Chapter 6 Extensions

Proximal Policy Optimization

PPO's Surrogate Loss

Why the Surrogate Loss Matches A2C at Collection Time

We stated earlier that before the first update the importance-weighted direction in Equation 6.5 reduces to the usual A2C actor direction. Let's verify this directly from the surrogate loss in Equation 6.6.

At the moment the rollout is collected, $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$, so the probability ratio from Equation 6.3 is

$$\rho_t(\boldsymbol{\theta}_{\text{old}}) = \frac{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid o_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid o_t)} = 1$$

for every timestep t. To compare gradients, we differentiate the surrogate loss with respect to $\boldsymbol{\theta}$. The advantage estimate $\hat{A}_t$ and the old log-probability ℓ_t^{old} from Equation 6.2 are fixed, so only the ratio depends on $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}} \text{loss}_{\text{actor}} = -\text{mean}_t\left[\hat{A}_t \, \nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta})\right].$$

As we showed earlier, because the denominator $\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid o_t)$ is constant in $\boldsymbol{\theta}$, the log-derivative identity gives

$$\nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta}) = \rho_t(\boldsymbol{\theta}) \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t \mid o_t).$$

Evaluating this at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$, where $\rho_t(\boldsymbol{\theta}_{\text{old}}) = 1$, leaves

$$\nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}} = \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t \mid o_t)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}}.$$

Substituting back into the gradient of the surrogate loss,

$$\nabla_{\boldsymbol{\theta}} \text{loss}_{\text{actor}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}} = -\text{mean}_t\left[\hat{A}_t \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t \mid o_t)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}}\right].$$

This gives us $-g$, the negative of the on-policy actor direction from Equation 4.13, evaluated at $\boldsymbol{\theta}_{\text{old}}$. So at the moment the rollout is collected, before any optimizer step has moved $\boldsymbol{\theta}$ away from $\boldsymbol{\theta}_{\text{old}}$, minimizing the PPO surrogate loss takes the same first step as the A2C actor update. The two differ only on later minibatch updates, once $\boldsymbol{\theta}$ has changed and the ratios are no longer all equal to one.

The same holds for the clipped loss in Equation 6.7. At $\rho_t(\boldsymbol{\theta}) = 1$, the ratio lies inside the interval $[1 - \epsilon, 1 + \epsilon]$, so the clipped and unclipped terms coincide, and clipping has no effect on the first update.