Exploration and Exploitation (Bandits)

Last Time

$(S, A, \cancel{T}, \cancel{R}, \gamma)$

- What is Reinforcement Learning?
- What are the main challenges in Reinforcement Learning?

  - Exploration + Exploitation
  - Credit Assignment
  - Generalization

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:

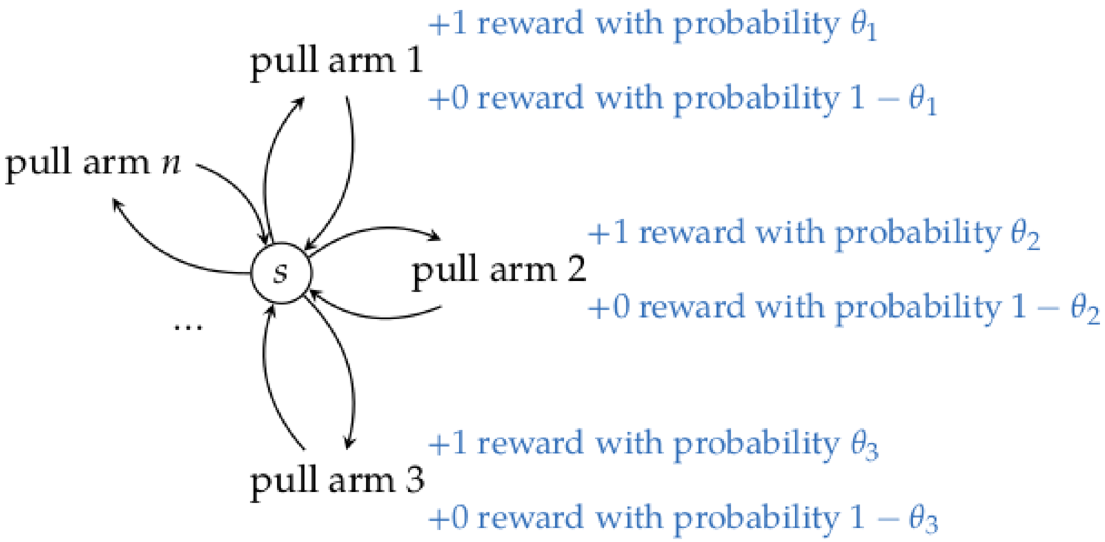
Tabular Maximum Likelihood Model-Based Reinforcement Learning

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 θ
- $\theta^* \equiv \max \theta$

"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."

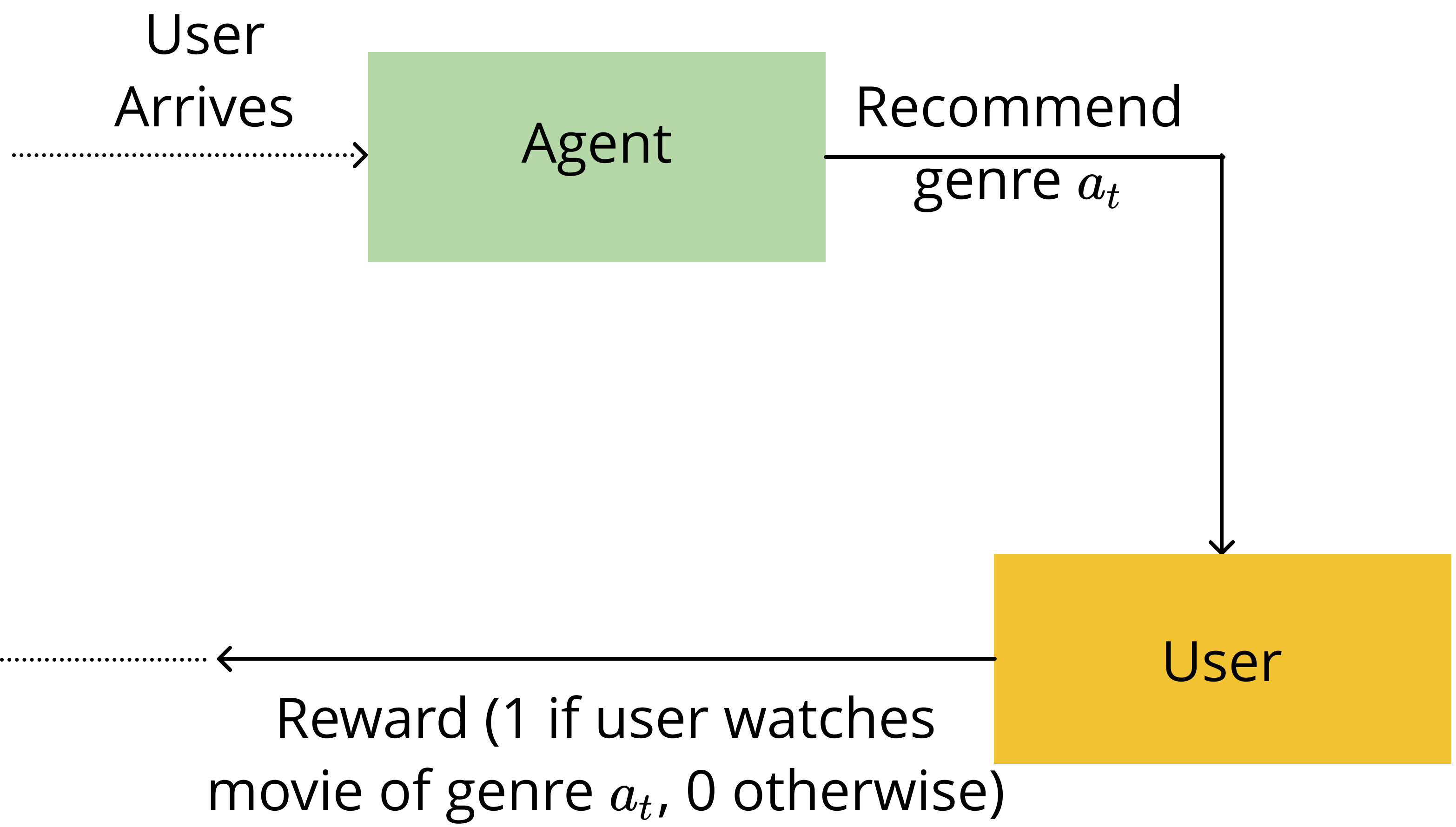
Bandits in the wild

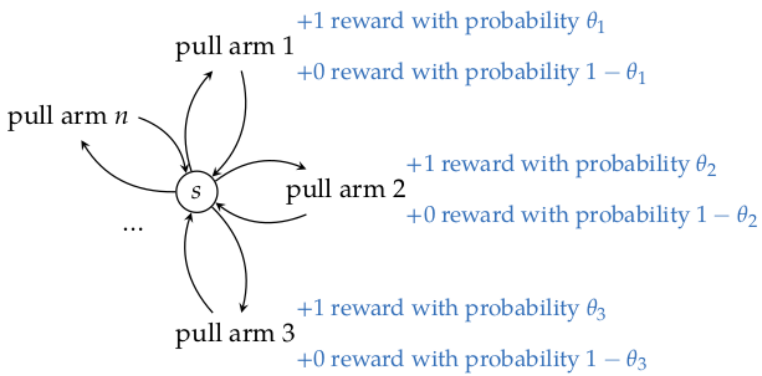
- Recommender systems (food, movies, activities)
- Allocation of clinical trials
- Satellite network optimization
- Spacecraft scheduling
- Motion planning
- Aircraft Part Maintenance

Recommendation System

- Recommend different genre of movies (e.g., action, adventure, comedy, romance, animation)
- User arrives at random
- Agent picks a genre to recommend to user
- User watches a movie
- Objective: maximize movies watched in recommended genre

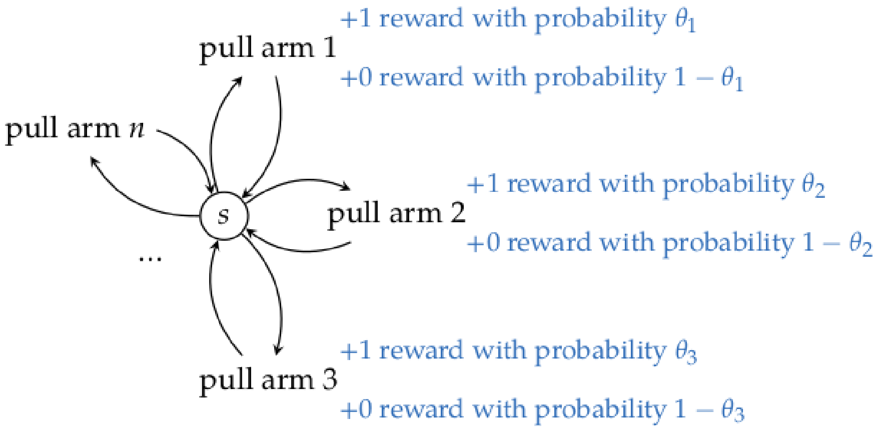
Recommender System as MAB



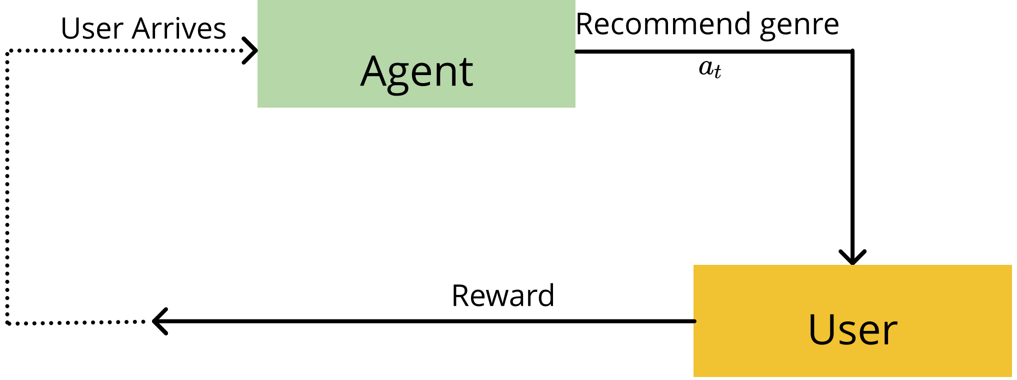


Recommender System as MAB

- θ_{a_t} is Bernoulli distribution
- $r_t \sim Bernoulli(\theta_{a_t})$ is a realization of the Bernoulli of genre a_t



Maximize sum of reward $\mathbb{E}[\sum_{t=1}^{n} r_t]$ = $\max \theta$



Bandits: Exploration/Exploitation

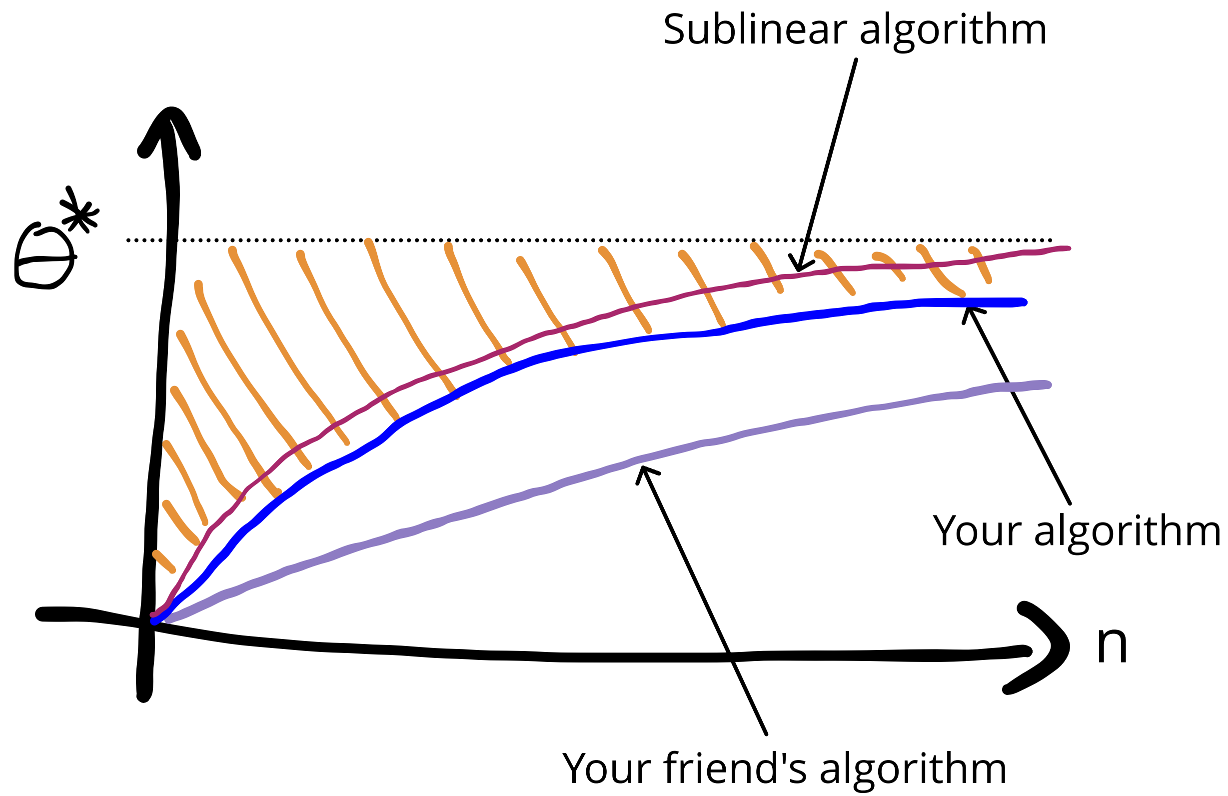
- Problem 1: Environment does not reveal reward of actions not selected
  - Agent should gain information by repeatedly selecting different actions => exploration
- Problem 2: Whenever agent selects a bad action, suffers regret
  - Agent should reduce regret by repeatedly selecting the best action => exploitation

Regret - how quickly to "warm up"

R(n) = $n\theta^* - \sum_{t=1}^{N} r_t$

Regret growth as n increases

- Worst case possible: O(n)
- Better: o(n): $\frac{R_n}{n} \to 0$
- Typical rates:
  - O(log N)
  - O($\sqrt{N}$)



Exploration Strategies

- Greedy
- Explore then Commit
- Epsilon-greedy
- Softmax
- Upper Confidence Bound (UCB)
- Bayesian Methods
- Dynamic Programming

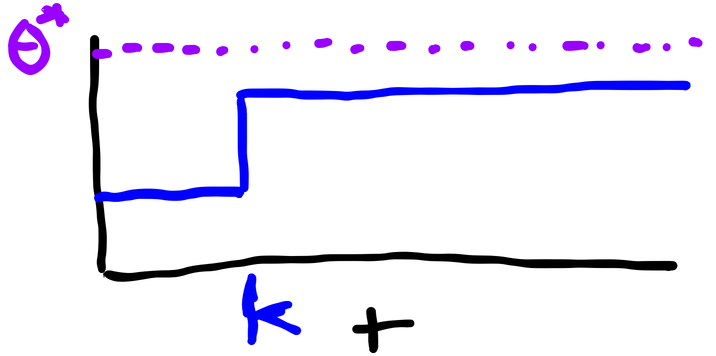
Greedy Strategy

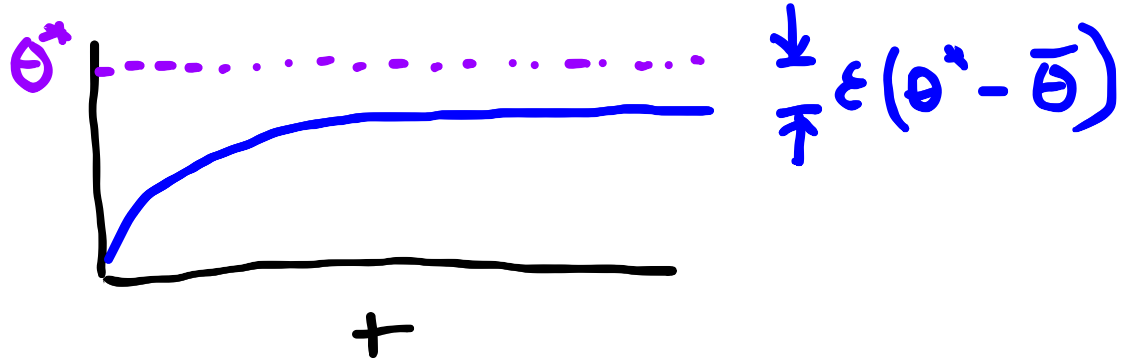
$$\rho_a = \frac{\text{number of wins}+1}{\text{number of tries}+1}$$

Choose $\underset{a}{\operatorname{argmax}}\, \rho_a$

Undirected Strategies

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



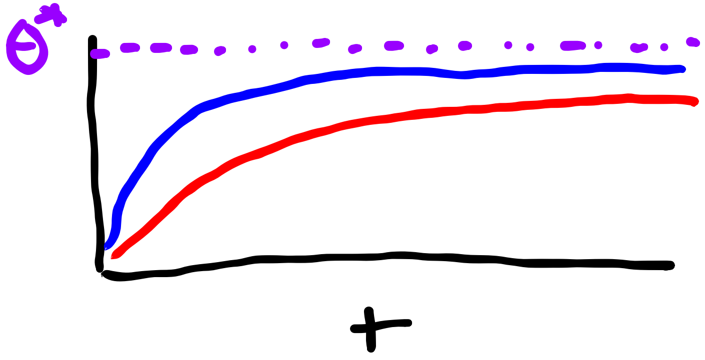


Directed Strategies

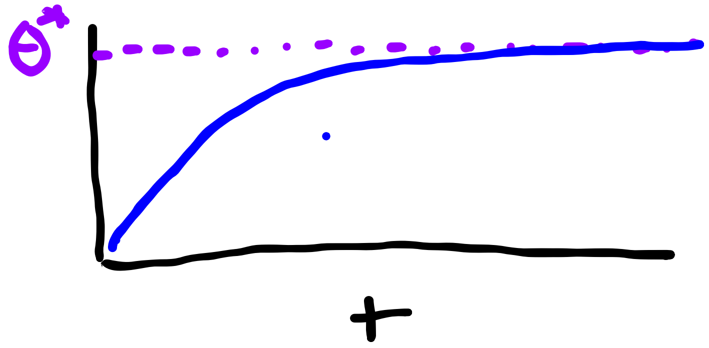
- Softmax
  Choose a with probability proportional to $e^{\lambda \rho_a}$

- Upper Confidence Bound (UCB)
  Choose $\underset{a}{\operatorname{argmax}}\, \rho_a + c\sqrt{\frac{\log N}{N(a)}}$

(remove gap with $\lambda \to \infty$)

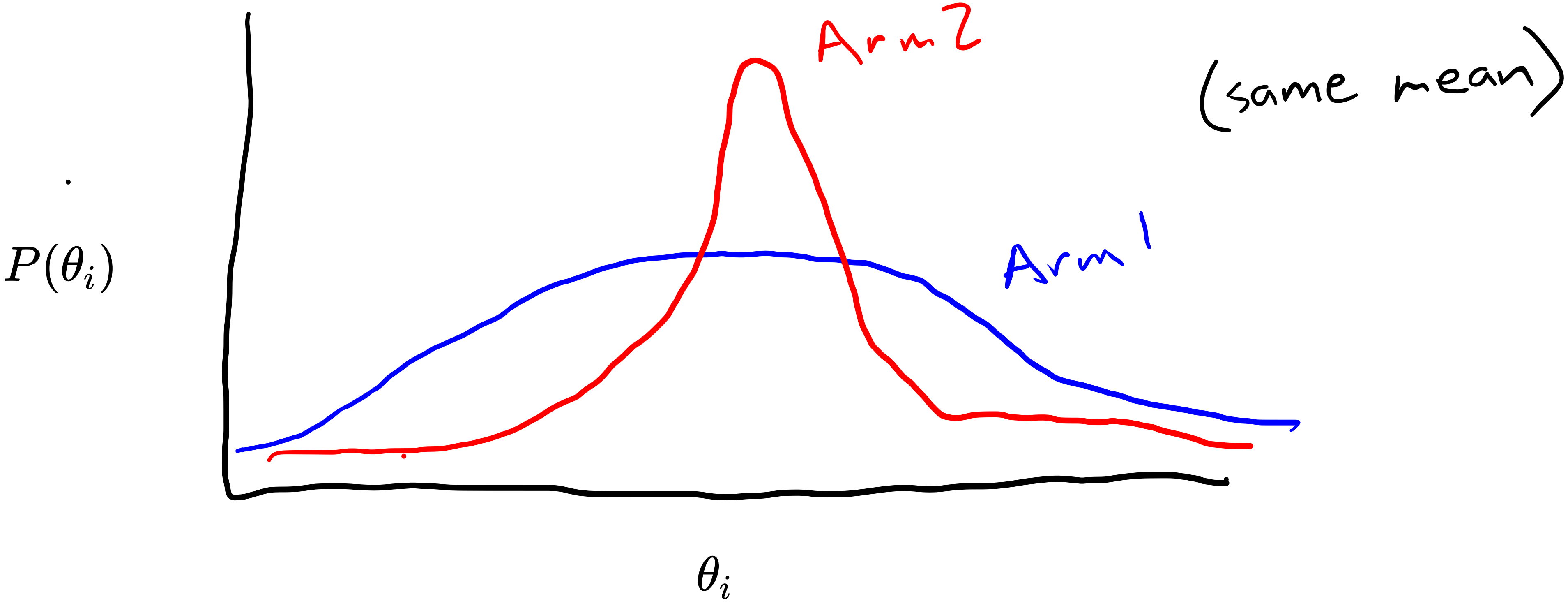


ε-greedy



Break

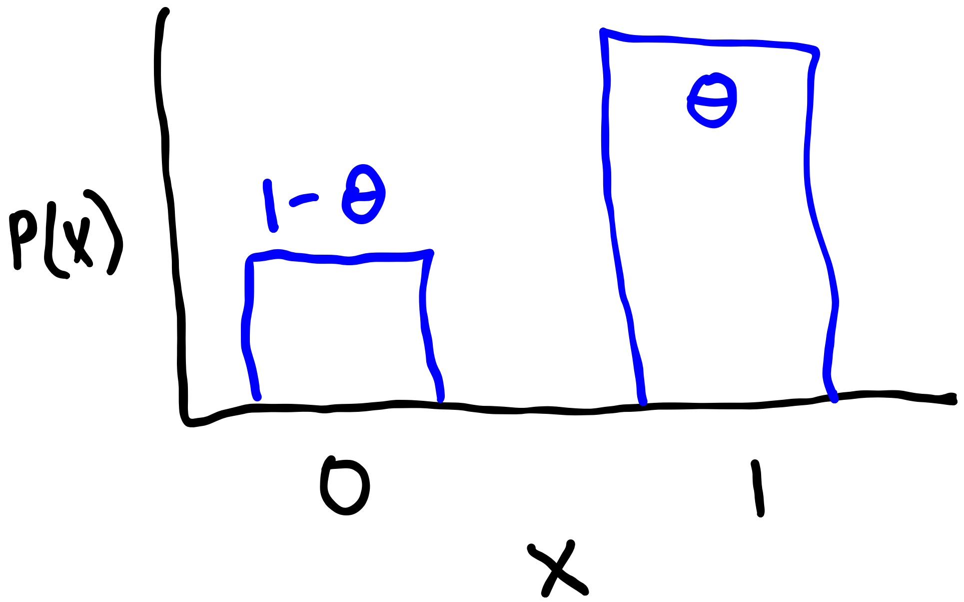
Discuss with your neighbor: Suppose you have the following *belief* about the parameters θ. Which arm should you choose to pull next?



Bayesian Estimation

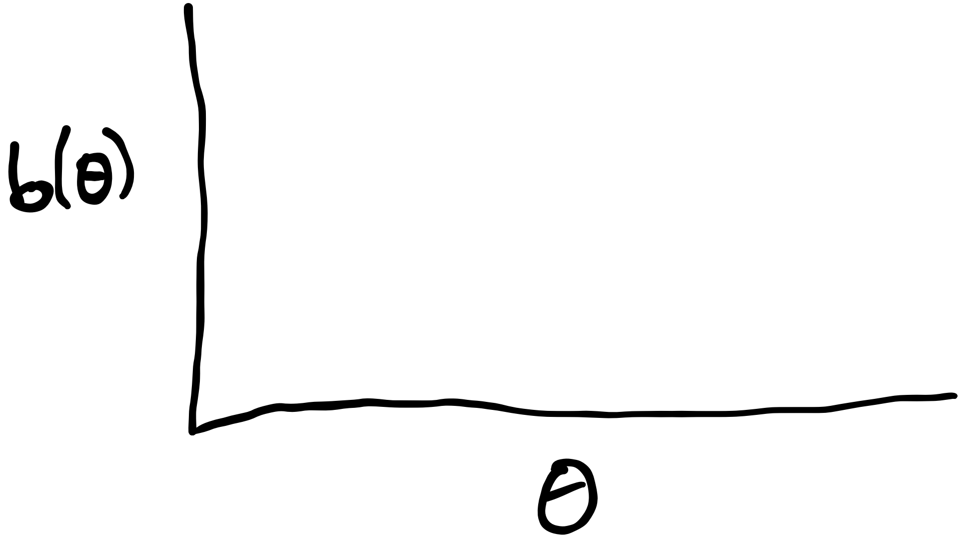
Bernoulli Distribution

$$\text{Bernoulli}(\theta)$$



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

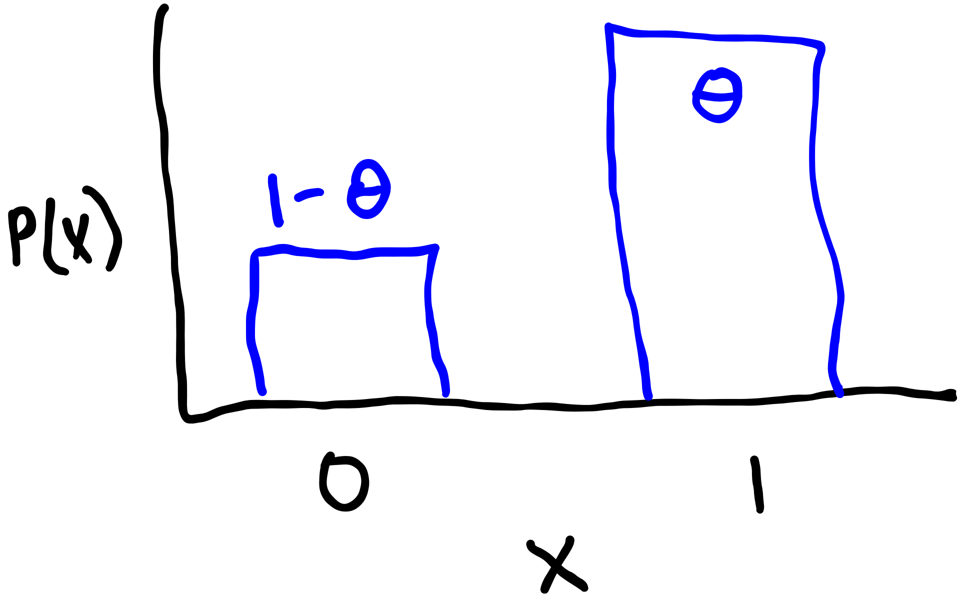
$w=4, l=1$



Bayesian Estimation

Bernoulli Distribution

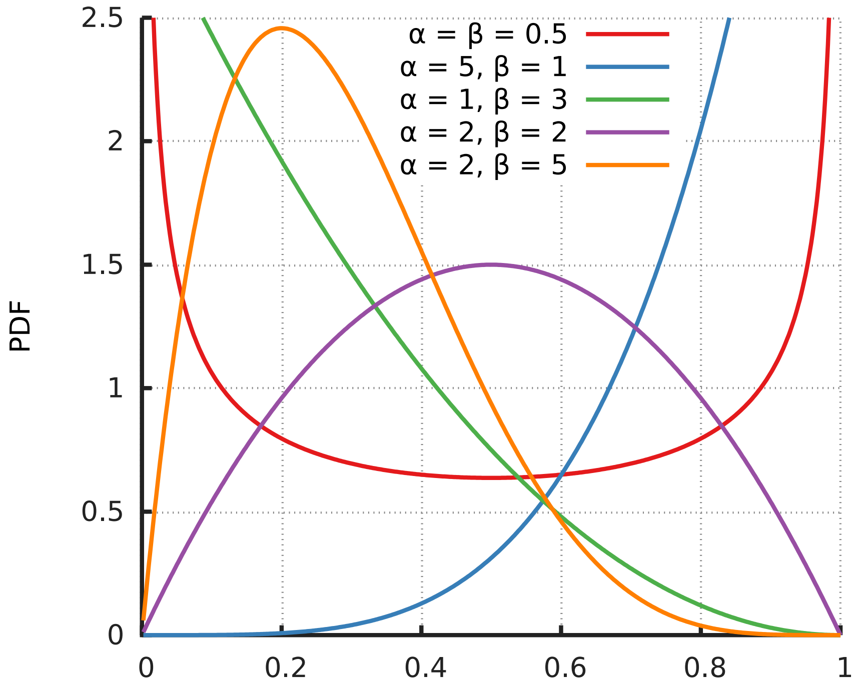
$$\text{Bernoulli}(\theta)$$



Beta Distribution

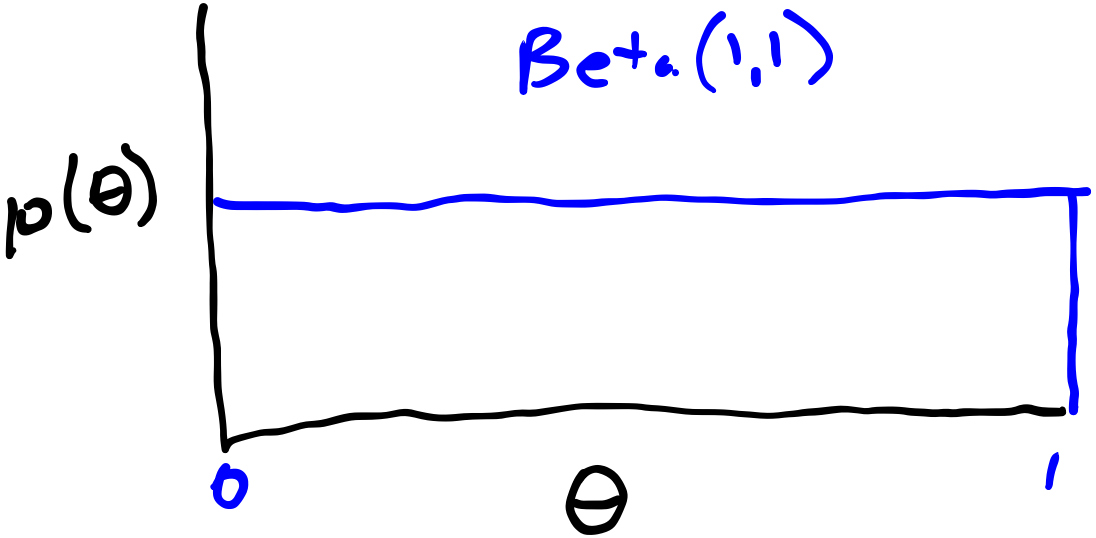
(distribution over Bernoulli distributions)

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



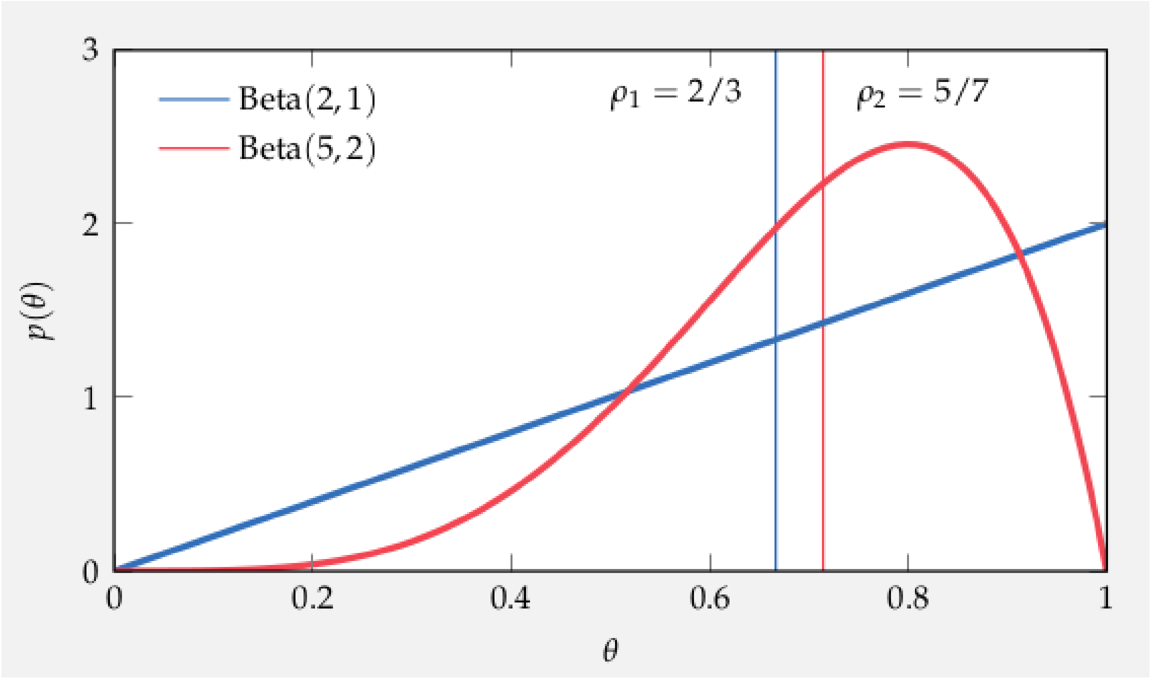
Bayesian Estimation

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

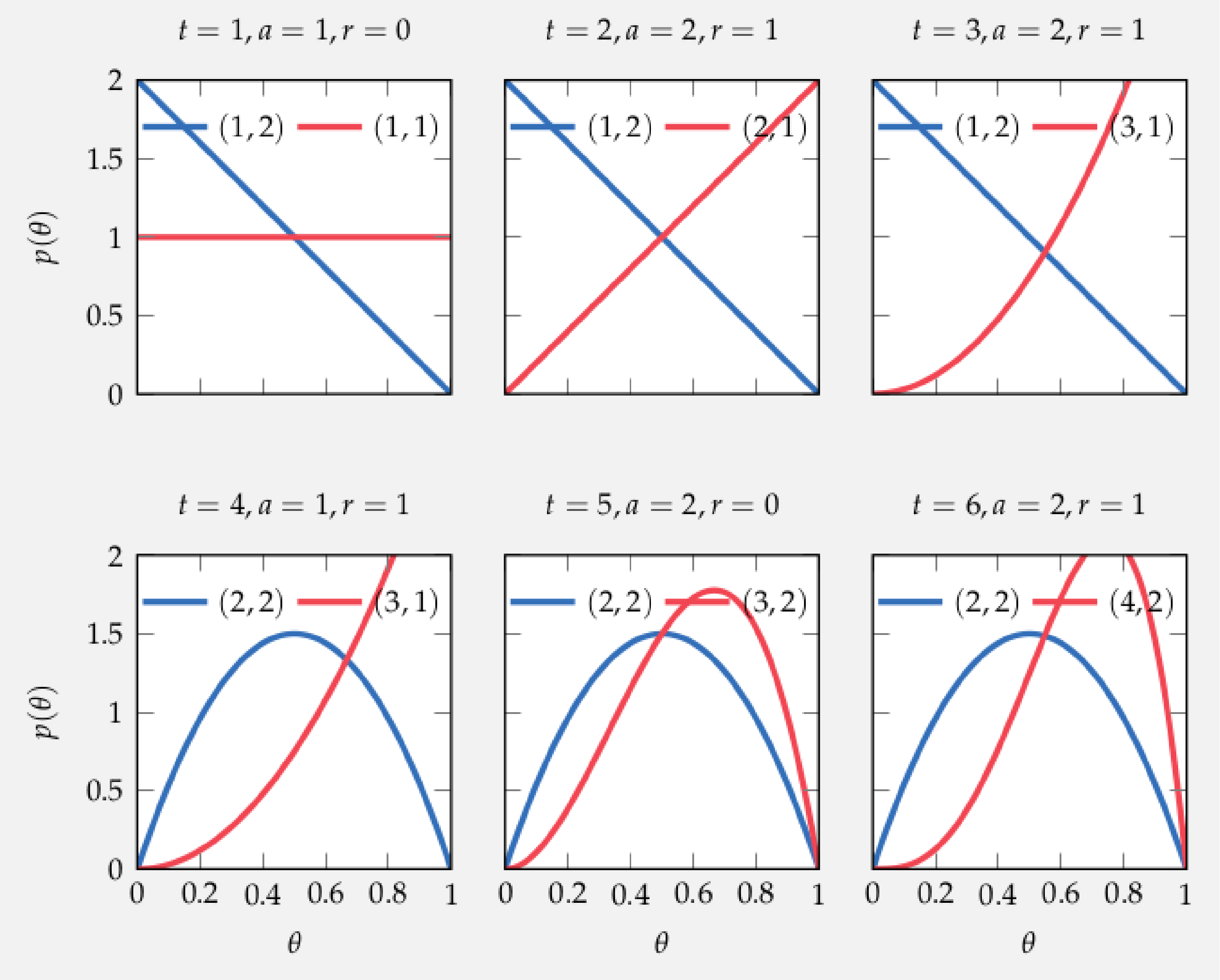


The posterior distribution of θ is

$$\mathrm{Beta}(w+1, l+1)$$



Bayesian Estimation



t = time

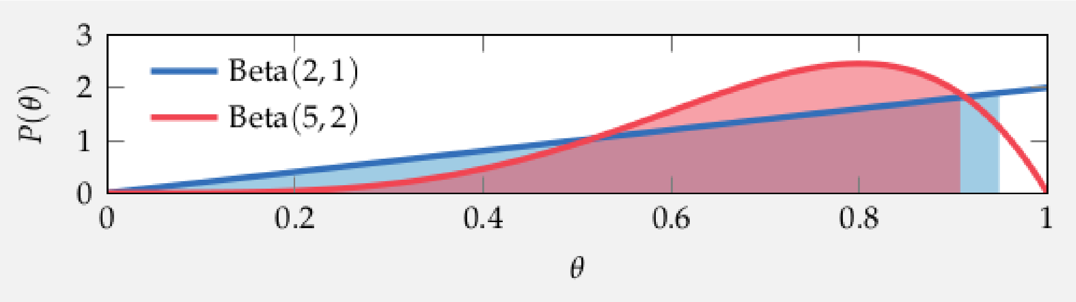
a = arm pulled

r = reward

Bayesian Bandit Algorithms

higher α = more optimistic

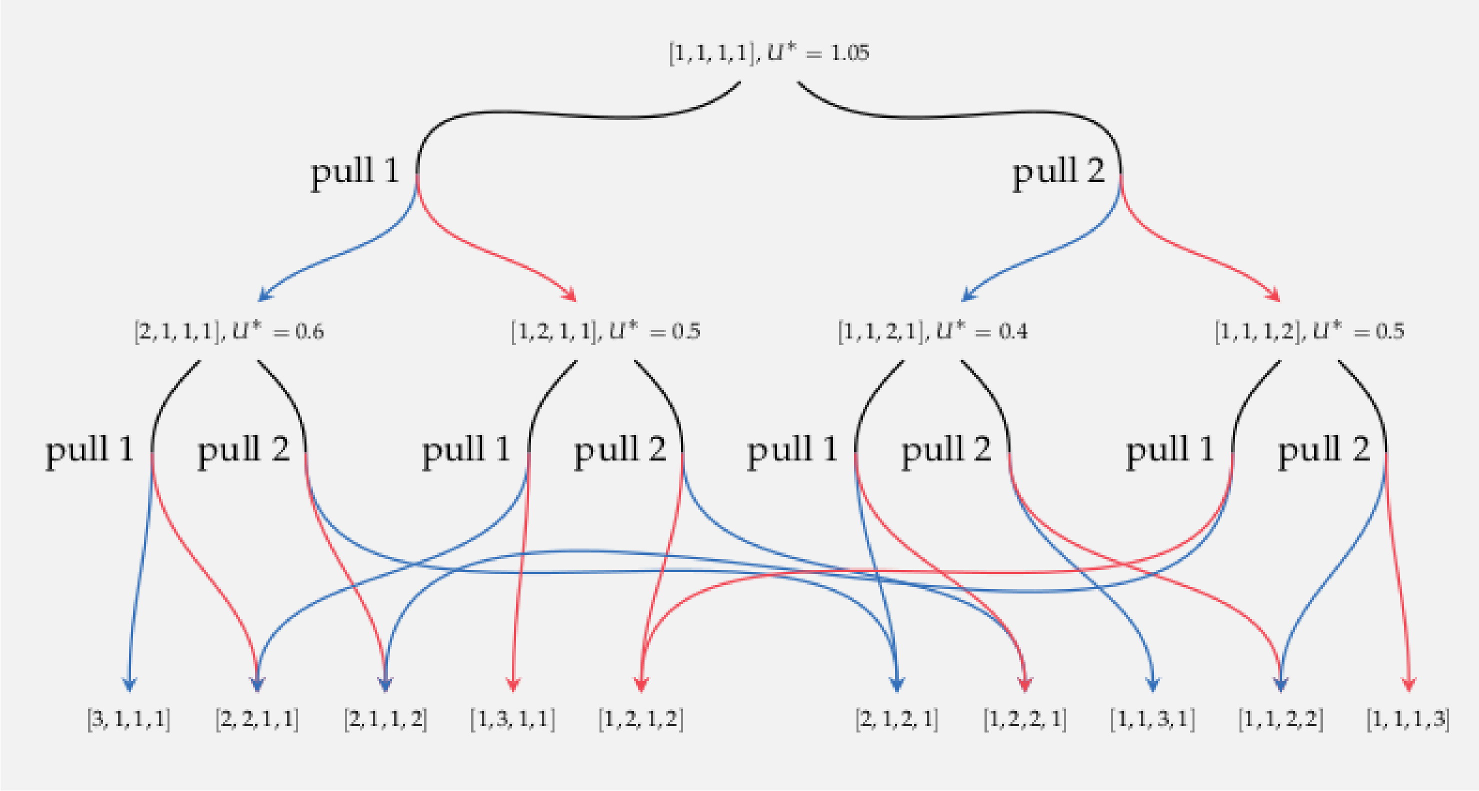
$\alpha = 0.9$



- Quantile Selection
  Choose a for which the α quantile of $p(\theta|data)$ is highest

- Thompson Sampling
  Sample $\hat{\theta}$ from $p(\theta|data)$
  Choose $\underset{a}{\operatorname{argmax}}\, \hat{\theta}_a$

Optimal Algorithm - Dynamic Programming



Easier to Implement

Faster

Review

Algorithm	**Optimal in Limit**	**Regret**
Greedy	No	O(N)
Epsilon-greedy	$\epsilon \to 0$	O(N)
Explore-commit	$k \to \infty$	O(N)
Softmax	$\lambda \to \infty$	O(N)
UCB	Yes	O(log(N))
Quantile Selection	Yes	O(log(N))
Thompson Sampling	Yes	O(log(N))
Dynamic Programming	Yes	

Less Regret

Guiding Questions

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