Spring Based locomotion of Bolt Robot using Reinforcement Learning

Alfred Cueva*
*Department of Mechanical Engineering, Seoul National University, Seoul, Korea
alfred11@snu.ac.kr

***Abstract*—In this work, I introduce deploying spring like joints to model human-like walking and solve problems regarding energy consumption. This approach exploits Potential Based Rewards (PBRS), torque-angle relations and PPO algorithm using Massively Deep Reinforcement Learning (MDRL). The hip pitch joints were modeled as parallel linear springs enabling the torques to account scaled actions and spring torques. The performance of PBRS the spring like joints and was compared in the 6 DOF Bolt Robot when commanded to move forward and sideways using IsaacGym. It was shown that this spring based joint modeling reduces the Cost of Transportation (COT) at high speeds and increases maximum commanded forward linear velocity compared to PBRS.**

I. Introduction

In recent years, Massive Deep Reinforcement Learning (MDRL) has revolutionized the field of robotics by enabling robots to learn complex tasks autonomously. By combining deep neural networks with reinforcement learning algorithms, MDRL allows robots to improve their performance through trial and error. This approach has led to more versatile and adaptive systems, as robots can learn from their own experiences and continuously enhance their skills [2], [1].

Bipedal locomotion presents a significant challenge in robotics, requiring precise control and coordination of actuators and sensors. Achieving robust and stable walking is crucial for bipedal robots, as disruptions in balance, energy inefficiencies, and unpredictable environments can hinder their performance [3]. To address these challenges, researchers have been exploring innovative approaches to improve gait robustness, with a specific focus on optimizing energy consumption.

Energy efficiency plays a vital role in robotic locomotion, as it directly impacts battery-powered robots' operational lifespan and reduces their reliance on external power sources. By minimizing energy consumption, robots can operate for longer durations, increasing autonomy and reducing the need for frequent recharging or battery replacements [3].

To optimize energy efficiency in robotic locomotion, researchers have integrated MDRL with novel techniques and control strategies. These approaches leverage the power of MDRL to learn optimal locomotion policies while considering energy consumption as a key objective. By formulating reward functions that encourage energy-efficient behaviors, robots can adapt their gait patterns, motor control, and actuator dynamics to minimize energy expenditure while maintaining stability and performance [2] [6].

Among the innovative approaches, one promising technique is the incorporation of parallel elastic elements in the robot's leg design. These elements act as passive mechanical springs, storing and releasing energy during locomotion. By effectively recycling energy, parallel elastic elements significantly reduce the overall energy expenditure required for walking, leading to improved energy efficiency and extended operation times [4].

Furthermore, potential-based reward shaping (PBRS) has shown promise in guiding the reinforcement learning process without altering the optimal policy. PBRS introduces reward terms based on potential functions, encouraging actions that lead to stable and energy-efficient gait patterns. By shaping the rewards, PBRS provides a mechanism for fine-tuning the learning process and improving the overall robustness of locomotion algorithms [5].

A. Related works

1) Parallel Elastic Elements Improve Energy Efficiency on the STEPPR Bipedal Walking Robot: Mazumdar et al. [4] presented a novel approach that incorporates parallel elastic elements into the leg design of the STEPPR bipedal walking robot. The researchers address the challenge of reducing energy consumption during locomotion by introducing passive elements that store and release energy during each step.

These elastic elements act as mechanical springs, absorbing energy during leg loading and releasing it during the push-off phase. By effectively recycling energy, the parallel elastic elements minimize the overall energy expenditure required for walking. The paper discusses the design considerations, materials selection, and stiffness values of the parallel elastic elements. The researchers also highlight the potential for further optimization and control strategies to maximize energy efficiency gains. The utilization of these mechanical springs reduces the robot's reliance on external power sources, leading to extended operation times and improved overall performance.

2) Benchmarking Potential Based Rewards for Learning Humanoid Locomotion: This paper [5] addresses the challenge of designing and fine-tuning reward functions in Reinforcement Learning (RL) pipelines. The researchers investigate the use of potential-based reward shaping (PBRS) as a means to guide the RL learning process without affecting the optimal policy.

While previous studies primarily focused on PBRS in grid worlds and low-dimensional systems, this paper benchmarks standard reward-shaping methods against PBRS in the context of a humanoid robot. The findings indicate that PBRS offers only marginal benefits in terms of convergence speed in high-dimensional systems. However, PBRS reward terms demonstrate greater robustness to scaling compared to typical reward shaping approaches, making them easier to tune. The paper provides empirical evidence and insights into the effectiveness of PBRS in RL for complex robotic tasks.

B. Contribution

This work proposes a positive energy oriented reward to guide the robot with less costly optimization of secondary task rewards. The energy-consumption based reward exploits the linear speed and torques for all joints in the robot. Potential-based rewards and regulations in joint space were used as for the locomotion baseline. Damping elements were introduced in the knee joints of the robot to test the policy for spring-like-legs, which are mostly found in nature. The proposed reward improved torque tracking and natural gait with lower energy consumption in flat terrain. The implementation of this work provides an approach to improve policy convergence and overall performance of bipedal walking.

II. Method

The commands were x, y linear velocities and yaw as inputs while joint torques were the actions. The observations were position and velocity for base and joints, commands, contacts, and projected gravity for a total of size of 33 with their respective scales and no noise. We used Deep Reinforcement Learning and all the training was done in simulation using 4096 environments for 1000 policy iterations and 100 Hz of control frequency. The simulator used was NVIDIA's Isaac Gym simulation environment which trains thousands of robots simultaneously using curriculum learning and Proximal Policy Optimization (PPO) [1]. The locomotion policy is illustrated in Algorithm 1.

A. Rewards

We used standard tracking rewards for joint linear and angular velocities using squared-exponential functions. For regulation in joint space, the limit and magnitude for all torques was penalized. A termination reward was established for orientation and base height z too. We also utilized PBRS rewards to regularize orientation, height, joint [5].

A novel PBRS for energy and joint regularization was implemented. The joint regularization penalizes the separation between the same joint for each leg, keeping the yaw at zero encouraging symmetric gait.

$$R_{\text{regularization}} = \frac{1}{\text{noise\_scale}} \times (0.5 \times \exp\left((q_0 - q_3) + (q_1 - q_4)\right))$$

$$R_{\text{energy}} = c \times \max\left(0, \sum_{i=1}^{6} \left\|\tau_i^{\top} q_i\right\|\right)$$

The original energy reward uses the robot's joint velocities and torques, using the sum over all joint's individual energies, a scaling factor, and the exponential function. This reward was then turned into the direct shaping reward based on the potential approach.

The general reward expression for the baseline was:

$$\begin{aligned} r_{\text{baseline}} =& c_1 |\tau|^2 + c_2 \exp\left(-\frac{|v_{x,y} - base_{x,y}|^2}{\sigma_{xy}}\right) \\ &+ c_3 \exp\left(-\frac{|w_z - base_z|^2}{\sigma_z}\right) \\ &+ c_4 \left(\frac{\tau_k \pi - \tau_{k-1} \pi}{\Delta t}\right)^2 \\ &+ c_5 \max\left(|\tau| - \beta_\tau \tau_{\max}, 0\right) \\ &+ c_6 \max\left(|q| - \beta_q q_{\max}, 0\right) \\ &+ R_{\text{regularization}} \\ &+ R_{\text{energy}} \end{aligned} \tag{1}$$

Algorithm 1 Locomotion Learning with PBRS

```
Require: Robot dynamics model $M$, Potential function $U(s)$,
    Discount factor $\gamma$, Exploration rate $\epsilon$, Learning rate $\alpha$
Ensure: Learned gait policy $\pi$
Initialize the value function $V(s)$ and gait policy $\pi(a|s)$ randomly
Initialize the replay buffer $D$
while not converged do
    Initialize the state $s$
    while episode not finished do
        Choose an action $a$ based on the current policy $\pi(a|s)$ with exploration rate $\epsilon$
        Take action $a$ and observe the next state $s'$ and energy cost $c$
        Store the transition $(s, a, c, s')$ in the replay buffer $D$
        Update $s = s'$
        Sample a minibatch of transitions $(s, a, c, s')$ from the replay buffer $D$
        Compute the potential-based reward $\tilde{c} = c + \gamma \cdot U(s') - U(s)$
        Update the value function $V(s)$ using the TD error:
        $\delta \leftarrow \tilde{c} + \gamma \cdot V(s') - V(s)$
        $V(s) \leftarrow V(s) + \alpha \cdot \delta$
        Update the gait policy $\pi(a|s)$ using the policy gradient theorem:
        $\nabla J(\theta) \leftarrow E_\pi [\nabla \log \pi(a|s) \cdot Q(s, a)]$
        $\theta \leftarrow \theta + \alpha \cdot \nabla J(\theta)$
    end while
end while
=0
```

III. Experiment

In this section, the proposed approach is tested empirically on the 6 DOF Bolt Robot. The Bolt Robot has 6DOF due to its six joints. Joint 1-3 refer to the left leg while Joint 4-6 are located in the right left shown in Fig. 1. Initially standard tracking task-oriented rewards, then PBRS novel rewards were implemented and lastly compared using Curriculum Learning and COT.

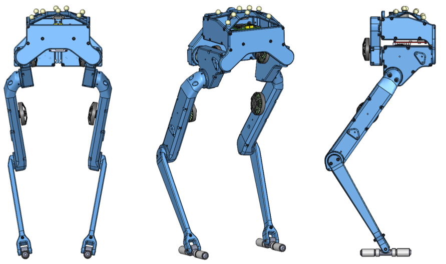
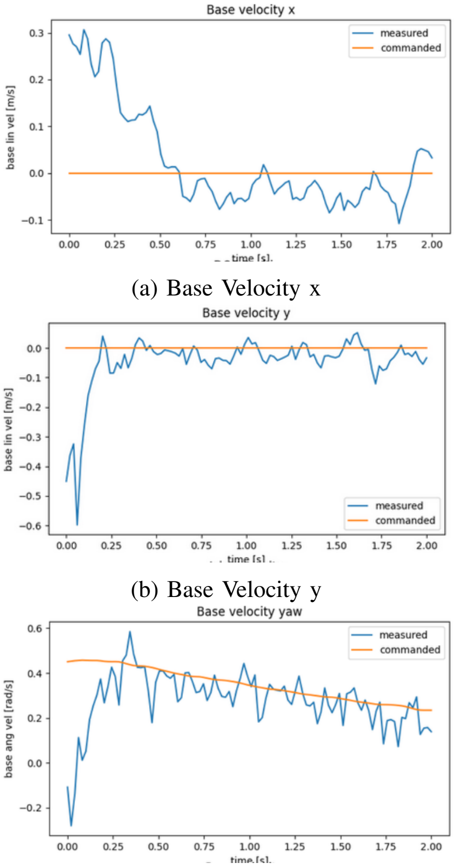


Fig. 1: Bolt Robot

IV. RESULTS

A. Reward shaping

Initial results focusing solely on positive energy have shown promise, generating a more stable gait compared to traditional energy penalties. As shown in Figure 2. The forward velocity (x direction) and the yaw velocity in the base of the robot both converges to the commanded velocity after 0.3 time step. Meanwhile, the sideways velocity converges faster, before 0.25s. However, all of the above measure velocities don't actually match the commanded values. Similarly, the torques also don't follow a uniform behaviour as seen in Figure 2 with a maximum achievable torque of 0.6.



(a) Base Velocity x

(b) Base Velocity y

(c) Base Velocity yaw

Fig. 2: Baseline Velocities

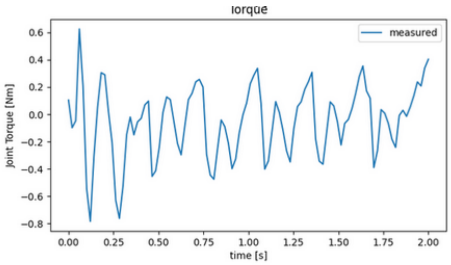
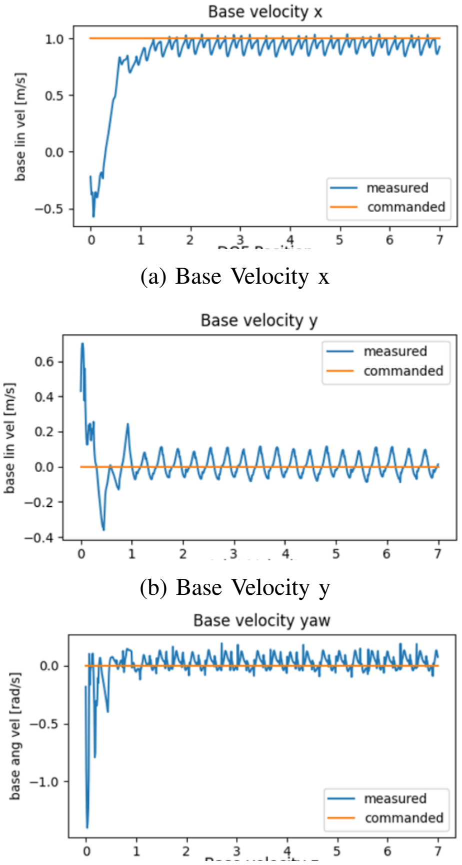


Fig. 3: Baseline Torque

On the other side when implementing the potential based rewards for forward motion we see more consistent overall results. As seen in Figure 3, the robot fully converges to the commanded velocity after 1s. Although it took more than the baseline the measure velocity is more consistent as clearly seen for Base Velocity y of Fig. 4. It takes time to reach the forward and sideways commanded velocities but for the yaw it happens instantaneously. This is because the potential based rewards in theory do encourage faster convergence compared to standard velocity/torque tracking rewards, specially if joint regularization is used. Moreover, as seen in Fig. 5 the new mean measured torque has a peak over 1 Nm if positive and well over 1.5 if negative, both bigger than the baseline torques.



(a) Base Velocity x

(b) Base Velocity y

(c) Base Velocity yaw

Fig. 4: PBRS Velocities

TABLE I: Joint Torque-Angle Relations

	Slope (Nm/rad)	Regression error	Joint Initial Angle (rad)
L Hip	-0.3390	0.3583	0.0153
L Knee	-1.5853	0.1970	0.3543
L Ankle	0.7933	0.7650	-0.6973
R Hip	-0.9900	0.4213	0.0233
R Knee	-1.2047	0.1753	0.3307
R Ankle	0.4438	0.6356	-0.4947

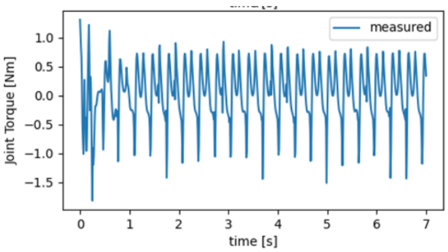


Fig. 5: PBRS Torque

B. Torque-Angle Joint Relations

With better tracking thanks to PBRS rewards, to model human-like motions the linear relations between torque and angles for all joints was analyzed. The new PBRS policy was used for different velocities in the forward motion. The joints with the lowest MSE error of the torque-angle plots were chosen as the best candidates at which virtual springs can be implemented. The velocities chosen were 0.5, 1, 1.5, 2, 3 and the average is presented in TABLE I.

From Table 1 the Left and Right Hip both show the smallest MSE error which is why we chose to model the second and fourth joints with virtual springs in the PPO training policy. With these new springs we would have new actions (torques) and can establish and new control type ”S” as seen in Algorithm 2.

Algorithm 2 Compute Torques

```
Input: $actions$
Output: $joint\ torques$
Initialize $actions\_scaled \leftarrow actions \times$ action_scale
Initialize $control\_type \leftarrow$ control_type
if $control\_type$ is ”P” then
    $torques \leftarrow k_p \times (actions\_scaled + q_{\text{default}} - \text{q}) - k_d \times \dot{q}$
else if $control\_type$ is ”V” then
    $torques \leftarrow k_p \times (actions\_scaled - \dot{q}) - k_d \times (\dot{q} - \dot{q}_{t-1})/\text{t}$
else if $control\_type$ is ”T” then
    $torques \leftarrow actions\_scaled$
else if $control\_type$ is ”S” then
    Set $q\_init \leftarrow$ [0, 0.3543, 0, 0, 0.3543, 0]
    Set $k \leftarrow$ [0, -1.5853, 0, 0, -1.5853, 0]
    $spring\_torque \leftarrow (k \times (actions - q\_init))$
    $torques \leftarrow actions\_scaled + spring\_torque$
    break
end if=0
```

C. Curriculum Learning

When adding this new spring torque to the policy’s actions the forward velocity under curriculum learning. In 10 000 iterations for training the policy the baseline converged to 2.5083 m/s while our new policy converged to 2.9876 m/s. This is a increase of 19.11% in forward velocity. The number of iterations was chosen that way as past this point the velocity didn’t significantly change.

D. Cost of Transportation

To test the performance of the approach we compared the Cost of Transportation (COT) defined as the total power over an interval of time while the positive COT refers to considering only the positive powers produced by the total sum of power from joints [7].

$$\text{COT} = \frac{\Delta W}{\text{mg}\Delta x} = \frac{1}{\text{mg}\Delta x} \sum_{j \in r,l} \int_{t_0}^{t_{\text{end}}} \left(F_{a,j}(t) \dot{h}_j(t) \right) dt \tag{2}$$

$$\text{COT}_{\text{positive}} = \frac{1}{\text{mg}\Delta x} \sum_{j \in r,l} \int_{t_0}^{t_{\text{end}}} \max \left(F_{a,j}(t) \dot{h}_j(t), 0 \right) dt \tag{3}$$

In equation (2), ΔW is the total power in a time interval, m is the robot mass, g is the acceleration due to gravity, Δx is the displacement, $F_{a,j}(t)$ represents the force of each actuator, and $\dot{h}_j(t)$ denotes the velocity of displacement for each unit of space. Equation (3) depicts the COT formula, modified to consider only positive power values.

We run the policy on 50 parallel robots and averaged the COT and positive COT at various velocities. From Fig. 6 and Figure 7. it is clear that the new policy with PBRS rewards and the found spring stiffness lead to higher COT at slow velocities until the turning point, at 1m/s for regular COT and at 2 m/s for Positive COT. After said turning points our policy leads to lower COT, the residual torque that the motor needs to produce is smaller with our method. That is because the robot can rely on the designed springs to bounce and encourage forward motion which otherwise is an additional effort at slow speeds.

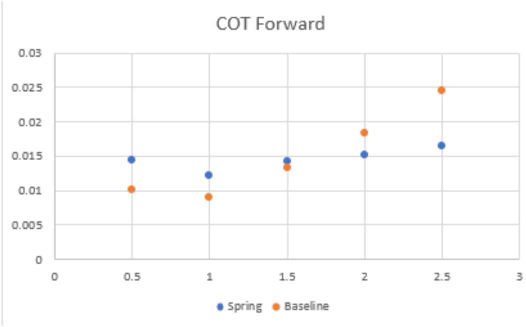


Fig. 6: Cost of Transportation

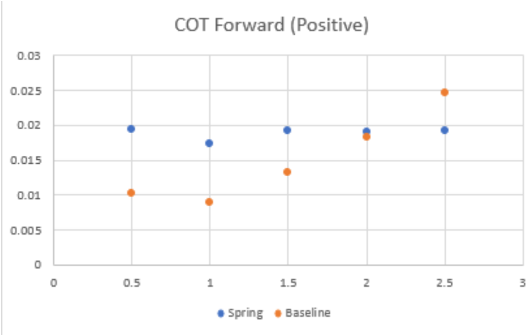


Fig. 7: Positive Cost of Transportation

V. CONCLUSION

In this work, we proposed an approach to reduce energy consumption of robots using virtual stiffness and potential based rewards on energy and joint regularization. To deal with randomness of the policy when commanded a velocity we found appropriate stiffness over various forward velocities using linear regression. Additionally, to evaluate performance we compared our new policy using curriculum learning for max forward velocity and compared the cost of transportation. We also provide considerations in why Potential Based Rewards encourage velocity convergence. Using this new control applying joint stiffness Bolt Robot successfully could move move forwards faster and reduce its energy consumption at high speeds. Future works can use sample trajectories from model based methods or include vision to model the gait of Bolt when facing an obstacle and still make full use of virtual joints in various directions.

REFERENCES

[1] Rudin, N., Hoeller, D., Reist, P., Hutter, M. (2020). Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning. *arXiv preprint arXiv:2008.10342*.

[2] Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., Hutter, M. (2019). Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), eaau5872. doi: 10.1126/scirobotics.aau5872

[3] Kormushev, P., Ugurlu, B., Calinon, S., Tsagarakis, N. G., Caldwell, D. G. (2011). Bipedal walking energy minimization by reinforcement learning with evolving policy parameterization. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2962-2968). doi: 10.1109/IROS.2011.6094427

[4] A. Mazumdar et al., "Parallel Elastic Elements Improve Energy Efficiency on the STEPPR Bipedal Walking Robot," in *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 2, pp. 898-908, April 2017, doi: 10.1109/TMECH.2016.2631170.

[5] Jeon, S. H., Heim, S., Khazoom, C., and Kim, S. (2023). "Benchmarking Potential Based Rewards for Learning Humanoid Locomotion." In *2023 IEEE International Conference on Robotics and Automation (ICRA 2023)*.

[6] Park, S., & Park, J. (2019). Vertical COM Motion Generation to Reduce Slipping and Mechanical Work during Walking. Retrieved from: http://dyros.snu.ac.kr/wp-content/uploads/2019/11/Vertical-COMmotion.pdf

[7] Koseki S, Kutsuzawa K, Owaki D and Hayashibe M (2023) Multimodal bipedal locomotion generation with passive dynamics via deep reinforcement learning. Front. Neurorobot. 16:1054239. doi: 10.3389/fnbot.2022.1054239