HomeostaticEnv: A Multi-Resource Environment for Homeostatic Reinforcement Learning Experiments

Gilzamir Gomes
gilzamir_gomes@uvanet.br
Department of Computer Science
State University of the Acaraú Valey

January 31, 2026

Abstract

Homeostatic regulation - the ability to maintain internal physiological variables within a viable range - is a fundamental property of living organisms. In the context of Artificial Intelligence, Homeostatic Reinforcement Learning (HRL) applies these biological principles to create agents capable of autonomous self-regulation. However, a significant challenge in HRL research is the lack of standardized environments that specifically test an agent's ability to manage high-dimensional physiological state spaces under constant decay. Critically, high-dimensional resource management implies an equally **high-dimensional continuous action space**, creating a complex control problem. This paper introduces `HomeostaticEnv`, a continuous control environment designed to simulate multi-resource homeostatic regulation. We detail the environment's dynamics and formalize three distinct reward mechanisms: Default (quadratic penalty), Euclidean (potential-based), and Operational Regimes (discrete-logic based). We demonstrate the environment's utility through Proximal Policy Optimization (PPO) experiments on high-dimensional tasks (up to 1000 simultaneous resources), highlighting the impact of reward signal design on survival and stability in large action spaces.

1 Introduction

Reinforcement Learning (RL) [1] has achieved remarkable success in domains ranging from game playing to robotics. Traditionally, RL agents maximize a cumulative extrinsic reward signal. However, in biological systems, behavior is often driven by *homeostasis*: the imperative to maintain critical physiological variables (such as temperature, glucose, or hydration) within a tight survival range against natural entropy [2].

Homeostatic Reinforcement Learning (HRL) integrates this concept, formulating the objective not as infinite maximization, but as the minimization of deviation from a setpoint. A core difficulty in HRL is the "curse of dimensionality." This challenge is two fold:

1. **State Space**: The agent must monitor N distinct physiological variables simultaneously.
2. **Action Space**: Regulation requires a corresponding high-dimensional continuous action space ($\mathcal{A} \in \mathbb{R}^N$).

An agent must often balance multiple conflicting resources simultaneously. For example, an action that restores one resource might deplete another, or simply time spent foraging for one resource allows others to decay. To facilitate research in this domain, we present `HomeostaticEnv`, a gymnasium-compatible [3] environment. Unlike standard grid-worlds or MuJoCo tasks, `HomeostaticEnv` focuses specifically on the dynamics of resource management, scaling to massive action spaces ($N = 1000$) where standard exploration strategies often fail.

2 Foundations

The theoretical underpinning of `HomeostaticEnv` relies on Drive-Reduction Theory [4]. We define an agent's internal state as a vector $S_t = [s_1, s_2, ..., s_N]$, representing N distinct physiological variables.

2.1 Zones of Stability

For each variable s_i, we define three critical regions relative to a target setpoint (typically 0.0):

1. **Target** (T_i): the ideal homeostatic equilibrium (set to 0.0).
2. **Safety Zone** (Z_{safe}): a range $[-z_{safe}, z_{safe}]$ where the organism operates optimally.
3. **Survival Zone** (Z_{surv}): a critical limit $[-z_{surv}, z_{surv}]$. If any $|s_i| > z_{surv}$, the agent "dies," and the episode terminates immediately.

3 The Environment Description

`HomeostaticEnv` models a continuous control problem where the agent must continuously counteract the natural decay of resources.

3.1 State and Action Space

The observation space is a continuous vector $\mathcal{O} \in \mathbb{R}^{3N}$, consisting of:

- Current physiological state values (s_t).
- Survival zone thresholds.
- Safety zone thresholds.

The action space is a continuous vector $\mathcal{A} \in [-1, 1]^N$. **High Dimensionality Note:** It is important to emphasize that the action space scales linearly with the number of resources. For a simulation with $N = 100$ variables, the agent must output a 100-dimensional vector at every step, deciding the exact intensity of correction for every single resource simultaneously.

3.2 Dynamics

The environment follows a linear difference equation. At each time step t, the state evolves according to:

$$s_{t+1,i} = s_{t,i} - \delta_i + \alpha \cdot a_{t,i} \tag{1}$$

Where:

- $delta_i$ is the decay ratio (fixed at 0.01).
- α is the action scaler (fixed at 0.2).
- $a_{t,i}$ is the clipped action $[-1, 1]$.

3.3 Reward Functions

A critical contribution of this environment is the implementation of three distinct homeostatic reward functions for studying multi-objective regulation. The two primary current approaches are based on Keramati and Gutkin [2]. Other research has implemented these functions in different domains [5, 6].

3.3.1 Default (Quadratic Penalty)

This function penalizes the agent based on the sum of squared distances from the target. It mimics a pure minimization of error.

$$R_{default} = -\sum_{i=1}^{N}(s_{t,i} - T_i)^2 \cdot k \tag{2}$$

where k is the reward scale.

3.3.2 Euclidean (Differential)

This is a potential-based reward function. It rewards the *improvement* in the total system state rather than the absolute state.

$$R_{euclidian} = (D_{t-1} - D_t) \cdot k \tag{3}$$

where $D_t = \sum(s_{t,i} - T_i)^2$. This formulation provides dense gradients for moving towards the center, regardless of the current distance.

3.3.3 Operational Regimes

This function implements a logic-based reward that prioritizes safety and penalizes worsening conditions when in danger. For each variable i, let $d_t = |s_{t,i} - T_i|$. We define binary indicators:

- $u_i = 1$ if $d_t \leq Z_{safe}$, else 0 (Inside Safety Zone).
- $v_i = 1$ if $d_t \geq d_{t-1}$, else 0 (Condition Worsening).

The reward is calculated as:

$$R_{regimes} = \sum_{i=1}^{N} [u_i + (1 - u_i)(1 - 3v_i)] \cdot k \tag{4}$$

This logic dictates:

- **Inside Safety:** $+1$ reward.
- **Outside Safety but Improving:** $+1$ reward.
- **Outside Safety and Worsening:** -2 penalty $(1 - 3(1))$.

4 Experiments

To validate the environment and compare the efficacy of the reward functions, we utilized the Proximal Policy Optimization (PPO) algorithm from the Stable-Baselines3 library [7].

4.1 Experimental Setup

We evaluated the environment across three distinct complexity levels to assess scalability:

1. **Low-Dim:** $N = 20$ variables.
2. **Mid-Dim:** $N = 100$ variables.
3. **High-Dim**: $N = 1000$ variables.

The $N = 1000$ scenario presents a significant challenge, requiring the policy network to map observations to a 100-dimensional continuous action vector while managing 1000 concurrent homeostatic loops.

The reward function based on regimes was tested in two versions. In version A, a fixed safety zone of 0.5 was used. In version B, a safety zone was used by sampling uniformly from the real interval $[0.5, 0.9]$.

- **Algorithm:** PPO (MlpPolicy)
- **Total Timesteps:** 250,000
- **Environments:** 8 parallel environments (SubprocVecEnv)
- **Max Steps per Episode:** 1000

Three separate training regimes were executed corresponding to the three reward functions: `default (quadratic)`, `euclidian (differential)`, and `operational_regimes`.

5 Results

The main metric in these experiments is the time the episode can be kept active, known as episode length. We also show that the best-performing methods are those with the lowest standard deviation.

In the simulation with 20 variables, the result in terms of episode length is shown in Figure 1, while the standard deviation is shown in Figure 2.

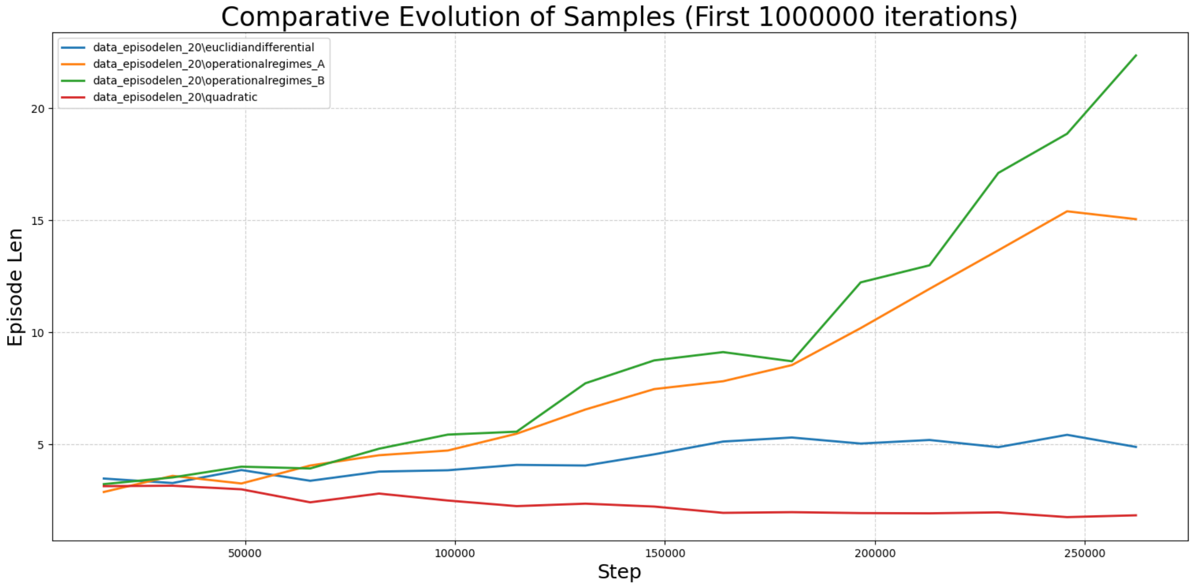

Figure 1: Comparison of Episode Length over 250k timesteps. Episode length of the different reward functions used in the experiment with 20 variables.

In the simulation with 100 variables, the result in terms of episode length is shown in Figure 3, while the standard deviation is shown in Figure 4.

In simulations with 1,000 variables, none of the reward functions converged, with agents lasting an average of only one time step.

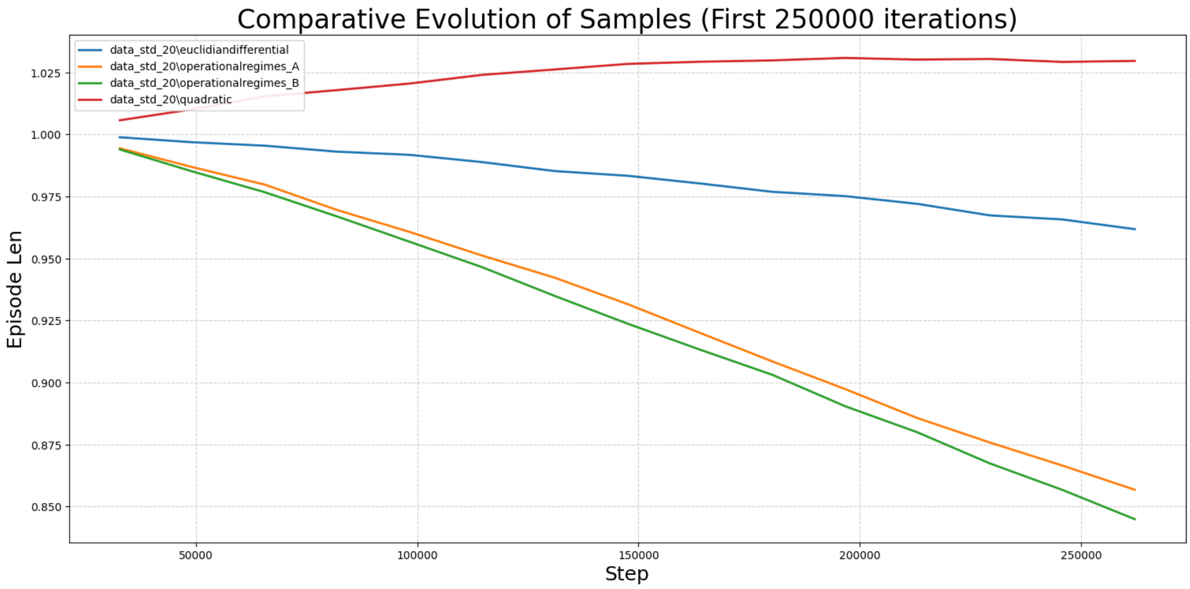

Figure 2: Comparison of Reward Stadard Deviation over 250k timesteps. Standard deviation of the different reward functions used in the experiment with 20 variables.

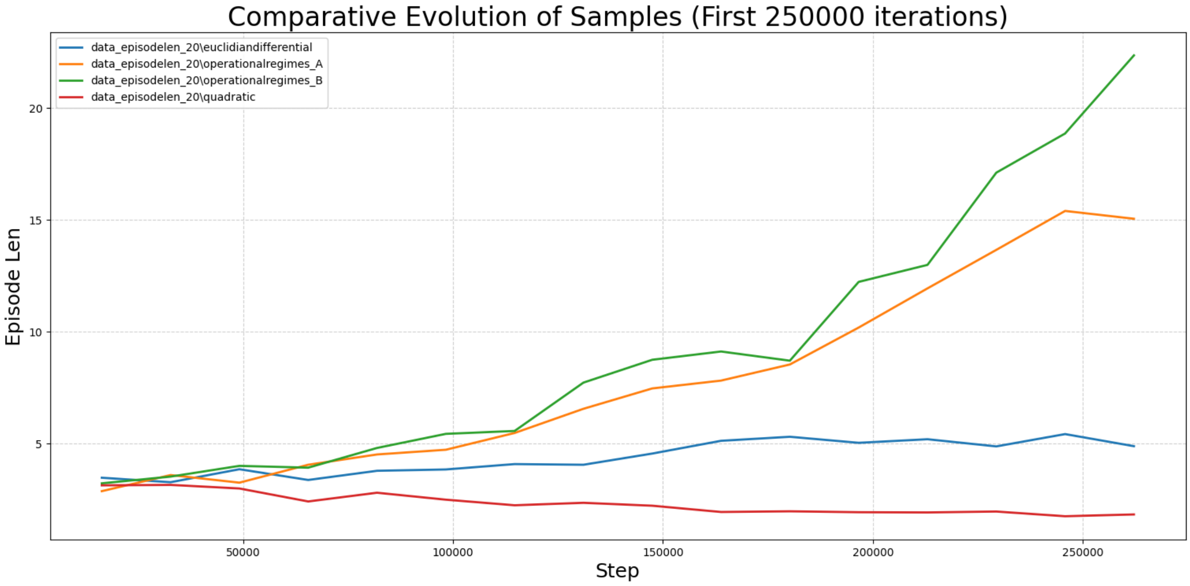

Figure 3: Comparison of Episode Length evolution over 250k timesteps. Episode length of the different reward functions used in the experiment with 100 variables.

6 Discussion of Results

The experimental results validate *HomeostaticEnv* as a challenging testbed for high-dimensional homeostatic regulation. The environment poses a scaling difficulty that directly impacts the agent's ability to maintain internal stability.

In the low-dimensional setting ($N = 20$), the *Operational Regimes* reward function demonstrates a slight advantage in maintaining episode length compared to other methods. Conversely, the *Default* quadratic penalty performs poorly even at this scale, indicating its insufficiency for multi-variable control.

As the complexity increases to $N = 100$, the dimensionality of the action space exposes further weaknesses. Both the *Euclidean Differential* and *Operational Regimes* functions yield

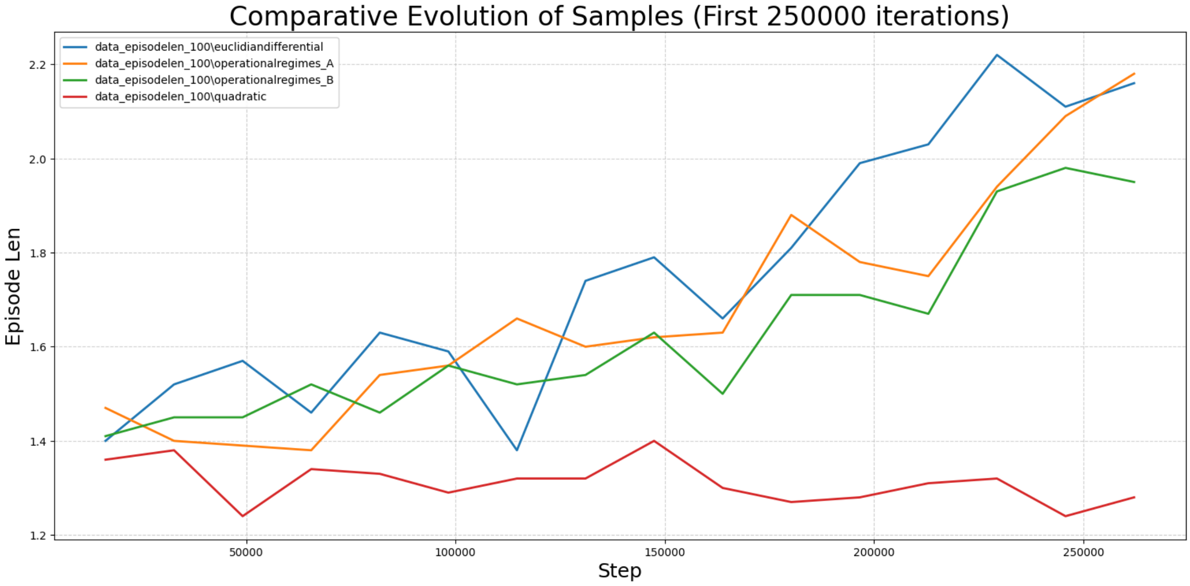

Figure 4: Comparison of Reward Stadard Deviation over 250k timesteps. Standard Deviation of the different reward functions used in the experiment with 100 variables.

nearly identical performance, though both exhibit significant degradation compared to their $N = 20$ results. This drop in performance confirms that the increased dimensionality creates a more complex mapping problem for the PPO algorithm. Meanwhile, the quadratic penalty continues to show poor results, as the aggregate signal from 100 variables becomes too noisy for effective gradient derivation.

The extreme case of $N = 1000$ variables presents a threshold where all tested reward functions fail to converge, with agents dying within the first episode. This outcome underscores the environment's utility as a benchmark for extreme continuous control tasks. The failure of standard PPO configurations at this level suggests that simply refining reward signals may be insufficient for massive state spaces, pointing toward the necessity of exploring Hierarchical Reinforcement Learning (HRL) and more resilient exploration strategies.

7 Conclusion

This paper introduced *HomeostaticEnv*, a multi-resource environment designed to evaluate Reinforcement Learning agents in the task of maintaining internal physiological equilibrium. Through a series of experiments, we demonstrated that the environment provides a scalable challenge, where the complexity of the control task increases significantly with the number of variables.

Our analysis of different reward mechanisms showed that the *Operational Regimes* function offers a slight advantage in lower dimensions ($N = 20$). However, as the system scales to $N = 100$, both the *Operational Regimes* and the *Euclidean Differential* functions exhibit similar performance, with both suffering from the increased difficulty of the state-action mapping. In contrast, the *Default* quadratic penalty proved insufficient for maintaining stability even in mid-dimensional settings.

The fact that no reward function achieved convergence in the 1000-variable setup confirms that *HomeostaticEnv* is a rigorous testbed for high-dimensional continuous control. These results underscore the limitations of standard PPO configurations and suggest that future work should focus on exploring Hierarchical Reinforcement Learning (HRL) and more sophisticated exploration strategies to manage massive physiological state spaces. We hope that *HomeostaticEnv* will serve as a valuable tool for the community to advance the development of more autonomous and resilient homeostatic agents.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning.* Cambridge, MA: MIT Press, 2nd ed., 2018.

[2] M. Keramati and B. Gutkin, "Homeostatic reinforcement learning for integrating reward collection and physiological stability," *eLife*, vol. 3, p. e04811, dec 2014.

[3] M. Towers *et al.*, "Gymnasium: A standard interface for reinforcement learning environments," *arXiv preprint arXiv:2106.00000*, 2023.

[4] C. L. Hull, *Principles of Behavior: An Introduction to Behavior Theory.* New York: Appleton-Century-Crofts, 1943.

[5] N. Yoshida, H. Kanazawa, and Y. Kuniyoshi, "Homeostatic reinforcement learning through soft behavior switching with internal body state," in 2023 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), (2023, Gold Coast, Australia. **Proceedings [...]**. [S.l.]), pp. 1–8, IEEE, 2023.

[6] N. Yoshida and Y. Kuniyoshi, "Unexpected capability of homeostasis for open-ended learning," in *2025 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–8, 2025.

[7] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.