Policy Representation, Policy Update, and Hyperparameters

1 Policy Representation, Policy Update, and Hyperparameters

1.1 Policy Representation

The policy in this context is represented by a softmax function parameterized by θ, which maps states to action probabilities. Mathematically, the policy $\pi(a|s;\theta)$ for a given state s and action a is defined as:

$$\pi(a|s;\theta) = \frac{\exp(\theta^T s)}{\sum_{a'} \exp(\theta^T s)}$$

where:

- s is the state vector.
- θ is the parameter matrix of dimensions [state_size $\times$ action_size].
- $\pi(a|s;\theta)$ is the probability of taking action a given state s.

1.2 Policy Update

The REINFORCE algorithm updates the policy parameters θ using the gradient of the expected reward. The update rule is:

$$\theta \leftarrow \theta + \alpha \nabla_\theta E[R_t|\pi_\theta]$$

For each episode:

1. **Policy Gradient**: Compute the gradient of the log-probability of the action taken:

$$\nabla_\theta J(\theta) = E_\pi \left[\sum_{t=0}^{T} \nabla_\theta \log \pi(a_t|s_t;\theta) G_t \right]$$

The parameter update for each time step t is:

$$\theta \leftarrow \theta + \alpha \sum_{t=0}^{T} \nabla_\theta \log \pi(a_t|s_t;\theta) G_t$$

For the Baseline REINFORCE algorithm, a value function $V(s;w)$ parameterized by w is used to reduce the variance of the policy gradient. The update rule becomes:

$$\theta \leftarrow \theta + \alpha \sum_{t=0}^{T} \nabla_\theta \log \pi(a_t|s_t;\theta)(G_t - V(s_t;w))$$

where $V(s_t;w)$ is the predicted value of state s_t.

1.3 Hyperparameters

- **Episodes** (N): Number of training episodes. Example: $N = 1000$.
- **Discount Factor** (γ): Determines the importance of future rewards. Example: $\gamma = 0.99$.
- **Learning Rate** for Policy Network (α): Step size for updating policy parameters. Example: $\alpha = 0.01$.
- **Learning Rate** for Value Network (β): Step size for updating value network parameters in the baseline algorithm. Example: $\beta = 0.01$.