

# **Exploration and Exploitation (Bandits)**

# Last Time

- What is Reinforcement Learning?
- What are the main challenges in Reinforcement Learning?
- How do we categorize RL approaches?

# Last Time

First RL Algorithm:

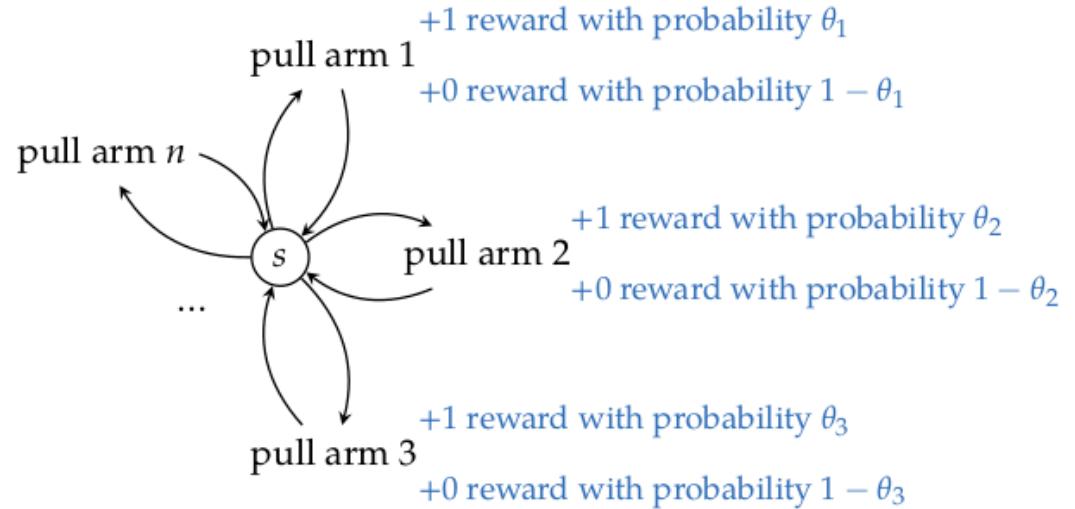
Tabular Maximum Likelihood Model-Based Reinforcement Learning

loop  
    choose action  $a$   
    gain experience  
    estimate  $T, R$   
    solve MDP with  $T, R$

# Guiding Questions

- What are the best ways to trade off Exploration and Exploitation?

# Bandits



- Bernoulli Bandit with parameters  $\theta$
- $\theta^* \equiv \max \theta$

*“According to Peter Whittle, “efforts to solve [bandit problems] so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany as the ultimate instrument of intellectual sabotage.”*

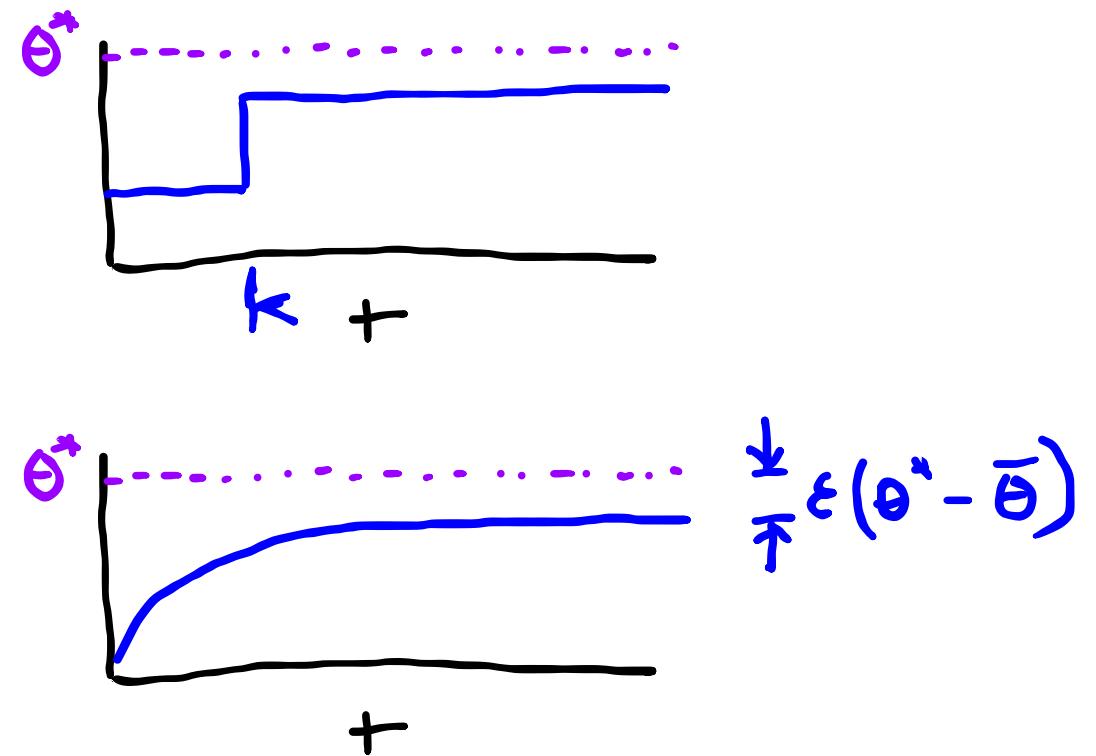
# Greedy Strategy

$$\rho_a = \frac{\text{number of wins}+1}{\text{number of tries}+1}$$

Choose  $\operatorname{argmax}_a \rho_a$

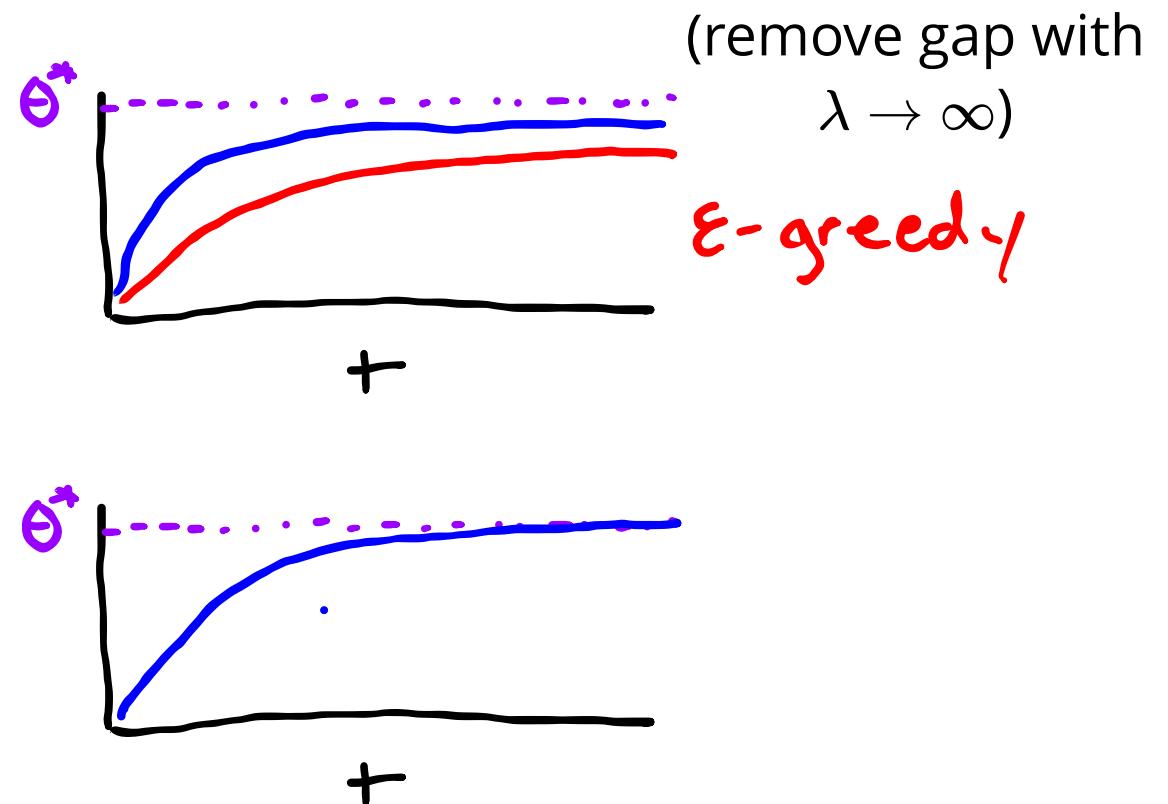
# Undirected Strategies

- Explore then Commit  
Choose  $a$  randomly for  $k$  steps  
Then choose  $\operatorname{argmax}_a \rho_a$
- $\epsilon$  - greedy  
With probability  $\epsilon$ , choose randomly  
Otherwise choose  $\operatorname{argmax}_a \rho_a$



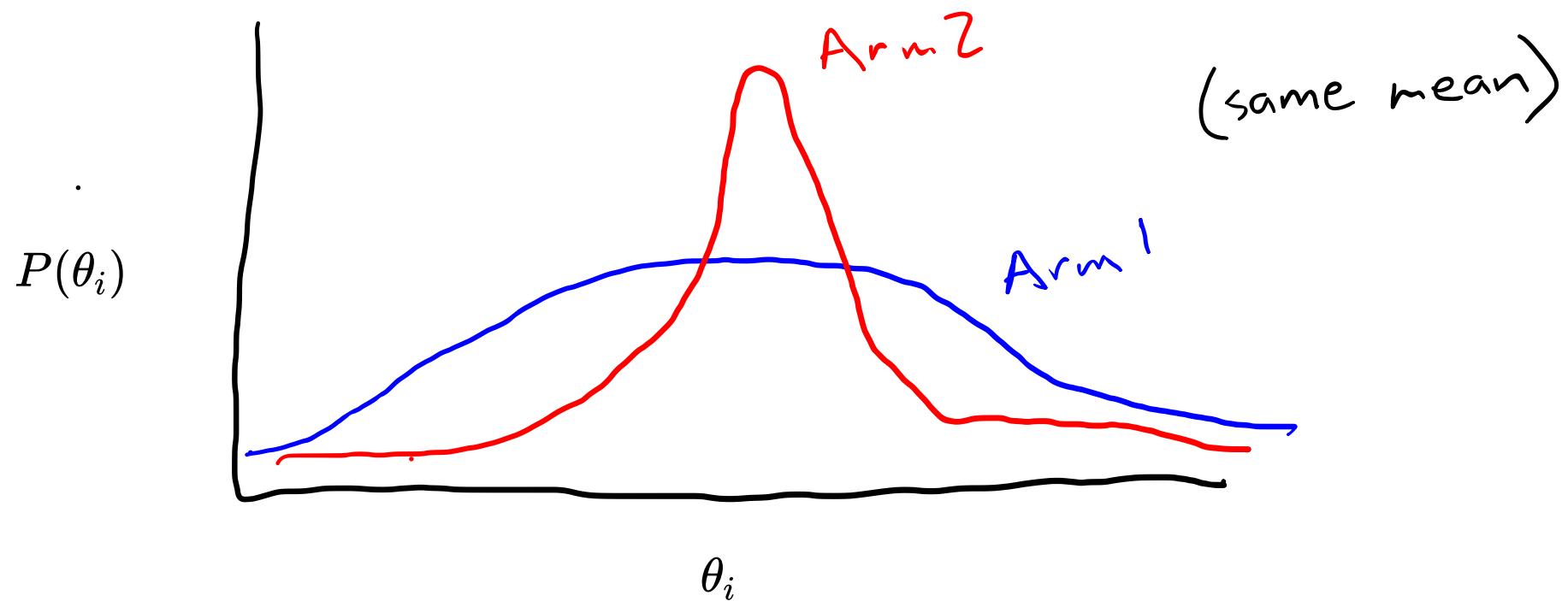
# Directed Strategies

- Softmax  
Choose  $a$  with probability proportional to  $e^{\lambda \rho_a}$
- Upper Confidence Bound (UCB)  
Choose  $\underset{a}{\operatorname{argmax}} \rho_a + c \sqrt{\frac{\log N}{N(a)}}$



# Break

Discuss with your neighbor: Suppose you have the following *belief* about the parameters  $\theta$ . Which arm should you choose to pull next?

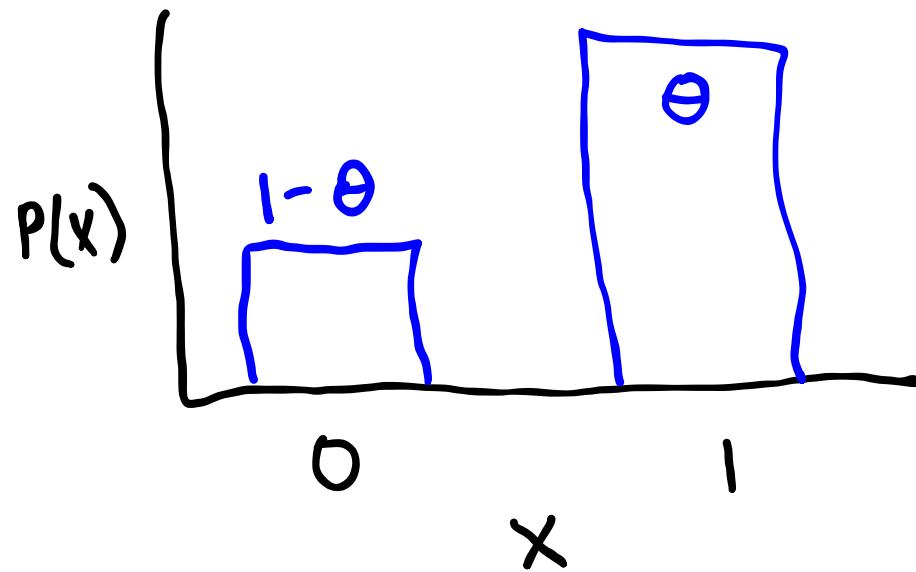


# Bayesian Estimation

Bernoulli Distribution

$\text{Bernoulli}(\theta)$

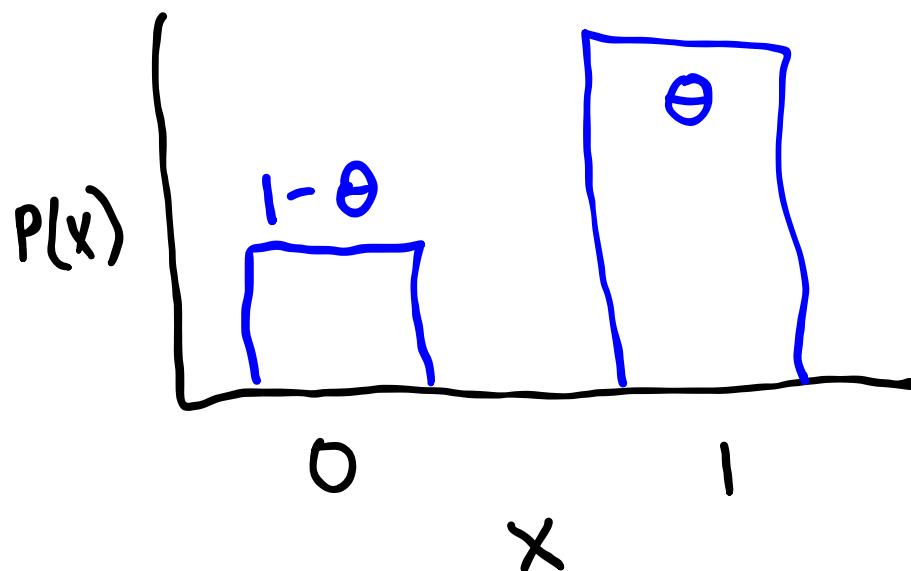
Discussion: Given that I have received  $w$  wins and  $l$  losses, what should my belief (probability distribution) about  $\theta$  look like?



# Bayesian Estimation

Bernoulli Distribution

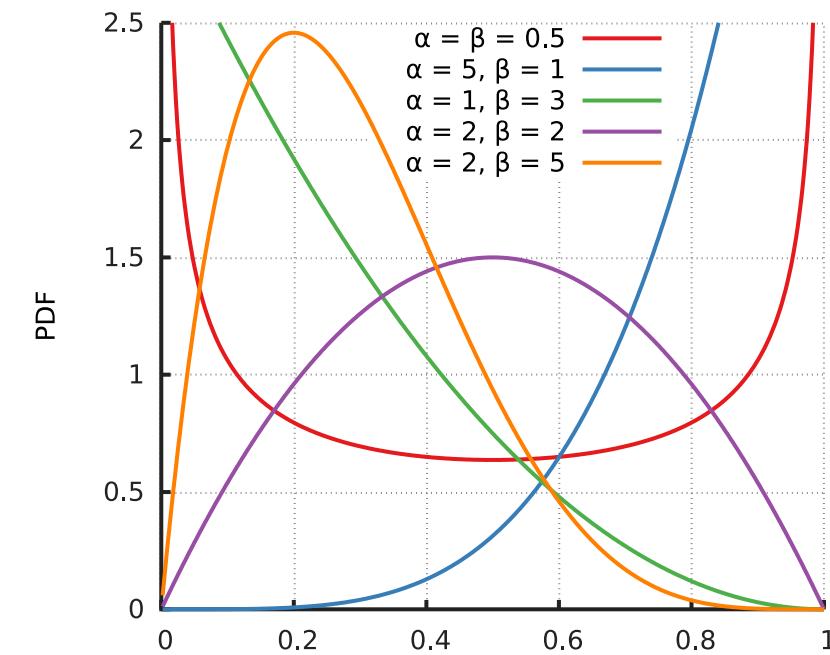
$\text{Bernoulli}(\theta)$



Beta Distribution

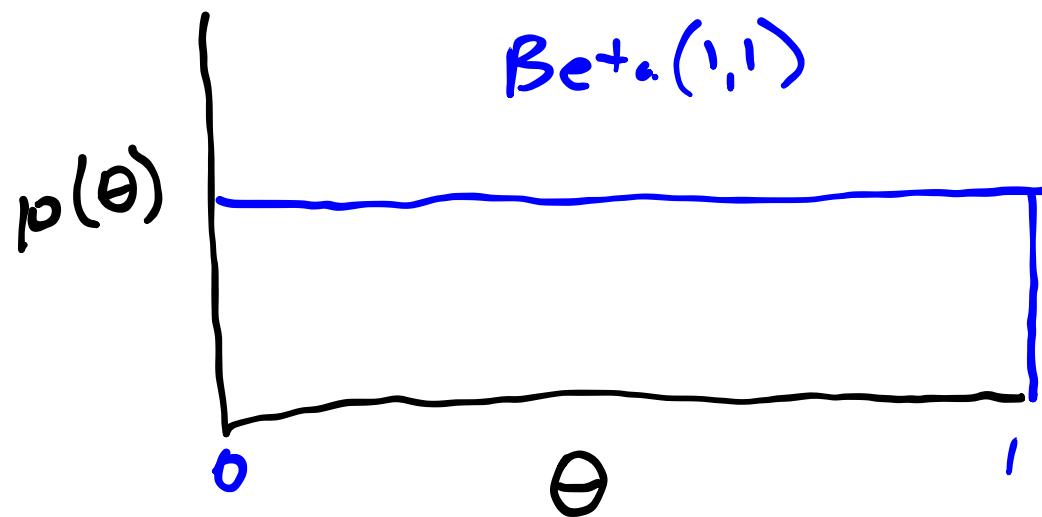
(distribution over Bernoulli distributions)

$\text{Beta}(\alpha, \beta)$

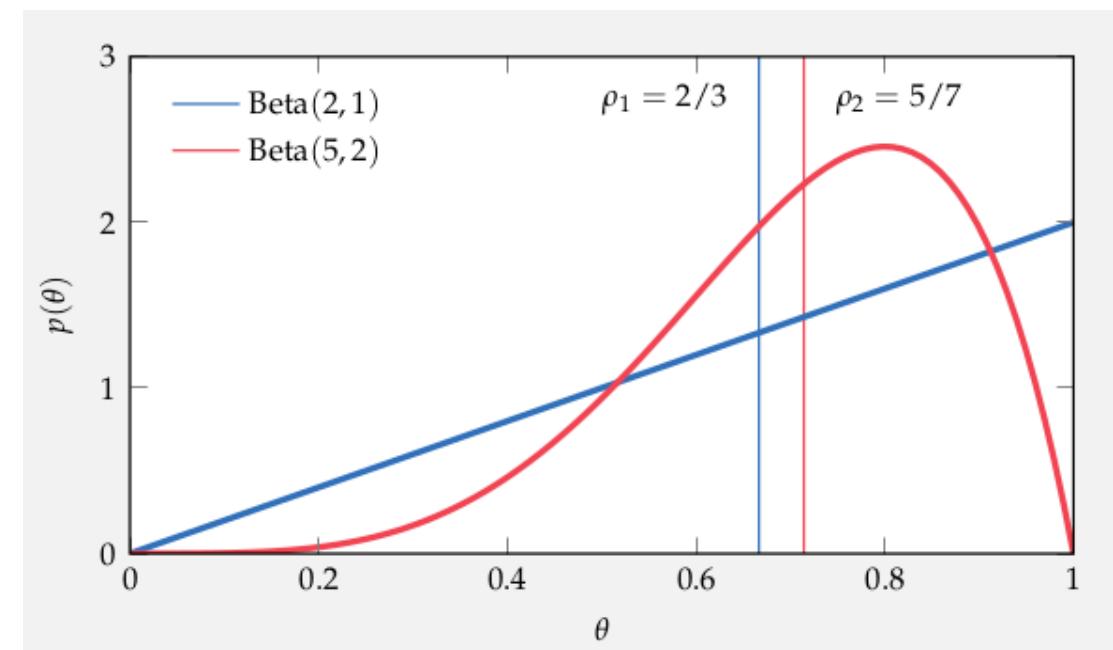


# Bayesian Estimation

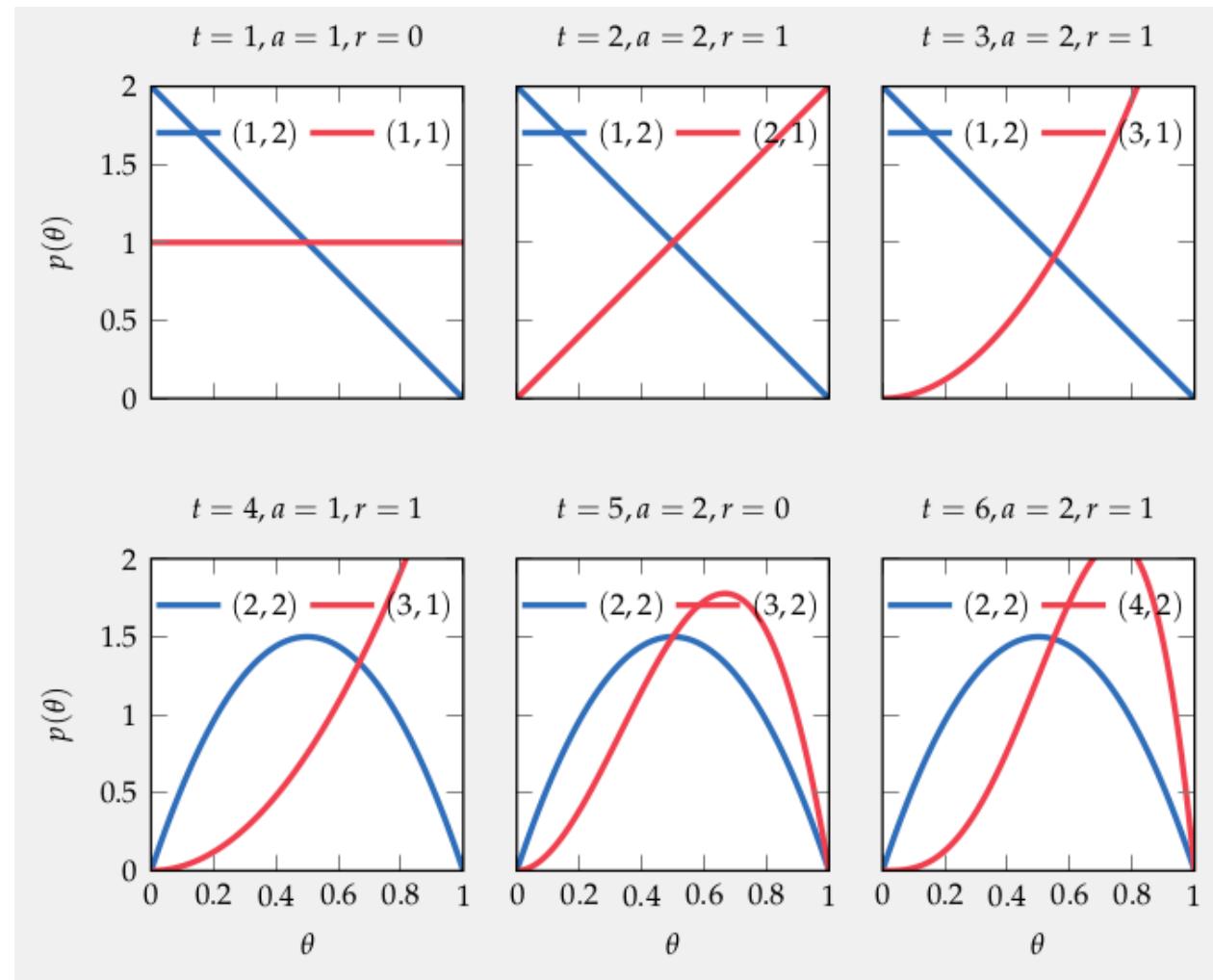
Given a  $\text{Beta}(1, 1)$  prior distribution



The posterior distribution of  $\theta$  is  
 $\text{Beta}(w + 1, l + 1)$



# Bayesian Estimation



$t$  = time

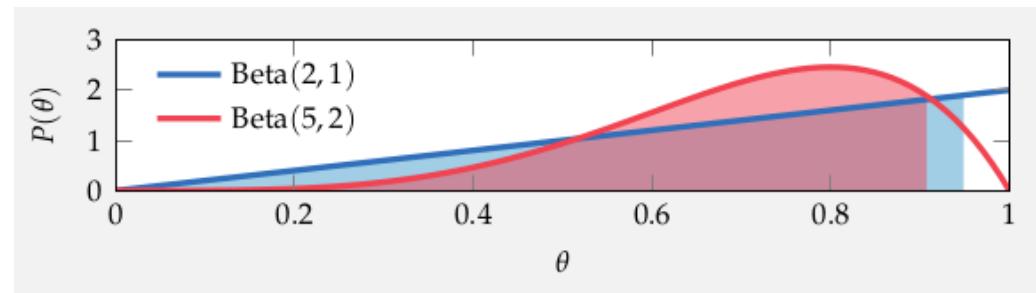
$a$  = arm pulled

$r$  = reward

# Bayesian Bandit Algorithms

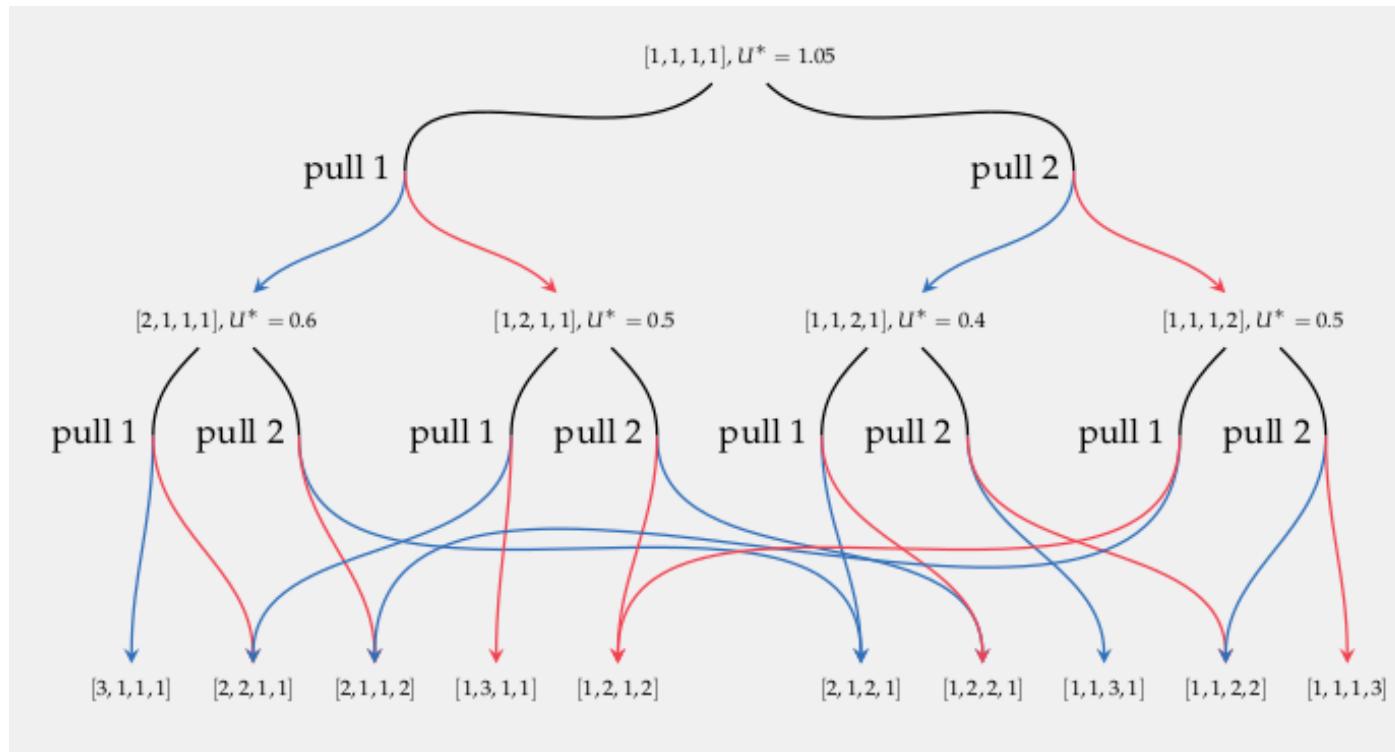
- Quantile Selection  
Choose  $a$  for which the  $\alpha$  quantile of  $b(\theta)$  is highest

$$\alpha = 0.9$$



- Thompson Sampling  
Sample  $\hat{\theta}$   
Choose  $\operatorname{argmax}_a \hat{\theta}_a$

# Optimal Algorithm - Dynamic Programming



# Review

# Guiding Questions

- What are the best ways to trade off Exploration and Exploitation