Proximal Policy Optimization Algorithms

[Schulman et al., 2017]

Seminar Computational Intelligence A
WS 2019
Recap Policy Gradient Methods

Policy gradient (PG) methods aim to maximize the expected return

\[ G(\pi_\theta) = \mathbb{E}_{\rho_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{\pi_\theta, \rho_0} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right] \]

by learning a parametrized policy \( \pi_\theta \), using gradient-based optimization, e.g.

\[ \theta_{t+1} = \theta_t + \alpha \nabla_\theta G(\pi_{\theta_t}) \]
Recap Policy Gradient Methods

Vanilla policy gradient:

\[ \theta_{t+1} = \theta_t + \alpha \gamma^t \sum_{k=0}^{T} \gamma^k r_{t+k} \frac{\nabla_{\theta} \pi(a_t|s_t, \theta_t)}{\pi(a_t|s_t, \theta_t)} \]
Recap Policy Gradient Methods

Vanilla policy gradient:

\[ \theta_{t+1} = \theta_t + \alpha \gamma^t \sum_{k=0}^{T} \gamma^k r_{t+k} \nabla_{\theta} \frac{\pi(a_t | s_t, \theta_t)}{\pi(a_t | s_t, \theta) \pi(a_t | s_t, \theta_t)} \]

Advantage Actor-Critic (A2C) [Mnih et al., 2016]:

\[ \theta_{t+1} = \theta_t + \alpha \gamma^t A_{\pi_\theta}(s_t, a_t) \nabla_{\theta} \frac{\pi(a_t | s_t, \theta_t)}{\pi(a_t | s_t, \theta) \pi(a_t | s_t, \theta_t)} \]
Advantage Function Estimation

The advantage function is defined as

$$A_{\pi \theta} (s_t, a_t) = Q_{\pi \theta} (s_t, a_t) - V_{\pi \theta} (s_t)$$

$$= Q_{\pi \theta} (s_t, a_t) - \mathbb{E}_{\pi \theta} \left( Q_{\pi \theta} (s_t, \cdot) \right)$$
Advantage Function Estimation

The advantage function is defined as

$$A_{\pi_\theta}(s_t, a_t) = Q_{\pi_\theta}(s_t, a_t) - V_{\pi_\theta}(s_t)$$

$$= Q_{\pi_\theta}(s_t, a_t) - \mathbb{E}_{\pi_\theta}\left(Q_{\pi_\theta}(s_t, \cdot)\right)$$

and can be estimated as

$$\hat{A}_{\pi_\theta}(s_t, a_t) = \sum_{k=0}^{T-1} \gamma^k r_{t+k} + \gamma^T \hat{V}(s_{t+T}) - \hat{V}(s_t)$$
Motivation for Proximal Policy Optimization

Vanilla policy gradient methods suffer from

- Poor data efficiency

- Poor robustness properties (hyperparameter tuning)
The PPO Algorithm

Algorithm 1 PPO, Actor-Critic Style

\begin{algorithm}
\begin{algorithmic}
\FOR {iteration = 1, 2, \ldots}
    \FOR {actor = 1, 2, \ldots, N}
        Run policy $\pi_{\theta_{old}}$ in environment for $T$ timesteps
        Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
    \ENDFOR
    Optimize surrogate $L$ \textit{wrt} $\theta$, with $K$ epochs and minibatch size $M \leq NT$
    $\theta_{old} \leftarrow \theta$
\ENDFOR
\end{algorithmic}
\end{algorithm}
Conservative Policy Iteration

[Kakade & Langford, 2002] showed, that:

\[
G(\pi) = G(\pi_{\text{old}}) + \sum_s \mu_{\pi}(s) \sum_a \pi(a|s) A_{\pi_{\text{old}}}(s, a)
\]
Conservative Policy Iteration

[Kakade & Langford, 2002] showed, that:

\[ G(\pi) = G(\pi_{old}) + \sum_s \mu_\pi(s) \sum_a \pi(a|s) A_{\pi_{old}}(s, a) \]

\[ \approx G(\pi_{old}) + \sum_s \mu_{\pi_{old}}(s) \sum_a \pi(a|s) A_{\pi_{old}}(s, a) \]
Conservative Policy Iteration

[Kakade & Langford, 2002] showed, that:

\[
G(\pi) = G(\pi_{old}) + \sum_{s} \mu_{\pi}(s) \sum_{a} \pi(a|s) A_{\pi_{old}}(s, a)
\]

\[
\approx G(\pi_{old}) + \sum_{s} \mu_{\pi_{old}}(s) \sum_{a} \pi(a|s) A_{\pi_{old}}(s, a)
\]

Policy update $\pi_{old} \rightarrow \pi$ increases performance for positive expected advantage!
Conservative Policy Iteration

For parametrized policies [Schulman et al., 2017]:

\[
\max_{\theta} G(\pi_{\theta_{old}}) + \sum_{s} \mu_{\pi_{\theta_{old}}}(s) \sum_{a} \pi_{\theta}(a|s) A_{\pi_{\theta_{old}}}(s, a)
\]

\[
\leftrightarrow \max_{\theta} \mathbb{E}_{\pi_{\theta}, \rho_{0}} \left( \frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\theta_{old}}(s, a) \right)
\]
Conservative Policy Iteration

For parametrized policies [Schulman et al., 2017]:

\[
\max_{\theta} G(\pi_{\theta_{old}}) + \sum_{s} \mu_{\pi_{\theta_{old}}}(s) \sum_{a} \pi_{\theta}(a|s) A_{\pi_{\theta_{old}}}(s, a)
\]

\[
\leftrightarrow \max_{\theta} \mathbb{E}_{\pi_{\theta}, \rho_{0}} \left( \frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\theta_{old}}(s, a) \right)
\]

Destructively large policy updates!
The Kullback-Leibler Divergence

The Kullback-Leibler Divergence (KLD) between two probability distributions \( p(x) \) and \( q(x) \) is defined as

\[
\bar{D}(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}
\]

and serves as a similarity measure (\( \bar{D}(p||q) = 0 \) indicates equality).
Solution A: Trust Region Policy Optimization

Trust Region Policy Optimization (TRPO) [Schulman et al., 2017]:

\[
\max_{\theta} \mathbb{E}_{\pi_{\theta}, \rho_0} \left[ \frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\theta_{old}}(s, a) \right]
\]

subject to \( \mathbb{E}_{\rho_0}[D(\pi_{\theta}(.|s) || \pi_{\theta_{old}}(.|s))] \leq \delta \)
Solution A: Trust Region Policy Optimization

Trust Region Policy Optimization (TRPO) [Schulman et al., 2017]:

\[
\max_{\theta} \mathbb{E}_{\pi, \rho_0} \left[ \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\theta_{old}}(s, a) \right]
\]

subject to \( \mathbb{E}_{\rho_0}[\bar{D}(\pi_\theta(\cdot|s) \mid \mid \pi_{\theta_{old}}(\cdot|s))] \leq \delta \)

Complicated implementation & costly optimization!
Solution B: Clipped Surrogate Objective

Let $\xi(\theta) := \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}$ (i.e. $\xi(\theta_{old}) = 1$), then

$$L_{CLIP} := \mathbb{E}_{\pi_\theta, \rho_0}\left[\min\{\xi(\theta) \cdot A_{\theta_{old}}(s, a), \text{clip}(\xi(\theta), 1 - \epsilon, 1 + \epsilon) \cdot A_{\theta_{old}}(s, a)\}\right]$$
Solution B: Clipped Surrogate Objective

Let \( \xi(\theta) := \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} \) (i.e. \( \xi(\theta_{old}) = 1 \)), then

\[
L^{CLIP} := \mathbb{E}_{\pi_\theta, \rho_0} \left[ \min \left\{ \xi(\theta) \cdot A_{\theta_{old}}(s, a), \text{clip}(\xi(\theta), 1 - \epsilon, 1 + \epsilon) \cdot A_{\theta_{old}}(s, a) \right\} \right]
\]
Solution B: Clipped Surrogate Objective

\[ L^{\text{CLIP}} \]

For \( A > 0 \):
- Graph showing a linear increase from 0 to 1, then a flat line from 1 to \( 1 + \epsilon \).

For \( A < 0 \):
- Graph showing a linear decrease from 0 to \( 1 - \epsilon \), then a flat line from \( 1 - \epsilon \) to 1.
Solution B: Clipped Surrogate Objective

\[
\nabla_\theta L^{\text{CLIP}} = \begin{cases} 
0 & \text{if } A > 0 \\
\nabla_\theta \pi_\theta(a|s) \left( \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} \right) A_{\theta_{\text{old}}}(s, a) & \text{if } A < 0
\end{cases}
\]
The PPO Algorithm

**Algorithm 1** PPO, Actor-Critic Style

```plaintext
for iteration=1, 2, ... do
    for actor=1, 2, ..., N do
        Run policy \( \pi_{\theta_{old}} \) in environment for \( T \) timesteps
        Compute advantage estimates \( \hat{A}_1, \ldots, \hat{A}_T \)
    end for

    Optimize surrogate \( L \) wrt \( \theta \), with \( K \) epochs and minibatch size \( M \leq NT \)

    \( \theta_{old} \leftarrow \theta \)
end for
```
The PPO Algorithm

Optimize surrogate $L$ wrt $\theta$, with $K$ epochs and minibatch size $M \leq NT$

$\theta \leftarrow \theta_{old}$

for epochs $k = 1, 2, ..K$ do

for each minibatch do

$\hat{g} \leftarrow 0$

for each sample $(a_i, s_i, \hat{A}_i), i = 1, 2, ..M$ in minibatch do

$\hat{g} \leftarrow \hat{g} + \nabla_\theta L^{CLIP}(a_i, s_i, \hat{A}_i)$

end for

$\theta \leftarrow \theta + \frac{\alpha}{M}\hat{g}$

end for

end for

$\theta_{old} \leftarrow \theta$
Performance on Continuous Domain Problems
## Performance on Atari Games

| Game            | A2C   | ACER  | PPO    | Jamesbond | Kangaroo | Krull    | KungFuMaster | MontezumaRevenge | MsPacman | NameThisGame | Pitfall | Pong | PrivateEye | Qbert | Riverraid | RoadRunner | Robotank | Seaquest | SpaceInvaders | StarGunner | Tennis | TimePilot | Tutankham | UpNDown | Venture | VideoPinball | WizardOfWor | Zaxxon |
|-----------------|-------|-------|--------|-----------|----------|---------|-----------|---------------|------------------|----------|--------------|---------|-----|------------|-------|----------|------------|---------|----------|-------------|------------|--------|----------|------------|---------|--------|------------|------------|--------|
| Alien           | 1141.7| 1655.4| **1850.3** |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Amidar          | 380.8 | **827.6** | 674.6  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Assault         | 1562.9| 4653.8 | **4971.9** |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Asterix         | 3176.3| **6801.2** | 4532.5  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Asteroids       | 1653.3| **2389.3** | 2097.5  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Atlantis        | 729265.3| 1841376.0 | **2311815.0** |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| BankHeist       | 1095.3| 1177.5 | **1280.6** |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| BattleZone      | 3080.0| 8983.3 | **17366.7** |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| BeamRider       | 3031.7| **3863.3** | 1590.0  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Bowling         | 30.1  | 33.3  | **40.1**   |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Boxing          | 17.7  | 98.9  | 94.6     |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Breakout        | 303.0 | **456.4**  | 274.8   |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Centipede       | 3496.5| 8904.8 | 4386.4  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| ChopperCommand  | 1171.7| **5287.7** | 3516.3  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| CrazyClimber    | 107770.0| **132461.0** | 110202.0 |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| DemonAttack     | 6639.1| **38808.3** | 11378.4 |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| DoubleDunk      | -16.2 | **-13.2** | -14.9   |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Enduro          | 0.0   | 0.0   | **358.3**  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| FishingDerby    | 20.6  | **34.7**  | 17.8    |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Freeway         | 0.0   | 0.0   | **32.5**   |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Frostbite       | 261.8 | 285.6  | **314.2**  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Gopher          | 1500.9| **37802.3** | 2932.9  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| Gravitar        | 194.0 | 225.3  | **737.2**  |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
| IceHockey       | -6.4  | -5.9  | **-4.2**   |          |          |         |           |                |                  |          |              |         |     |            |       |          |            |         |          |              |          |        |            |          |        |
### Performance on Atari Games

<table>
<thead>
<tr>
<th>Description</th>
<th>A2C</th>
<th>ACER</th>
<th>PPO</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) avg. episode reward over all of training</td>
<td>1</td>
<td>18</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>(2) avg. episode reward over last 100 episodes</td>
<td>1</td>
<td>28</td>
<td>19</td>
<td>1</td>
</tr>
</tbody>
</table>