Bootstrapped Fitness Critics with Bidirectional Temporal Difference

Golden Rockefeller
rockefeg@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Kagan Tumer
kagan.tumer@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Abstract

Evolutionary algorithms (EAs) are well suited for solving many real-world multiagent coordination problems with global, long-term feedback. However, EAs struggle when the feedback becomes sparse and uninformative. In such cases, a system designer can use *Fitness Critics*, which are functional models that estimate the value of an agent's contribution to transform the sparse domain feedback into a dense reward signal. However, existing methods for updating fitness critics do not leverage the temporal information about when a reward is received. Ideally, temporal difference (TD) methods can leverage temporal information about the sparse feedback signal to bootstrap Fitness Critics. Yet, due to the structure Fitness Critics, direct application of TD methods coevolutionary algorithms result in Fitness Critics that under-represent the rewards that are received earlier in the episode. This paper introduces *Bidirectional Fitness Critics* (BFCs), which makes use of a novel, bidirectional temporal difference method, to successfully bootstrap the training of fitness critics with temporal reward information, without under-representing early rewards. The paper demonstrates a significant increase in the converged performance of agents coevolved with BFCs on a multiagent coordination domain.

CCS Concepts: • **Computing methodologies** → Multiagent systems; **Partially-observable Markov decision processes**; **Reinforcement learning**.

Keywords: fitness critics, evolutionary algorithms, temporal difference learning

ACM Reference Format:
Golden Rockefeller and Kagan Tumer. 2022. Bootstrapped Fitness Critics with Bidirectional Temporal Difference. In *Genetic and Evolutionary Computation Conference Companion (GECCO '22 Companion), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3520304.3528999

GECCO '22 Companion, July 9–13, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9268-6/22/07.
https://doi.org/10.1145/3520304.3528999

1 INTRODUCTION

This paper introduces the bidirectional temporal difference method to bootstraps the training of Bidirectional Fitness Critics (BFCs), while keeping equal temporal representation in the approximated fitness evaluation. BFCs introduce a *reverse value* concept, which is a measure of the sum of past rewards, to counterbalance the future-focused Q values found in TD methods. The paper demonstrates a significant increase in the converged performance of agents trained with BFCs on a multiagent coordination domain.

2 Related Works

Approximating the fitness function [1, 4, 6] in evolutionary algorithms (EAs) can be a useful strategy in many difficult problems. Fitness Critics [5] (in Section 2.3) extend the idea of fitness approximation to partially-observable Markov decision processes (in Section 2.1). However, unlike with other methods, the approximated fitness of Fitness Critics isn't a direct function of the policy parameters that are being evolved. Instead, the approximated fitness is based on the expression of that policy though the agent's observations and actions.

2.1 Partially-Observable Markov Decision Processes

In a partially-observable Markov decision process (POMDP), an agent receives an observation $o \in \Omega$ every time-step, which is based on the environment's current state $s \in S$. The agent then responds to each observation with an action $a \in A$. After the agent executes the action, the environment reaches a new state $s' \in S$. Subsequently, the agent receives a reward $r \in \mathbb{R}$ to reflect its performance in the environment. This agent interacts with the environment repeatedly for multiple time-steps, in response to new states and observations. Typically, agents determine their actions by following a trainable policy.

2.2 Policy Search with Evolutionary Algorithms

Evolutionary algorithms (EAs) improve the quality of candidate solutions through multiple cycles of the evolutionary operations: evaluation, selection, recombination, mutation and reinsertion [2]. EAs can be applied to find high performing agent policies [8, 10].

The cross-entropy method is an evolutionary strategy that operates on a parameterized distribution of candidate solutions. The cross-entropy $H(p, q)$ is a measure of how one probability distribution q is different from another probability distribution p. Every generation, the cross-entropy method update a policy-sampling distribution $\mathcal{Q}$ to a new distribution u that minimizes the cross-entropy to the distribution $\mathcal{P}$ that is formed by the set of elite candidate policies:

$$\mathcal{Q} \leftarrow \arg\min_{u} \big(H(\mathcal{P}, u) + sH(\mathcal{Q}, u)\big) \tag{1}$$

A cross-entropy divergence penalty $s \geq 0$ is introduced to the cross-entropy method as a *smoothing parameter*, to prevent the prematurely convergence of $\mathcal{Q}$ by reducing the divergence from u to $\mathcal{Q}$.

2.3 Fitness Critics

The Fitness Critic $F : (T; \vec{p}) \mapsto f$ is a function approximator that maps an agent's trajectory $T = \langle (o_1, a_1), ..., (o_n, a_n) \rangle$ to a fitness evaluation $f \in \mathbb{R}$, and is parameterized by the vector $\vec{p} \in \mathbb{R}^k$ of size k. The Fitness Critic makes use of an interior functional model called the *step-wise critic* $C : (o, a; t, \vec{p}) \mapsto c$, which is updated alongside the training of the agent's policy. These updates allow the Fitness Critic to learn the relationship between the experienced agent trajectories and the sampled feedback for those trajectories. The Fitness Critic then uses the step-wise critic to evaluate each observation-action pair in the agent's trajectory. The Fitness Critics aggregates these evaluations (e.g. by taking the mean) into a single fitness score that the evolutionary algorithm can use:

$$F(T; \vec{p}) = \frac{1}{n} \sum_{t=1}^{n} C(o_t, a_t; t, \vec{p}) \tag{2}$$

2.3.1 Trajectory-Wise Updates using Sampled Fitness Score.

The Fitness Critic parameters $\vec{p}$ can be updated with gradient descent to regress the Fitness Critic towards a sampled fitness score $\mathcal{R}$:

$$\vec{p} \leftarrow \vec{p} + \alpha[\mathcal{R} - F(T; \vec{p})] \nabla_{\vec{p}} F(T; \vec{p}) \tag{3}$$

where α is the learning rate. $\mathcal{R}$ is typically the sum of all rewards experienced throughout an episode.

2.3.2 Step-Wise Updates using Sampled Fitness Score.

Alternatively, $\vec{p}$ can be updated with gradient descent performed in a step-wise manner to regress each individual step-wise evaluation to the sampled fitness score $\mathcal{R}$:

$$\begin{aligned} \vec{p} &\leftarrow \vec{p} + \alpha[\mathcal{R} - c_t] \nabla_{\vec{p}} c_t \\ c_t &:= C(o_t, a_t; t, \vec{p}) \end{aligned} \tag{4}$$

2.4 Temporal Difference Methods

In reinforcement learning, temporal difference methods [10, 11] can bootstrap the learning of the value estimates by leveraging the values estimates of future time-steps. Many reinforcement learning methods involve training a function approximator $\hat{Q} : (o, a; t, \vec{p_q}) \mapsto q$ [9] to approximate a Q value function for guiding the policy optimization process. The Q function approximator is parameterized by the vector $\vec{p_q}$. For brevity, the $\hat{Q}$ value term q_t that is used throughout this paper is defined as:

$$q_t := \hat{Q}(o_t, a_t; t, \vec{p}_q) \tag{5}$$

State-Action-Reward-State-Action (SARSA) is a common temporal difference learning technique where the parameters of the Q function approximator is updated as follows:

$$\vec{p_q} \leftarrow \vec{p_q} + \alpha[r_t + \gamma q_{t+1} - q_t] \nabla_{\vec{p_q}} q_t \tag{6}$$

where α is the learning rate; γ is the discount factor; and r_t is the reward received at time-step t. The advantage function method [3] is another temporal difference method that uses an advantage function that addresses the problem of bias in the $\hat{Q}$ value estimate.

TD-λ is an temporal difference method that uses *eligibility traces* [7]. The method use a trace factor $\lambda \in [0, 1]$ to determine the balance between pure temporal difference methods ($\lambda = 0$) and Monte Carlo methods ($\lambda = 1$). The temporal difference target for one step can be calculated efficiently from the temporal difference target for the next step:

$$z_t = r_t + \gamma[(1 - \lambda) q_{t+1} + \lambda z_{t+1}] \tag{7}$$

3 Overview: Bidirectional Fitness Critics

Bidirectional Fitness Critics (BFCs) combine many concepts in order to provide agents with informative feedback:

- BFCs build upon **Fitness Critics** to turn sparse feedback signal into dense feedback that the agent can better learn from.
- BFCs use **bidirectional temporal difference** to leverage temporal reward information that can bootstrap the training of the Fitness Critic.
- The bidirectional temporal difference update extends temporal difference by introducing a **reverse value** to ensure that all rewards are weighted equally.
- Similarly, BFC use **eligibility traces** to fine-tune the temporal difference update.
- BFCs use **Ensemble Fitness Critics** models to avoids self-referential problems in the temporal difference update.

4 Bidirectional Temporal Difference Methods For Fitness Critics

Training Fitness Critics with temporal difference methods can result in Fitness Critics that underrepresent early rewards. This is due to the Q value being a function of future rewards. That is, the first Q value will represent all future results, the second Q value will represent all potential rewards after the first step, the third Q value will represent

all potential rewards after the second step and so on. Aggregating these Q values will lead to an underrepresentation of early rewards, as the rewards for those early steps are only represented in early Q values while later rewards are represented in early and later Q values.

The bidirectional temporal difference method addresses the underrepresentation of early rewards by introducing a *reverse value* term that represents the sum of *past* rewards. BFCs sum the reverse value with the Q value to equalize the temporal representation of agent rewards. In BFCs, reverse value u is approximated with the function approximator $\hat{V}_{rev} : (o; t, \vec{p}_u) \mapsto u$. These parameters $\vec{p}_u$ for this function approximator can be updated as such:

$$\begin{aligned} \vec{p}_u &\leftarrow \vec{p}_u + \alpha[r_{t-1} + u_{t-1} - u_t]\nabla_{\vec{p}_u} u_t \\ u_t &:= \hat{V}_{rev}(o_t; t, \vec{p}_u) \end{aligned} \tag{8}$$

By adding the results of the Q function approximator and the reverse value function approximator $\hat{V}_{rev}$ together, the resulting step-wise critic $\hat{C}_{Bidi} : (o, a; t, \vec{p}_{qu}) \mapsto c$ addresses the problem of the underrepresentation of early rewards:

$$\hat{C}_{Bidi}(o_t, a_t; t, (\vec{p}_q, \vec{p}_u)) = \hat{Q}(o_t, a_t; t, \vec{p}_q) + \hat{V}_{rev}(o_t; t, \vec{p}_u) \tag{9}$$

Reverse eligibility traces combines the reverse value concept with eligibility traces. With reverse eligibility traces, the approximator update is based on the value of past steps:

$$z_{rev,t} = r_{t-1} + [(1 - \lambda)u_{t-1} + \lambda z_{rev,t-1}] \tag{10}$$

5 Ensemble Fitness Critics

With the temporal difference (TD) method, updating the approximator for one time-step could invalidate the value approximation for other time-steps. To address the TD self-referential problem, the *Ensemble Fitness Critics* is made up of multiple independent functional submodels, one for each time-step. Now, updating the value function associated with one time-step will not affect the value function associated with another time-steps. In this paper, these submodels are refered to as *sub-step critics*.

For evaluating an observation-action pair, the ensemble step-wise critic can take the weighted average of each sub-step critics evaluation. The ensemble step-wise critic is defined as:

$$C_{Ens}(o, a; \vec{p}) = \sum_{j=1}^{n} w_j C_j(o, a; \vec{p}_j) \tag{11}$$

where C is the ensemble step-wise critic; C_j is the sub-step critic for the time-step j; w_j is the weighting that reflects the certainty in C_j's current evaluation; $\vec{p}_j$ is the parameter vector for C_j.

Due to the double summation over time-steps, a simple implementation of Ensemble Fitness Critic methods can be slow compared to non-ensemble Fitness Critic implementations. However, if the sub-step critics are implemented as lookup tables (i.e. associated arrays), then critic values can be stores and updated with lazy evaluations in a method that somewhat resembles the weighted cumulative moving average method. This resulting method is called Efficient Ensemble Fitness Critics with Lookup Tables (EEFCLT).

6 Experiments and Results

This paper compares the performance of the agents trained with different variations of the Fitness Critics, including the Bidirectional Fitness Critics. Agents are trained on a multiagent coordination domain. Agents trained without Fitness Critics (No Critic) use the sum of rewards as the policy's fitness score. All Ensemble Fitness Critics are some modification of Efficient Ensemble Fitness Critics with Lookup Tables (EEFCLT). Non-Ensemble Fitness Critics use cumulative moving average schemes that are conceptually comparable to EEFCLT. For the Fitness Critics that use eligibility traces, a schedule determines what the eligibility trace factor λ should be for a given evolutionary generation. The agents' policies are evolved with cross-entropy optimization. The agents' policies are modelled as lookup tables that map an observation to an action-sampling categorical distribution.

The tightly coupled multi-rover domain is a multiagent coordination domain with sparse feedback. In this domain, there are 4 rovers are operating in a 10 by 10 gridworld environment. There are also 4 points of interest (POIs). The rovers have limited sensing capabilities and are tasked with capturing as many POIs as possible within a time-frame. For a POI to be captured, two rovers must occupy that POI's cell at the same time. Once a POI is captured, that POI is removed from the environment and another POI is generated at a random location; additionally, the rovers' team performance score is increased by 1 with every captured POI. The rovers start at the middle of the gridworld at the beginning of each episode. The initial POIs positions are random. There are 100 time-steps for each episode. The rover sensing capabilities are illustrated in Figure 1.

The agents in the tightly coupled multi-rover domain were trained for 6000 generations. The results for these experiments are present in Table 1 that shows the converged performance scores for the agents.

Bidirectional Ensemble TD(λ) Fitness Critics achieved the highest performance. Bidirectional Ensemble Fitness Critics with pure temporal difference methods (TD(0))) were slow to train, probably due to large delays in the reward signal. Due to the sparsity of the reward signal, agents trained without Fitness Critics, and those trained with Trajectory-wise Update Fitness Critics achieved a low performance. Due to the underrepresentation of early rewards, the SARSA and Advantage Fitness Critics were not able to reach the performance of the Bidirectional Ensemble Fitness Critics, despite all three method using some version of TD learning. The application of ensemble methods does not show any significant impact

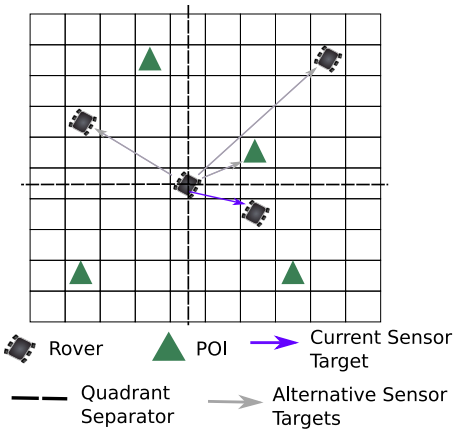

Figure 1. The tightly coupled multi-rover domain. Rovers are able to target either the closest point of interest (POI), or other rovers. The rover's observation is based one of four target types, the distance to the target, and which quadrant the target is in.

Table 1. Converged Performance Scores for Various Fitness Critics on the Tightly Coupled Multi-Rover Domain

Fitness Critic Variation	Score
Bidirectional Ensemble TD(λ)	**5.2 ± 0.23**
Bidirectional Ensemble TD(0)	0.7 ± 0.07
Bidirectional Ensemble TD(1)	3.6 ± 0.22
Bidirectional TD(λ)	1.3 ± 0.72
SARSA Ensemble TD(λ)	3.8 ± 0.12
SARSA TD(λ)	3.6 ± 0.13
Advantage Ensemble TD(λ)	3.6 ± 0.29
Step-Wise Ensemble	3.6 ± 0.16
Step-Wise	3.5 ± 0.15
Trajectory-Wise	0.7 ± 0.16
No Fitness Critic	0.9 ± 0.08

for the Fitness Critics with step-wise updates, as these Fitness Critics do not use TD learning; ensemble methods were introduced to improve TD methods. Ensemble methods were most effective when paired with the Bidirectional Fitness Critic method. All other Fitness Critic variations performed similarly and achieved a moderate performance score.

7 Conclusion and Future Work

This paper introduces Bidirectional Fitness Critics (BFC), which makes use of temporal reward information to successfully bootstrap the training of Fitness Critics. Existing methods for updating Fitness Critics are unable to leverage to temporal information about when the rewards were received throughout an episode. Using standard temporal difference (TD) methods to access this temporal reward information results in the Fitness Critic that underrepresent early rewards, which can negatively impact the evolution of good agent policies. The bidirectional temporal difference method extends TD concepts to enable BFCs to maintain equal temporal representation of rewards. This paper shows the effectiveness of Ensemble BFCs in a representative multiagent coordination problem.

Acknowledgments

This work was partially supported by the National Science Foundation under grant No. IIS-1815886 and by the Air Force Office of Scientific Research under grant No. FA9550-19-1-0195.

References

[1] Jesus S. Aguilar-Ruiz, Daniel Mateos, and Domingo S. Rodriguez. 2003. Evolutionary Neuroestimation of Fitness Functions. In *Progress in Artificial Intelligence*, Fernando Moura Pires and Salvador Abreu (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 74–83.

[2] Thomas Back, David B. Fogel, and Zbigniew Michalewicz (Eds.). 1999. *Basic Algorithms and Operators* (1st ed.). IOP Publishing Ltd., Bristol, UK, UK.

[3] Leemon C Baird. 1994. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Vol. 4. IEEE, 2448–2453.

[4] I. Paenke, J. Branke, and Yaochu Jin. 2006. Efficient Search for Robust Solutions by Means of Evolutionary Algorithms and Fitness Approximation. *Trans. Evol. Comp* 10, 4 (Aug. 2006), 405–420. https://doi.org/10.1109/TEVC.2005.859465

[5] Golden Rockefeller, Shauharda Khadka, and Kagan Tumer. 2020. Multi-Level Fitness Critics for Cooperative Coevolution. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1143–1151.

[6] Yasuhito Sano and Hajime Kita. 2002. Optimization of noisy fitness functions by means of genetic algorithms using history of search with test of estimation. In *Computational Intelligence, Proceedings of the World on Congress on*, Vol. 1. IEEE Computer Society, Los Alamitos, CA, USA, 360–365. https://doi.org/10.1109/CEC.2002.1006261

[7] Satinder P Singh and Richard S Sutton. 1996. Reinforcement learning with replacing eligibility traces. *Machine learning* 22, 1 (1996), 123–158.

[8] Harold Soh and Yiannis Demiris. 2011. Evolving Policies for Multi-Reward Partially Observable Markov Decision Processes (MR-POMDPs). In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation* (Dublin, Ireland) *(GECCO '11)*. Association for Computing Machinery, New York, NY, USA, 713–720. https://doi.org/10.1145/2001576.2001674

[9] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller (Eds.), Vol. 12. MIT Press. https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf

[10] Matthew E. Taylor, Shimon Whiteson, and Peter Stone. 2007. Temporal Difference and Policy Search Methods for Reinforcement Learning: An Empirical Comparison. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2 (AAAI'07)*. AAAI Press, Vancouver, British Columbia, Canada, 1675–1678.

[11] Gerald Tesauro. 1995. Temporal Difference Learning and TD-Gammon. *Commun. ACM* 38, 3 (mar 1995), 58–68. https://doi.org/10.1145/203330.203343