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:

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

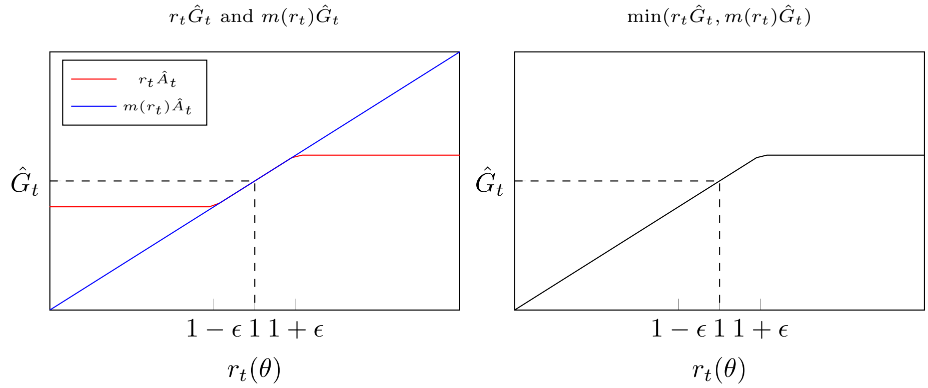
- The goal is to investigate replacements for the clipper function $\text{clip}(r_t, 1-\epsilon, 1+\epsilon)$. Let us refer to these replacements as "min-filters," and let $m(r_t)$ denote an arbitrary min-filter.

- In this experimental framework, we have the loss function:

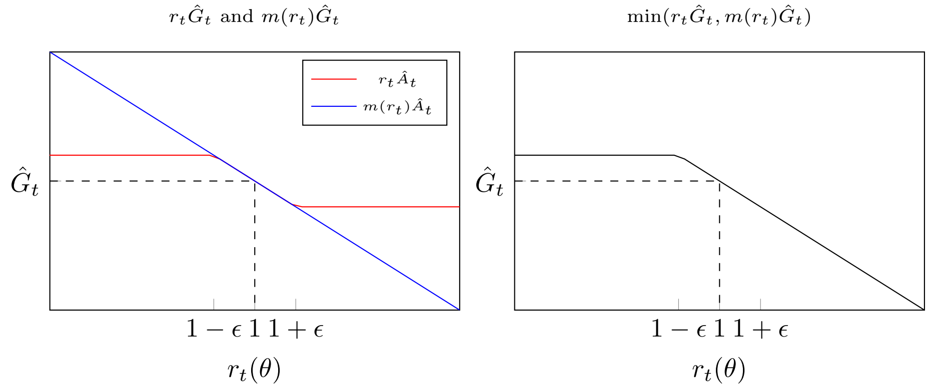
$$L^{m}(\theta) = \hat{\mathbb{E}}_t \left[\min\left(r_t(\theta)\hat{G}_t, m(r_t)\hat{G}_t\right)\right]$$

- 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:

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

- 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?

L^{CLIP} Penalty Differences for Different Estimate Models

- Consider the set of expectation-component parameters $((r_1, \hat{G}_1), (r_2, \hat{G}_2), \ldots (r_T, \hat{G}_T))$, where all r and $\hat{G}$ are uncorrelated.
- Under L^{CLIP}, a particular timestep t will be penalized in either of two cases:
  - $\hat{G}_t$ is positive and $r_t > 1 + \epsilon$.
  - $\hat{G}_t$ is negative and $r_t < 1 - \epsilon$.
- Therefore, by the assumption of independence of $\hat{G}_t$ and r_t, we have the expected number of penalized timesteps:
  $(p(\hat{G}_t > 0)p(r_t > 1 + \epsilon) + p(\hat{G}_t < 0)p(r_t < 1 - \epsilon))T$.
- Let $\hat{G}_{1,t}$ and $\hat{G}_{2,t}$ be two alterate estimators. To understand differences in the expected number of penalized timesteps as we modify the sign of $\hat{G}$, define the ratio:

$$
\begin{aligned}
r_{diff} &= \frac{(p(\hat{G}_{1,t} > 0)p(r_t > 1 + \epsilon) + p(\hat{G}_{1,t} < 0)p(r_t < 1 - \epsilon))T}{(p(\hat{G}_{2,t} > 0)p(r_t > 1 + \epsilon) + p(\hat{G}_{2,t} < 0)p(r_t < 1 - \epsilon))T} \\
&= \frac{p(\hat{G}_{1,t} > 0)p(r_t > 1 + \epsilon) + p(\hat{G}_{1,t} < 0)p(r_t < 1 - \epsilon)}{p(\hat{G}_{2,t} > 0)p(r_t > 1 + \epsilon) + p(\hat{G}_{2,t} < 0)p(r_t < 1 - \epsilon)}
\end{aligned}
$$

- If $p(r_t > 1 + \epsilon) = p(r_t < 1 - \epsilon)$, this ratio degenerates to 1 regardless of the sign distributions of $\hat{G}_1$ and $\hat{G}_2$.
- Consider the example where $p(\hat{G}_{1,t} > 0) = p(\hat{G}_{1,t} < 0) = 0.5$ and $p(\hat{G}_{2,t} > 0) = 1$, $p(\hat{G}_{2,t} < 0) = 0$. Finding the conditions under which $r_{diff} > 1$:

$$
\begin{aligned}
r_{diff} &> 1 \\
\frac{0.5(p(r_t > 1 + \epsilon) + p(r_t < 1 - \epsilon))}{p(r_t > 1 + \epsilon)} &> 1 \\
\frac{p(r_t > 1 + \epsilon) + p(r_t < 1 - \epsilon)}{p(r_t > 1 + \epsilon)} &> 2 \\
p(r_t > 1 + \epsilon) + p(r_t < 1 - \epsilon) &> 2p(r_t > 1 + \epsilon) \\
p(r_t < 1 - \epsilon) &> p(r_t > 1 + \epsilon)
\end{aligned}
$$

- Consider a continuous action space and gaussian policies with trainable but state-independent standard deviations.
- In general, on a particular state s, the standard deviation encoded by θ will decrease as the agent becomes more certain of its actions, and the mean will get further from the mean encoded by θ_{old}. Visualizing the overlaid gaussians, both of these actions will make it more likely that the above condition is true.
- Therefore, as training progresses in a single iteration, we expect that the $\hat{G}_1$ estimate will begin to induce penalties on more timesteps than the $\hat{G}_2$ estimate.

- Following similar logic as above, we have that if $p(\hat{G}_{2,t} > 0) = 0$, $p(\hat{G}_{2,t} < 0) = 1$, the condition $r_{diff} > 1$ requires that $p(r_t < 1-\epsilon) < p(r_t > 1+\epsilon)$. However, because we have just reasoned that the opposite relation tends to be true as an iteration progresses, it must be be the case that $r_{diff} < 1$ - that is, using such an estimator results in more penalized timesteps.

- Testing this theory empirically on the InvertedPendulum-v2 environment with a standard PPO agent and an advatage estimate $\hat{G}_t$, I observed that, in a single iteration, the number of $(\hat{G}_t, r_t)$ where ($\hat{G}_t > 0$ and $r_t > 1+\epsilon$) or ($\hat{G}_t < 0$ and $r_t < 1-\epsilon$) became consitently greater than the number of r_t where $r_t > 1+\epsilon$, and consistently less than the number of r_t where $r_t < 1-\epsilon$. This is in agreement with the above theoretical results.

Expected Loss Contributions

- Assume a gaussian action space with fixed standard deviations and clipping min-filter.
- It can be shown that the point at which $r_t = 1 + \epsilon$ is:

$$x^{+} = \frac{(\mu^2 - \mu_{old}^2) + 2\sigma^2 \ln(1+\epsilon)}{2(\mu - \mu_{old})}$$

- Similarly, it can be shown that the point at which $r_t = 1 - \epsilon$ is:

$$x^{-} = \frac{(\mu^2 - \mu_{old}^2) + 2\sigma^2 \ln(1-\epsilon)}{2(\mu - \mu_{old})}$$

- Let $p(\mu, x)$ be the probability of x given a gaussian distribution with fixed standard deviation σ and mean μ.
- Solving for the expected ratio coefficients for positive estimators:

$$\begin{aligned}
E[r_{t,CLIP}^{+}] &= \int_{-\infty}^{x^+} p(\mu_{old}, x) r_t(x) dx + \int_{x^+}^{\infty} p(\mu_{old}, x)(1+\epsilon) dx \\
&= \int_{-\infty}^{x^+} p(\mu_{old}, x) \frac{p(\mu, x)}{p(\mu_{old}, x)} dx + (1+\epsilon) \int_{x^+}^{\infty} p(\mu_{old}, x) dx \\
&= \int_{-\infty}^{x^+} p(\mu, x) dx + (1+\epsilon) \int_{x^+}^{\infty} p(\mu_{old}, x) dx
\end{aligned}$$

- Finding the expected penalty contribution:

$$\begin{aligned}
1 - E[r_{t,CLIP}^{+}] &= 1 - \int_{-\infty}^{x^+} p(\mu, x) dx - (1+\epsilon) \int_{x^+}^{\infty} p(\mu_{old}, x) dx \\
&= \int_{x^+}^{\infty} p(\mu, x) dx - (1+\epsilon) \int_{x^+}^{\infty} p(\mu_{old}, x) dx \\
&= \int_{x^+}^{\infty} p(\mu, x) - (1+\epsilon) p(\mu_{old}, x) dx
\end{aligned}$$

- Similarly, solving for the expected ratio coefficients for negative estimators:

$$\begin{aligned}
E[r_{t,CLIP}^{-}] &= \int_{-\infty}^{x^-} p(\mu_{old}, x)(1-\epsilon) dx + \int_{x^-}^{\infty} p(\mu_{old}, x) r_t(x) dx \\
&= (1-\epsilon) \int_{-\infty}^{x^-} p(\mu_{old}, x) dx + \int_{x^-}^{\infty} p(\mu, x) dx
\end{aligned}$$

– Finding the expected penalty contribution:

$$
\begin{aligned}
E[r^-_{t,CLIP}] - 1 &= \int_{-\infty}^{x^-} p(\mu_{old}, x)(1-\epsilon)dx + \int_{x^-}^{\infty} p(\mu, x)dx - 1 \\
&= -\left(1 - \int_{-\infty}^{x^-} p(\mu_{old}, x)(1-\epsilon)dx - \int_{x^-}^{\infty} p(\mu, x)dx\right) \\
&= -\left(-\int_{-\infty}^{x^-} p(\mu_{old}, x)(1-\epsilon)dx + \int_{-\infty}^{x^-} p(\mu, x)dx\right) \\
&= \int_{-\infty}^{x^-} p(\mu_{old}, x)(1-\epsilon)dx - \int_{-\infty}^{x^-} p(\mu, x)dx \\
&= \int_{-\infty}^{x^-} p(\mu_{old}, x)(1-\epsilon) - p(\mu, x)dx
\end{aligned}
$$

– Alternatively, this can be written as:

$$
\begin{aligned}
E[r^-_{t,CLIP}] - 1 &= \int_{-\infty}^{x^-} p(\mu_{old}, x)(1-\epsilon)dx - \int_{-\infty}^{x^-} p(\mu, x)dx \\
&= (1-\epsilon)\left(1 - \left(\int_{x^-}^{x^+} p(\mu_{old}, x)dx + \int_{x^+}^{\infty} p(\mu_{old}, x)dx\right)\right) - \\
&\left(1 - \left(\int_{x^-}^{x^+} p(\mu, x)dx + \int_{x^+}^{\infty} p(\mu, x)dx\right)\right) \\
&= (1-\epsilon)\left(1 - \int_{x^-}^{x^+} p(\mu_{old}, x)dx - \int_{x^+}^{\infty} p(\mu_{old}, x)dx\right) - \\
&\left(1 - \int_{x^-}^{x^+} p(\mu, x)dx - \int_{x^+}^{\infty} p(\mu, x)dx\right) \\
&= 1 - \epsilon - (1-\epsilon)\left(\int_{x^-}^{x^+} p(\mu_{old}, x)dx + \int_{x^+}^{\infty} p(\mu_{old}, x)dx\right) - 1 \\
&+ \int_{x^-}^{x^+} p(\mu, x)dx + \int_{x^+}^{\infty} p(\mu, x)dx \\
&= -\epsilon - (1-\epsilon)\left(\int_{x^-}^{x^+} p(\mu_{old}, x)dx + \int_{x^+}^{\infty} p(\mu_{old}, x)dx\right) \\
&+ \int_{x^-}^{x^+} p(\mu, x)dx + \int_{x^+}^{\infty} p(\mu, x)dx
\end{aligned}
$$

– Finding the differences between these two values

$$
\begin{aligned}
(1 - E[r^+_{t,CLIP}]) - (E[r^-_{t,CLIP}] - 1) &= \int_{x^+}^{\infty} p(\mu, x)dx - (1+\epsilon)\int_{x^+}^{\infty} p(\mu_{old}, x)dx \\
&+ \epsilon + (1-\epsilon)\left(\int_{x^-}^{x^+} p(\mu_{old}, x)dx + \int_{x^+}^{\infty} p(\mu_{old}, x)dx\right) \\
&- \int_{x^-}^{x^+} p(\mu, x)dx - \int_{x^+}^{\infty} p(\mu, x)dx \\
&= -(1+\epsilon)\int_{x^+}^{\infty} p(\mu_{old}, x)dx \\
&+ \epsilon + (1-\epsilon)\left(\int_{x^-}^{x^+} p(\mu_{old}, x)dx + \int_{x^+}^{\infty} p(\mu_{old}, x)dx\right) \\
&- \int_{x^-}^{x^+} p(\mu, x)dx \\
&= -\epsilon\int_{x^+}^{\infty} p(\mu_{old}, x)dx \\
&+ \epsilon + \left((1-\epsilon)\int_{x^-}^{x^+} p(\mu_{old}, x)dx - \epsilon\int_{x^+}^{\infty} p(\mu_{old}, x)dx\right) \\
&- \int_{x^-}^{x^+} p(\mu, x)dx \\
&= -2\epsilon\int_{x^+}^{\infty} p(\mu_{old}, x)dx \\
&+ \epsilon + (1-\epsilon)\int_{x^-}^{x^+} p(\mu_{old}, x)dx \\
&- \int_{x^-}^{x^+} p(\mu, x)dx \\
&= \epsilon + (1-\epsilon)\int_{x^-}^{x^+} p(\mu_{old}, x)dx \\
&- \left(\int_{x^-}^{x^+} p(\mu, x)dx + 2\epsilon\int_{x^+}^{\infty} p(\mu_{old}, x)dx\right)
\end{aligned}
$$

– Generalizing the results to two ϵ:

$$
\begin{aligned}
(1 - E[r^+_{t,CLIP}]) - (E[r^-_{t,CLIP}] - 1) &= \epsilon^- + (1-\epsilon^-)\int_{x^-}^{x^+} p(\mu_{old}, x)dx \\
&- \left(\int_{x^-}^{x^+} p(\mu, x)dx + (\epsilon^+ + \epsilon^-)\int_{x^+}^{\infty} p(\mu_{old}, x)dx\right)
\end{aligned}
$$

References

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

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

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