**Hyperparameter Tuning for PPO on LunarLander-v2**

**Introduction**

This document outlines the thought process and tests involved in tuning the hyperparameters for a PPO agent training on the LunarLander-v2 environment. The goal was to balance learning efficiency, stability, and performance.

**Hyperparameter Selection**

1. **Learning Rate** (**learning rate=0.0028**)
   * **Reasoning**: The learning rate controls how much the model updates its parameters with respect to the loss gradient. A slightly reduced learning rate from the typical 0.0032 helps in ensuring more stable learning, reducing the risk of overshooting optimal policies.
   * **Test**: Started with 0.0036 and gradually decreased to 0.0028 based on training stability and reward trends.
2. **Number of Steps** (**n\_steps=20480**)
   * **Reasoning**: The number of steps to run for each environment per update. A larger number of steps helps in more accurate gradient estimates, especially beneficial when doubling the total timesteps for better policy learning.
   * **Test**: Initially tested with 10240, observed learning curves and increased to 20480 to enhance stability and performance.
3. **Batch Size** (**batch size=512**)
   * **Reasoning**: The batch size used for each update. Larger batch sizes tend to provide more stable updates and are more efficient on modern hardware.
   * **Test**: Tested with smaller batch sizes like 256, but found 512 provided a good balance of stability and computational efficiency.
4. **Discount Factor** (**gamma=0.995**)
   * **Reasoning**: Discount factor determines how future rewards are weighted compared to immediate rewards. A value of 0.995 encourages the agent to consider slightly longer-term rewards, which is useful in the LunarLander-v2 environment.
   * **Test**: Tested standard values (0.99) and slightly increased to 0.995 for better long-term strategy learning.
5. **GAE Lambda** (**gae\_lambda=0.95**)
   * **Reasoning**: Lambda for generalized advantage estimation. Balances bias vs. variance in advantage estimation. Set to 0.95, which is a common choice that generally works well.
   * **Test**: Kept standard but evaluated in conjunction with **n\_steps** for optimal performance.
6. **Clip Range** (**clip\_range=0.2**)
   * **Reasoning**: Clipping range for PPO. Helps in maintaining the stability of policy updates by limiting the change in the policy at each step.
   * **Test**: Used the standard value of 0.2, as it typically provides a good balance of policy update constraint.
7. **Entropy Coefficient** (**ent\_coef=0.04**)
   * **Reasoning**: Coefficient for entropy bonus added to the policy loss. Encourages exploration by penalizing certainty in the policy. Adjusted to 0.04 to encourage more exploration.
   * **Test**: Started with default (0.01) and increased to 0.04 to encourage more exploration, which is critical in the initial stages of learning.
8. **Number of Epochs** (**n\_epochs=10**)
   * **Reasoning**: Number of epochs to train the model. This parameter was kept the same as it generally balances training time and convergence well.
   * **Test**: Maintained standard to ensure reliable training while avoiding overfitting.

**Training and Evaluation:**

* **Total Timesteps**: Trained for **1200000** timesteps to ensure the agent has enough interactions with the environment to learn effectively.
* **Progress Bar**: Enabled to monitor the training progress visually.

**Testing and Results**

| **Hyperparameter** | **Initial Value** | **Adjusted Value** | **Reasoning** | **Outcome** |
| --- | --- | --- | --- | --- |
| Learning Rate | 0.0036 | 0.0028 | Stability in updates | Improved learning stability |
| n\_steps | 10240 | 20480 | Accurate gradient estimates | Enhanced performance |
| Batch Size | 256 | 512 | Update stability and efficiency | More stable learning curves |
| Gamma | 0.99 | 0.995 | Long-term reward consideration | Better policy learning |
| GAE Lambda | 0.95 | 0.95 | Balance bias vs. variance | Optimal performance |
| Clip Range | 0.2 | 0.2 | Policy update constraint | Maintained stability |
| Entropy Coefficient | 0.01 | 0.04 | Encourage exploration | Increased exploration |
| Number of Epochs | 10 | 10 | Training reliability | Consistent performance |

**Conclusion**

The adjusted hyperparameters provided a stable and effective training process for the PPO agent on the LunarLander-v2 environment. The agent was able to learn a good policy that balances landing smoothly while avoiding crashes, indicating successful hyperparameter tuning. The final model was saved for potential future use or further tuning if necessary.