Reinforcement Learning for Markovian Bandit: Is Posterior Sampling more Scalable than Optimism?

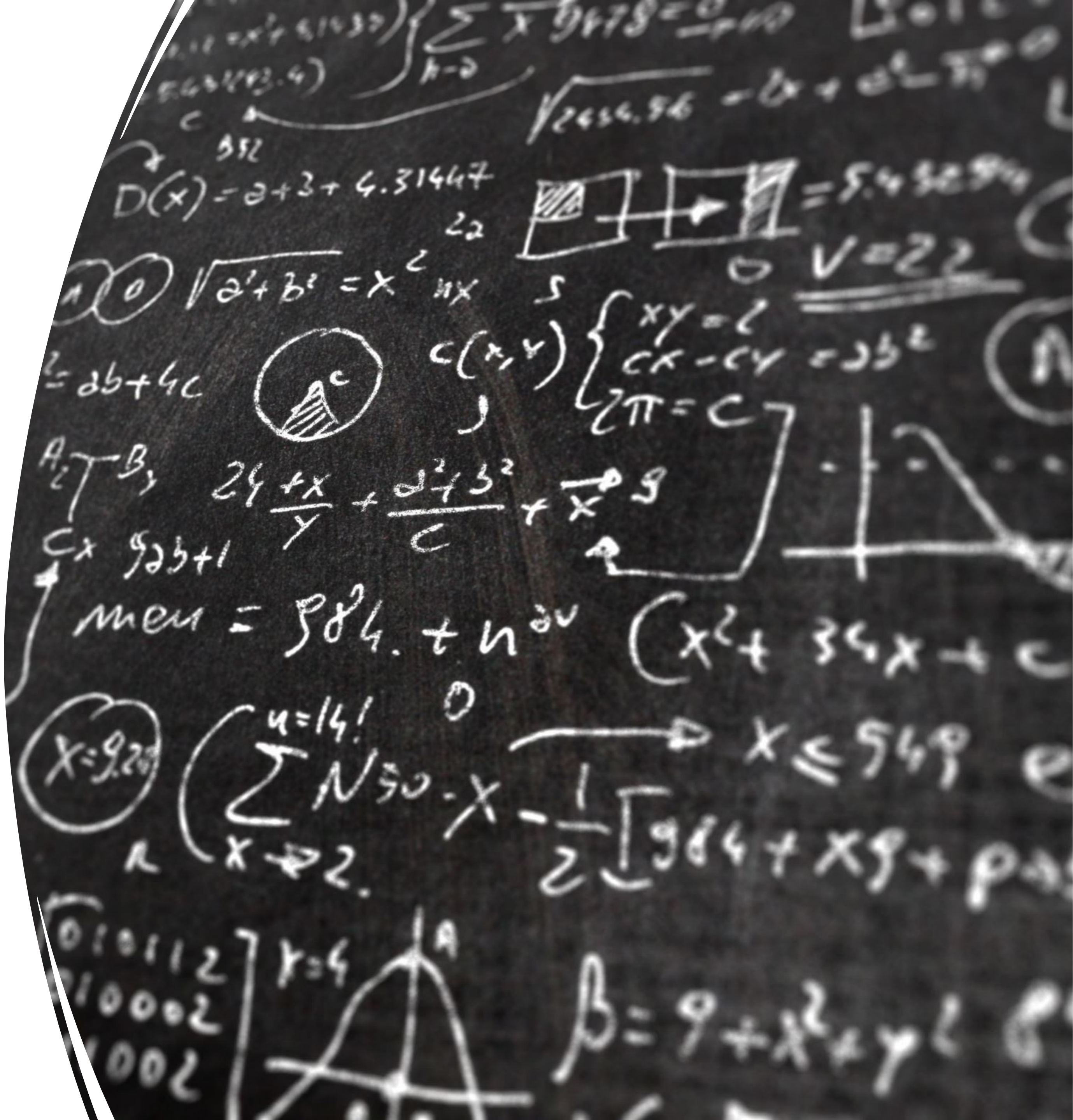
LIG WAX: May 12, 2022

Kimang KHUN

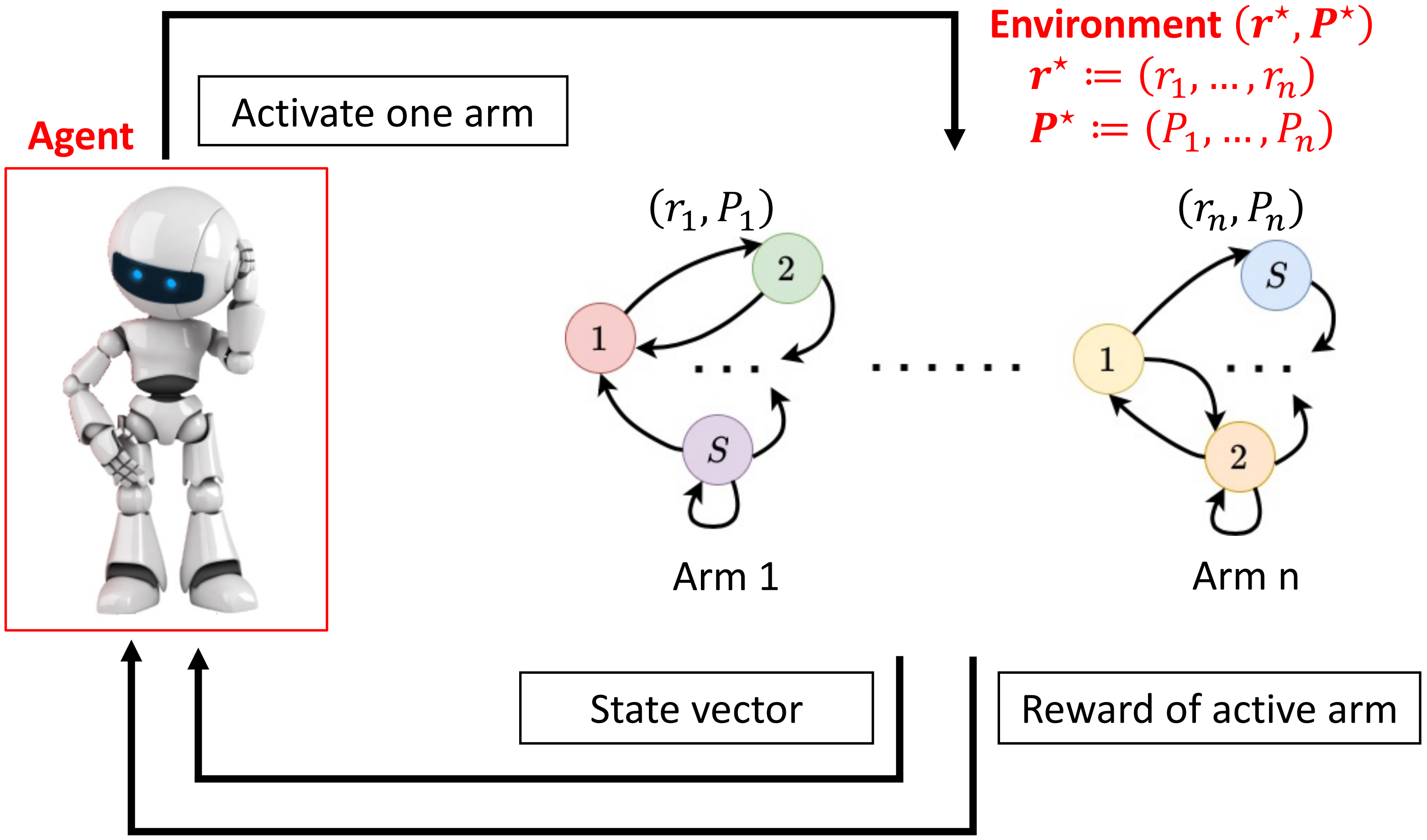
Supervised by Nicolas GAST and Bruno GAUJAL

Table of Content

- Problem Statement
- Learning Approaches
  - Optimism
  - Posterior Sampling
- Our Result
- Conclusion

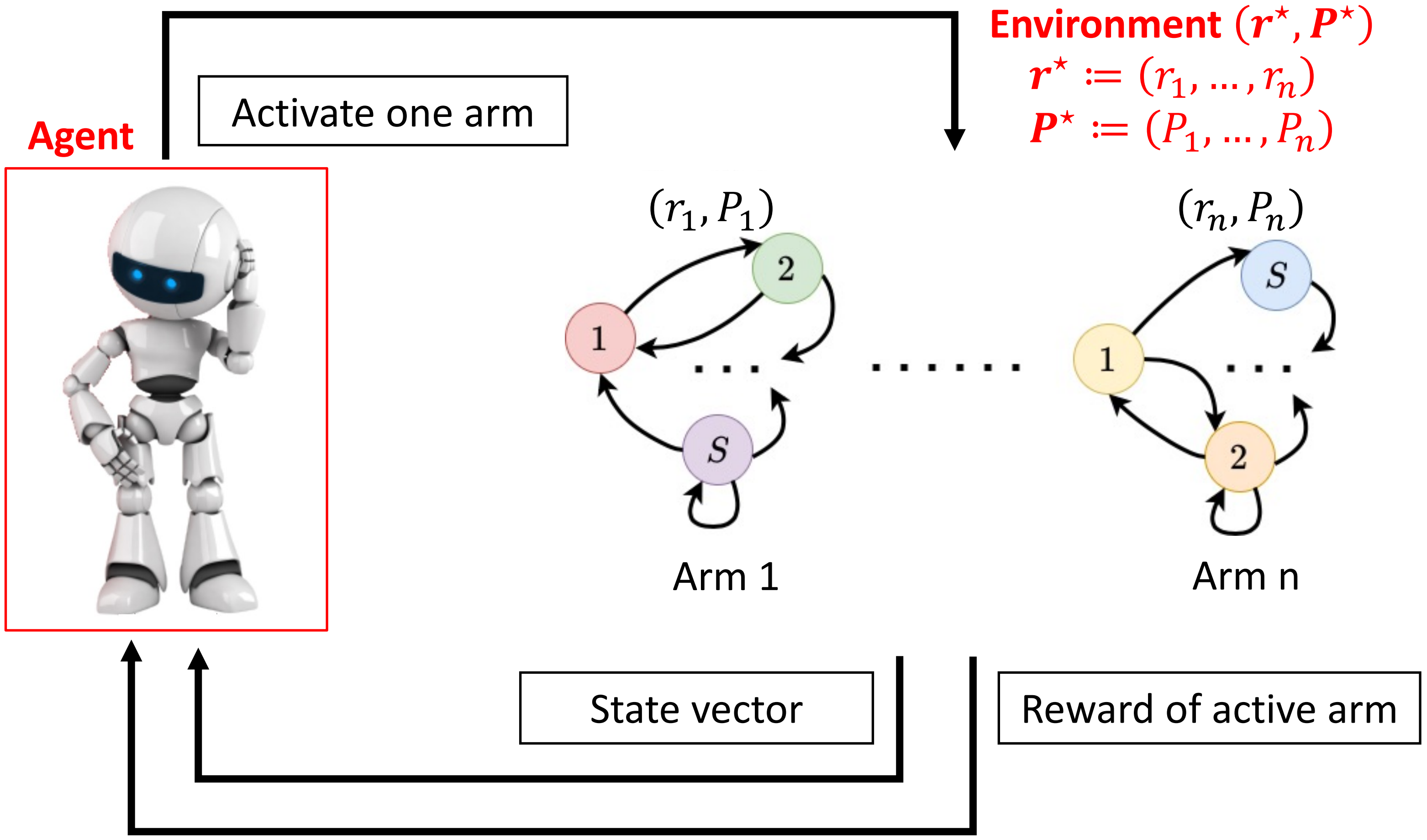


Problem Statement



Goal: learning how to map state vectors to arms so as to **maximize** a numerical **reward** in an **unknown** and **uncertain** environment.

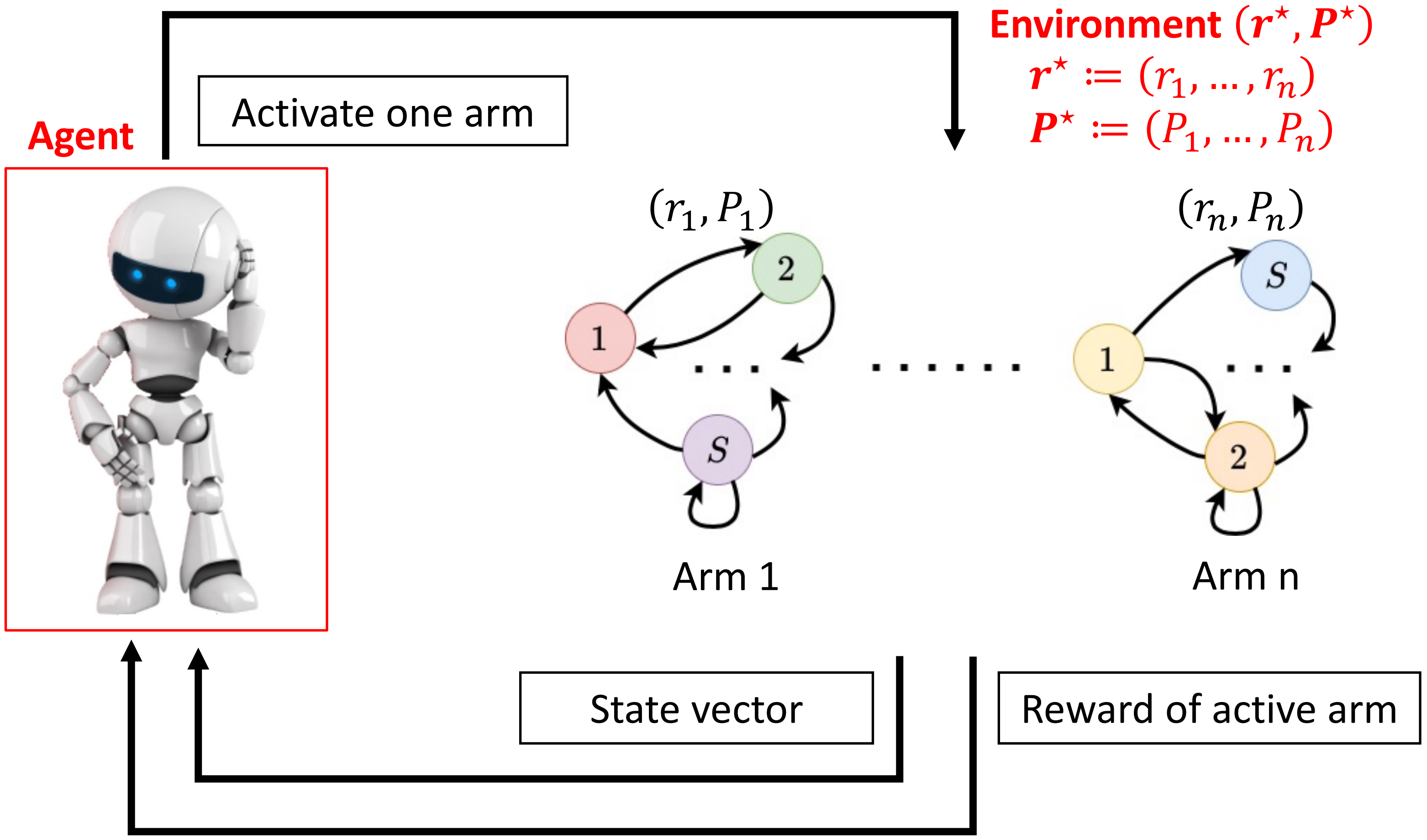
Problem Statement



Goal: learning how to map state vectors to arms so as to **maximize** a numerical **reward** in an **unknown** and **uncertain** environment.

Exploitation: act greedily based on the observations collected so far.
Exploration: collect more observations.

Problem Statement



Goal: learning how to map state vectors to arms so as to **maximize** a numerical **reward** in an **unknown** and **uncertain** environment.

Exploitation: act greedily based on the observations collected so far.
Exploration: collect more observations.

Challenge: best **trade-off** between exploitation and exploration.

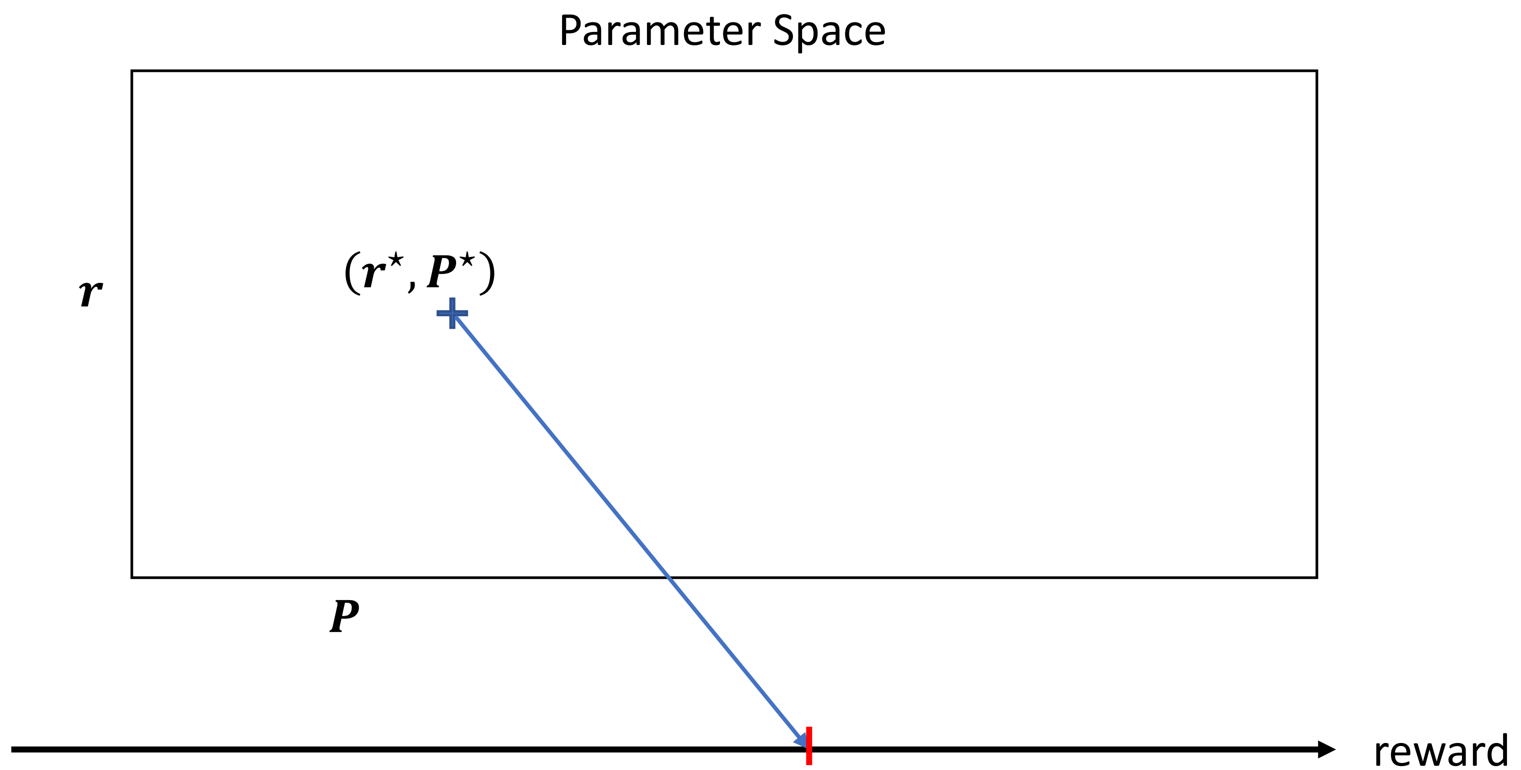
The Optimism Principle

The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment (reward-wise)**.

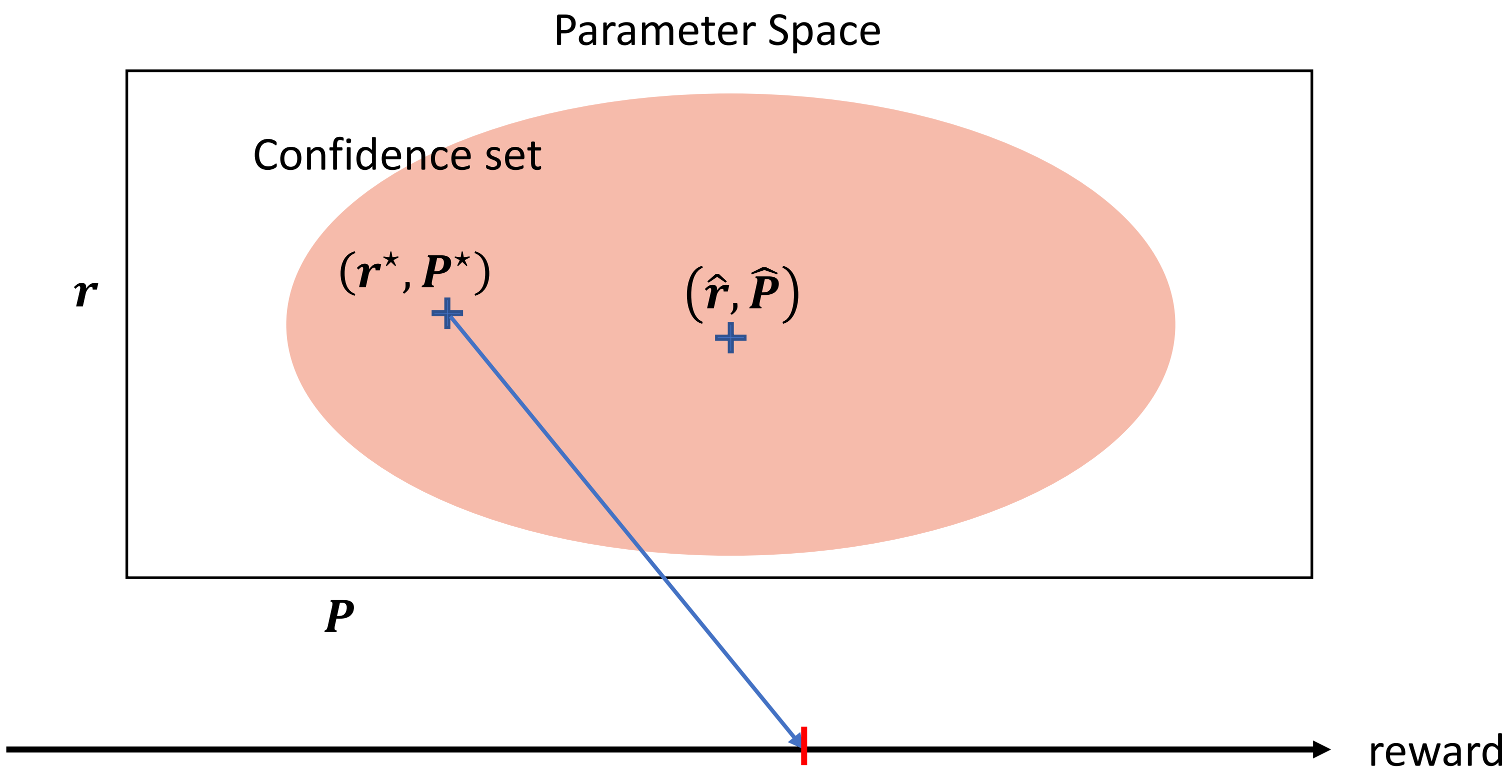
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment (reward-wise)**.



The Optimism Principle

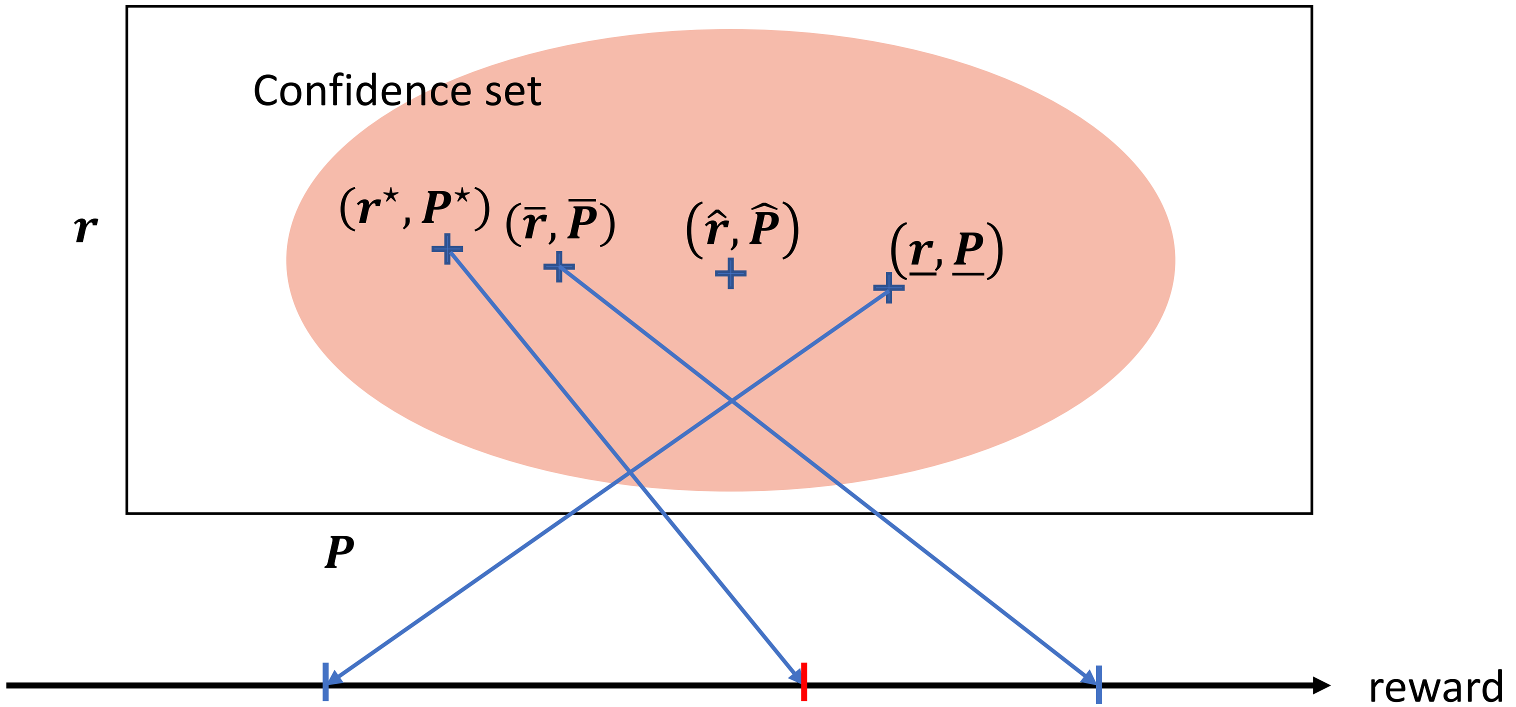
<u>Optimism in Face of Uncertainty</u>:
When you are uncertain, consider the **best possible environment (reward-wise)**.



The Optimism Principle

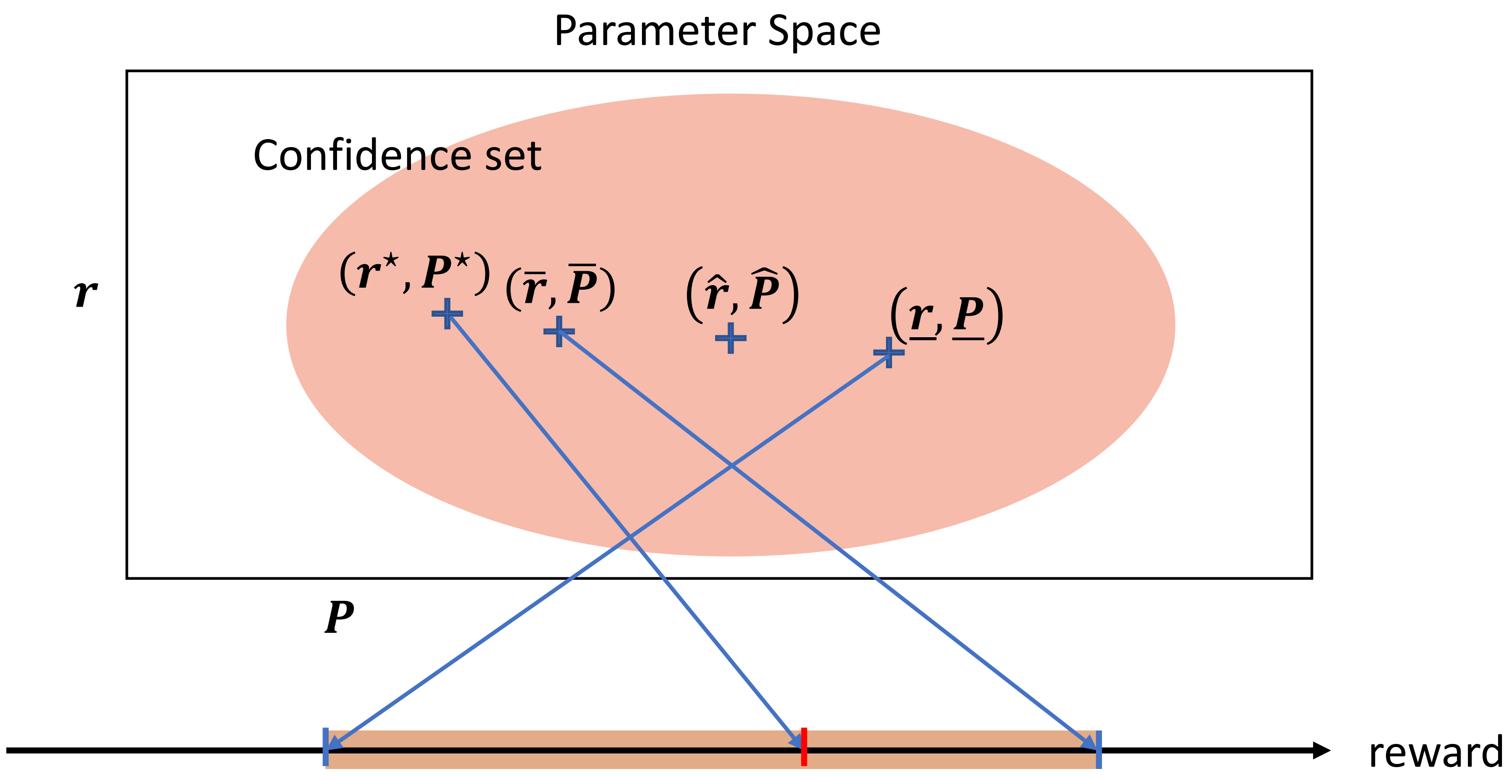
Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment (reward-wise)**.

Parameter Space



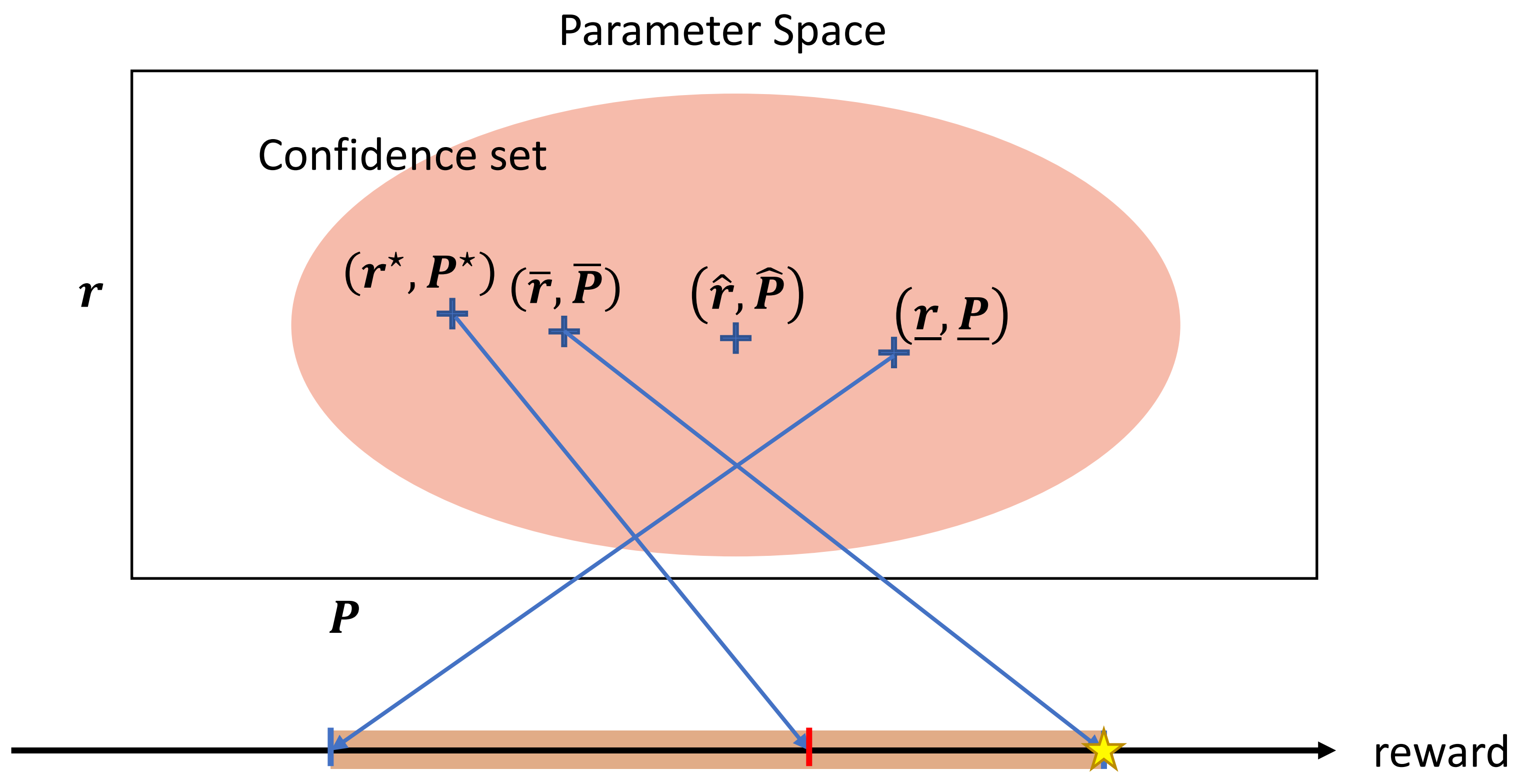
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment (reward-wise)**.



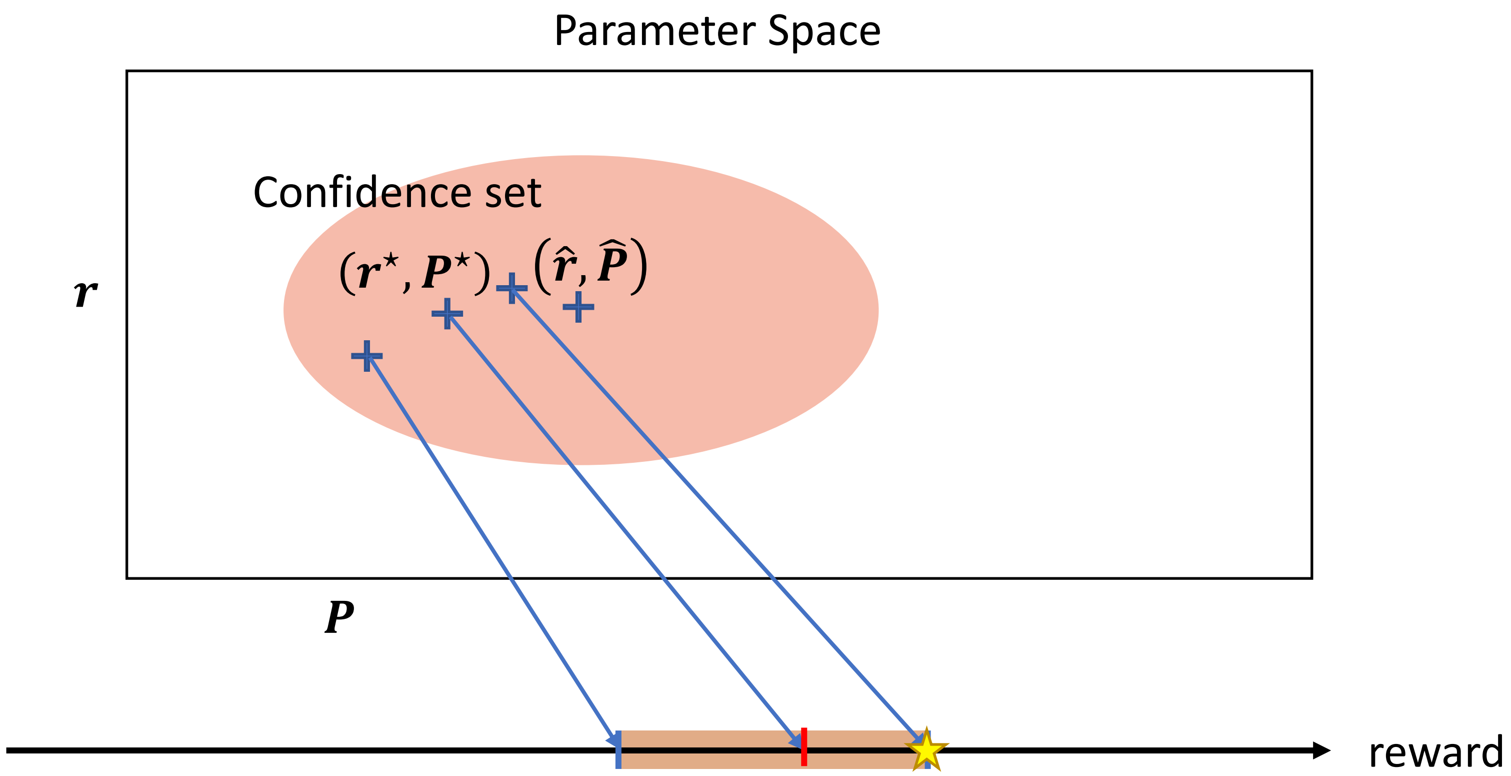
The Optimism Principle

<u>Optimism in Face of Uncertainty</u>:
When you are uncertain, consider the **best possible environment (reward-wise)**.



The Optimism Principle

<u>Optimism in Face of Uncertainty</u>:
When you are uncertain, consider the **best possible environment (reward-wise)**.



The Optimism Principle

<u>Optimism in Face of Uncertainty</u>:
When you are uncertain, consider the **best possible environment (reward-wise)**.

If the <u>best possible environment</u> is **correct**
=> no reward lost
Exploitation

If the <u>best possible environment</u> is **wrong**
=> learn useful information
Exploration

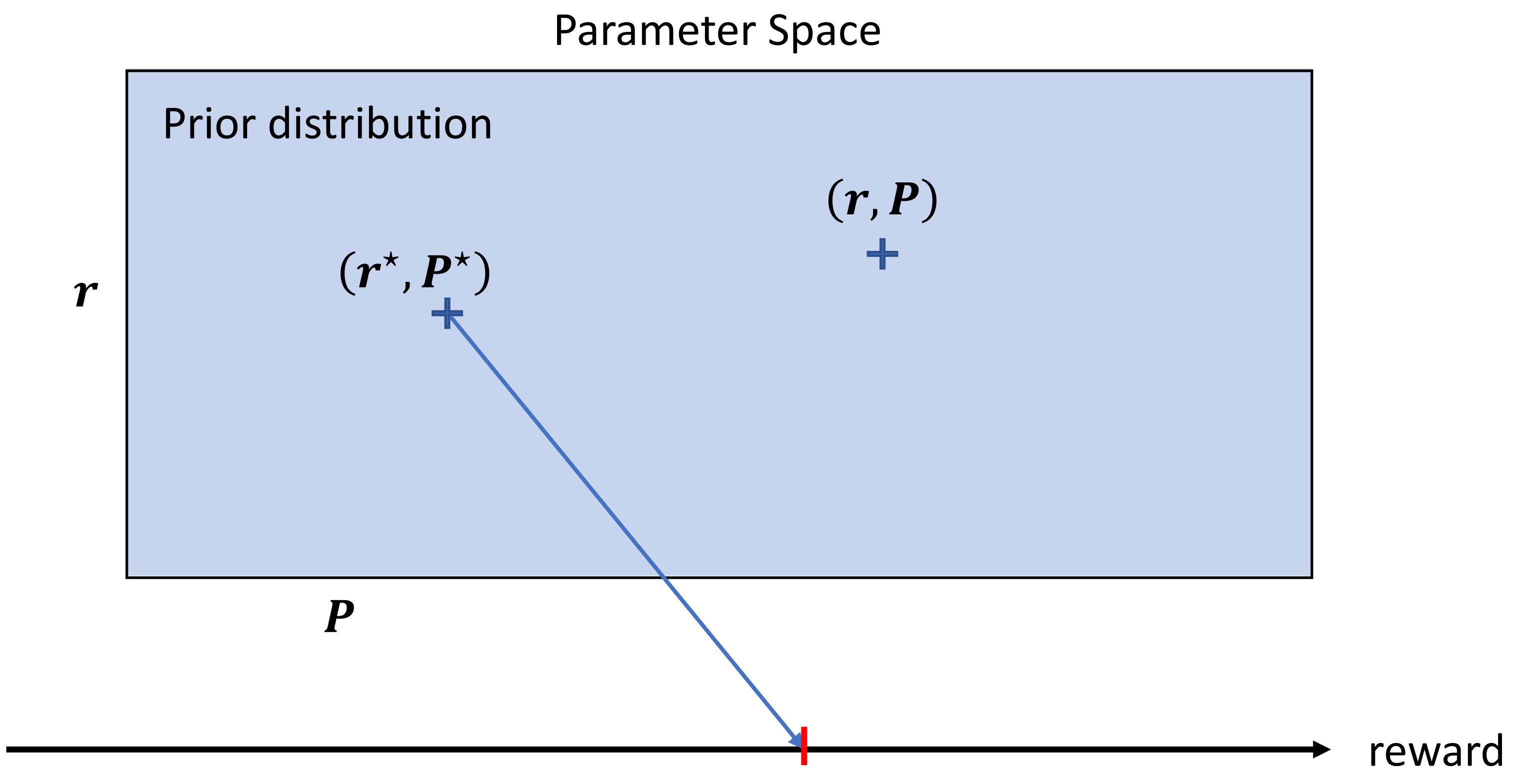
$\Rightarrow$ Build confidence set for each pair (r_i, P_i)
$\Rightarrow$ Choose $(\bar{r}_i, \bar{P}_i)_{i\in[n]}$ such that $(\bar{\boldsymbol{r}}, \bar{\boldsymbol{P}})$ is the **best possible environment**

Posterior Sampling

$$\mathbb{P}(\ H \mid O\ ) = \frac{\mathbb{P}(\ O \mid H\ )\ \mathbb{P}(H)}{\mathbb{P}(O)}$$

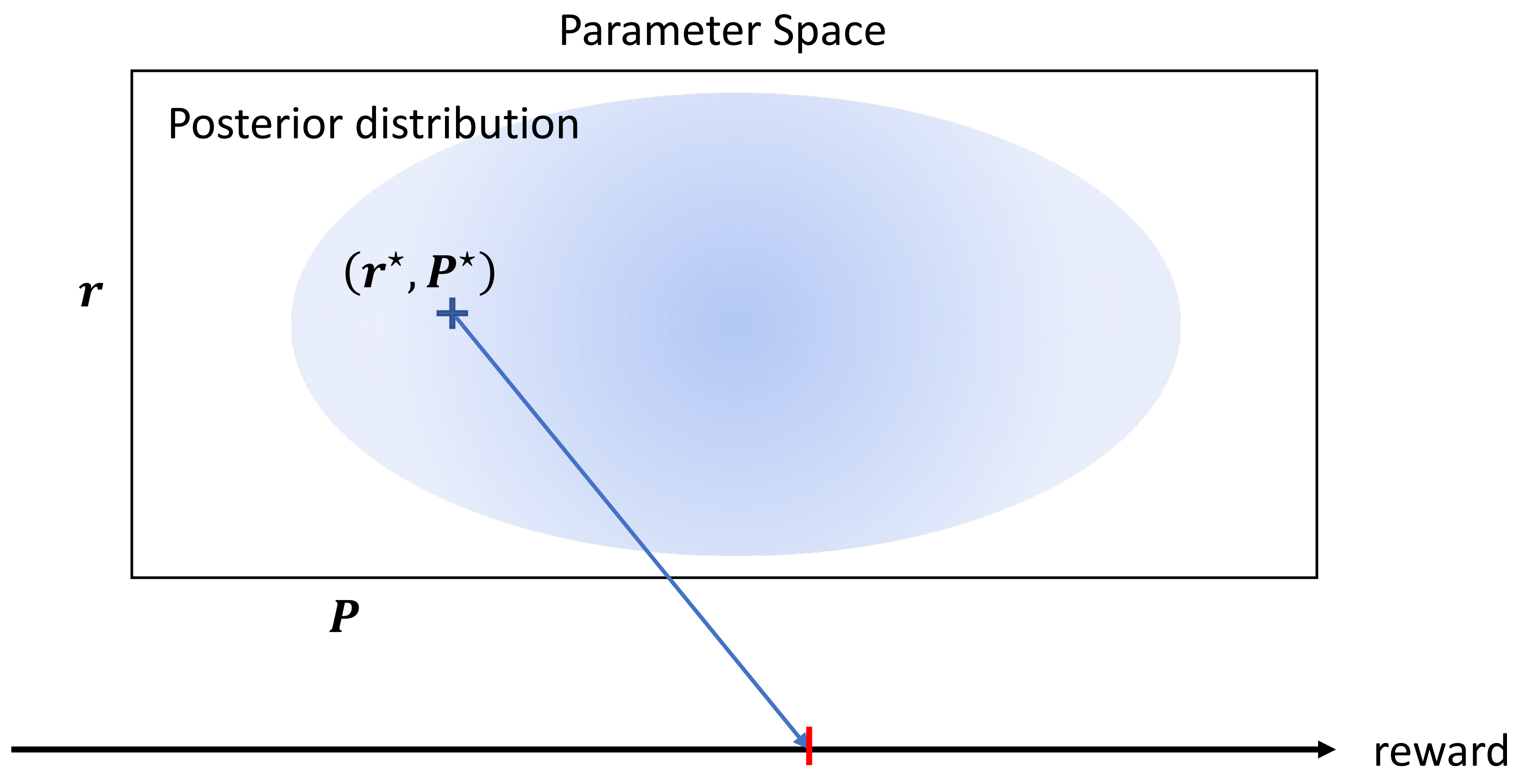
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the **environment** is sampled from a **certain distribution**.
Sample an environment from posterior distribution and **act greedily**.



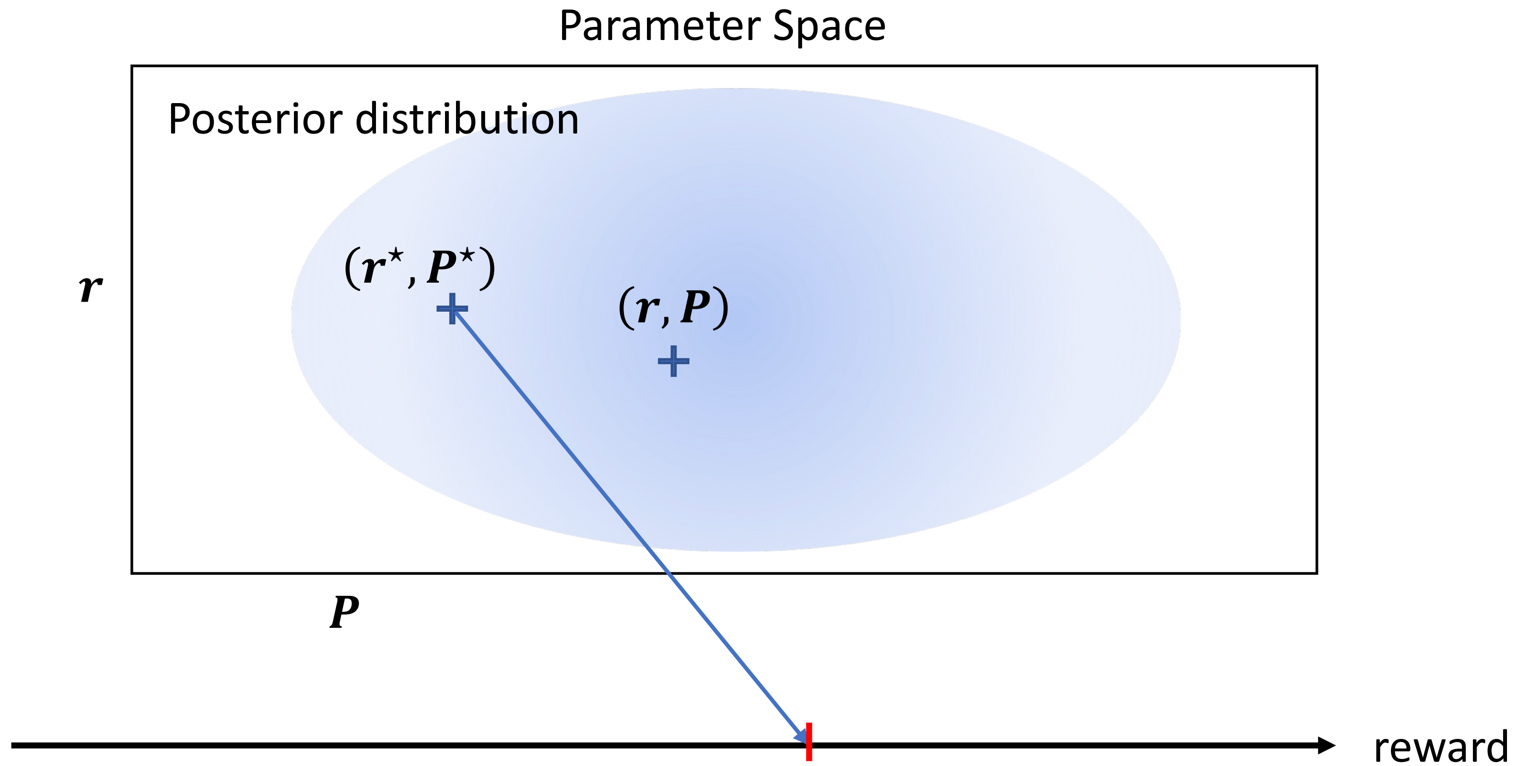
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

<u>Posterior Sampling</u>:
Hypothesis: the **environment** is sampled from a **certain distribution**.
Sample an environment from posterior distribution and **act greedily**.



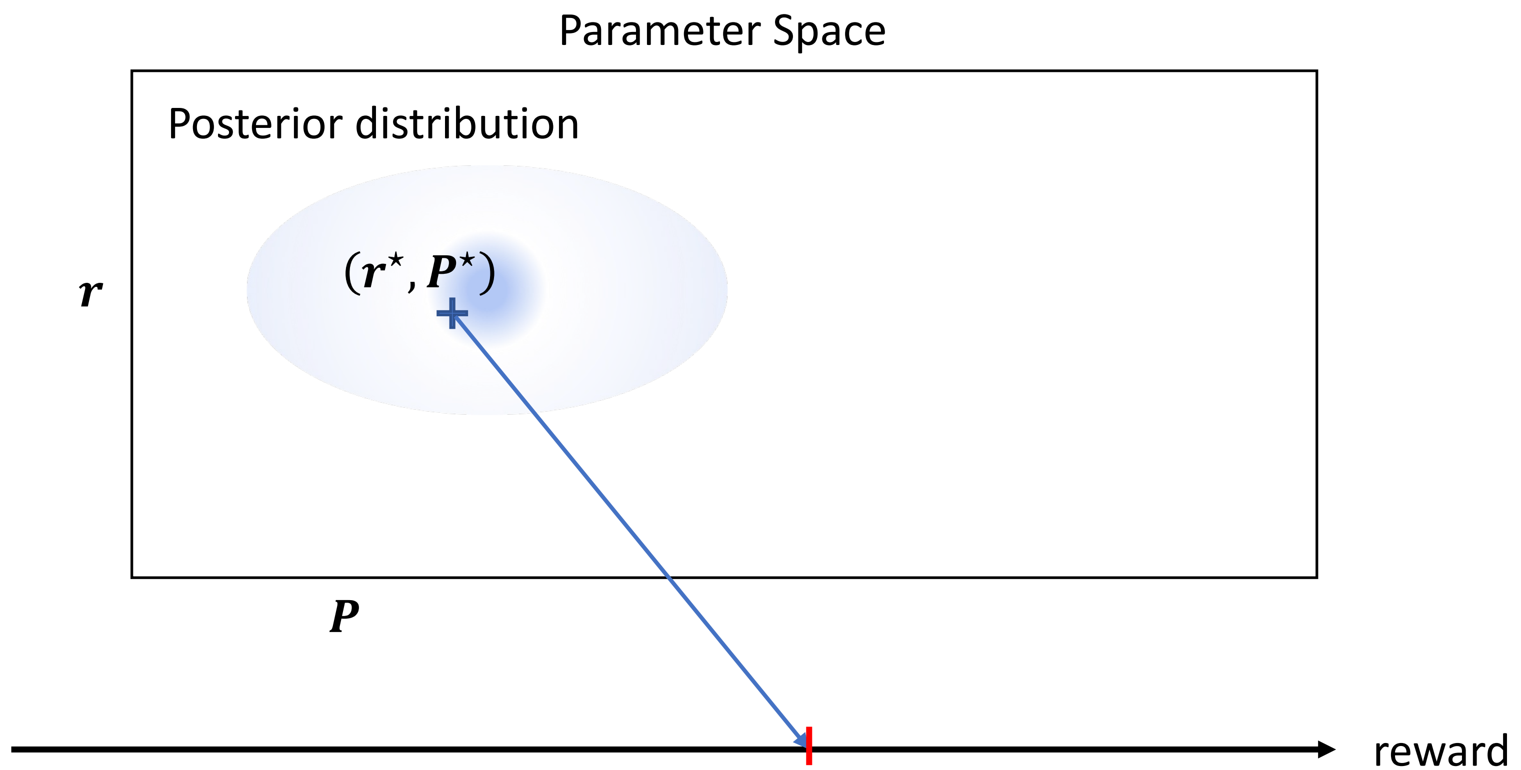
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

<u>Posterior Sampling</u>:
Hypothesis: the **environment** is sampled from a **certain distribution**.
Sample an environment from posterior distribution and **act greedily**.



Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the **environment** is sampled from a **certain distribution**.
Sample an environment from posterior distribution and **act greedily**.



Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the **environment** is sampled from a **certain distribution**.
Sample an environment from posterior distribution and **act greedily**.

More observations
=> posterior concentrates on the true environment
Exploitation

Few observations
=> uncertainty in the estimate
Exploration

$\Rightarrow$ Choose prior distribution ϕ_i for each arm i
$\Rightarrow$ Compute posterior $\phi_i(\cdot \mid O)$ and sample each pair $(r_i, P_i) \sim \phi_i(\cdot \mid O)$

Our Result

- Runtime:
  - When $(\boldsymbol{r}, \boldsymbol{P})$ is **given**, an optimal solution (Gittins index policy) can be computed in $(2/3)nS^3 + O(nS^2)$ [Gast et al., 2022]
  - The imaginary environment of both approaches is a Markovian bandit, Gittins index policy is **applicable**
- Learning Performance:
  - Keeping the estimate of $(r_i, P_i)_{i\in[n]}$ is linear in n

=> Both approaches are scalable.

Conclusion

- We show how the Optimism and Posterior Sampling approaches can be used to learn Markovian bandit problem.
- We conclude that both approaches are scalable in the number of arms.

Future Work

- What if the non-active arms also change state?