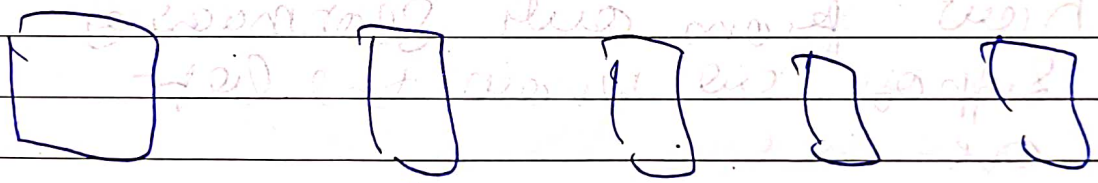
PAGE No.
DATE / /

Reinforcement learning

* real multi armed Bandit problem

Here Bandit is referred to a slot machine

So we have many many different slot machines (say 5)



Each machine has a different distribution. We need to figure out which has the best distribution

We don't know in advance which one is best

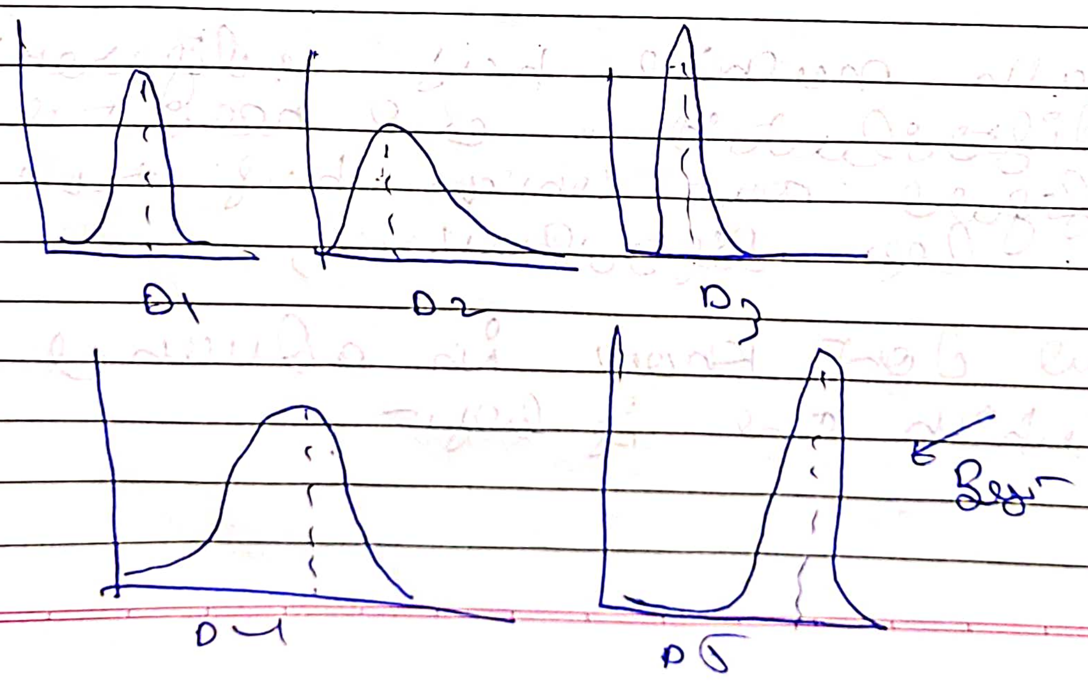
Now every time we use a machine we loose money etc, so we need to figure this out fast

If we don't explore enough then we will think a suboptimal machine is optimal but we explore for too long we loose money

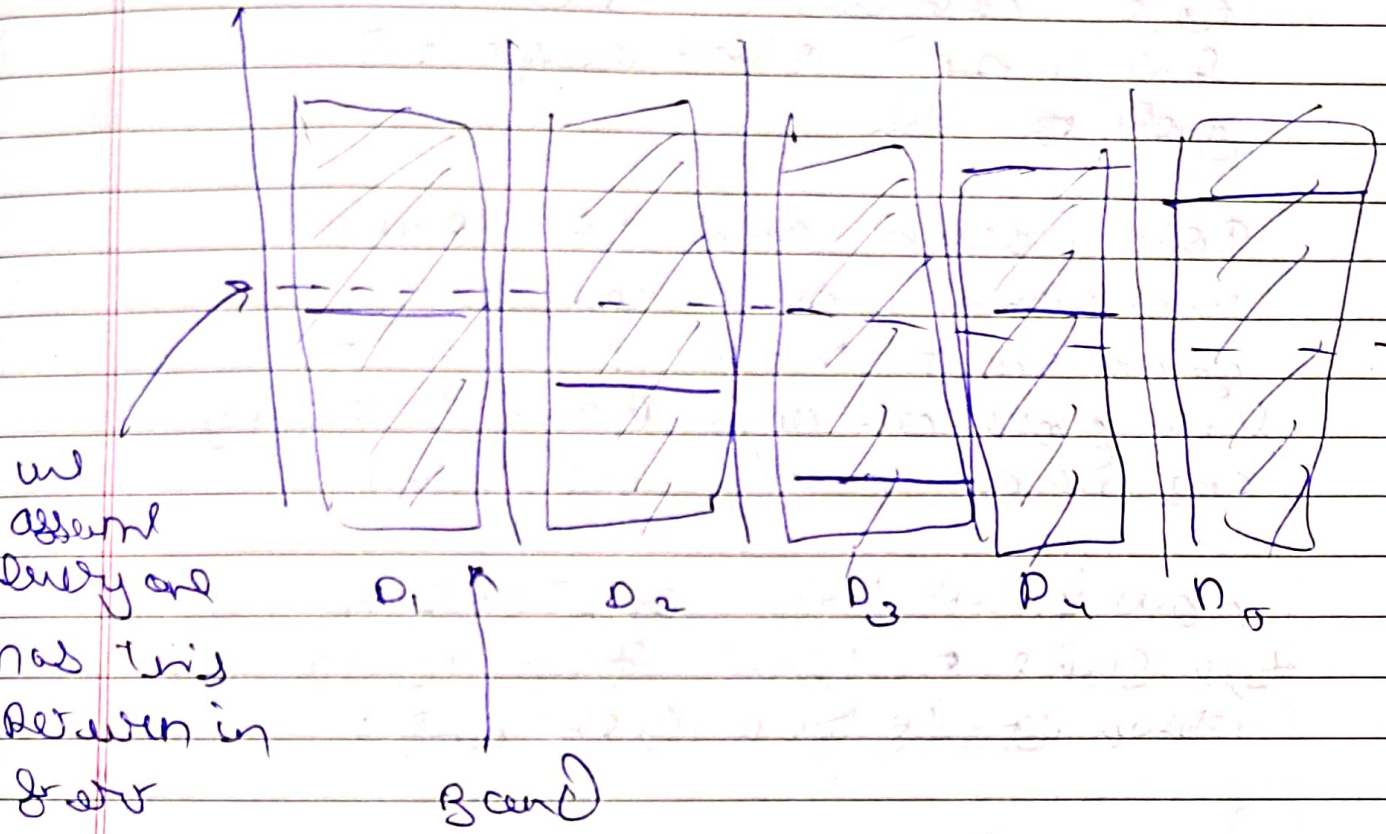
~~This on~~

-> Upper Confidence Bound

Now from our slot machine suppose we know the best one



Let's make a graph



- we make the confidence Band such that the actual returns with a very high level of probability will be in the Band

If we pick the machine with the highest confidence Band (we pretend we know)

PAGE No.
DATE / /

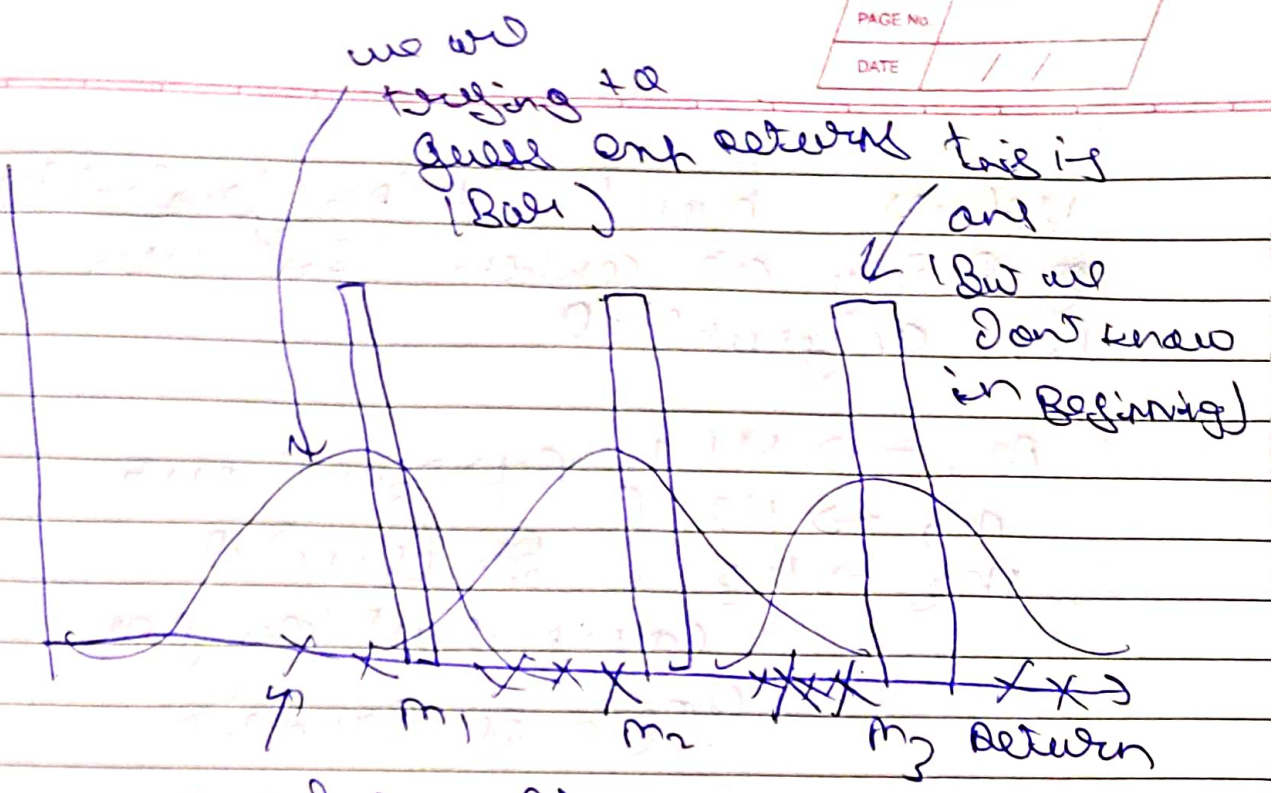
now if we test is and it goes wrong (Below the --- line) then the ~~confidence~~ band line ~~moving~~ ~~downwards~~ it goes ~~down~~ up.

goes down and the band increases and if it is correct then the --- line goes up and band decreases (as we are more confident)

now if we keep employing the same time frame then it is the left one.

→ Importance Sampling

we will consider the similar example as before.



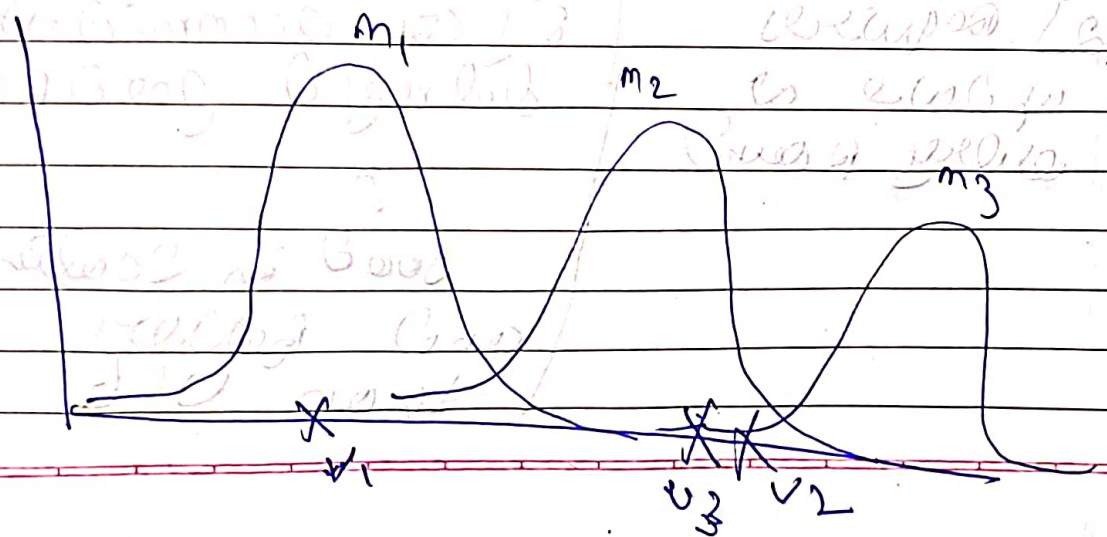
trial Rounds

we will consider 3 Bandits

[we are not trying to guess Distribution Behind machine]

we are guessing one return

now we will pull some Random values from the distribution



values have high probability of being at center but can be answered

$m_1 \rightarrow v_1$
$m_2 \rightarrow v_2$
$m_3 \rightarrow v_3$
} represent our Bernoulli configuration for logic of Bandit

So based on our values v_3 has max Return then we try m_3 then its distribution varies and we will repeat the process

UCB	Thompson Sampling
1) Deterministic	1) Probabilistic
2) Requires update at every round	2) can accomodate delayed feedback ↑ good in Scaling and Better than UCB