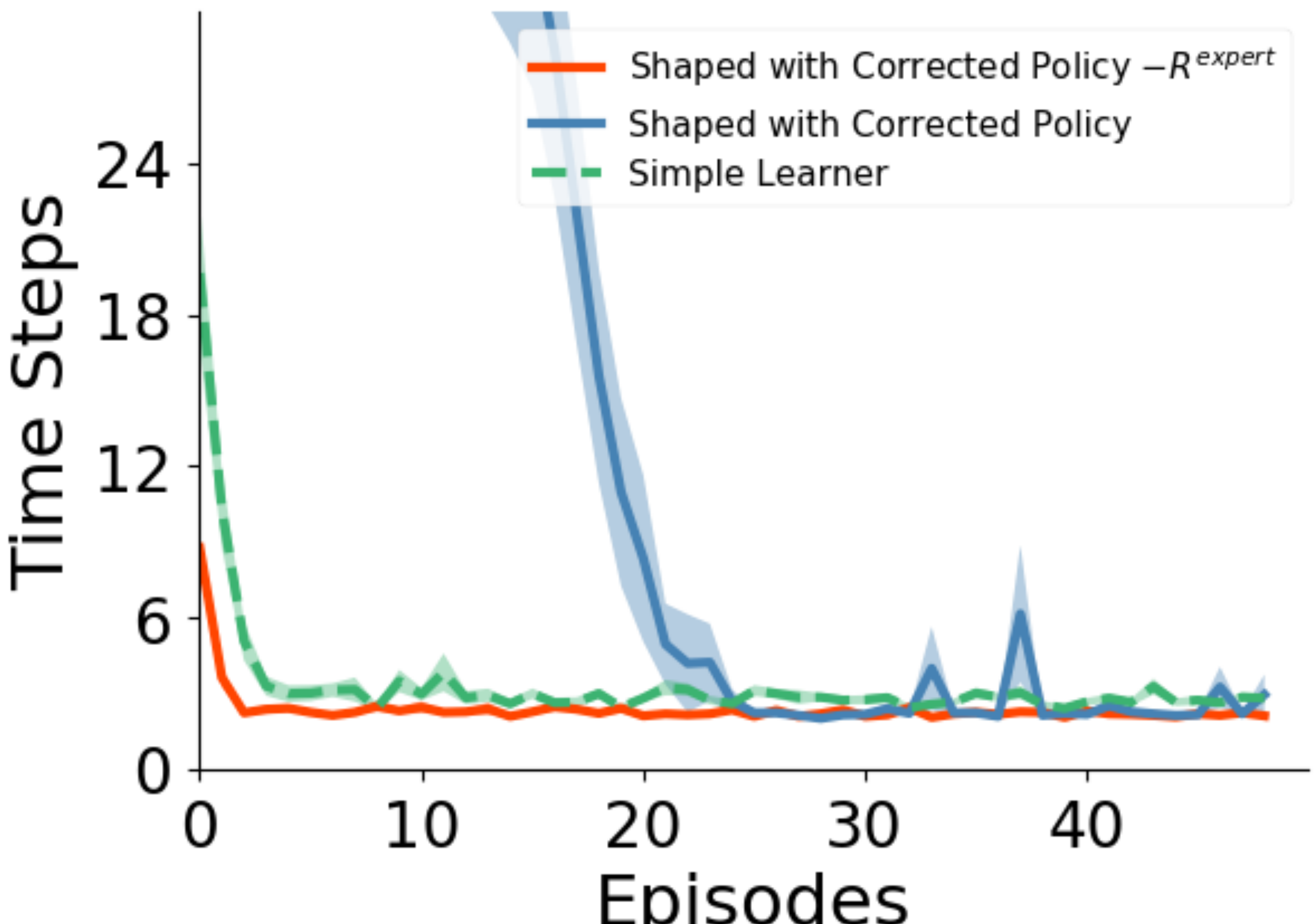
Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy

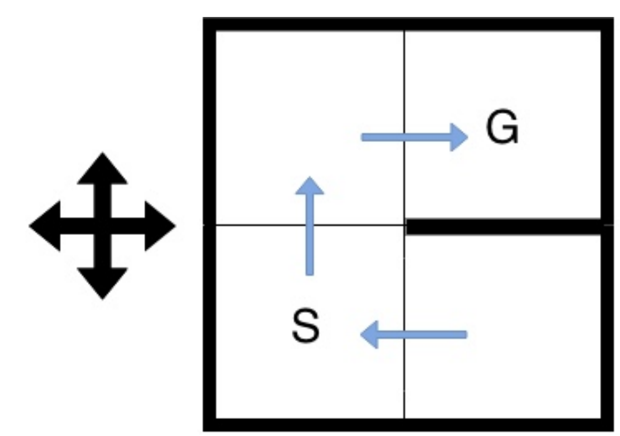




Experiments: Corrected Dynamic PBRS

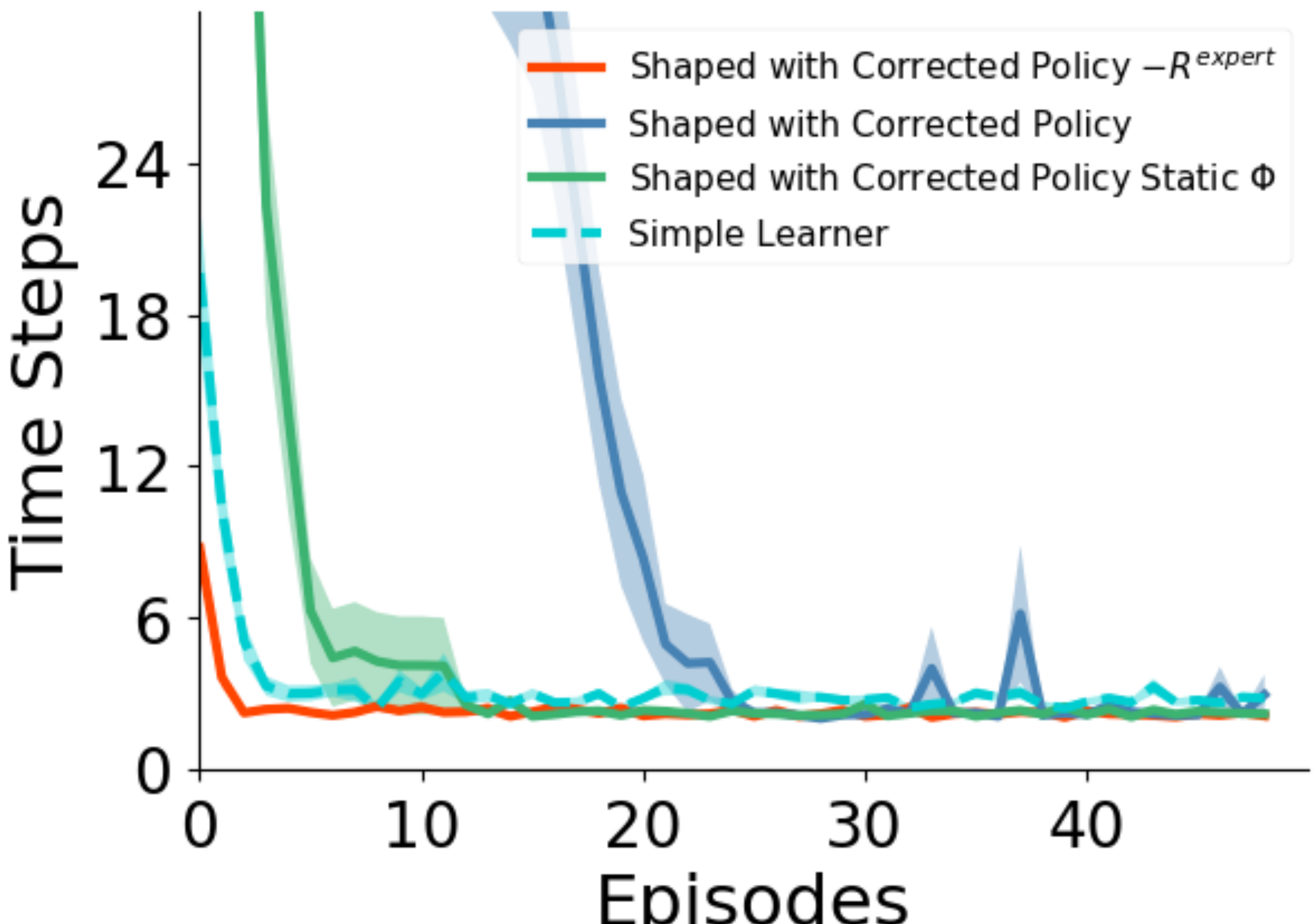
Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy

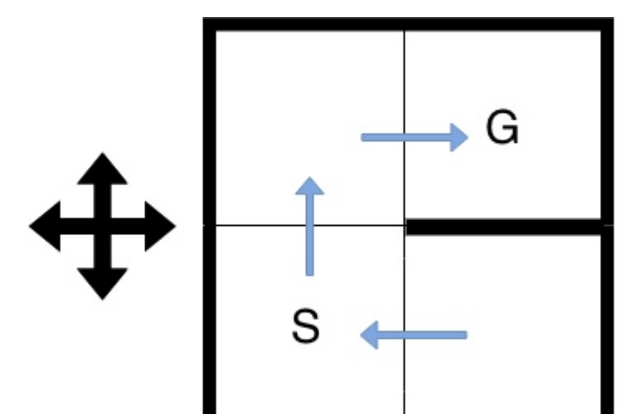




Experiments: Corrected Dynamic PBRS

Sarsa(0), $\gamma = 0.3$, ϵ-greedy policy





Another Look at Shaped Values

$$R' := R + \overbrace{F_t}^{\gamma\Phi_{t+1}(s',a')-\Phi_t(s,a)}$$

$$M = \langle S, A, P, \gamma, R\rangle \longrightarrow M' = \langle S, A, P, \gamma, R'\rangle$$

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi_t(s,a)$$

$$\pi^*_{M'} := argmax_a(Q^*_{M'}(s,a) + \Phi_t(s,a))$$

Another Look at Shaped Values

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi_t(s,a)$$

$$\pi^*_{M'} = argmax_a(Q^*_{M'}(s,a) + \eta\Phi_t(s,a))$$

Another Look at Shaped Values

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi_t(s,a)$$

$$\pi^*_{M'} = argmax_a(Q^*_{M'}(s,a) + \eta\Phi_t(s,a))$$

Starts from 0 and reaches 1

Another Look at Shaped Values

$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi_t(s,a)$$

$$\pi^*_{M'} = argmax_a(Q^*_M(s,a) - \Phi_t(s,a) + \eta\Phi_t(s,a))$$

$$\pi^*_{M'} = argmax_a(Q^*_M(s,a) - (1-\eta)\Phi_t(s,a))$$

Another Look at Shaped Values

$$Q^*_M(s, a) = Q^*_{M'}(s, a) + \Phi_t(s, a)$$

$$\pi^*_{M'} = argmax_a(Q^*_M(s, a) - \Phi_t(s, a) + \eta\Phi_t(s, a))$$

$$\pi^*_{M'} = argmax_a(Q^*_M(s, a) - \xi\Phi_t(s, a))$$

Soft-Shaped

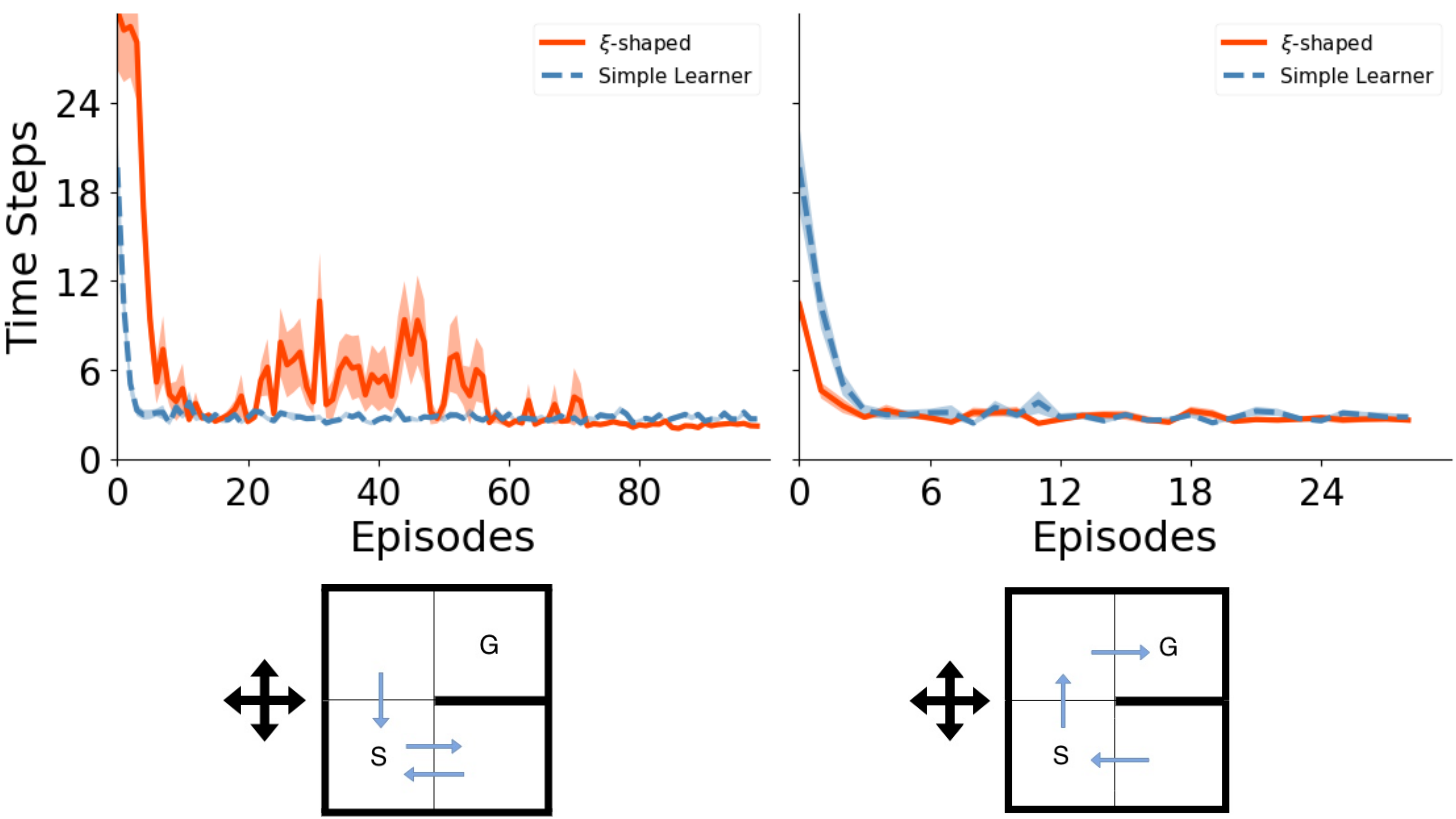
$$Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi_t(s,a)$$

$$\pi^*_{M'} = argmax_a(Q^*_M(s,a) - \Phi_t(s,a) + \eta\Phi_t(s,a))$$

$$\pi^*_{M'} = argmax_a(Q^*_M(s,a) - \xi\Phi_t(s,a))$$

Starts from 1 and reaches 0

Experiments: Soft-Shaped



Summary

- A brief overview of the RL shaping literature
- Pointing out the necessary correction of bias term in dynamic PBRS framework
- Supporting Experiments
- A possible solution

Thanks for attending!