Les Fondations Mathématiques de PPO

1 Ratio entre les politiques

Le ratio entre la nouvelle politique et l'ancienne est calculé comme suit:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$$

Ce ratio permet de comparer la probabilité qu'une action a_t soit choisie dans un état s_t entre la nouvelle politique π_θ et l'ancienne $\pi_{\theta_{\text{old}}}$.

2 Avantage estimé

L'avantage est une mesure de la qualité d'une action donnée par rapport à l'état. Il est défini comme :

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$$

Dans la pratique, l'avantage $A(s_t, a_t)$ est souvent approximé à l'aide des récompenses futures :

$$A(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} - V(s_t)$$

où γ est le facteur d'actualisation qui pondère les récompenses futures, et r_{t+k} représente la récompense obtenue après k étapes.

3 Fonction objectif avec clipping

La fonction principale de PPO maximise la récompense totale tout en limitant les changements brusques dans la politique :

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min\left(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right]$$

Le terme $\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)$ empêche le ratio $r_t(\theta)$ de s'éloigner trop de 1, limitant ainsi les changements excessifs.

4 Fonction de perte pour la valeur

Le réseau de la valeur (V_ϕ) est entraîné à minimiser l'erreur quadratique entre la valeur prédite et la valeur cible :

$$L^{\text{value}}(\phi) = \mathbb{E}_t\left[(V_\phi(s_t) - V_{\text{target}})^2\right]$$

Ici, V_{target} est la valeur cible calculée à partir des récompenses futures.

5 Pénalité d'entropie (optionnelle)

Pour encourager l'exploration, PPO peut ajouter un terme d'entropie à la fonction objectif :

$$L^{\text{entropy}}(\pi_\theta) = \mathbb{E}_t\left[\mathcal{H}(\pi_\theta(\cdot|s_t))\right]$$

où $\mathcal{H}(\pi_\theta)$ est l'entropie de la politique, définie comme :

$$\mathcal{H}(\pi_\theta) = -\sum_a \pi_\theta(a|s) \log \pi_\theta(a|s)$$

Ce terme favorise des politiques moins déterministes, permettant ainsi à l'agent d'explorer davantage.

6 Perte totale pour PPO

La perte totale combine trois termes:

1. La fonction objectif de la politique (L^{CLIP}).
2. La perte pour la valeur (L^{value}).
3. La régularisation d'entropie (L^{entropy}).

L'expression mathématique de la perte totale est:

$$L(\theta, \phi) = L^{\text{CLIP}}(\theta) - c_1 L^{\text{value}}(\phi) + c_2 L^{\text{entropy}}(\pi_\theta)$$

où c_1 et c_2 sont des coefficients qui pondèrent l'importance relative des termes.

7 Récompense totale attendue

L'objectif général de PPO est de maximiser les récompenses futures cumulées, actualisées avec un facteur γ:

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Ici, $\tau \sim \pi_\theta$ indique que les trajectoires τ sont générées en suivant la politique π_θ.

8 Valeur cible

La valeur cible utilisée pour entraîner le réseau de la valeur est calculée comme:

$$V_{\text{target}} = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

Cette valeur représente la somme pondérée des récompenses futures que l'agent peut espérer à partir de l'état s_t.