Flappy Bird
Frame-Based Policy Gradient

Fall 2022

Two New Knowledge

- Generalized Advantage Estimation (GAE)
  - UC Berkley CS285: Lecture 6, Part 4
    - Video
    - Slides
  - Original paper
- Proximal Policy Optimization (PPO)
  - University of Waterloo CS885: Lecture 15b
    - Video
    - Slides
  - Original paper

Outline

- Recap
- GAE
- PPO

Recap

- Policy gradient has 2 parts
  - Left part is a log probability of executing an action
  - Right part is an advantage term.

∇ log prob. of actions $\times$

1. $\sum_{t=0}^{\infty} r_t$: total reward of the trajectory.
2. $\sum_{t'=t}^{\infty} r_{t'}$: reward following action a_t.
3. $\sum_{t'=t}^{\infty} r_{t'} - b(s_t)$: baselined version of previous formula.
4. $Q^{\pi}(s_t, a_t)$: state-action value function.
5. $A^{\pi}(s_t, a_t)$: advantage function.
6. $r_t + V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$: TD residual.

Left part

Right part: Several formula can be chosen

Outline

- Recap
- GAE
- PPO

Generalized Advantage Estimation (GAE)

Eligibility traces & n-step returns

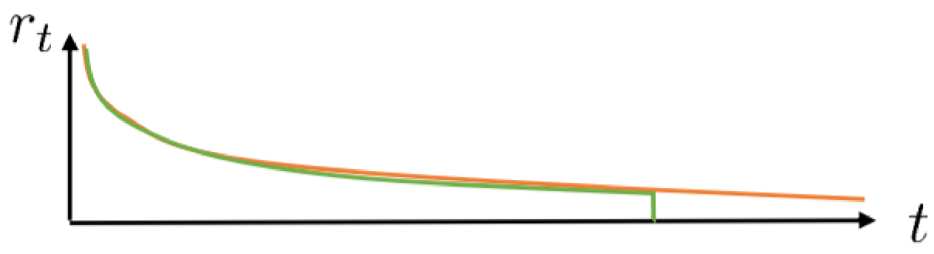
$$\hat{A}^{\pi}_{\mathrm{C}}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \hat{V}^{\pi}_{\phi}(\mathbf{s}_{t+1}) - \hat{V}^{\pi}_{\phi}(\mathbf{s}_t)$$

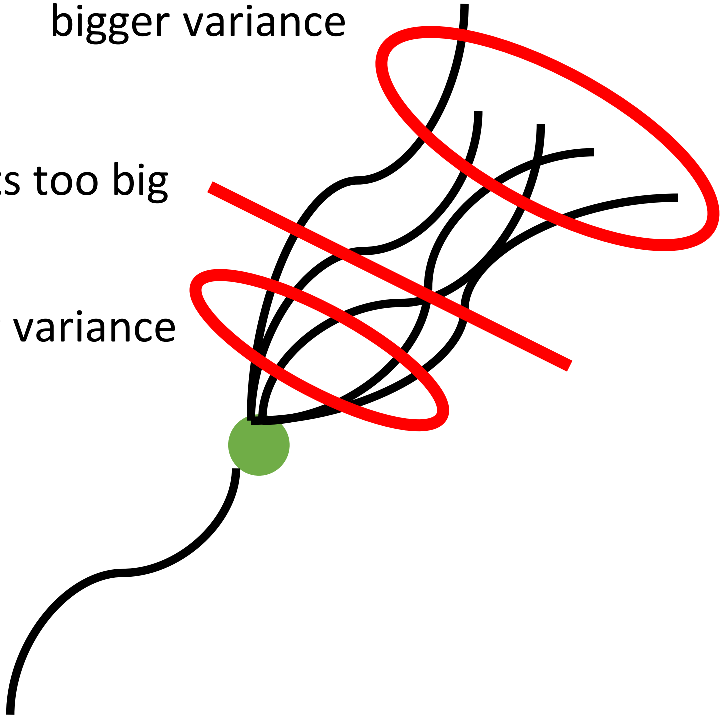
+ lower variance
- higher bias if value is wrong (it always is)

$$\hat{A}^{\pi}_{\mathrm{MC}}(\mathbf{s}_t, \mathbf{a}_t) = \sum_{t'=t}^{\infty} \gamma^{t'-t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) - \hat{V}^{\pi}_{\phi}(\mathbf{s}_t)$$

+ no bias
- higher variance (because single-sample estimate)

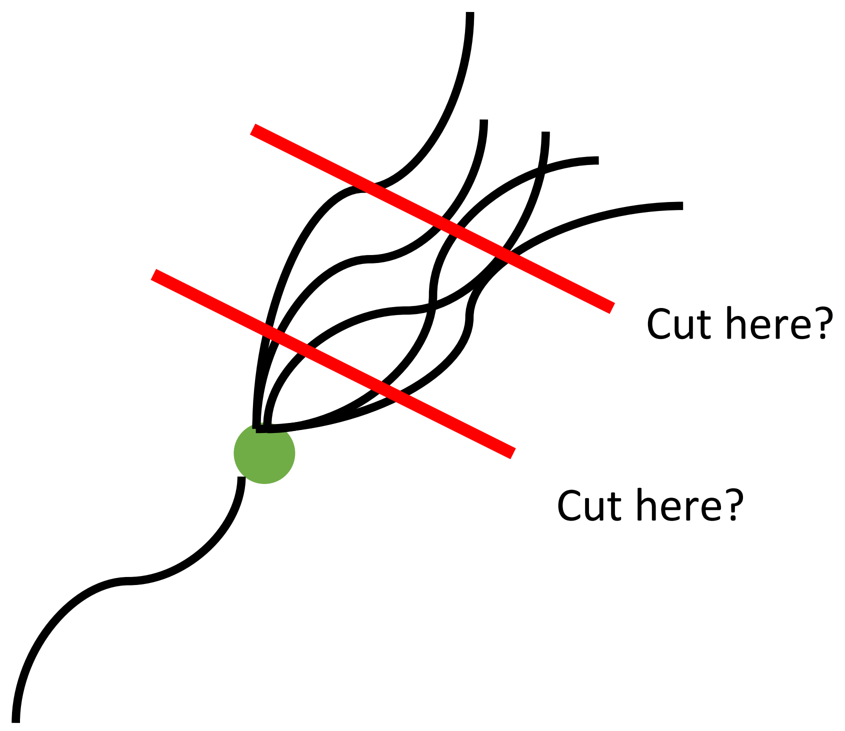
Can we combine these two, to control bias/variance tradeoff?





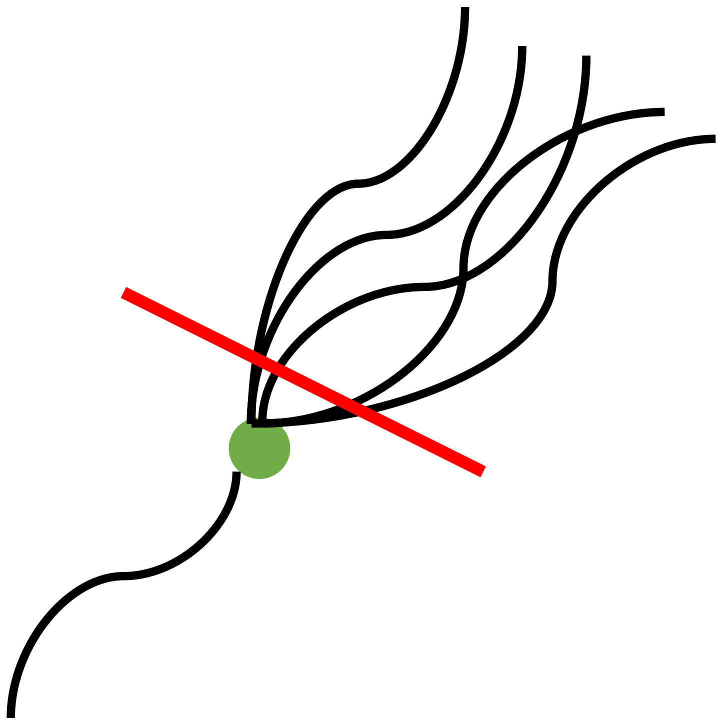
$$\hat{A}^{\pi}_{n}(\mathbf{s}_t, \mathbf{a}_t) = \sum_{t'=t}^{t+n} \gamma^{t'-t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) - \hat{V}^{\pi}_{\phi}(\mathbf{s}_t) + \gamma^{n} \hat{V}^{\pi}_{\phi}(\mathbf{s}_{t+n})$$

Do We Have to Choose Just One N?



Cut Everywhere All at Once

Cut everywhere all at once and use exponentially-weighted average to add up



The notations of this slide are from the original paper

The Derivative of GAE

$$\hat{A}_n^{\pi}(\mathbf{s}_t, \mathbf{a}_t) = \sum_{t'=t}^{t+n} \gamma^{t'-t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) - \hat{V}_\phi^\pi(\mathbf{s}_t) + \gamma^n \hat{V}_\phi^\pi(\mathbf{s}_{t+n})$$

Cut at t+1: $\hat{A}_t^{(1)} := \delta_t^V \qquad = -V(s_t) + r_t + \gamma V(s_{t+1})$

Cut at t+2: $\hat{A}_t^{(2)} := \delta_t^V + \gamma \delta_{t+1}^V \qquad = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2})$

Cut at t+3: $\hat{A}_t^{(3)} := \delta_t^V + \gamma \delta_{t+1}^V + \gamma^2 \delta_{t+2}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 V(s_{t+3})$

$$\hat{A}_t^{(k)} := \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}^V = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{k-1} r_{t+k-1} + \gamma^k V(s_{t+k})$$

$$\hat{A}_t^{(\infty)} = \sum_{l=0}^{\infty} \gamma^l \delta_{t+l}^V = -V(s_t) + \sum_{l=0}^{\infty} \gamma^l r_{t+l},$$

$\gamma^\infty V(s_{t+k})$ becomes zero

The Derivative of GAE (Con.)

exponential weighted average

$$A_t^{GAE} = \frac{A_t^{(1)} + \lambda A_t^{(2)} + \lambda^2 A_t^{(3)} + \lambda^3 A_t^{(4)} + \dots}{1+\lambda+\lambda^2+\lambda^3+\dots}$$

$$= \frac{A_t^{(1)} + \lambda A_t^{(2)} + \lambda^2 A_t^{(3)} + \lambda^3 A_t^{(4)} + \dots}{\frac{1\cdot(1-\lambda^k)\approx 1}{1-\lambda}}$$

$$= (1-\lambda)\left(A_t^{(1)} + \lambda A_t^{(2)} + \lambda^2 A_t^{(3)} + \lambda^3 A_t^{(4)} + \dots\right)$$

The notations of this slide are from the original paper

The Derivative of GAE (Con.)

The generalized advantage estimator GAE(γ, λ) is defined as the exponentially-weighted average of these k-step estimators:

$$
\begin{aligned}
\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} &:= (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \dots\right) \quad \text{Exponentially-weighted average} \\
&= (1-\lambda)(\delta_t^V + \lambda(\delta_t^V + \gamma\delta_{t+1}^V) + \lambda^2(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V) + \dots) \\
&= (1-\lambda)(\delta_t^V(1+\lambda+\lambda^2+\dots) + \gamma\delta_{t+1}^V(\lambda+\lambda^2+\lambda^3+\dots) \\
&\qquad + \gamma^2\delta_{t+2}^V(\lambda^2+\lambda^3+\lambda^4+\dots) + \dots) \\
&= (1-\lambda)\left(\delta_t^V\left(\frac{1}{1-\lambda}\right) + \gamma\delta_{t+1}^V\left(\frac{\lambda}{1-\lambda}\right) + \gamma^2\delta_{t+2}^V\left(\frac{\lambda^2}{1-\lambda}\right) + \dots\right) \\
&= \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V
\end{aligned}
\tag{16}
$$

Two Special Case

There are two notable special cases of this formula, obtained by setting $\lambda = 0$ and $\lambda = 1$.

$$\mathrm{GAE}(\gamma, 0): \quad \hat{A}_t := \delta_t \qquad = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{17}$$

$$\mathrm{GAE}(\gamma, 1): \quad \hat{A}_t := \sum_{l=0}^{\infty} \gamma^l \delta_{t+l} = \sum_{l=0}^{\infty} \gamma^l r_{t+l} - V(s_t) \tag{18}$$

Outline

- Recap
- GAE
- PPO

Proximal Policy Optimization Algorithms

Now We've Learned GAE

∇ log prob. of actions $\times$ GAE

Let's improve the left part

Efficiently Use Data

- We should drop all trajectory data after update the agent. Because the distribution of the agent's action shifts after update.
- Can't we use old data to update the agent more times?

TRPO/PPO is a method that we could leverage old data by simply multiplying a correction item when update the agent

Importance Sampling

- Importance sampling is a statistic technique to estimate one distribution by sampling from another distribution

$$
\begin{aligned}
E_{x\sim p}[f(x)] &= \int f(x)p(x)dx \\
&= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\
&= E_{x\sim q}[f(x)\frac{p(x)}{q(x)}] \\
&\approx \frac{1}{N}\sum_{i=1,x^i\in q}^{N} f(x^i)\frac{p(x^i)}{q(x^i)}
\end{aligned}
$$

Estimate p from q

Reference: cs885-spring18-lecture15b p.5

Surrogate Objective

$$E_{x\sim p}[f(x)] = E_{x\sim q}[f(x)\frac{p(x)}{q(x)}]$$

$$\nabla J(\theta) = E_{(s_t,\, a_t)\,\sim\,\pi_\theta}[\nabla \log \pi_\theta(a_t|s_t)\, A(s_t, a_t)]$$

$$= E_{(s_t,\, a_t)\sim\pi_{\theta_{old}}}[\frac{\pi_\theta(s_t, a_t)}{\pi_{\theta_{old}}(s_t, a_t)} \nabla \log \pi_\theta(a_t|s_t)\, A(s_t, a_t)]$$

$$J(\theta) = E_{(s_t,\, a_t)\sim\pi_{\theta_{old}}}[\frac{\pi_\theta(s_t, a_t)}{\pi_{\theta_{old}}(s_t, a_t)} A(s_t, a_t)]$$

Surrogate objective function

Reference: cs885-spring18-lecture15b p.7

TRPO objective

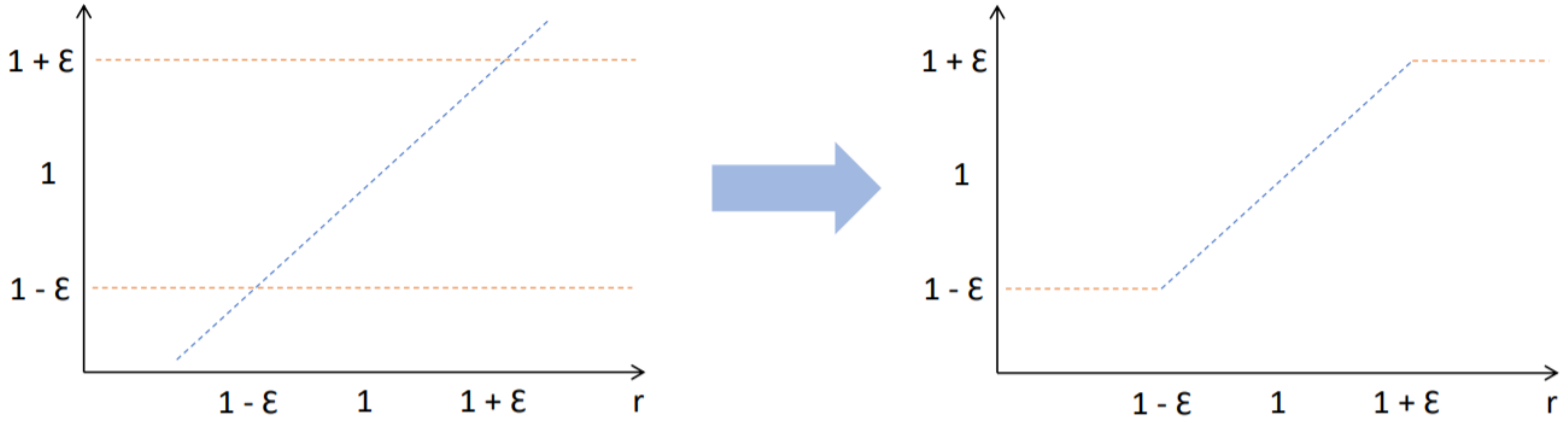
- TRPO use conjugate gradient algorithm
  - Slow because need to calculate Hessian matrix

$$\begin{aligned}
&\underset{\theta}{\text{maximize}} \quad \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] \\
&\text{subject to} \quad \hat{\mathbb{E}}_t[\text{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]] \leq \delta.
\end{aligned}$$

PPO with Clipped Objective

$$\underset{\theta}{\text{maximize}} \quad \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] \qquad r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

Fluctuation happens when r changes too quickly → limit r within a range?



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

Reference: cs885-spring18-lecture15b p.15

PPO in Practice

$$L_t^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t\left[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)\right]$$

Surrogate objective function

a squared-error loss for "critic"

$$(V_\theta(s_t) - V_t^{\text{targ}})^2$$

entropy bonus to ensure sufficient exploration

encourage "diversity"

* c1, c2: empirical values, in the paper, c1=1, c2=0.01

Reference: cs885-spring18-lecture15b p.17

Assignment

Running the code of PPO X GAE.

Writing a report about what you observe.

Deadline:

2022/12/22 11:59 p.m.