Interpretación de métricas en PPO

Durante el entrenamiento con `Proximal Policy Optimization (PPO)`, diversas métricas permiten diagnosticar la estabilidad y el progreso del aprendizaje. A continuación se describen los indicadores más relevantes y sus rangos típicos de valores.

1. Recompensa promedio por episodio (`ep_rew_mean`)

Mide la recompensa total promedio obtenida por episodio. Su escala depende del entorno, pero lo relevante es la tendencia.

- Si aumenta suavemente, el agente está mejorando.
- Si oscila, todavía está explorando.
- Si disminuye bruscamente, las recompensas pueden estar mal balanceadas o invertidas.

Con recompensas normalizadas, los valores suelen estar entre -10 y 10; con recompensas crudas, pueden ser mucho mayores.

2. Varianza explicada (`explained_variance`)

Indica cuánto de la varianza de los retornos futuros logra predecir el crítico.

- 0.0–0.3: el crítico aún no aprende correctamente.
- 0.3–0.6: aprendizaje intermedio, el modelo empieza a capturar la estructura.
- 0.6–0.9: entrenamiento maduro y estable.
- ¿0.9: posible sobreajuste.

3. Fracción de recorte (`clip_fraction`)

Representa el porcentaje de actualizaciones del gradiente recortadas por el parámetro `clip_range`.

- 0.05–0.20: rango óptimo.
- ≈ 0: la política no cambia (aprendizaje estancado).
- ¿0.30: actualizaciones demasiado agresivas.

4. Divergencia KL aproximada (`approx_kl`)

Mide la diferencia entre la política nueva y la anterior tras una actualización.

- 0.001–0.02: ideal.
- ¡0.001: cambios insignificantes.
- ¿0.03: cambios excesivos, posible inestabilidad.

5. Entropía de la política (`entropy_loss`)

Evalúa el grado de aleatoriedad o exploración del agente.

- Inicio: -3.5 a -2.5 (alta exploración).
- Fase media: -2.2 a -1.8 (equilibrio).
- Fase final: ≈ -1.5 (decisiones deterministas).

6. Pérdida del crítico (`value_loss`)

Error cuadrático medio entre el valor estimado y el retorno real.

- 0.01–1.0: estable.
- Valores del orden de 10^6 o mayores indican falta de normalización.

7. Pérdida del actor (`policy_gradient_loss`)

Magnitud del gradiente promedio que actualiza la política.

- -0.02 a -0.1: cambios moderados y estables.
- ≈ 0: gradientes pequeños, aprendizaje lento.
- < -0.3: actualizaciones inestables.

8. Pérdida total (`loss`)

Suma ponderada de las pérdidas de política y valor. Su escala depende de la configuración, pero debe permanecer estable. Pérdidas que crecen o explotan indican inestabilidad numérica.

9. Longitud promedio de episodios (`ep_len_mean`)

Promedio de pasos que tarda un episodio en completarse.

- Si aumenta al inicio y luego se estabiliza, el agente está explorando correctamente.
- Si crece indefinidamente, el agente podría estar atascado.

10. Métricas de rendimiento

Variables como `fps`, `iterations`, `time_elapsed` o `n_updates` solo reflejan el rendimiento computacional del entrenamiento y no afectan la calidad del aprendizaje.

Resumen de rangos típicos

Métrica	**Rango típico**	**Interpretación**
`ep_rew_mean`	tendencia ascendente	mejora del agente
`explained_variance`	0.3–0.9	crítico aprende
`clip_fraction`	0.05–0.20	cambios sanos en política
`approx_kl`	0.001–0.02	actualizaciones estables
`entropy_loss`	-3 a -1.5	de exploración a explotación
`value_loss`	0.01–1.0	crítico estable
`policy_gradient_loss`	-0.02 a -0.1	actor ajustando correctamente
`loss`	pequeño y estable	entrenamiento coherente
`ep_len_mean`	estable	episodios consistentes

En resumen, un entrenamiento de PPO se considera saludable cuando las métricas se mantienen dentro de estos rangos y muestran tendencias suaves y coherentes en lugar de fluctuaciones extremas o divergencias numéricas.