Nightmares of Policy Optimization

Sanjiban Choudhury
Switch from costs to rewards

All optimal control / planning literature written as costs

All RL literature written as rewards
We assumed black-box policies ...
Have we redacted too much?

SUBJECT: ______________

I. ______________ is being s_________ work and biological study at __________. The procedure and biological study at __________.

2. This ______________ continues to be __________ and ______________ and ______________.

3. The ______________ estimated to in the}
Black-box vs White-box vs Gray-box

**Black Box**

\[ s_0 \xrightarrow{} \pi_\theta \xrightarrow{} J(\theta) \]

**White Box**

\[ s_0 \xrightarrow{} \pi_\theta \xrightarrow{} a_0 \xrightarrow{} f(\cdot) \xrightarrow{} s_1 \xrightarrow{} \pi_\theta \xrightarrow{} a_1 \xrightarrow{} f(\cdot) \xrightarrow{} \ldots \xrightarrow{} J(\theta) \]
Black-box vs White-box vs Gray-box
How can we take gradients if we don’t know the dynamics?
The Likelihood Ratio Trick!
Algorithm 20: The REINFORCE algorithm.

Start with an arbitrary initial policy $\pi_\theta$

while not converged do

Run simulator with $\pi_\theta$ to collect $\{\xi^{(i)}\}_{i=1}^N$

Compute estimated gradient

$$\tilde{\nabla}_\theta J = \frac{1}{N} \sum_{i=1}^N \left[ \left( \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta \left( a^{(i)}_t \mid s^{(i)}_t \right) \right) R(\xi^{(i)}) \right]$$

Update parameters $\theta \leftarrow \theta + \alpha \tilde{\nabla}_\theta J$

return $\pi_\theta$
Causality: Can actions affect the past?
The Policy Gradient Theorem

\[ \nabla_\theta J = E_{p(\xi|\theta)} \left[ \sum_{t=0}^{T-1} \left( \nabla_\theta \log \pi_\theta(a_t|s_t) \left( \sum_{t'=0}^{t-1} r(s_{t'}, a_{t'}) + \sum_{t'=t}^{T-1} r(s_{t'}, a_{t'}) \right) \right] \]

\[ = E_{p(\xi|\theta)} \left[ \sum_{t=0}^{T-1} \left( \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{T-1} r(s_{t'}, a_{t'}) \right) \right], \]

\[ \nabla_\theta J = E_{p(\xi|\theta)} \left[ \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t, a_t) \right] \]

\[ \nabla_\theta J = E_{s \sim d^{\pi_\theta}(s), a \sim \pi_\theta(a|s)} \left[ \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s, a) \right] \]
Life is good!

This solves everything ...
The Three Nightmares of Policy Optimization
Nightmare 1:

Variance
When Q values for all rollouts in a batch are high?

\[ \nabla_{\theta} J = E_{s \sim d^{\pi_{\theta}}(s), a \sim \pi_{\theta}(a|s)} \left[ \nabla_{\theta} \log \pi_{\theta}(a|s) Q^{\pi_{\theta}}(s, a) \right] \]

Recall that one of the reasons for the high variance is that the algorithm does not know how well the trajectories perform compared to other trajectories. Therefore, by introducing a baseline for the total reward (or reward to go), we can update the policy based on how well the policy performs compared to a baseline.
Solution: Subtract a baseline!

\[ \nabla_\theta J = E_{d^{\pi_\theta}(s)} E_{\pi_\theta(a|s)} \left[ \nabla_\theta \log(\pi_\theta(a|s)) \left( Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) \right) \right]. \]

We can prove that this does not change the gradient

\[ \nabla_\theta J = E_{d^{\pi_\theta}(s)} E_{\pi_\theta(a|s)} \left[ \nabla_\theta \log(\pi_\theta(a|s)) A^{\pi_\theta}(s,a) \right] \]

But turns Q values into advantage (which is lower variance)

Justify the move to advantage using PDL!
Nightmare 2:
Distribution Shift
What happens if your step-size is large?

\[ \nabla_\theta J = E_{d^{\pi_\theta(s)}} E_{\pi_\theta(a|s)} \left[ \nabla_\theta \log(\pi_\theta(a|s)A^{\pi_\theta}(s,a)) \right] \]
The problem of distribution shift
The problem of distribution shift

![Graph showing the problem of distribution shift. The x-axis represents the state-action pair (s,a), the y-axis represents advantage. The graph shows two curves, one labeled $A^s$ and the other labeled $A^\pi$, with an overestimate indicated by arrows.](image-url)
The problem of distribution shift
The problem of distribution shift
How does distribution shift manifest?

The true performance difference

\[ J(\pi') - J(\pi) = \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d_t^\pi} A^\pi(s, \pi'(s)) \]

(New) (Old)

What our estimator currently approximates

\[ \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d_t^\pi} A^\pi(s, \pi'(s)) \]
Slowly change policies

Keep $d^t_\pi$ close to $d^t_{\pi'}$
Idea: Update distributions slowly

Does this simply mean do gradient descent with a small step size?
Does gradient descent keep distribution change small?

Gradient Descent is simply Steepest Descent with L2 norm

$$\max_{\Delta \theta} J(\theta + \Delta \theta) \quad s.t. \quad \|\Delta \theta\| \leq \epsilon$$

Does this ensure $d_{\pi_{\theta + \Delta \theta}}$ and $d_{\pi_{\theta}}$ are close??
What if we change norms?

Gradient Descent is simply Steepest Descent with L2 norm

$$\max_{\Delta \theta} J(\theta + \Delta \theta) \quad s.t. \quad \|\Delta \theta\| \leq \epsilon \quad \rightarrow \quad \Delta \theta = \nabla_{\theta} J(\theta)$$

What would update look like for another norm?

$$\max_{\Delta \theta} J(\theta + \Delta \theta) \quad s.t. \quad \Delta \theta^\top G(\theta) \Delta \theta \leq \epsilon \quad \rightarrow \quad \Delta \theta = \frac{1}{2\lambda} G^{-1}(\theta) \nabla_{\theta} J(\theta)$$
What’s a good norm for distributions?
What is a good norm for distributions?

\[
\max_{\Delta \theta} J(\theta + \Delta \theta)
\]

s.t. \( KL(P(\theta + \Delta \theta) \left| \left| P(\theta) \right) \leq \epsilon \)
What is a good norm for distributions?

$$\max_{\Delta \theta} J(\theta + \Delta \theta)$$

s.t. $$KL(P(\theta + \Delta \theta) || P(\theta)) \leq \epsilon$$

s.t. $$\Delta \theta^T G(\theta) \Delta \theta \leq \epsilon$$

Fischer Information Matrix

$$G(\theta) = \mathbb{E}_{p_\theta} \left[ \nabla_\theta \log(p_\theta) \nabla_\theta \log(p_\theta)^\top \right]$$
“Natural” Gradient Descent

Start with an arbitrary initial policy $\pi_\theta$

while not converged do

Run simulator with $\pi_\theta$ to collect $\{\xi^{(i)}\}_{i=1}^N$

Compute estimated gradient

$$\hat{\nabla}_\theta J = \frac{1}{N} \sum_{i=1}^N \left[ \left( \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta \left( a^{(i)}_t | s^{(i)}_t \right) \right) R(\xi^{(i)}) \right]$$

$$\hat{G}(\theta) = \frac{1}{N} \sum_{i=1}^N \left[ \nabla_\theta \log \pi_\theta(a_i|s_i) \nabla_\theta \log \pi_\theta(a_i|s_i) \right]^T$$

Update parameters $\theta \leftarrow \theta + \alpha \hat{G}^{-1}(\theta) \hat{\nabla}_\theta J$.

return $\pi_\theta$

Modern variants are TRPO, PPO, etc
Nightmare 3:
Local Optima
The Ring of Fire

+1

+100

-10
The Ring of Fire

Get’s sucked into a local optima!!
Idea: What if we had a “good reset distribution?”
Idea: What if we had a “good reset distribution?”
Idea: What if we had a “good reset distribution?”

Run REINFORCE from different start states
Idea: What if we had a “good reset distribution?”

Run REINFORCE from different start states
Idea: What if we had a “good reset distribution?”

Run REINFORCE from different start states
Solution: Use a good “reset” distribution

Choose a reset distribution $\mu(s)$ instead of start state distribution

Try your best to “cover” states the expert will visit

Justify using the PDL!
The Policy Gradient Theorem

\[ \nabla J = E_p (\nabla \log \pi(a|s)) \left( \sum_{t=0}^{T-1} r(s_t, a_t) + \sum_{t=0}^{T-1} r(s_{t'}, a_{t'}) \right) \]

\[ = E_p \left( \sum_{t=0}^{T-1} \nabla \log \pi(a|s) \sum_{t'=t}^{T-1} r(s_{t'}, a_{t'}) \right) \]

\[ \nabla J = E_p (\nabla \pi(a|s)) \left[ \sum_{t=0}^{T-1} \nabla \pi(a|s) Q^\pi(s_t, a_t) \right] \]

1. High Variance: Subtract baseline
2. Distribution Shift: *Natural* Gradient Descent
3. Local Optima: Use Reset Distribution