Proximal Policy Optimization (PPO)

2023-04-04

Actor Critic módszer

Actor Critic (A2C) egy hibrid architektúrával, amely értékalapú és policy-alapú módszereket kombinál, és amely a variancia csökkentésével segít stabilizálni a tanulást:

- Egy Actor szabályozza, hogyan viselkedjen az ágensünk (policy-alapú módszer).
- Egy Critic, amely azt méri, hogy a végrehajtott cselekvés mennyire jó (értékalapú módszer).

Proximal Policy Optimization (PPO)

A Proximal Policy Optimization (PPO) architektúra a túl nagy Policy-frissítések elkerülése révén javítja az agens képzési stabilitását.

Ehhez egy arányt használunk, amely a jelenlegi és a régi policy közötti különbséget (arányt) jelzi.

Ezt az arányt egy adott tartományban [1-ϵ,1+ϵ] vágjuk el.

Ez biztosítja, hogy a policy frissítése ne legyen túl nagy, és hogy a képzés stabilabb legyen.

A PPO mögött meghúzódó intuíció

A Proximal Policy Optimization (PPO) lényege az, hogy javítani akarjuk a policy képzési stabilitását azáltal, hogy korlátozzuk a policy-ben minden egyes képzési lépésben végrehajtott változtatásokat.

El akarjuk kerülni a túl nagy policy frissítéseket.

A PPO mögött meghúzódó intuíció

Ennek két oka van:

- Empirikusan tudjuk, hogy a kisebb policy-frissítések a képzés során nagyobb valószínűséggel konvergálnak az optimális megoldáshoz.

- Egy túl nagy lépés a policy-frissítésben azt eredményezheti, hogy "leesünk a szikláról", és hosszú időbe telik (vagy nincs lehetőségünk) ezt helyreállítani.

A PPO mögött meghúzódó intuíció

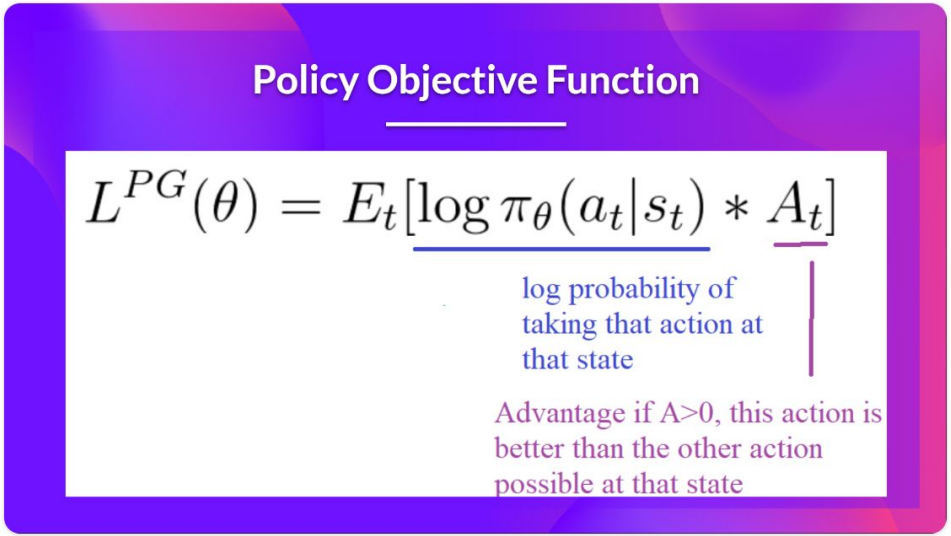
A PPO esetében tehát “konzervatív” módon frissítjük a policy-t.

Ehhez meg kell mérnünk, hogy a jelenlegi és a korábbi policy között (arány számítással) mennyi változás történt.

Ezt az arányt pedig egy [1-ϵ,1+ϵ] tartományban vágjuk majd el.

A vágást helyettesítő függvény

Emlékezzünk vissza, a cél, amit a Reinforce-ban optimalizálni kell:

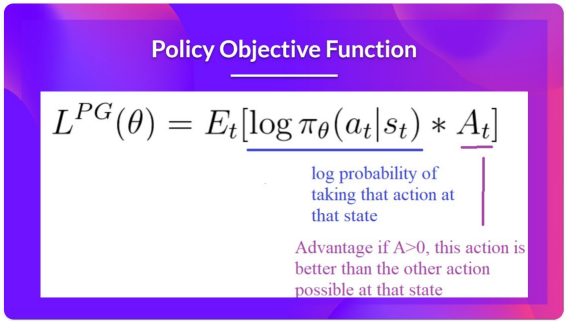


A vágást helyettesítő függvény

A függvény a gradiens emelkedéssel az ágensünket olyan cselekvésekre ösztönözi, amelyek magasabb jutalomhoz vezetnek, és elkerüljük a káros cselekvéseket.

Ennek azonban van 2 hibája:

- Túl kicsi - a képzési folyamat túl lassú volt
- Túl nagy - túl nagy volt a variabilitás a képzésben.

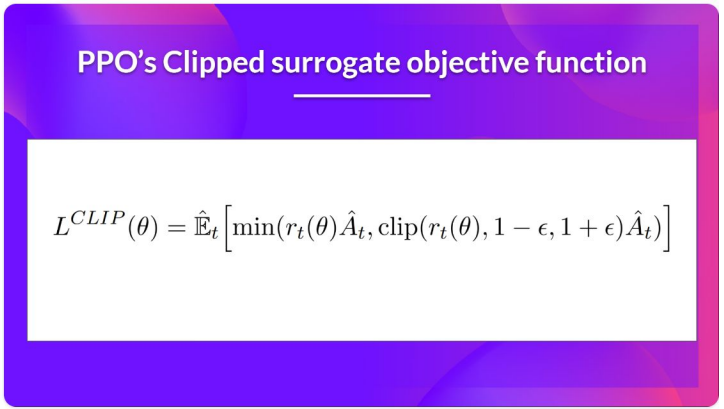


A vágást helyettesítő függvény

A PPO esetében az ötlet az, hogy a policy frissítését egy célfüggvénnyel korlátozzuk, amelyet **Clipped surrogate objective function**-nek nevezünk.

A függvény a policy változást egy kis tartományban korlátozza egy **vágás** segítségével.

Elkerüli a destruktív nagy súlyok frissítését.



A Ratio funkció

Az aktuális policy s_t állapotában történő cselekvés valószínűsége osztva az előzővel.

The ratio function

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

The ratio function

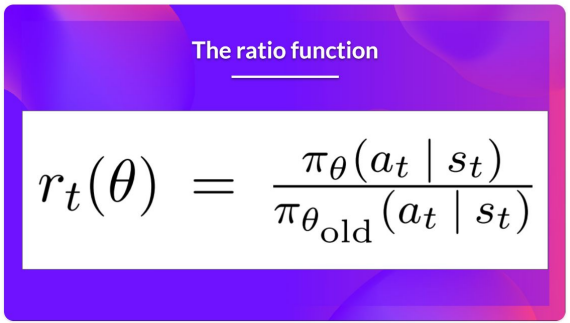
$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

A Ratio függvény

- Ha $r_t(\theta)>1$, akkor az s_t állapotbeli cselekvés valószínűbb az aktuális policyben, mint a régi policyben.

- Ha $r_t(\theta)$ 0 és 1 között van, akkor a cselekvés kevésbé valószínű a jelenlegi policyben, mint a régiben.

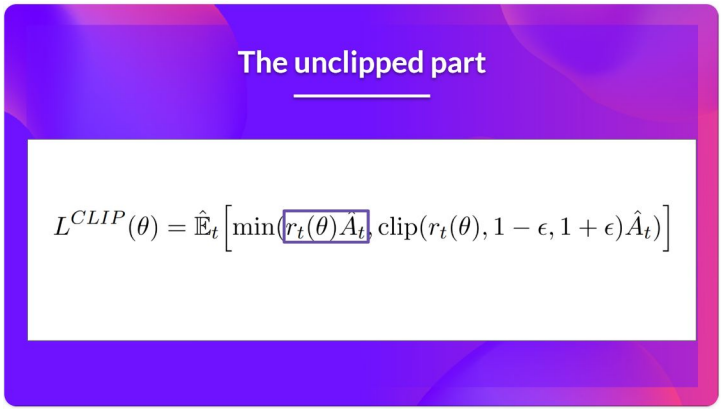
Ez a valószínűségi arány tehát egy egyszerű módja a régi és a jelenlegi policy közötti eltérés becslésének.

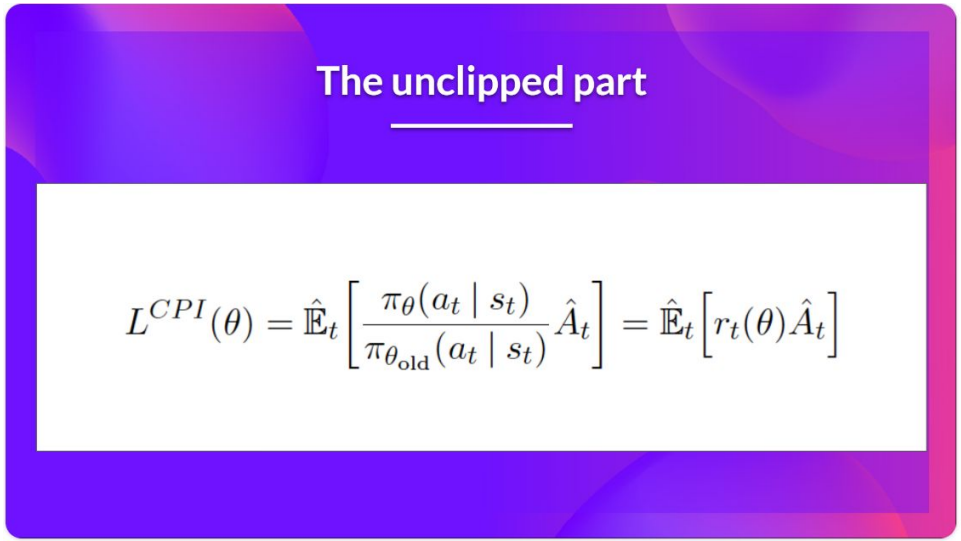


A vágást helyettesítő függvény

Ez az arány helyettesítheti a policy cél függvényében használt logaritmikus valószínűséget.

Így megkapjuk az új célfüggvény bal oldali részét: az arányt megszorozzuk az előnnyel.

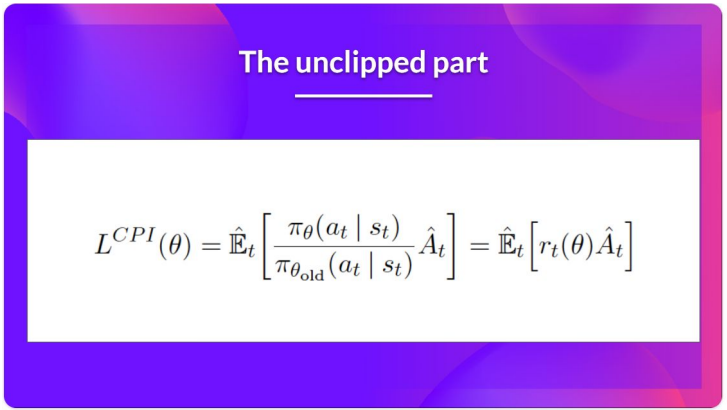




A vágást helyettesítő függvény

Mi történik, ha az aktuális policy-ban a új intézkedés sokkal valószínűbb, mint a korábbi policy-ban?

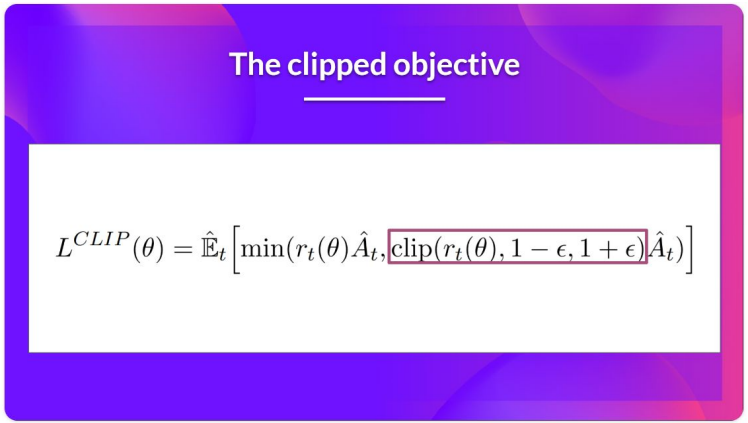
Akkor ez egy jelentős policy-gradiens lépéshez, és így ez egy túlzott policy-frissítéshez vezetne.



A vágást helyettesítő függvény

A célfüggvényt úgy kell korlátozni, hogy büntetjük azokat a változásokat, amelyek az arányt az 1-től távolabbra viszik.

Az arány megkurtításával biztosítjuk, hogy ne legyen túl nagy a policy frissítése, mivel az aktuális policy nem különbözhet túlságosan a régitől.



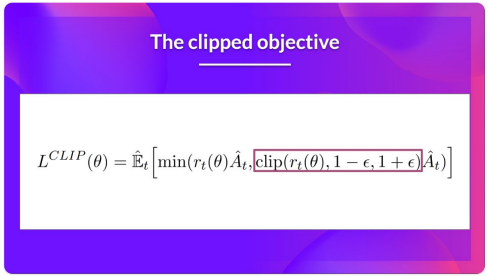
A vágást helyettesítő függvény

Ez a levágott rész egy olyan $r_t(\theta)$ amely [1-ϵ,1+ϵ] közé van vágva.

A Clipped Surrogate objektív függvénnyel két valószínűségi arányt kapunk,

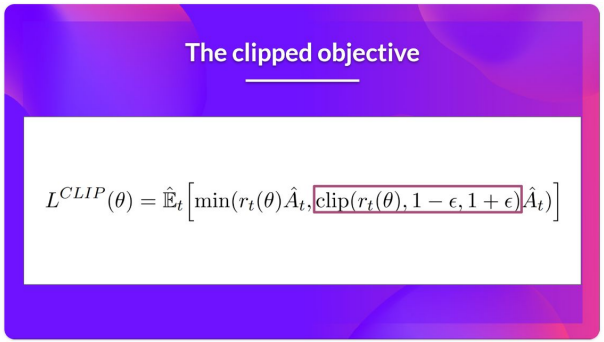
- egy le nem vágott
- egy levágott tartományban

[1-ϵ,1+ϵ] - az epsilon egy hiperparaméter, amely segít meghatározni ezt a clipped tartományt.



A vágást helyettesítő függvény

Ezután a levágott és a le nem vágott célkitűzés minimumát vesszük, így a végső célkitűzés a le nem vágott célkitűzés alsó korlátja (pesszimista korlátja).



A vágást helyettesítő függvény

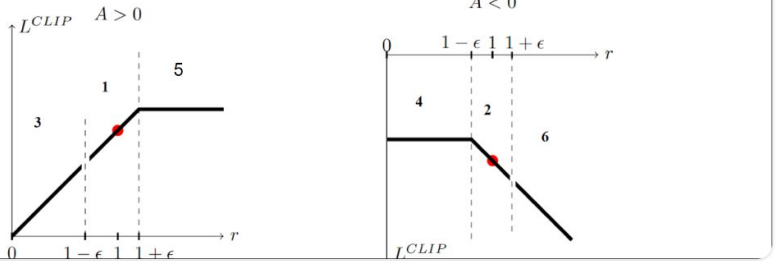
Ezután a levágott és a nem levágott célkitűzés minimumát vesszük, így a végső célkitűzés a nem levágott célkitűzés alsó korlátja (pesszimista korlátja).

Hogyan néz ki ez a Clipped Surrogate Objective Function

Pl.: Hat különböző helyzetünk van.

Először is ne feledjük, hogy a ”levágott” és a ”le nem levágott” célkitűzések közötti minimumot vesszük.

	$p_t(\theta)\ > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓

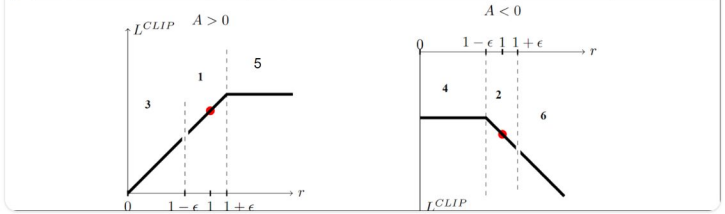


A levágott helyettesítő objektív függvény vizualizálása

Az 1. és a 2. helyzetben a vágás nem alkalmazható, mivel az arány a [1-ϵ,1+ϵ] tartomány között van.

1. pozitív előnyünk van: az akció jobb, mint az összes akció átlaga az adott állapotban. Ezért ösztönöznünk kell a jelenlegi policyt, hogy növeljük az adott akció végrehajtásának valószínűségét abban az állapotban.

	$p_t(\theta) > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓

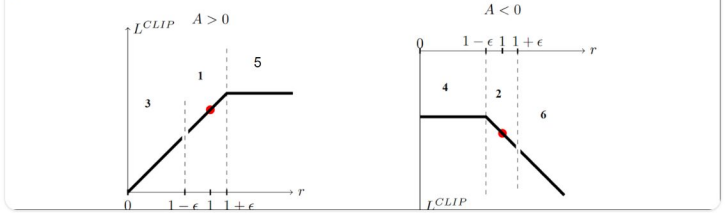


A levágott helyettesítő objektív függvény vizualizálása

A 2. negatív előnyünk van: az akció rosszabb, mint az összes akció átlaga abban az állapotban. Ezért a jelenlegi policyt el kell tántorítani attól, hogy abban az állapotban ezt az akciót megtegyük.

Mivel az arány intervallumok között van, csökkenthetjük annak a valószínűségét, hogy a politikánk abban az állapotban végrehajtja az adott akciót.

	$p_t(\theta) > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓

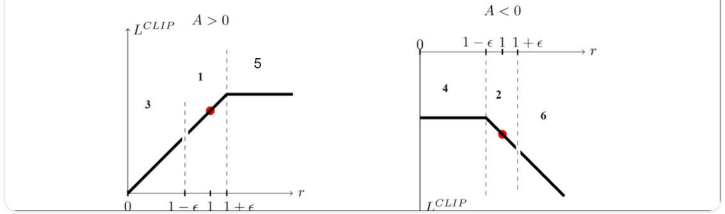


A levágott helyettesítő objektív függvény vizualizálása

3. és 4. eset: az arány a tartomány alatt van.

Ha a valószínűségi arány kisebb, mint [1-ϵ], akkor a cselekvés végrehajtásának valószínűsége az adott állapotban sokkal kisebb, mint a régi policy esetén.

	$p_t(\theta)\ > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓

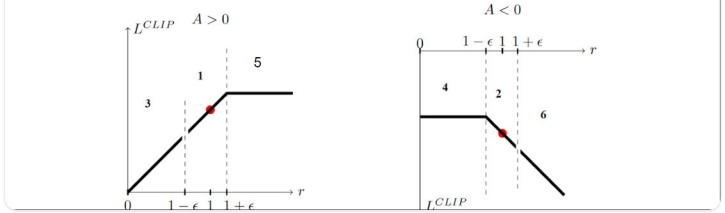


A levágott helyettesítő objektív függvény vizualizálása

A 3. szituációban, az előnybecslés pozitív (A>0), akkor növelni akarjuk az adott cselekvés végrehajtásának valószínűségét abban az állapotban.

A 4. szituációban, az előnybecslés negatív, akkor nem akarjuk tovább csökkenteni az adott cselekvés végrehajtásának valószínűségét abban az állapotban. Ezért a gradiens = 0 (mivel egy lapos vonalon vagyunk), így nem frissítjük a súlyainkat.

	$p_t(\theta) > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓

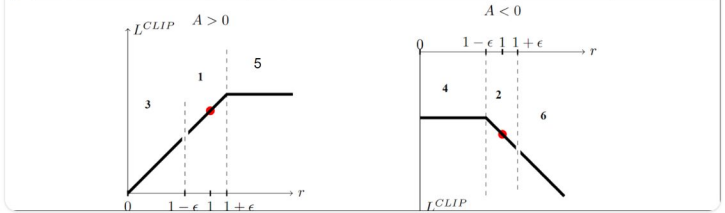


A levágott helyettesítő objektív függvény vizualizálása

5. és 6. eset: az arány meghaladja a tartományt.

Ha a valószínűségi arány nagyobb, mint [1+ϵ], akkor a jelenlegi policyben sokkal nagyobb a valószínűsége annak, hogy az adott akciót az adott állapotban végrehajtjuk, mint a korábbi policyban.

	$p_t(\theta)\ > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓

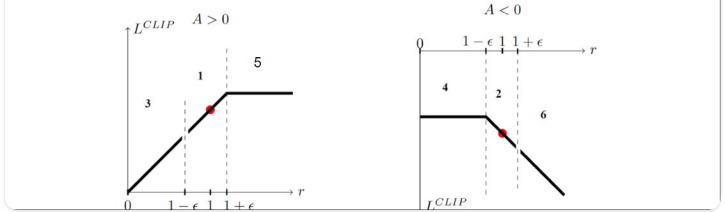


A levágott helyettesítő objektív függvény vizualizálása

Az 5. szituációban, az előny pozitív, akkor nem akarunk túl mohók lenni.

Már nagyobb a valószínűsége annak, hogy abban az állapotban megtesszük ezt a cselekvést, mint a korábbi policyben. Ezért a gradiens = 0 (mivel egy lapos vonalon vagyunk), így nem frissítjük a súlyainkat.

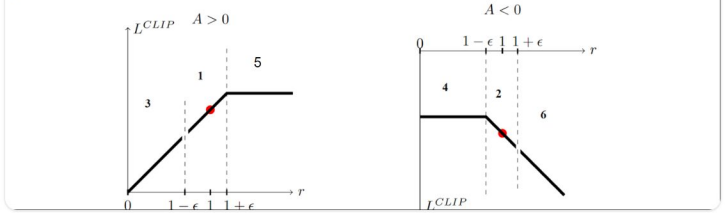
	$p_t(\theta) > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓



A levágott helyettesítő objektív függvény vizualizálása

A 6. szituációban, az előny negatív, akkor csökkenteni akarjuk az adott cselekvés végrehajtásának valószínűségét abban az állapotban.

	$p_t(\theta) > 0$	A_t	Return Value of *min*	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	−	$p_t(\theta)A_t$	no	−	✓
3	$p_t(\theta) < 1-\epsilon$	+	$p_t(\theta)A_t$	no	+	✓
4	$p_t(\theta) < 1-\epsilon$	−	$(1-\epsilon)A_t$	yes	−	**0**
5	$p_t(\theta) > 1+\epsilon$	+	$(1+\epsilon)A_t$	yes	+	**0**
6	$p_t(\theta) > 1+\epsilon$	−	$p_t(\theta)A_t$	no	−	✓



A levágott helyettesítő objektív függvény vizualizálása

A policy-t frissítjük a UNCLIPPED résszel.

Ha a minimum a levágott objektív rész, akkor nem frissítjük a policy-nk súlyait, mivel a gradiens egyenlő lesz 0-val.

Csak akkor frissítjük a policy-t, ha:

- Az arányunk a [1-ϵ,1+ϵ] tartományban van.

- Az arányunk a tartományon kívül van, de az előny a tartományhoz való közeledéshez vezet.

- Az arány alatt van, de az előny > 0

- Az arány felett vagyunk, de az előny < 0

A levágott helyettesítő objektív függvény vizualizálása

Ha a minimum a levágott arány, akkor a gradiens miért 0?

Ha az arány le van vágva, akkor a derivált ebben az esetben nem az $r_t(\theta)*A_t$ deriváltja lesz, hanem vagy az $(1-\epsilon)*A_t$ deriváltja, vagy az $(1+\epsilon)*A_t$ deriváltja, amelyek értéke= 0.

A levágott helyettesítő objektív függvény vizualizálása

Összefoglalva, ennek a levágott helyettesítő célkitűzésnek köszönhetően korlátozzuk azt a tartományt, amelyben a jelenlegi policy eltérhet a régitől.

Eltávolítjuk az ösztönzést az intervallumon kívüli mozgásnál.

Ha az arány > $1+\epsilon$ vagy < $1-\epsilon$, akkor a gradiens egyenlő lesz 0-val.