Research Plan

Area of Focus

- In their paper on Proximal Policy Optimization, Schulman et. al. [1] propose the clipped surrogate loss function for a fixed parameter ϵ:

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

  where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ and $\hat{A}_t$ is the generalized advantage estimator. For simplicity, let $r_t(\theta) = r_t$. $\hat{A}_t$ can be replaced with a number of other "γ-just" estimators that must satisfy certain conditions [2] . Generalizing $\hat{A}_t$ to these estimators, which will be denoted $\hat{G}_t$, yields:

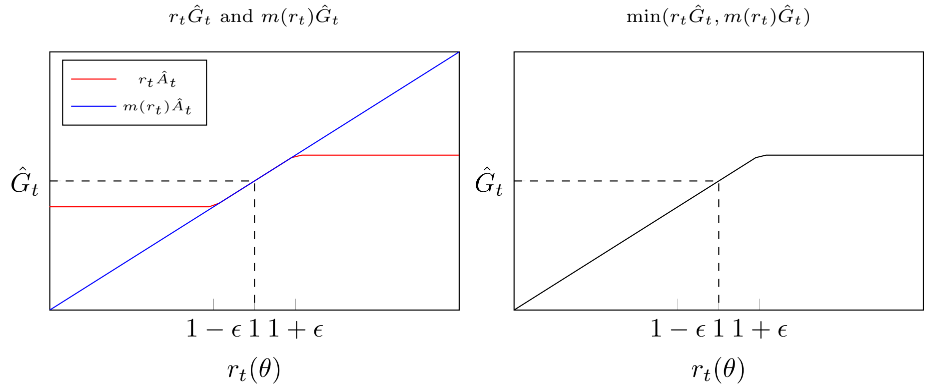
$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t\hat{G}_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{G}_t \right) \right]$$

- The goal is to investigate replacements for the clipper function $\text{clip}(r_t, 1-\epsilon, 1+\epsilon)$. Let us refer to these replacements as "min-filters," and let $m(r_t)$ denote an arbitrary min-filter.
- In this experimental framework, we have the loss function:

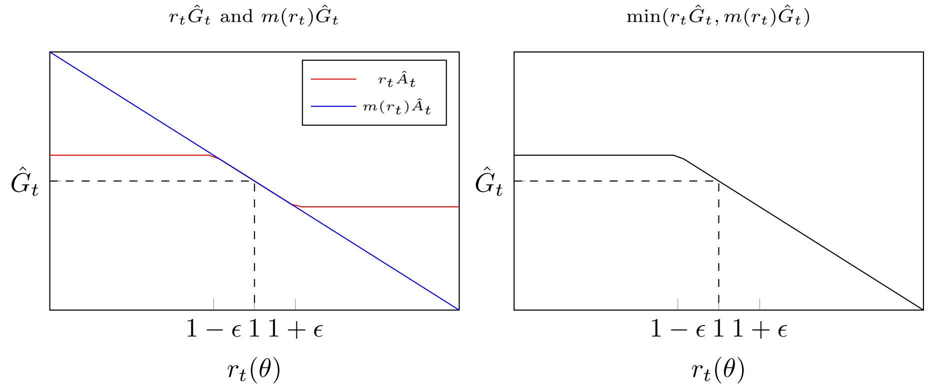
$$L^{m}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{G}_t, m(r_t)\hat{G}_t \right) \right]$$

- L^{CLIP} is simply an instance of this where $m(r_t) = \text{clip}(r_t, 1-\epsilon, 1+\epsilon)$.

– Illustrating minimization under L_{CLIP} on individual expectation components:



Expectation component, $\hat{G}_t > 0$



Expectation component, $\hat{G}_t < 0$

– The paper on Trust Region Policy Optimization by Schulman et. al. [3] proposes a target function whose maximization guarantees monotonic improvement:

$$targ(\theta) = L_{\theta_{old}}(\theta) - CD_{KL}^{max}(\theta, \theta_{old})$$

where C is a fixed positive constant (see paper for specifics) and it is shown that

$$L_{\theta_{old}}(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim p_{\theta_{old}}, a\sim\theta_{old}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}A_{\theta_{old}}(s,a)\right]$$

where $p_{\theta_{old}}$ is the normalized discounted visitation frequency distribution.

– Assuming that the on-policy distribution matches the normalized discounted visitation frequency distribution, we can write:

$$L_{\theta_{old}}(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{t\in(1,\dots\infty)}\left[r_t\hat{A}_t\right]$$

- By definition, any γ-just estimator can replace $\hat{A}_t$ because doing so only adds a constant to $targ(\theta)$. Therefore, we can redefine $L_{\theta_{old}}(\theta)$ as:

$$L_{g,\theta_{old}}(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{t\in(1,\dots\infty)}\left[r_t\hat{G}_t\right]$$

- Plugging into the target function, multiplying by $1-\gamma$, and absorbing $1-\gamma$ into C leaves us with the gradient-equivalent target function:

$$\begin{aligned} targ_g(\theta) &= \mathbb{E}_t\left[r_t\hat{G}_t\right] - CD_{KL}^{max}(\theta,\theta_{old}) \\ \nabla_\theta targ_g(\theta) &= \nabla_\theta targ(\theta) \end{aligned}$$

- Consider the case where $\forall t \in (1,\dots\infty)$, $\hat{G}_t > 0$ and $r_t < 1+\epsilon$ and let $\theta \neq \theta_{old}$. In this case, no penalty is applied and the clipped loss is a strict overestimate without the same gradient:

$$\begin{aligned} L^{CLIP}(\theta) &= \hat{\mathbb{E}}_t\left[\min\left(r_t\hat{G}_t, \mathrm{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{G}_t\right)\right] \\ &= \hat{\mathbb{E}}_t\left[r_t\hat{G}_t\right] \\ &\geq \mathbb{E}_t\left[r_t\hat{G}_t\right] - CD_{KL}^{max}(\theta,\theta_{old}) \\ &= targ_g(\theta) \\ \nabla_\theta L^{CLIP}(\theta) &= \nabla_\theta\hat{\mathbb{E}}_t\left[r_t\hat{G}_t\right] \\ &\neq \nabla_\theta\mathbb{E}_t\left[r_t\hat{G}_t\right] - C\nabla_\theta D_{KL}^{max}(\theta,\theta_{old}) \\ &= \nabla_\theta targ_g(\theta) \end{aligned}$$

- Removing the assumption that $\hat{G}_t > 0$, the above still holds only if, for all positive $\hat{G}_t$, $r_t < 1+\epsilon$, and for all negative $\hat{G}_t$, $r_t > 1-\epsilon$.
- If r_t is independent of the sign of $\hat{G}_t$, this is generally a harder condition to meet. Experimentally, I found that, on almost every batch, the number of timesteps t where $r_t < 1+\epsilon$ was greater than the number of timesteps where ($\hat{A}_t < 0$ and $r_t > 1-\epsilon$) or ($\hat{A}_t > 0$ and $r_t < 1+\epsilon$). This means that, if $\hat{G}_t$ can be both positive and negative, penalties become more possible, allowing $L^{CLIP}(\theta)$ to better approximate $targ_g(\theta)$, better guaranteeing monotonic improvement.
- This reasoning could explain the preference for advantage estimators over value estimators, because the condition that $\mathbb{E}_t(\hat{A}_t) = 0$ requires that advantage estimators be negative half the time, while value functions are typically always positive or always negative.
- Research question: In some cases, it is simpler to implement a value estimator than an advantage estimator. Can we design a min-filter that specifically addresses the above concerns to make it more feasable to use a value estimator in Proximal Policy Optimization?

References

[1] https://arxiv.org/abs/1707.06347

[2] https://arxiv.org/abs/1506.02438

[3] https://arxiv.org/abs/1502.05477