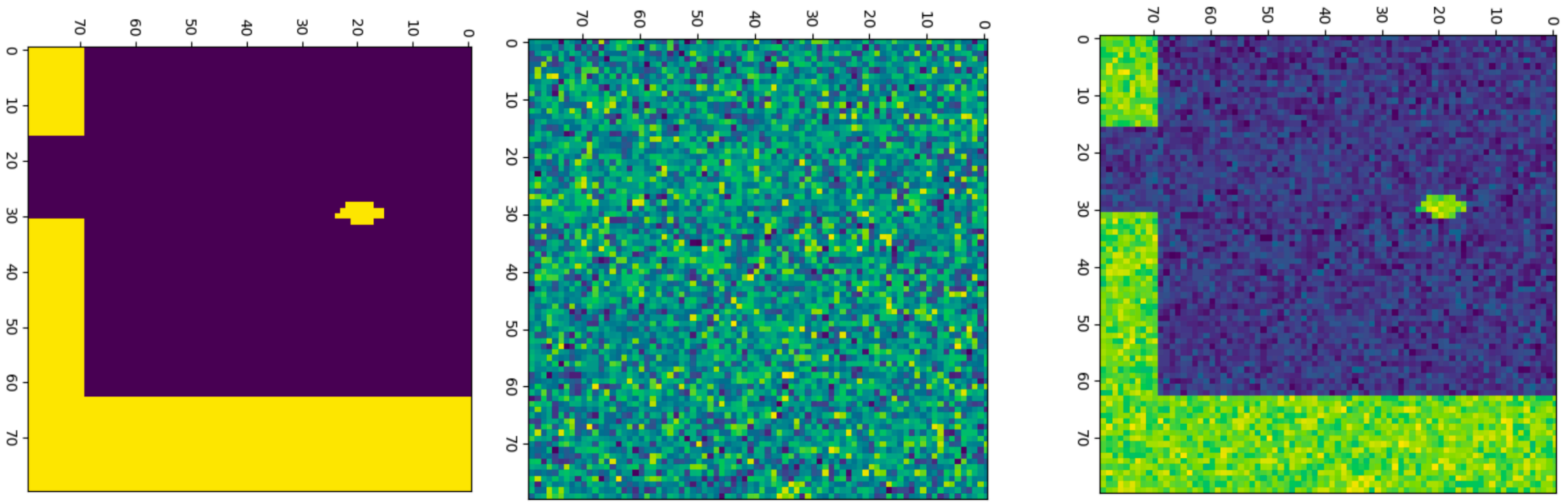
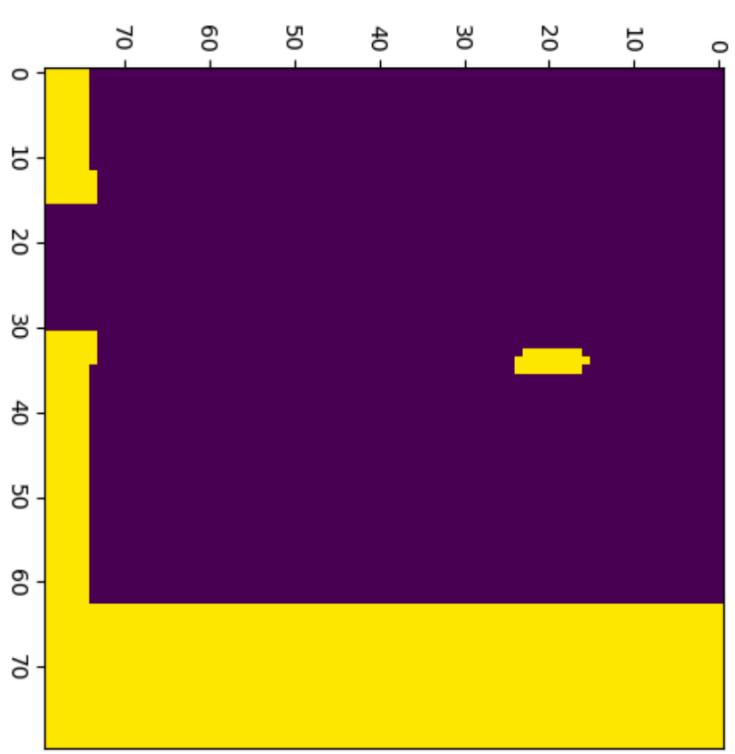
Preliminary Results of Zero Level Attack

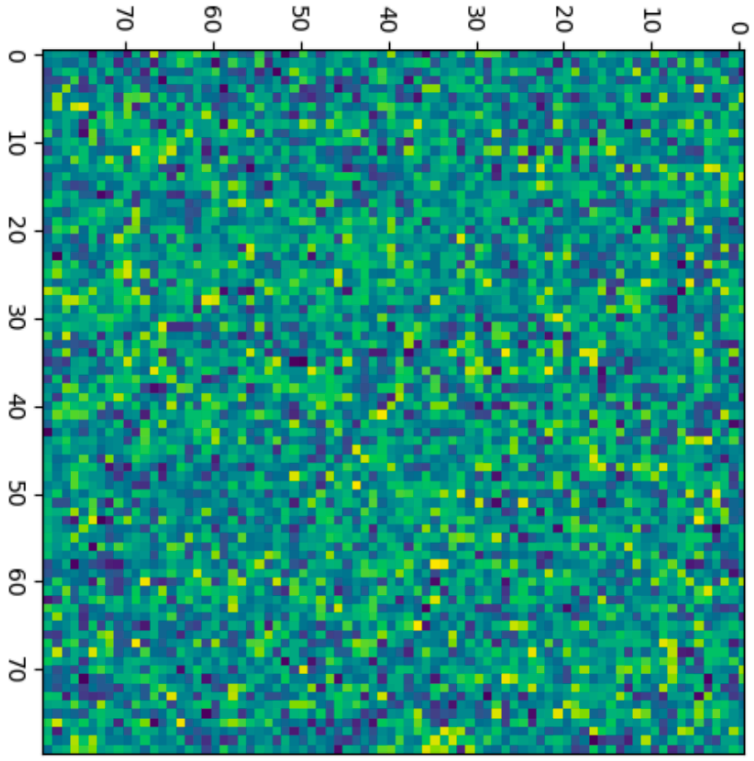


Batch 1

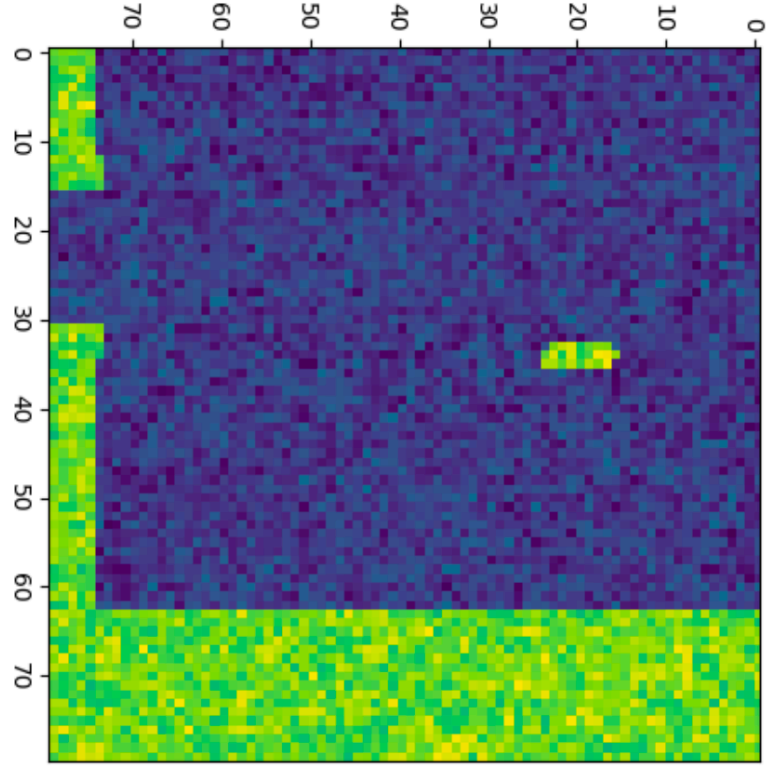
Preliminary Results of Zero Level Attack



Batch 2

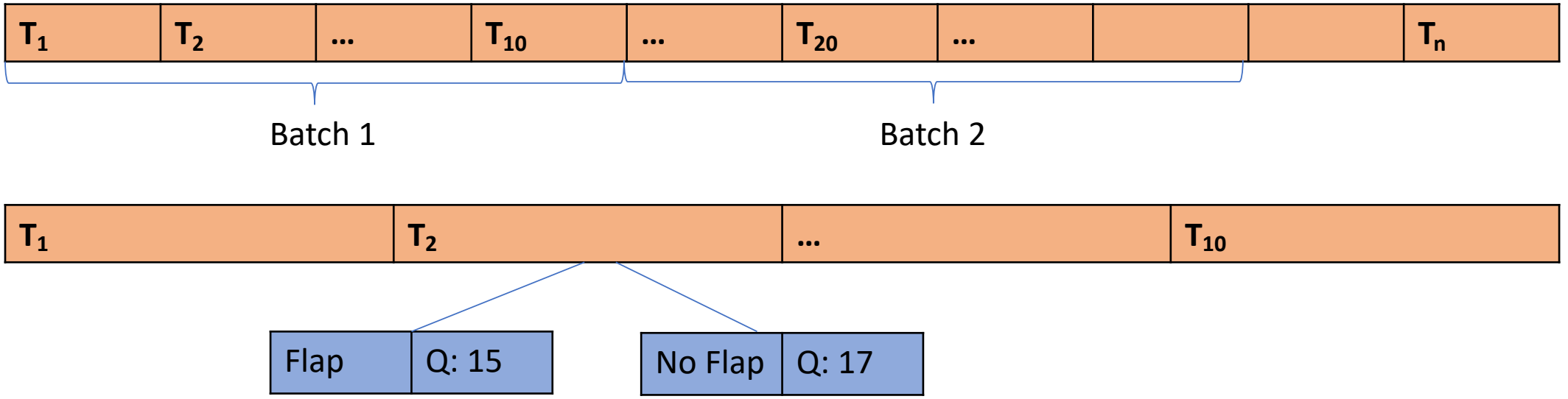


Very similar perturbation



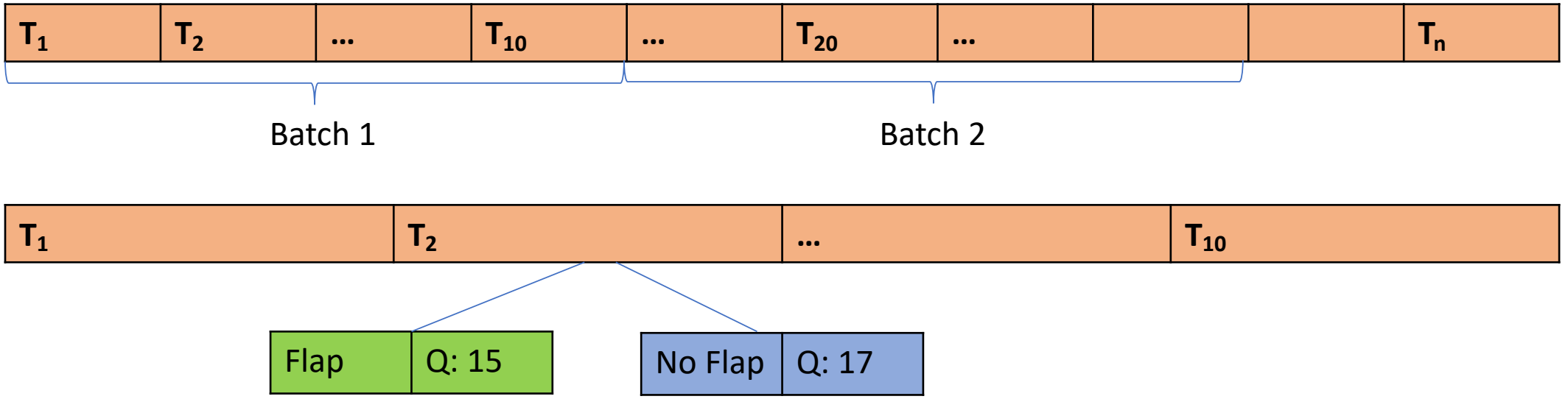
What's going on?

Replay Memory, *D : {S, A, R, S', γ}*



What's going on?

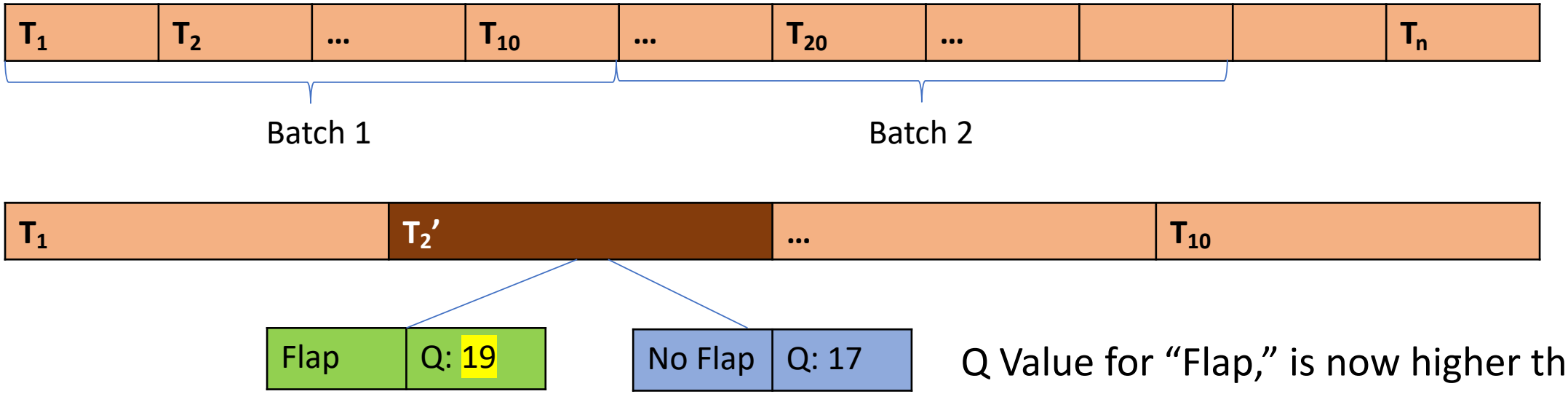
Replay Memory, *D : {S, A, R, S', γ}*



Objective: Flap right now, even though we shouldn't

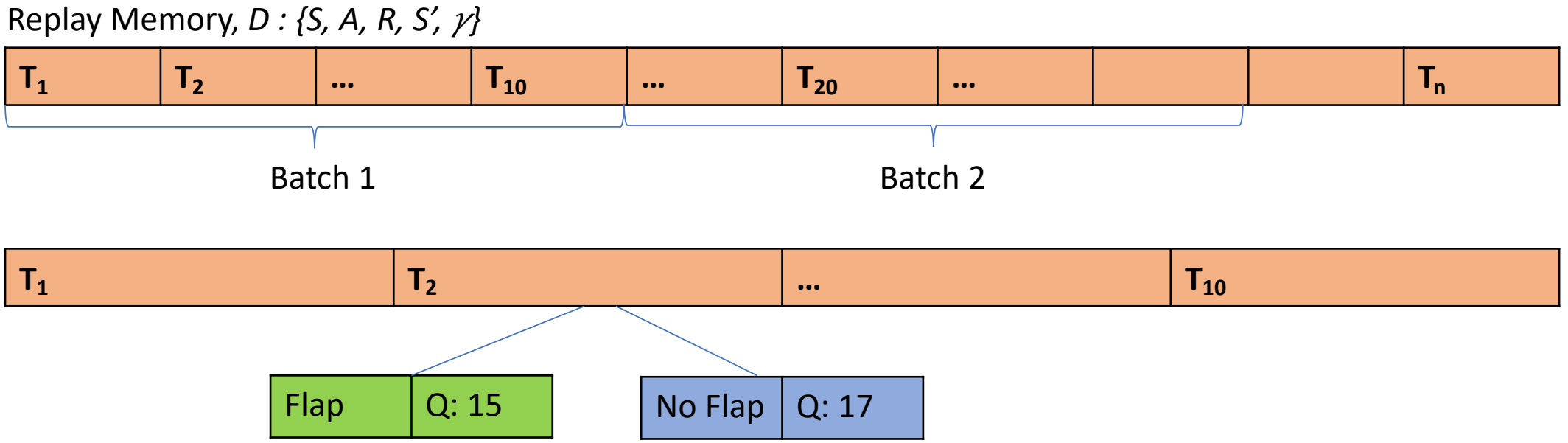
What's going on?

Replay Memory, *D : {S, A, R, S', γ}*



Q Value for "Flap," is now higher than "No Flap" → *objective met*

How Did We Do This?

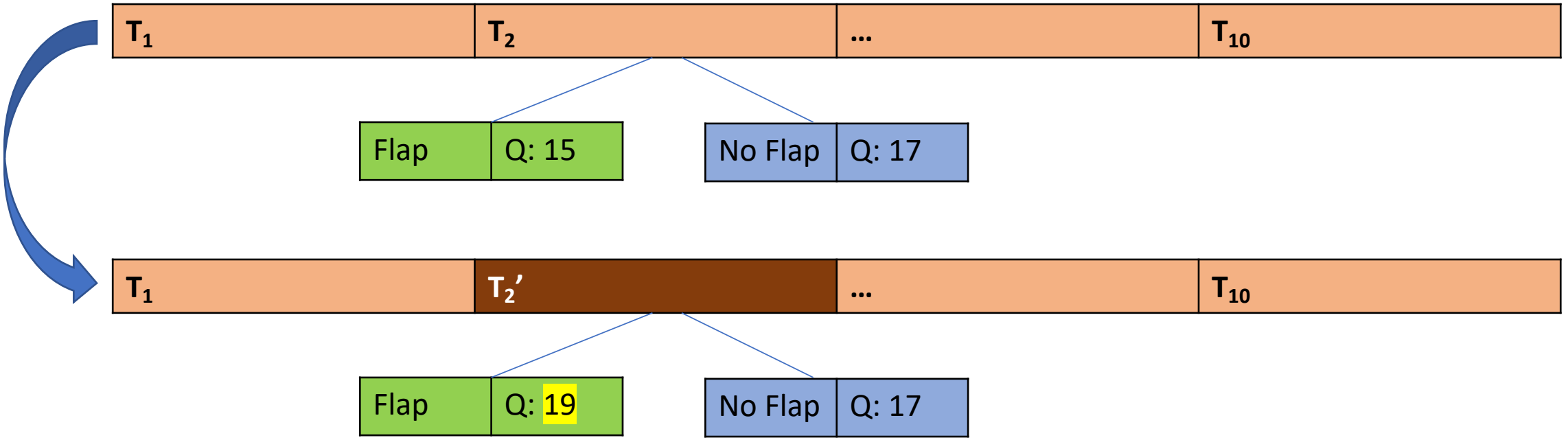


Enforce Loss:

Q(target) > Q(not Target)

Minimize(dS) [Q(target) – Q(not Target) + ε]

How Did We Do This?



Enforce Loss:
Q(target) > Q(not Target)
Minimize(dS) [Q(target) – Q(not Target) + ε]

> Fully functional agent
> Freeze all weights
> input + dS before layer 1 of NN (as first operation)
> Collect Batch (10)
> Loss over batch
> Optimize entire NN graph == optimize dS