# A2C Agent

L'algoritmo Advantage Actor-Critic (A2C) è un metodo di apprendimento per rinforzo che combina i vantaggi dei metodi di attore-critico (actor-critic). In questo contesto, l'agente è composto da due parti principali: l'attore (actor) e il critico (critic).

## Componenti

1. Rete di Policy (Attore): la rete di policy è parametrizzata da *theta* e definisce una distribuzione probabilistica delle azioni.

2. Rete di Valore (Critico): questa rete stima il valore di un dato stato, cioè la somma attesa delle ricompense future partendo da quello stato. La rete di valore è parametrizzata *theta\_v*.

## Funzionamento

- Per ogni episodio, l'agente resetta l'ambiente e raccoglie esperienze.

- Ad ogni passo temporale:

- L'agente osserva lo stato corrente ***s\_t***.

- Seleziona un'azione ***a\_t*** campionata dalla distribuzione della policy.

- Esegue l'azione ***a\_t***e riceve la ricompensa ***r\_{t+1}*** e il prossimo stato ***s\_{t+1}*.**

- Memorizza la transizione (***s\_t, a\_t, r\_{t+1}, s\_{t+1****}*).

- Dopo aver raccolto un batch di esperienze, l'agente calcola i ritorni ***G\_t*** e i vantaggi ***A\_t***.

- Il ritorno ***G\_t*** è la somma scontata delle ricompense future

- Il vantaggio ***A\_t*** è la differenza tra il ritorno e il valore stimato dello stato

- Calcola la perdita della policy usando i vantaggi. La perdita della policy è:

***L = -Log(policy)\*Vantaggio.***

Minimizzando questa perdita, l'agente aggiorna i parametri per migliorare le probabilità delle azioni che portano a vantaggi positivi.

- Calcola la perdita del valore come l'errore quadratico medio tra i ritorni e i valori stimati. La loss è:

***L = (G\_t – V(S\_t))^2***

# PPO Agent

Proximal Policy Optimization (PPO) è un algoritmo di apprendimento per rinforzo che migliora l'approccio Actor-Critic utilizzando un metodo di ottimizzazione più stabile ed efficiente. PPO utilizza una tecnica di clipping per limitare l'aggiornamento dei parametri della politica, evitando cambiamenti troppo drastici.

## Componenti

1. Rete di Policy (Attore): la rete di policy è parametrizzata da *theta* e definisce una distribuzione probabilistica delle azioni.

2. Rete di Valore (Critico): questa rete stima il valore di un dato stato, cioè la somma attesa delle ricompense future partendo da quello stato. La rete di valore è parametrizzata *theta\_v*.

## Funzionamento

- Per ogni episodio, l'agente resetta l'ambiente e raccoglie esperienze.

- Ad ogni passo temporale:

- L'agente osserva lo stato corrente ***s\_t***.

- Seleziona un'azione ***a\_t*** campionata dalla distribuzione della policy.

- Esegue l'azione ***a\_t***e riceve la ricompensa ***r\_{t+1}*** e il prossimo stato ***s\_{t+1}*.**

- Memorizza la transizione (***s\_t, a\_t, r\_{t+1}, s\_{t+1****}*).

- Dopo aver raccolto un batch di esperienze, l'agente calcola i ritorni ***G\_t*** e i vantaggi ***A\_t***.

- Il ritorno ***G\_t*** è la somma scontata delle ricompense future

- Il vantaggio ***A\_t*** è la differenza tra il ritorno e il valore stimato dello stato

- Calcola il rapporto di probabilità tra la nuova politica e la politica vecchia per ogni azione. La perdita è ottenuta con la tecnica di clipping per limitare l'aggiornamento. Minimizzando questa perdita, l'agente aggiorna i parametri ***theta*** mantenendo i cambiamenti nella policy entro limiti ragionevoli.

- Calcola la perdita del valore come l'errore quadratico medio tra i ritorni e i valori stimati. La loss è:

***L = (G\_t – V(S\_t))^2***