World Value Functions in Continuous Control

Student: Phenyo Modiba, Supervisor: Geraud Nangue Tasse
University of the Witwatersrand, Johannesburg, South Africa

WITS

Do world value functions improve sample efficiency in continuous control?

What We Want

A method to scale WVFs to **infinite continuous goal spaces**.
Improved **sample efficiency and generalization.**
Build goal-conditioned policies capable of efficient exploration and transfer.

Continuous-control agents struggle with **sample efficiency**, **exploration**, and **goal specification**.

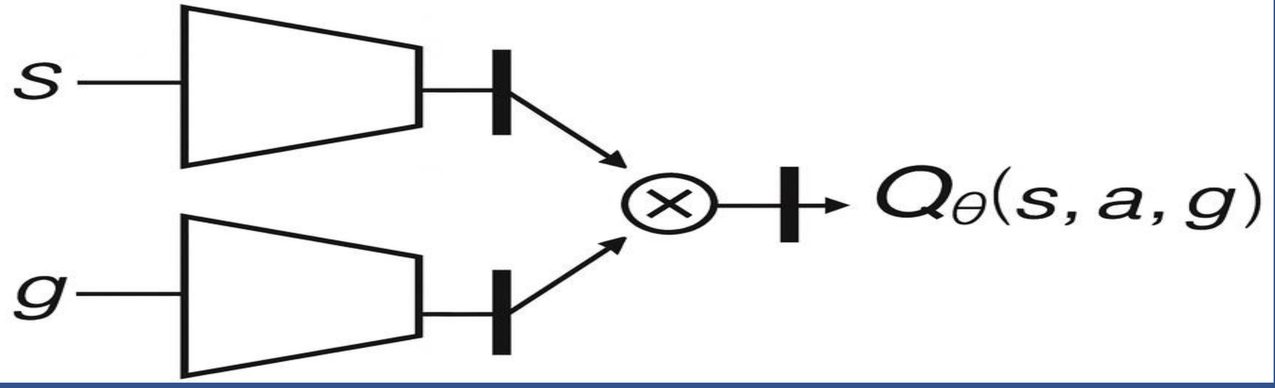
WVFs generalize well in discrete spaces but require new architectures to function in **continuous, high-dimensional goal spaces**.

WVFs in Continuous Control: Challenges

- **Continuous goal spaces are infinite:** WVF theory assumes maximizing over a small, discrete goal set. In continuous control the goal space is unbounded and non-enumerable.
- **High-dimensional state and goal spaces:** High-dimensional continuous states/goals force the WVF to rely on neural approximators instead of tables, increasing learning difficulty.
- **Goal selection becomes ambiguous:** The goal space cannot be enumerated, making it unclear which states should serve as goals or how to incorporate them into WVF learning.

[Tasse et al. 2022] Geraud Nangue Tasse, Benjamin Rosman, and Steven James. **World value functions: Knowledge representation for learning and planning.** arXiv preprint arXiv:2206.11940, 2022.

Solution: UVFA



Solution: SAC-WVF

- A **shared embedding encoder** for states/goals
- A dual-critic **WVFQCritic** that estimates $Q(s, a, g)$
- A goal-conditioned **Actor** producing $\pi(a \mid s, g)$
  - Agent stores termination states when it chooses the "done" action in **HER Buffer**
  - Some of these states become **discovered internal goals**
  - Goals sampled using **Langevin dynamics**

$$g_{t+1} = g_t - \eta \nabla_g(-Q(s, g)) + \sqrt{2\eta}\epsilon$$

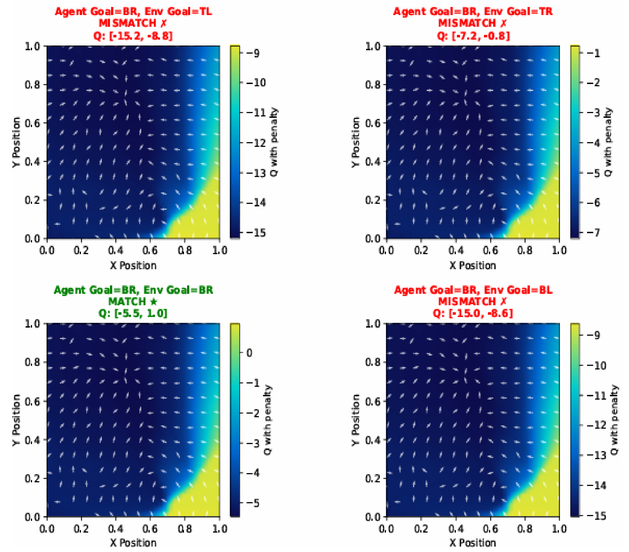
Reward Shaping: Sparse

Reward shaping penalises unintended absorbing states and adding a small step cost encodes the intended goal and ensures every transition contributes information to the world value function.

$$\hat{R}(s, g, a, s') = \begin{cases} +1.0, & \text{if } s' \text{ is absorbing and } s' = g, \\ -10.0, & \text{if } s' \text{ is absorbing and } s' \neq g, \\ -0.1, & \text{otherwise.} \end{cases}$$

Intended Goal

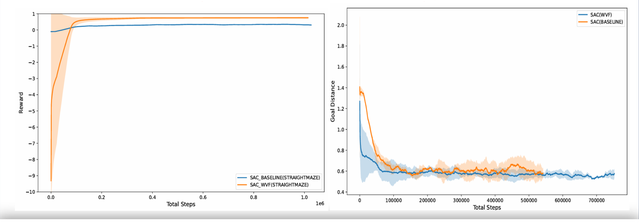
The value of achieving internal goal g from state s via action a



"WVF value fields peak when the agent's internal goal matches the environment's goal — as WVF theory predicts, intended goals yield the highest world value."

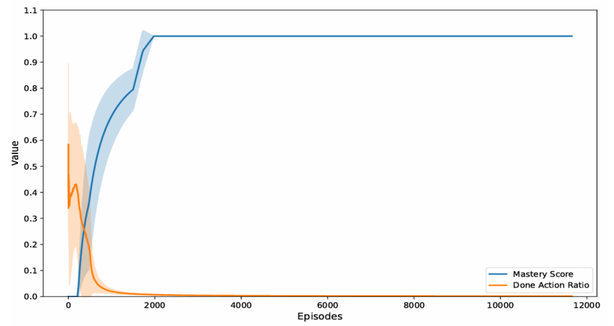
Sampling Efficiency

WVF accelerates learning: SAC-WVF reaches optimal reward far faster than standard SAC



Mastery Emerges as WVFs Learn All Goals

"The agent's ability to reach all internal goals converges to 1.0, while the done action becomes unnecessary once the world value structure is learned and goals become predictable."



Zero-shot transfer performance of SAC vs. SAC-WVF across training environments

"Zero-shot performance arises because WVFs learn how to reach all internal goals during training. This mastery of the environment's goal structure enables immediate transfer to new goal configurations, with evaluation reported as mean ± standard deviation."

Training Env/Method	FetchReach	FetchPush
FetchReach-SAC	0.660 +- 0.474	0.440 +- 0.458
FetchReach-SACWVF	1.00 +- 0.00	0.633 +- 0.240