Lecture 8: Policy Gradient II. Advanced policy gradient section slides from Joshua Achiam (OpenAI)’s slides, with minor modifications

Emma Brunskill

CS234 Reinforcement Learning.

Winter 2023
Select all that are true about policy gradients:

1. \( \nabla_{\theta} V(\theta) = \mathbb{E}_{\pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(s, a) Q_{\pi_{\theta}}(s, a)] \) ✔
2. \( \theta \) is always increased in the direction of \( \nabla_{\theta} \ln(\pi(S_t, A_t, \theta)) \). ❌
3. State-action pairs with higher estimated \( Q \) values will increase in probability on average ✔
4. Are guaranteed to converge to the global optima of the policy class ❌
5. Not sure
Select all that are true about policy gradients:

1. \( \nabla_\theta V(\theta) = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) Q^{\pi_\theta}(s, a)] \)
2. \( \theta \) is always increased in the direction of \( \nabla_\theta \ln(\pi(S_t, A_t, \theta)) \).
3. State-action pairs with higher estimated \( Q \) values will increase in probability on average
4. Are guaranteed to converge to the global optima of the policy class
5. Not sure

1 and 3 are true. The direction of \( \theta \) also depends on the Q-values /returns. We are only guaranteed to reach a local optima
Vanilla Policy Gradient
"Vanilla" Policy Gradient Algorithm

Initialize policy parameter $\theta$, baseline $b$

\begin{algorithmic}
  \FOR {iteration=1, 2, \ldots}
    \STATE Collect a set of trajectories by executing the current policy $\pi^i$
    \STATE At each timestep $t$ in each trajectory $\tau^i$, compute
    \STATE \hspace{0.5cm} Return $G^i_t = \sum_{t'=t}^{T-1} r^i_{t'}$, and
    \STATE \hspace{0.5cm} Advantage estimate $\hat{A}^i_t = G^i_t - b(s_t)$.
    \STATE Re-fit the baseline, by minimizing $\sum_i \sum_t \| b(s_t) - G^i_t \|^2$,
    \STATE Update the policy, using a policy gradient estimate $\hat{g}$,
    \STATE Which is a sum of terms $\nabla_\theta \log \pi(a_t|s_t, \theta) \hat{A}_t$.
    \STATE (Plug $\hat{g}$ into SGD or ADAM)
  \ENDFOR
\end{algorithmic}
Likelihood Ratio / Score Function Policy Gradient

- Policy gradient:
  \[ \nabla_\theta \mathbb{E}[R] \approx \left( \frac{1}{m} \right) \sum_{i=1}^{m} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t, s_t)(G_t^{(i)} - b(s_t)) \]

- Fixes that improve simplest estimator
  - Temporal structure (shown in above equation)
  - Baseline (shown in above equation)
  - **Alternatives to using Monte Carlo returns** \( G_t^i \) **as estimate of expected discounted sum of returns for the policy parameterized by** \( \theta \)?
Choosing the Target

- $G_t^i$ is an estimation of the value function at $s_t$ from a single roll out
- Unbiased but high variance
- Reduce variance by introducing bias using bootstrapping and function approximation
  - Just like in we saw for TD vs MC, and value function approximation
Actor-critic Methods

- Estimate of $V/Q$ is done by a critic
- **Actor-critic** methods maintain an explicit representation of policy and the value function, and update both
- A3C (Mnih et al. ICML 2016) is a very popular actor-critic method
Recall:

$$\nabla_\theta \mathbb{E}_\tau [R] = \mathbb{E}_\tau \left[ \sum_{t=0}^{T-1} \nabla_\theta \log \pi(a_t|s_t; \theta) \left( \sum_{t'=t}^{T-1} r_{t'} - b(s_t) \right) \right]$$

$$\nabla_\theta \mathbb{E}_\tau [R] \approx \mathbb{E}_\tau \left[ \sum_{t=0}^{T-1} \nabla_\theta \log \pi(a_t|s_t; \theta) \left( Q(s_t, a_t; w) - b(s_t) \right) \right]$$

Letting the baseline be an estimate of the value $V$, we can represent the gradient in terms of the state-action advantage function

$$\nabla_\theta \mathbb{E}_\tau [R] \approx \mathbb{E}_\tau \left[ \sum_{t=0}^{T-1} \nabla_\theta \log \pi(a_t|s_t; \theta) \hat{A}_\pi(s_t, a_t) \right]$$

where the advantage function $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$

$$Q^\pi(s, \omega, w) - V^\pi(s, \omega)$$
Advanced Policy Gradients
Outline

Theory:
1. Problems with Policy Gradient Methods
2. Policy Performance Bounds
3. Monotonic Improvement Theory

Algorithms:
1. Proximal Policy Optimization
The Problems with Policy Gradients
Policy Gradients Review

Policy gradient algorithms try to solve the optimization problem

$$\max_{\theta} J(\pi_\theta) \doteq \mathbb{E}_{\tau \sim \pi_\theta} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

by taking stochastic gradient ascent on the policy parameters $\theta$, using the policy gradient

$$g = \nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)A^{\pi_\theta}(s_t, a_t) \right].$$

Limitations of policy gradients:

- Sample efficiency is poor
- Distance in parameter space $\neq$ distance in policy space!
  - What is policy space? For tabular case, set of matrices
    $$\Pi = \left\{ \pi : \pi \in \mathbb{R}^{|S| \times |A|}, \sum_a \pi_{sa} = 1, \pi_{sa} \geq 0 \right\}$$
  - Policy gradients take steps in parameter space
  - Step size is hard to get right as a result
Sample efficiency for vanilla policy gradient methods is poor.

Discard each batch of data immediately after **just one gradient step**.

Why? PG is an **on-policy expectation**.

Two main approaches to obtaining an unbiased estimate of the policy gradient:

- Run policy in environment and collect sample trajectories, then form sample estimate. (More stable)
- Use trajectories from other policies with **importance sampling**. (Less stable)
Importance sampling is a technique for estimating expectations using samples drawn from a different distribution.

$$E_{x \sim P} [f(x)] = \int_x p(x) f(x) \, dx$$

$$= \int_x p(x) q(x) \frac{f(x)}{q(x)} \, dx$$

$$= \int_x q(x) \left[ \frac{p(x)}{q(x)} f(x) \right] \, dx$$

$$= E_{x \sim q} \left[ \frac{p(x)}{q(x)} f(x) \right]$$
Importance sampling is a technique for estimating expectations using samples drawn from a different distribution.

\[
E_{x \sim P} [f(x)] = E_{x \sim Q} \left[ \frac{P(x)}{Q(x)} f(x) \right] \approx \frac{1}{|D|} \sum_{x \in D} \frac{P(x)}{Q(x)} f(x), \quad D \sim Q
\]

The ratio \(P(x)/Q(x)\) is the importance sampling weight for \(x\).

\[
\text{var} \left( f^2 \right) = \frac{1}{N} \text{var} \left( \frac{P(x)}{Q(x)} f(x) \right) = \frac{1}{N} \left( E_{x \sim Q} \left( \frac{P(x)}{Q(x)}^2 f(x)^2 \right) - E_{x \sim Q} \left( \frac{P(x)}{Q(x)} f(x) \right)^2 \right)
\]
Importance Sampling

Importance sampling is a technique for estimating expectations using samples drawn from a different distribution.

$$
E_{x \sim P} \left[ f(x) \right] = E_{x \sim Q} \left[ \frac{P(x)}{Q(x)} f(x) \right] \approx \frac{1}{|D|} \sum_{x \in D} \frac{P(x)}{Q(x)} f(x), \quad D \sim Q
$$

The ratio $P(x)/Q(x)$ is the importance sampling weight for $x$.

What is the variance of an importance sampling estimator?

$$
\text{var}(\hat{\mu}_Q) = \frac{1}{N} \text{var} \left( \frac{P(x)}{Q(x)} f(x) \right)
$$

$$
= \frac{1}{N} \left( E_{x \sim Q} \left[ \left( \frac{P(x)}{Q(x)} f(x) \right)^2 \right] - E_{x \sim Q} \left[ \frac{P(x)}{Q(x)} f(x) \right]^2 \right)
$$

$$
= \frac{1}{N} \left( E_{x \sim P} \left[ \frac{P(x)}{Q(x)} f(x)^2 \right] - E_{x \sim P} [f(x)]^2 \right)
$$

The term in red is problematic—if $P(x)/Q(x)$ is large in the wrong places, the variance of the estimator explodes.
Importance Sampling for Policy Gradients

Here, we compress the notation $\pi_\theta$ down to $\theta$ in some places for compactness.

$$g = \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \theta} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t | s_t) A^\theta(s_t, a_t) \right]$$

$$= \sum_{\tau} \sum_{t=0}^{\infty} \gamma^t P(\tau_t | \theta) \nabla_\theta \log \pi_\theta(a_t | s_t) A^\theta(s_t, a_t)$$

$$= \mathbb{E}_{\tau \sim \theta'} \left[ \sum_{t=0}^{\infty} \frac{P(\tau_t | \theta)}{P(\tau_t | \theta')} \gamma^t \nabla_\theta \log \pi_\theta(a_t | s_t) A^\theta(s_t, a_t) \right]$$

$$\frac{P(\tau_t | \theta)}{P(\tau_t | \theta')} = \frac{\mu(s_0)}{\mu(s_0)} \prod_{t=1}^{x} \pi_{\theta'}(a_t | s_t) P(s_{t+1} | s_t, a_t)$$

$$= \prod_{t=1}^{T} \frac{\pi_{\theta'}(a_t | s_t)}{\pi_\theta(a_t | s_t)}$$
Importance Sampling for Policy Gradients

Here, we compress the notation $\pi_\theta$ down to $\theta$ in some places for compactness.

\[
g = \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \theta} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)A^\theta(s_t, a_t) \right]
\]

\[
= \sum_{\tau} \sum_{t=0}^{\infty} \gamma^t P(\tau_t|\theta) \nabla_\theta \log \pi_\theta(a_t|s_t)A^\theta(s_t, a_t)
\]

\[
= \mathbb{E}_{\tau \sim \theta'} \left[ \sum_{t=0}^{\infty} \frac{P(\tau_t|\theta)}{P(\tau_t|\theta')} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)A^\theta(s_t, a_t) \right]
\]
Importance Sampling for Policy Gradients

Here, we compress the notation $\pi_\theta$ down to $\theta$ in some places for compactness.

$$g = \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \theta} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)A^\theta(s_t, a_t) \right]$$

$$= \sum_{\tau} \sum_{t=0}^{\infty} \gamma^t P(\tau|\theta) \nabla_\theta \log \pi_\theta(a_t|s_t)A^\theta(s_t, a_t)$$

$$= \mathbb{E}_{\tau \sim \theta'} \left[ \sum_{t=0}^{\infty} \frac{P(\tau|\theta)}{P(\tau|\theta')} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)A^\theta(s_t, a_t) \right]$$

Challenge? **Exploding or vanishing importance sampling weights.**

$$P(\tau_t|\theta) = \frac{\mu(s_0) \prod_{t'=0}^{t} P(s_{t'+1}|s_{t'}, a_{t'}) \pi_\theta(a_{t'}|s_{t'})}{\mu(s_0) \prod_{t'=0}^{t} P(s_{t'+1}|s_{t'}, a_{t'}) \pi_{\theta'}(a_{t'}|s_{t'})} = \prod_{t'=0}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\pi_{\theta'}(a_{t'}|s_{t'})}$$

Even for policies only slightly different from each other, **many small differences multiply to become a big difference.**

Big question: how can we make efficient use of the data we already have from the old policy, while avoiding the challenges posed by importance sampling?
Policy gradient algorithms are stochastic gradient ascent:

$$\theta_{k+1} = \theta_k + \alpha_k \hat{g}_k$$

with step $\Delta_k = \alpha_k \hat{g}_k$.

- If the step is too large, **performance collapse** is possible (Why?)
Choosing a Step Size for Policy Gradients

Policy gradient algorithms are stochastic gradient ascent:

\[ \theta_{k+1} = \theta_k + \alpha_k \hat{g}_k \]

with step \( \Delta_k = \alpha_k \hat{g}_k \).

- If the step is too large, **performance collapse** is possible (Why?)
- If the step is too small, progress is unacceptably slow
- “Right” step size changes based on \( \theta \)

Automatic learning rate adjustment like advantage normalization, or Adam-style optimizers, can help. But does this solve the problem?

**Figure:** Policy parameters on x-axis and performance on y-axis. A bad step can lead to performance collapse, which may be hard to recover from.
The Problem is More Than Step Size

Consider a family of policies with parametrization:

\[ \pi_\theta(a) = \begin{cases} 
\sigma(\theta) & a = 1 \\
1 - \sigma(\theta) & a = 2 
\end{cases} \]

Figure: Small changes in the policy parameters can unexpectedly lead to **big** changes in the policy.

Big question: how do we come up with an update rule that doesn’t ever change the policy more than we meant to?
Policy Performance Bounds
Relative Performance of Two Policies

In a policy optimization algorithm, we want an update step that
- uses rollouts collected from the most recent policy as efficiently as possible,
- and takes steps that respect **distance in policy space** as opposed to distance in parameter space.

To figure out the right update rule, we need to exploit relationships between the performance of two policies.

**Performance difference lemma**: In CS234 HW2 you proved that for any policies $\pi, \pi'$

$$J(\pi') - J(\pi) = \mathbb{E}_{\tau \sim \pi'} \left[ \sum_{t=0}^{\infty} \gamma^t A(\pi)(s_t, a_t) \right]$$

(1)

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d, a \sim \pi'} [A(\pi)(s, a)]$$

(2)

where

$$d(\pi)(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s\mid \pi)$$
Can we use this for policy improvement, where $\pi'$ represents the new policy and $\pi$ represents the old one?

$$\max_{\pi'} J(\pi') = \max_{\pi'} J(\pi') - J(\pi)$$

$$= \max_{\pi'} \mathbb{E}_{\tau \sim \pi'} \left[ \sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t) \right]$$

This is suggestive, but not useful yet.

Nice feature of this optimization problem: defines the performance of $\pi'$ in terms of the advantages from $\pi$!

But, problematic feature: still requires trajectories sampled from $\pi'$...
Looking at it from another angle...

In terms of the discounted future state distribution $d^\pi$, defined by

$$d^\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s | \pi),$$

we can rewrite the relative policy performance identity:

$$J(\pi') - J(\pi) = E_{\tau \sim \pi'} \left[ \sum_{t=0}^{\infty} \gamma^t A^\pi(s_t, a_t) \right]$$

$$= \frac{1}{1 - \gamma} E_{s \sim d^\pi', a \sim \pi'} [A^\pi(s, a)]$$

$$= \frac{1}{1 - \gamma} E_{s \sim d^\pi', a \sim \pi} \left[ \frac{\pi'(a|s)}{\pi(a|s)} A^\pi(s, a) \right]$$

...almost there! Only problem is $s \sim d^\pi'$. [s \sim d^\pi', unknown]
A Useful Approximation

What if we just said \( d_{\pi'} \approx d_{\pi} \) and didn’t worry about it?

\[
J(\pi') - J(\pi) \approx \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\pi}, a \sim \pi} \left[ \frac{\pi'(a|s)}{\pi(a|s)} A^{\pi}(s, a) \right]
\]

\[
d_{\pi'} \Rightarrow L_{\pi}(\pi')
\]

Turns out: this approximation is pretty good when \( \pi' \) and \( \pi \) are close! But why, and how close do they have to be?

**Relative policy performance bounds:**

\[
\left| J(\pi') - (J(\pi) + L_{\pi}(\pi')) \right| \leq C \sqrt{\mathbb{E}_{s \sim d_{\pi}} [D_{KL}(\pi' || \pi)[s]]}
\]  \hspace{1cm} (3)

If policies are close in KL-divergence—the approximation is good!

---

\(^1\) Achiam, Held, Tamar, Abbeel, 2017
What is KL-divergence?

For probability distributions $P$ and $Q$ over a discrete random variable, 

$$D_{KL}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Properties:
- $D_{KL}(P \| P) = 0$
- $D_{KL}(P \| Q) \geq 0$
- $D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$ — Non-symmetric!

What is KL-divergence between policies?

$$D_{KL}(\pi' \| \pi)[s] = \sum_{a \in A} \pi'(a|s) \log \frac{\pi'(a|s)}{\pi(a|s)}$$
What did we gain from making that approximation?

\[ J(\pi') - J(\pi) \approx \mathcal{L}_\pi(\pi') \]

\[ \mathcal{L}_\pi(\pi') = \frac{1}{1 - \gamma} \mathbb{E}_{a \sim \pi} \left[ \frac{\pi'(a|s)}{\pi(a|s)} A^\pi(s, a) \right] \]

\[ = \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t \frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)} A^\pi(s_t, a_t) \right] \]

- This is something we can optimize using trajectories sampled from the old policy \( \pi \)!
- Similar to using importance sampling, but because weights only depend on current timestep (and not preceding history), they don’t vanish or explode.
Recommended Reading

- “Approximately Optimal Approximate Reinforcement Learning,” Kakade and Langford, 2002
- “Trust Region Policy Optimization,” Schulman et al. 2015
- “Constrained Policy Optimization,” Achiam et al. 2017

---

Monotonic Improvement Theory
From the bound on the previous slide, we get

\[ J(\pi') - J(\pi) \geq L_\pi(\pi') - C \sqrt{\mathbb{E}_{s \sim d^\pi} [D_{KL}(\pi' || \pi)(s)]}. \]

- If we maximize the RHS with respect to \( \pi' \), we are guaranteed to improve over \( \pi \).
- This is a majorize-maximize algorithm w.r.t. the true objective, the LHS.
- And \( L_\pi(\pi') \) and the KL-divergence term can both be estimated with samples from \( \pi \)!
Proof of improvement guarantee: Suppose \( \pi_{k+1} \) and \( \pi_k \) are related by

\[
\pi_{k+1} = \arg \max_{\pi'} L_{\pi_k}(\pi') - C \sqrt{\frac{E}{s \sim d_{\pi_k}}} \left[ D_{KL}(\pi' || \pi_k)[s] \right].
\]

\( \Pi_k \) is feasible

\[ L_{\pi_k}(\pi_k) = \mathbb{E}_{s,a \sim d_{\pi_k}} Q_{\pi_k}(s,a) \]

\[ D_{KL}(\pi' || \pi_k) = 0 \]

\[ 1 = 0 \]

\[ 2 = 0 \]

\[ 1 - 2 \geq 0 \]

\[ 1 - 2 > 0 \]

\[ J(\pi_{k+1}) - J(\pi_k) \geq L_{\pi_k}(\pi_{k+1}) - C \sqrt{E_{\cdots \cdots}} > 0 \]
Monotonic Improvement Theory

Proof of improvement guarantee: Suppose $\pi_{k+1}$ and $\pi_k$ are related by

$$
\pi_{k+1} = \arg \max_{\pi'} \mathbb{L}_{\pi_k}(\pi') - C \sqrt{\mathbb{E}_{s \sim d_{\pi_k}} \left[ D_{KL}(\pi' || \pi_k)[s] \right]}.
$$

- $\pi_k$ is a feasible point, and the objective at $\pi_k$ is equal to 0.
  - $\mathbb{L}_{\pi_k}(\pi_k) \propto \mathbb{E}_{s, a \sim d_{\pi_k, \pi_k}} [A_{\pi_k}(s, a)] = 0$
  - $D_{KL}(\pi_k || \pi_k)[s] = 0$
- $\implies$ optimal value $\geq 0$
- $\implies$ by the performance bound, $J(\pi_{k+1}) - J(\pi_k) \geq 0$

This proof works even if we restrict the domain of optimization to an arbitrary class of parametrized policies $\Pi_\theta$, as long as $\pi_k \in \Pi_\theta$. 
Approximate Monotonic Improvement

\[ \pi_{k+1} = \arg \max_{\pi'} \mathcal{L}_{\pi_k}(\pi') - C \sqrt{\mathbb{E}_{s \sim d_{\pi_k}} [D_{KL}(\pi' || \pi_k)[s]]}. \]  

(4)

Problem:
- \( C \) provided by theory is quite high when \( \gamma \) is near 1
- \( \implies \) steps from (4) are too small.

Potential Solution:
- Tune the KL penalty
- Use KL constraint (called trust region).
Proximal Policy Optimization (PPO) is a family of methods that approximately penalize policies from changing too much between steps. Two variants:

- **Adaptive KL Penalty**
  - Policy update solves unconstrained optimization problem
    \[
    \theta_{k+1} = \arg\max_{\theta} L_{\theta_k}(\theta) - \beta_k \bar{D}_{KL}(\theta||\theta_k)
    \]
  - Penalty coefficient \(\beta_k\) changes between iterations to approximately enforce KL-divergence constraint
**Algorithm 1** PPO with Adaptive KL Penalty

Input: initial policy parameters $\theta_0$, initial KL penalty $\beta_0$, target KL-divergence $\delta$

for $k = 0, 1, 2, \ldots$ do

Collect set of partial trajectories $D_k$ on policy $\pi_k = \pi(\theta_k)$

Estimate advantages $\hat{A}_t^{\pi_k}$ using any advantage estimation algorithm

Compute policy update

$$\theta_{k+1} = \arg \max_\theta L_{\theta_k}(\theta) - \beta_k \bar{D}_{KL}(\theta||\theta_k)$$

by taking $K$ steps of minibatch SGD (via Adam)

if $\bar{D}_{KL}(\theta_{k+1}||\theta_k) \geq 1.5\delta$ then

$$\beta_{k+1} = 2\beta_k$$

else if $\bar{D}_{KL}(\theta_{k+1}||\theta_k) \leq \delta/1.5$ then

$$\beta_{k+1} = \beta_k/2$$

end if

end for

- Initial KL penalty not that important—it adapts quickly
- Some iterations may violate KL constraint, but most don’t
Proximal Policy Optimization (PPO) is a family of methods that approximately enforce KL constraint \textbf{without computing natural gradients}. Two variants:

- **Adaptive KL Penalty**
  - Policy update solves unconstrained optimization problem
    \[
    \theta_{k+1} = \arg \max_{\theta} \mathcal{L}_{\theta_k}(\theta) - \beta_k \bar{D}_{KL}(\theta||\theta_k)
    \]
  - Penalty coefficient $\beta_k$ changes between iterations to approximately enforce KL-divergence constraint

- **Clipped Objective**
  - New objective function: let $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_k}(a_t|s_t)$. Then
    \[
    \mathcal{L}^{CLIP}_{\theta_k}(\theta) = \mathbb{E}_{\tau \sim \pi_k} \left[ \sum_{t=0}^{T} \min\left( r_t(\theta)\hat{A}^{\pi_k}_t, \text{clip} \left( r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}^{\pi_k}_t \right) \right]
    \]
    where $\epsilon$ is a hyperparameter (maybe $\epsilon = 0.2$)
  - Policy update is $\theta_{k+1} = \arg \max_{\theta} \mathcal{L}^{CLIP}_{\theta_k}(\theta)$
Algorithm 2 PPO with Clipped Objective

Input: initial policy parameters $\theta_0$, clipping threshold $\epsilon$

for $k = 0, 1, 2, \ldots$ do
  Collect set of partial trajectories $D_k$ on policy $\pi_k = \pi(\theta_k)$
  Estimate advantages $\hat{A}_{t}^{\pi_k}$ using any advantage estimation algorithm
  Compute policy update
  $$\theta_{k+1} = \arg \max_{\theta} L_{\theta_k}^{CLIP}(\theta)$$
  by taking $K$ steps of minibatch SGD (via Adam), where
  $$L_{\theta_k}^{CLIP}(\theta) = \mathbb{E}_{\tau \sim \pi_k} \left[ \sum_{t=0}^{T} \min (r_t(\theta)\hat{A}_{t}^{\pi_k}, \text{clip} (r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{t}^{\pi_k}) \right]$$

end for

- Clipping prevents policy from having incentive to go far away from $\theta_{k+1}$
- Clipping seems to work at least as well as PPO with KL penalty, but is simpler to implement
But how does clipping keep policy close? By making objective as pessimistic as possible about performance far away from $\theta_k$:

**Figure:** Various objectives as a function of interpolation factor $\alpha$ between $\theta_{k+1}$ and $\theta_k$ after one update of PPO-Clip.

---

Schulman, Wolski, Dhariwal, Radford, Klimov, 2017
Empirical Performance of PPO

Figure: Performance comparison between PPO with clipped objective and various other deep RL methods on a slate of MuJoCo tasks.  

- Wildly popular, and key component of ChatGPT

---

6Schulman, Wolski, Dhariwal, Radford, Klimov, 2017
Recommended Reading

PPO

- “Proximal Policy Optimization Algorithms,” Schulman et al. 2017
  [7]  
- OpenAI blog post on PPO, 2017  
  [8]

---

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation Matters in Deep RL: A Case Study on PPO and TRPO. ICLR 2020
https://openreview.net/forum?id=r1etN1rtPB

Reward scaling, learning rate annealing, etc. can make a significant difference
Theory:
1. Problems with Policy Gradient Methods
2. Policy Performance Bounds
3. Monotonic Improvement Theory

Proximal Policy Optimization:
1. Approximately constraints policy steps
2. Relatively simple to implement
3. Good empirical success and very widely used