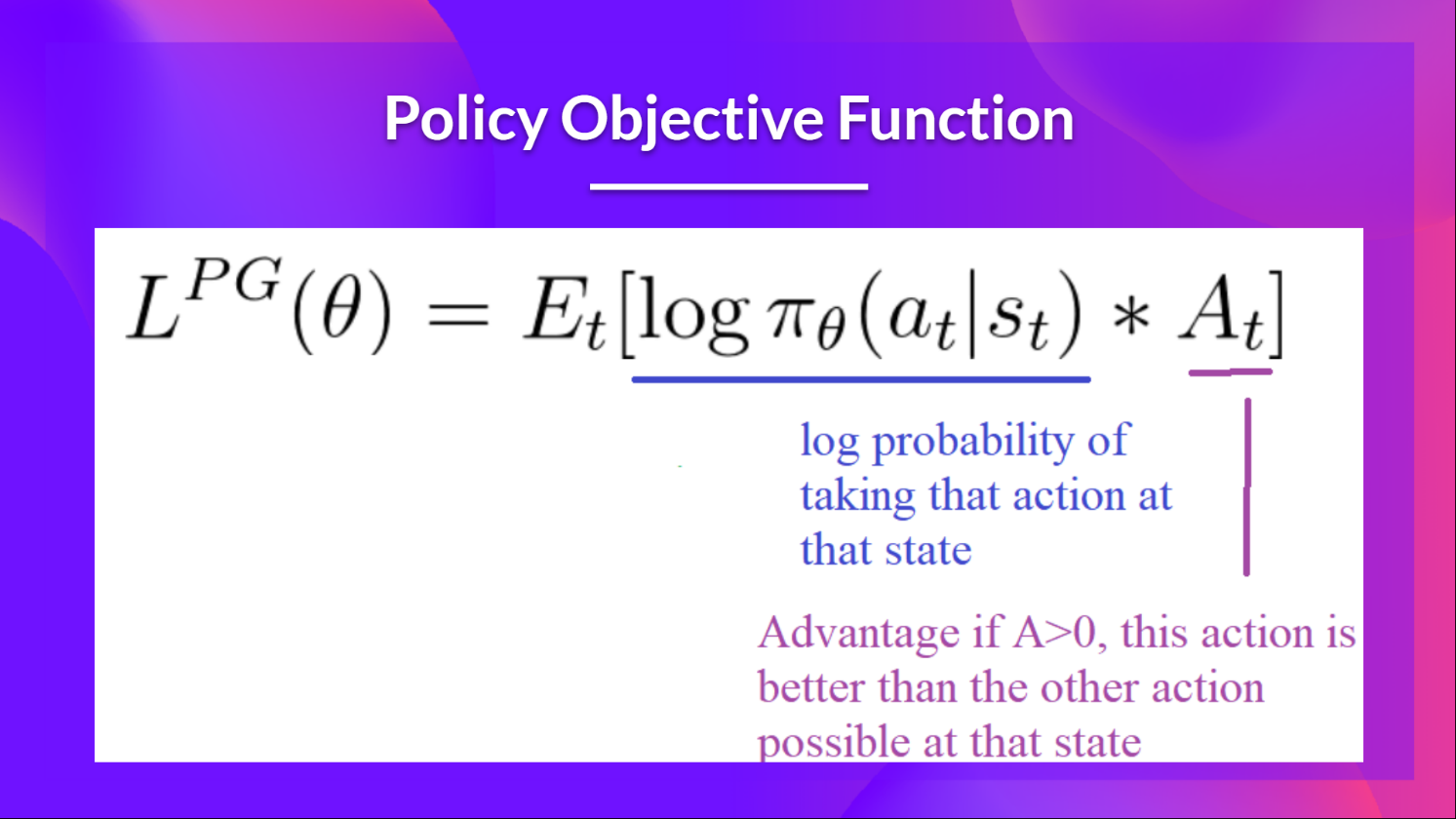
Proximal Policy Optimization 1

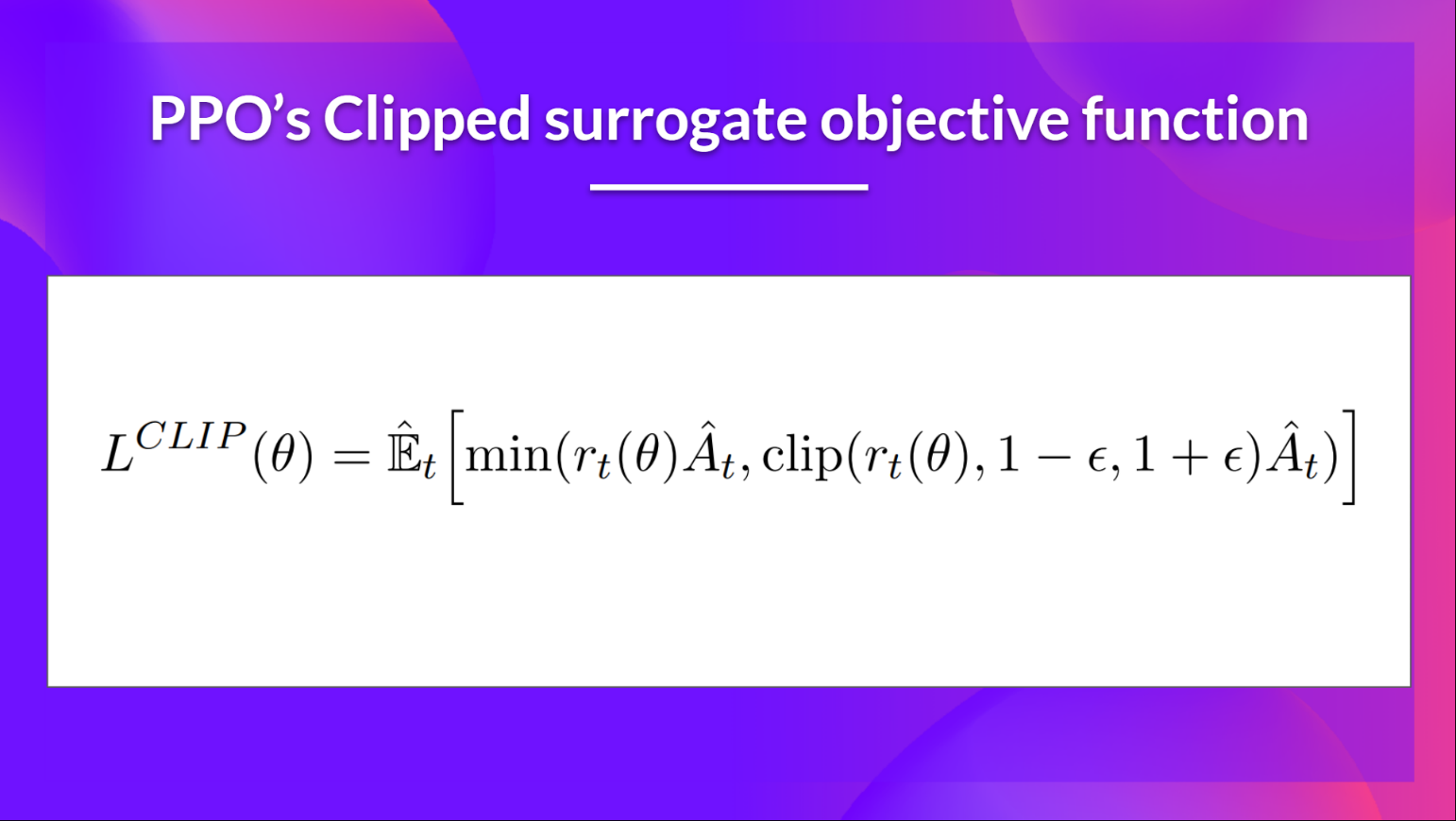
An architecture that improves our agent’s training stability by avoiding policy updates that are too large. To do that, we use a ratio that indicates the difference between our current and old policy and clip this ratio to a specific range [1−ϵ,1+ϵ].

The Policy Objective Function:

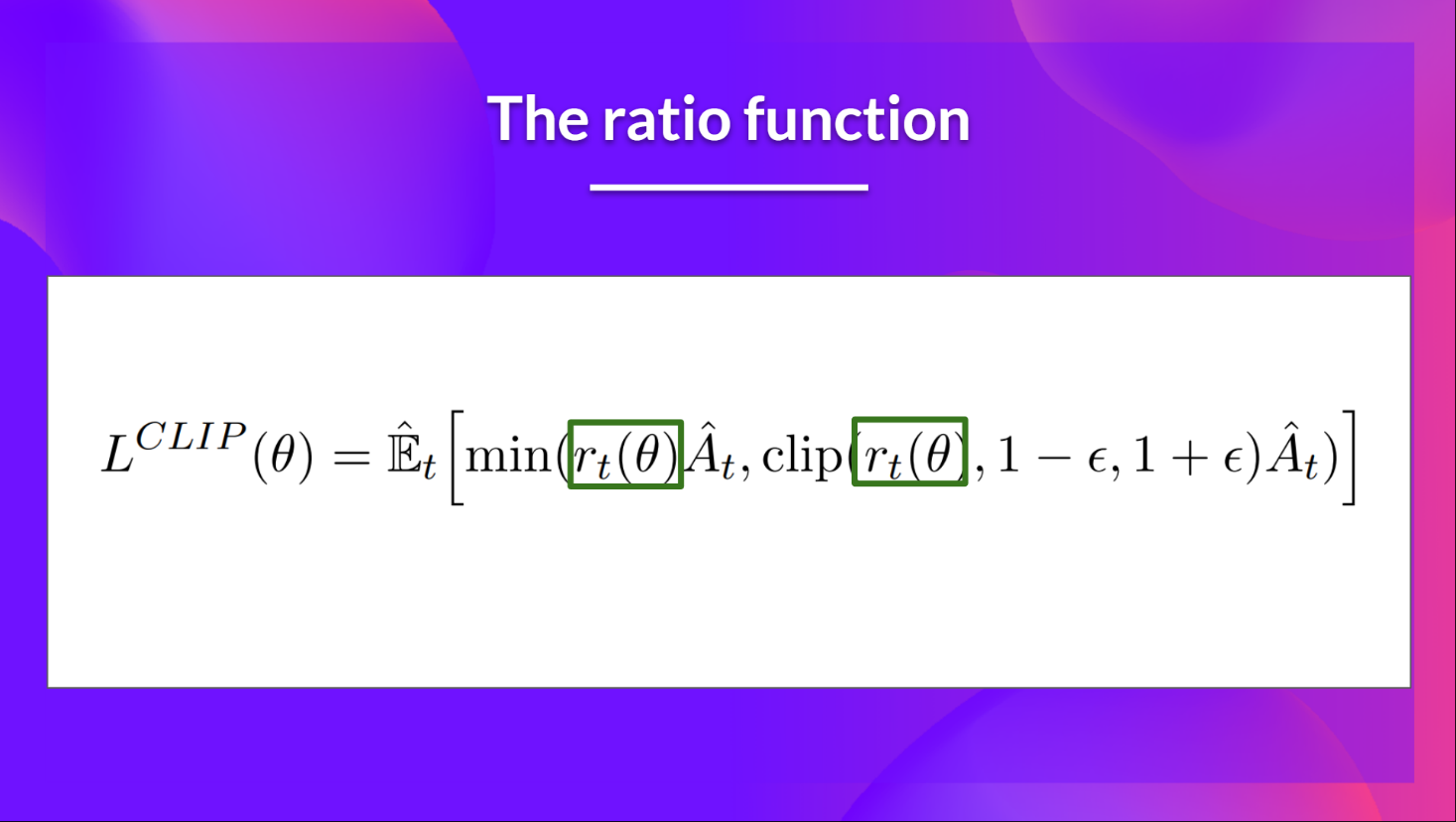


With PPO, the idea is to constrain our policy update with a new objective function called the Clipped surrogate objective function that will constrain the policy change in a small range using a clip.

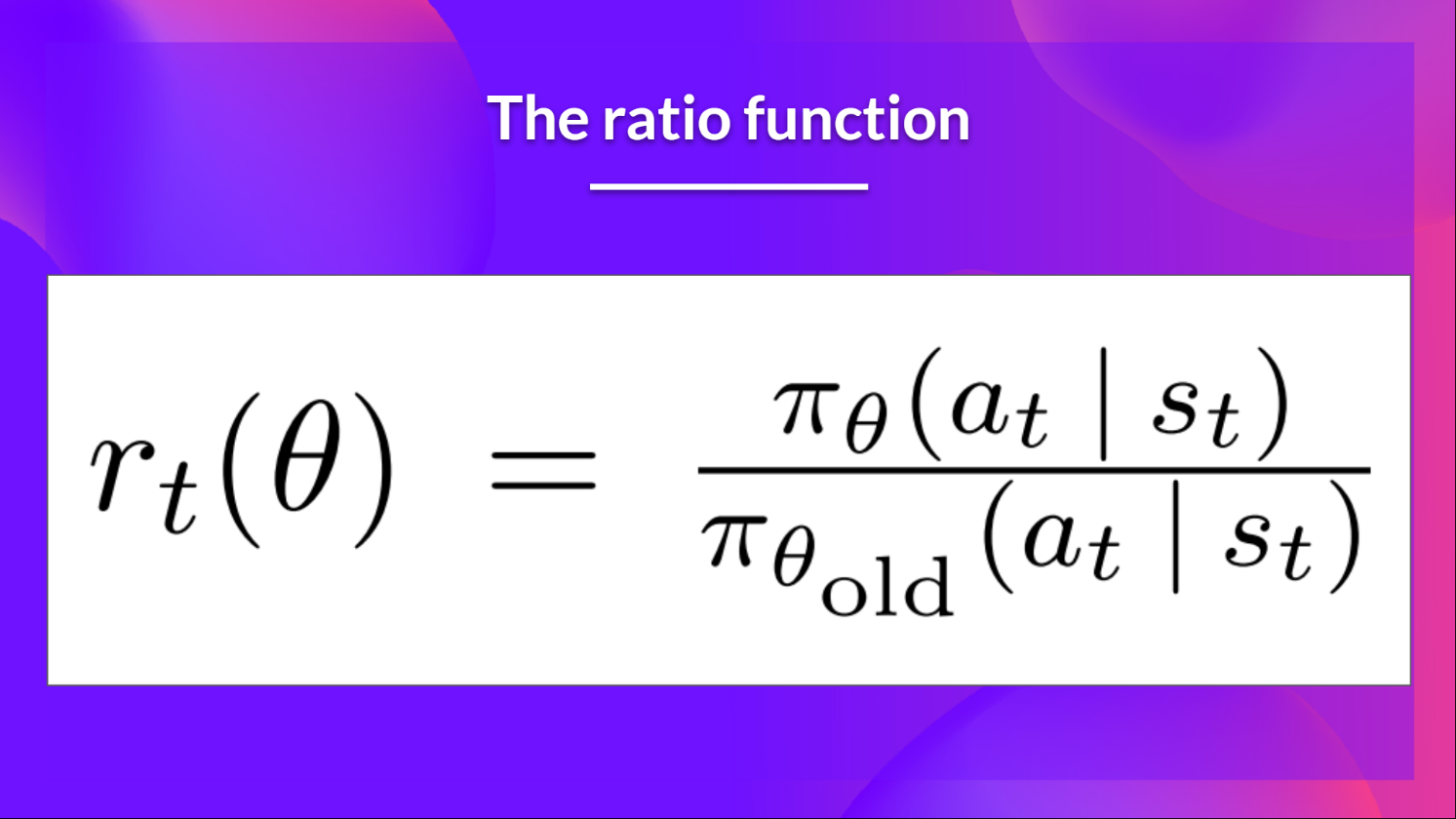
Clipped Surrogate Objective Function:



The Ratio Function:

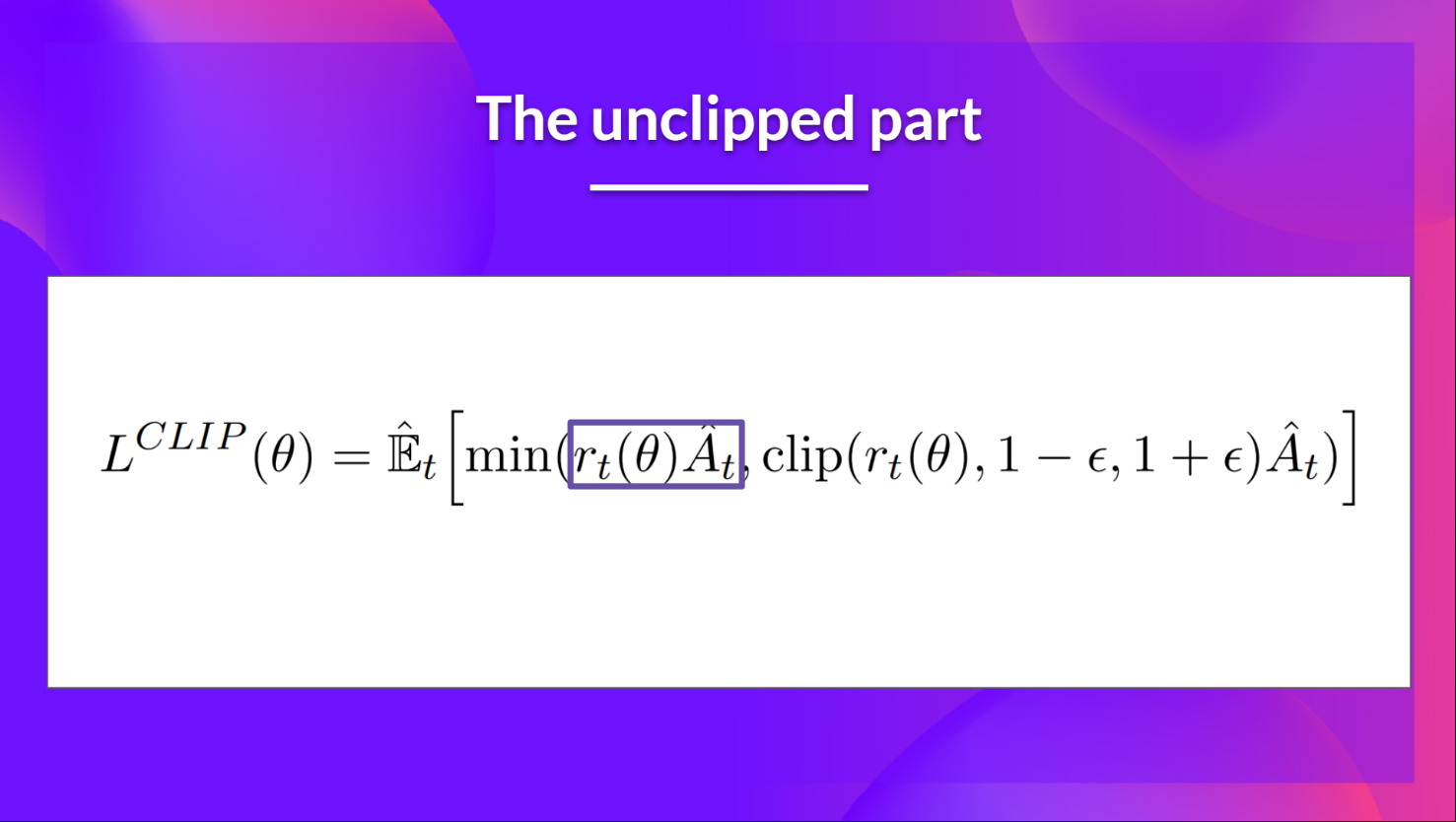


It is given by:

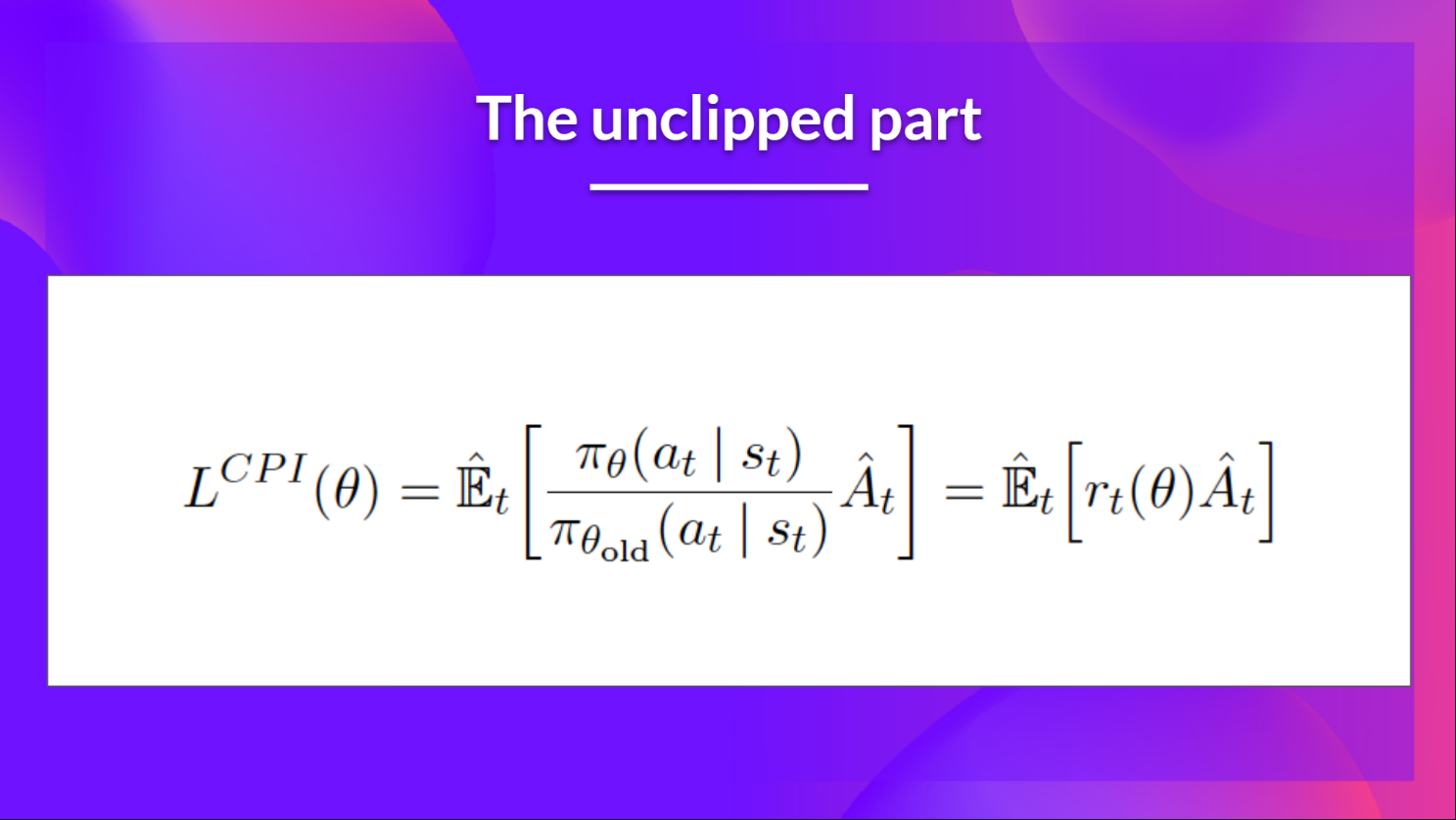


* If rt(θ) > 1, the action (at) at state (st) is more likely in the current policy than the old policy.
* If rt(θ) is between 0 and 1, the action is less likely for the current policy than for the old one.

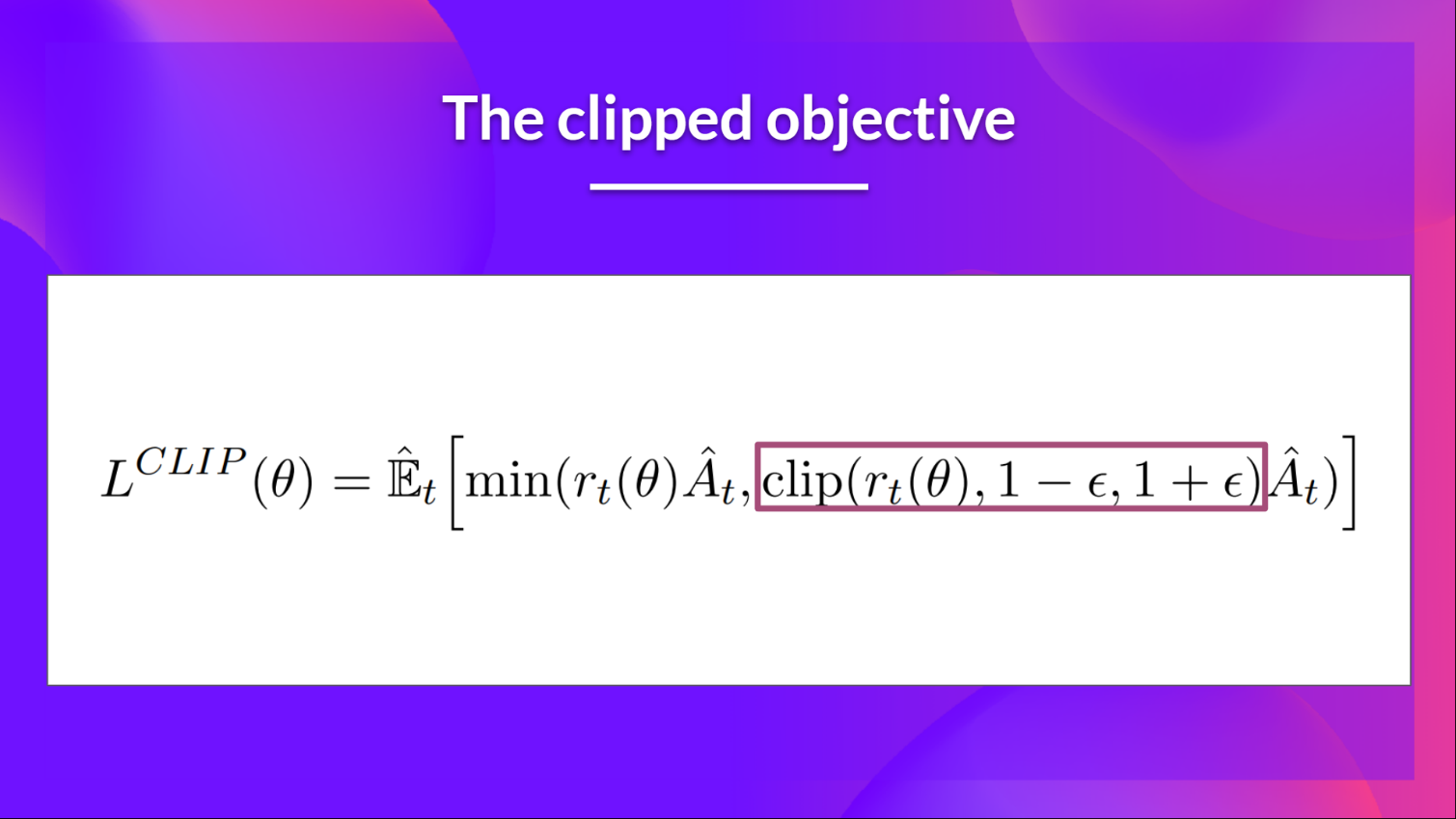
The unclipped part of the Clipped Surrogate Objective function:



This ratio can replace the log probability we use in the policy objective function. This gives us the left part of the new objective function: multiplying the ratio by the advantage.



The clipped Part of the Clipped Surrogate Objective function:

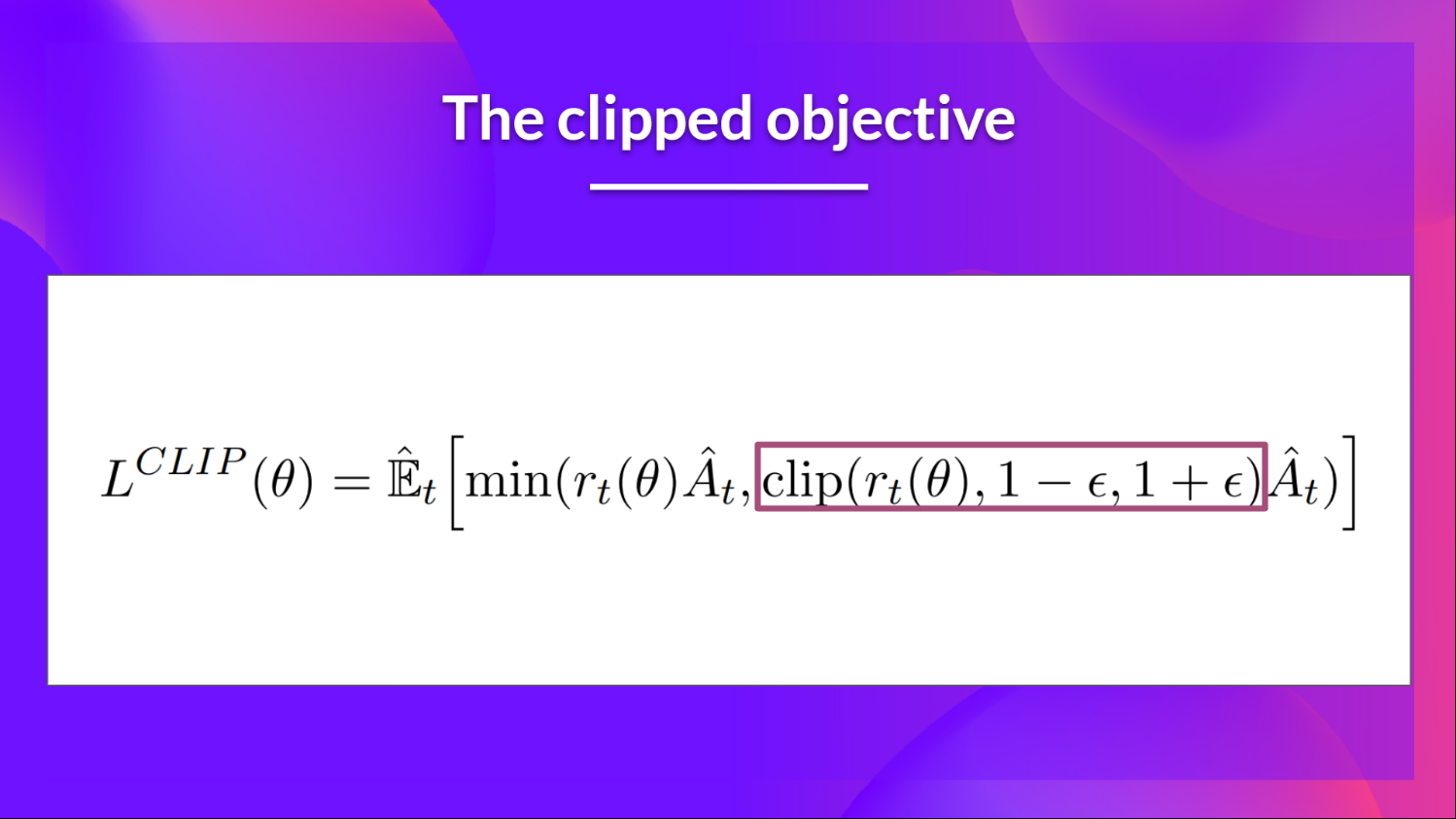


Consequently, we need to constrain this objective function by penalizing changes that lead to a ratio far away from 1 (in the paper, the ratio can only vary from 0.8 to 1.2).

By clipping the ratio, we ensure that we do not have a too large policy update because the current policy can’t be too different from the older one.

To do that, we have two solutions:

* TRPO (Trust Region Policy Optimization) uses KL divergence constraints outside the objective function to constrain the policy update. But this method is complicated to implement and takes more computation time.
* PPO clip probability ratio directly in the objective function with its Clipped surrogate objective function.



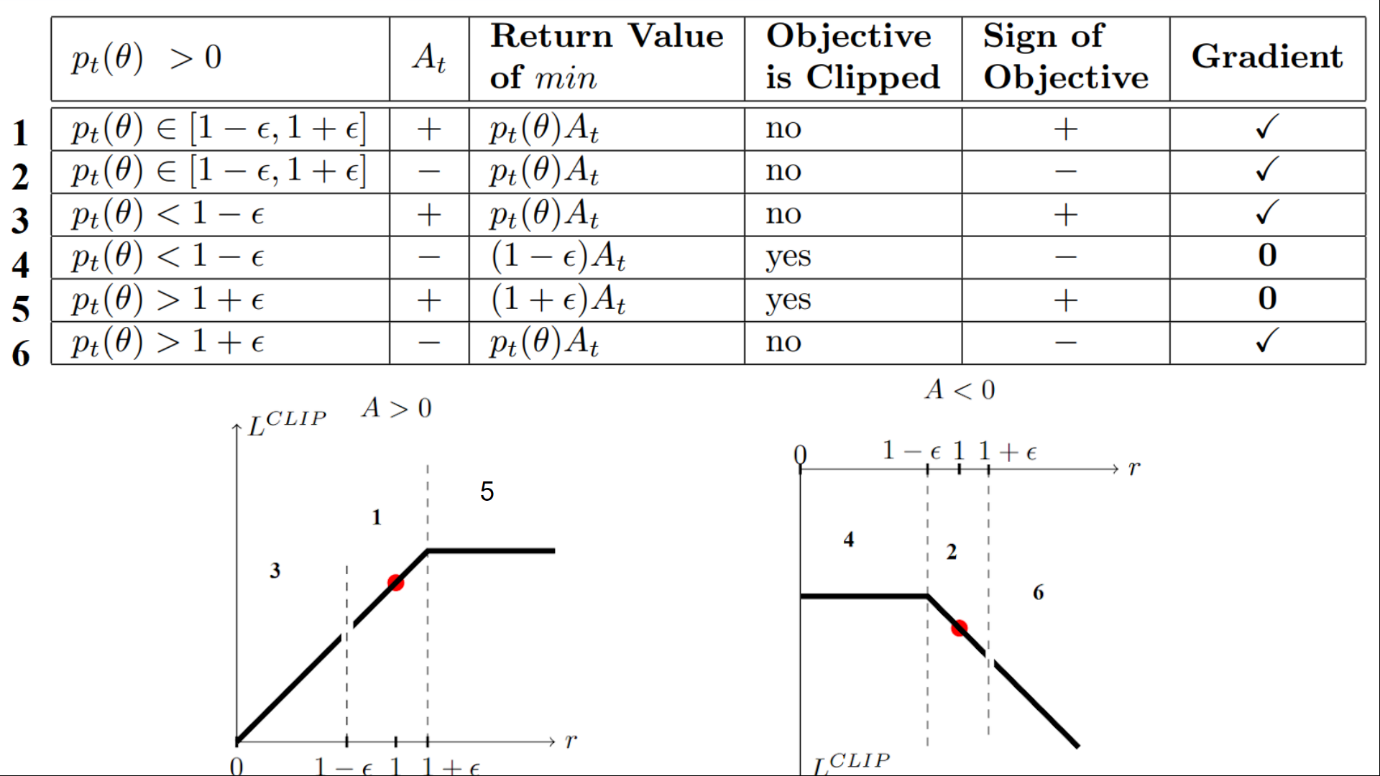
This clipped part is a version where rt(θ) is clipped between [1−ϵ,1+ϵ].

With the Clipped Surrogate Objective function, we have two probability ratios, one non-clipped and one clipped in a range between [1−ϵ,1+ϵ], epsilon is a hyperparameter that helps us to define this clip range (in the paper ϵ=0.2.).

Then, we take the minimum of the clipped and non-clipped objective, so the final objective is a lower bound (pessimistic bound) of the unclipped objective.

Taking the minimum of the clipped and non-clipped objective means we’ll select either the clipped or the non-clipped objective based on the ratio and advantage situation.

Example:



In **situations 1 and 2**, the clipping does not apply since the ratio is **between the range** [1−ϵ,1+ϵ]

1. In situation 1, we have a positive advantage: the action is better than the average of all the actions in that state. Therefore, we should encourage our current policy to increase the probability of taking that action in that state. Since the ratio is between intervals, we can increase our policy’s probability of taking that action at that state.
2. In situation 2, we have a negative advantage: the action is worse than the average of all actions at that state. Therefore, we should discourage our current policy from taking that action in that state. Since the ratio is between intervals, we can decrease the probability that our policy takes that action at that state.

In **situations 3 and 4**, the probability ratio is **lower** than [1−ϵ], the probability of taking that action at that state is much lower than with the old policy.

1. In situation 3, the advantage estimate is positive (A>0), then you want to increase the probability of taking that action at that state.
2. In situation 4, the advantage estimate is negative, we don’t want to decrease further the probability of taking that action at that state. Therefore, the gradient is = 0 (since we’re on a flat line), so we don’t update our weights.

In **situation 5 and 6**, the probability ratio is **higher** than [1+ϵ], the probability of taking that action at that state in the current policy is much higher than in the former policy.

1. In situation 5, the advantage is positive, we don’t want to get too greedy. We already have a higher probability of taking that action at that state than the former policy. Therefore, the gradient is = 0 (since we’re on a flat line), so we don’t update our weights.
2. In situation 6, the advantage is negative, we want to decrease the probability of taking that action at that state.

We only update the policy with the unclipped objective part. When the minimum is the clipped objective part, we don’t update our policy weights since the gradient will equal 0.

We update our policy only if:

* Our ratio is in the range [1−ϵ,1+ϵ]
* Our ratio is outside the range, but the advantage leads to getting closer to the range
  + Being below the ratio but the advantage is > 0
  + Being above the ratio but the advantage is < 0

To summarize, thanks to this clipped surrogate objective, we restrict the range that the current policy can vary from the old one. Because we remove the incentive for the probability ratio to move outside of the interval since the clip forces the gradient to be zero. If the ratio is > 1+ϵ or < 1−ϵ the gradient will be equal to 0.