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:

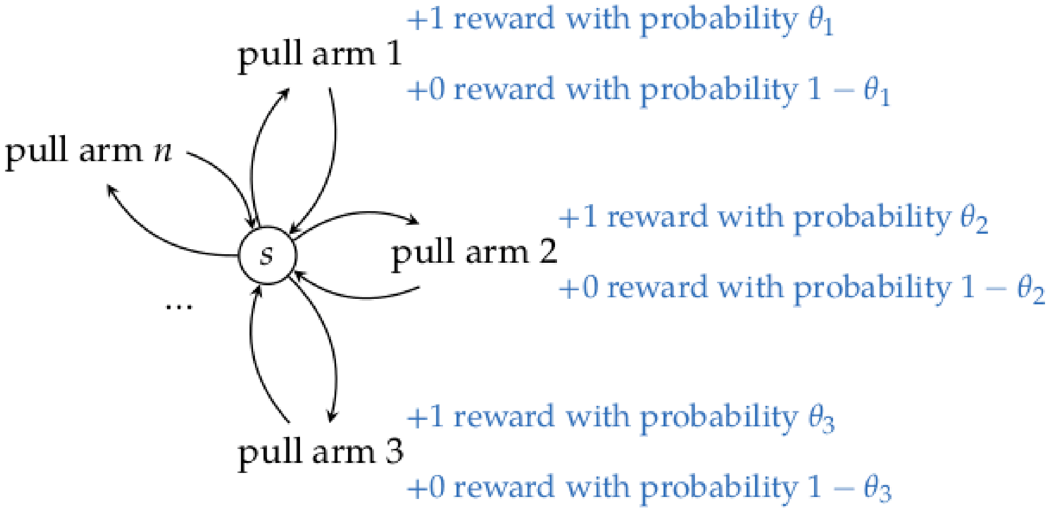
Tabular Maximum Likelihood Model-Based Reinforcement Learning

```
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



+1 reward with probability θ_1
pull arm 1
+0 reward with probability $1-\theta_1$
pull arm n
+1 reward with probability θ_2
s
pull arm 2
+0 reward with probability $1-\theta_2$
...
+1 reward with probability θ_3
pull arm 3
+0 reward with probability $1-\theta_3$

- 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."*

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

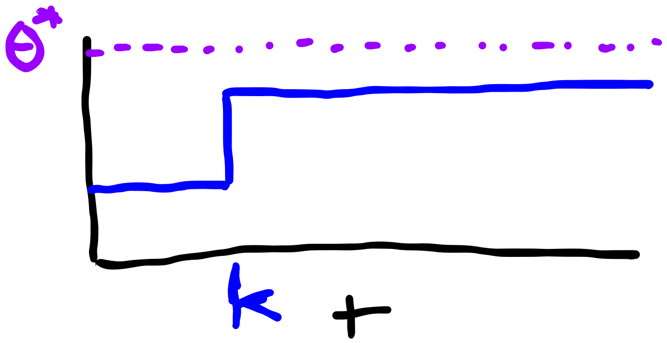
  Choose a randomly for k steps

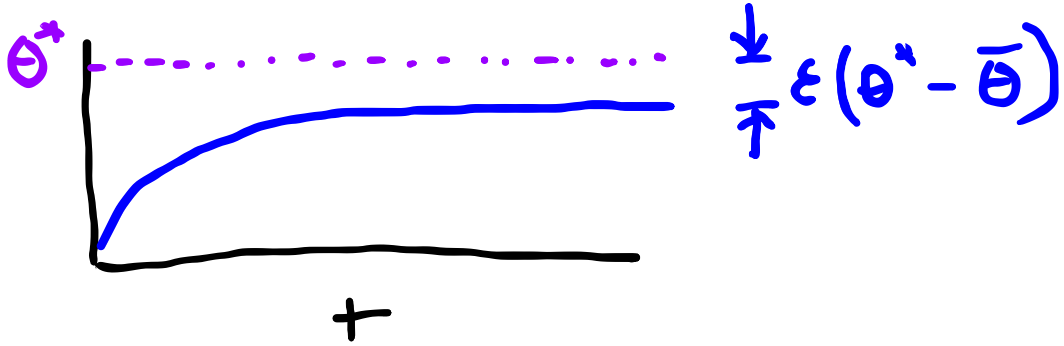
  Then choose $\operatorname*{argmax}_{a} \rho_a$

- ϵ - greedy

  With probability ϵ, choose randomly

  Otherwise choose $\operatorname*{argmax}_{a} \rho_a$

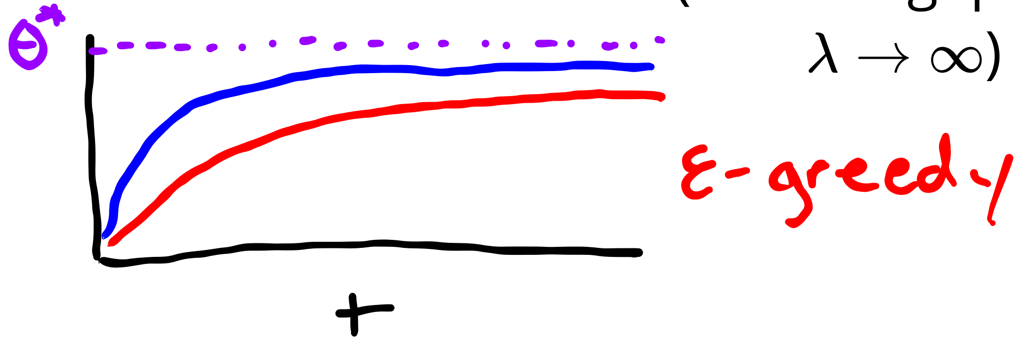




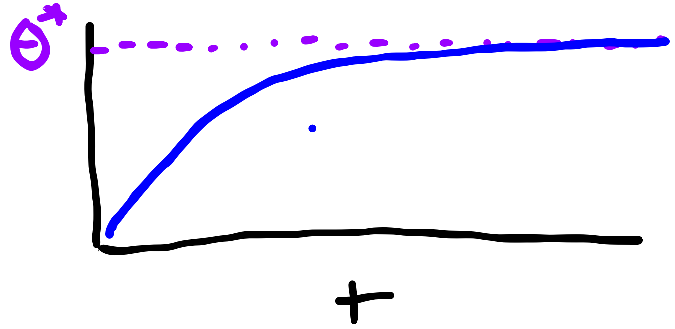
Directed Strategies

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

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

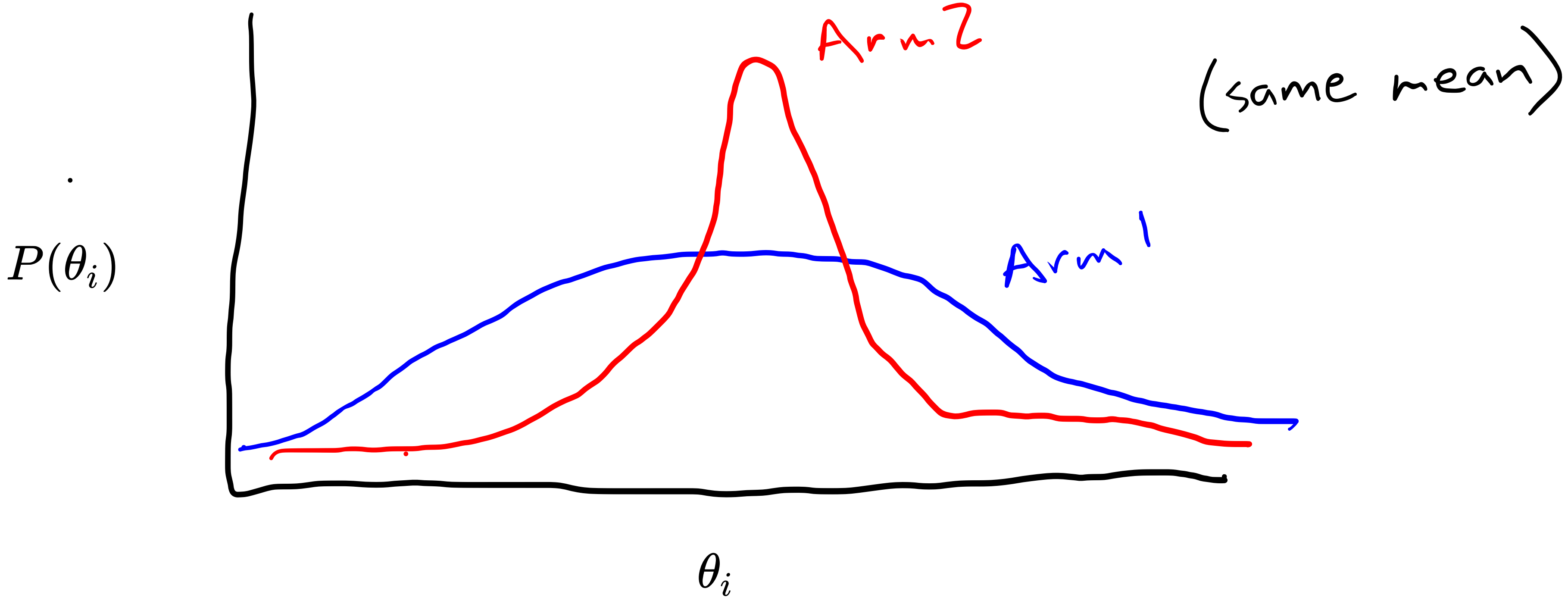


(remove gap with $\lambda \to \infty$)



Break

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

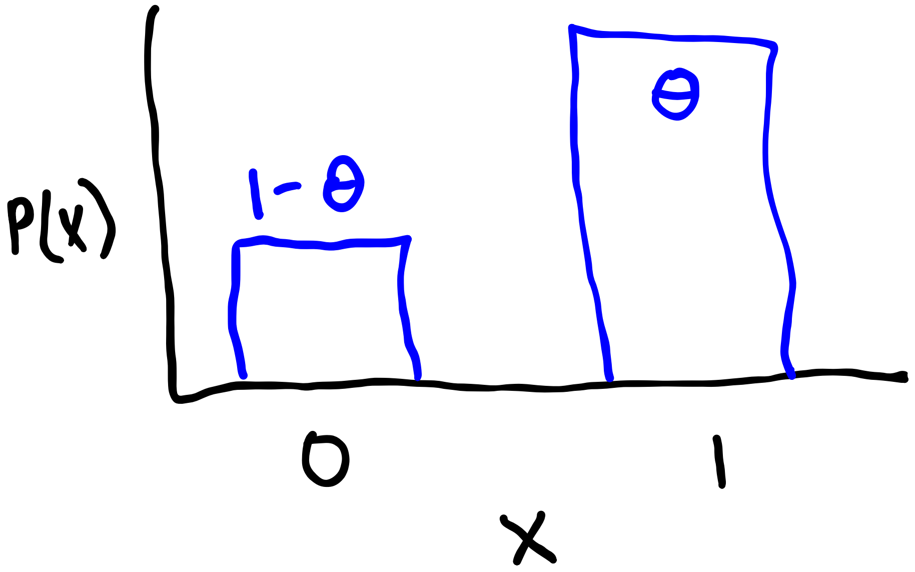


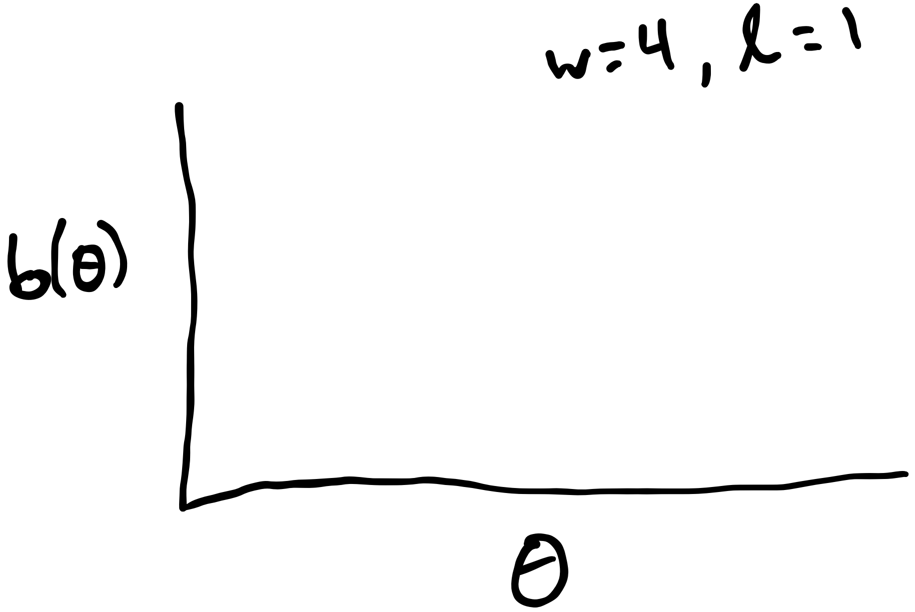
Bayesian Estimation

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?

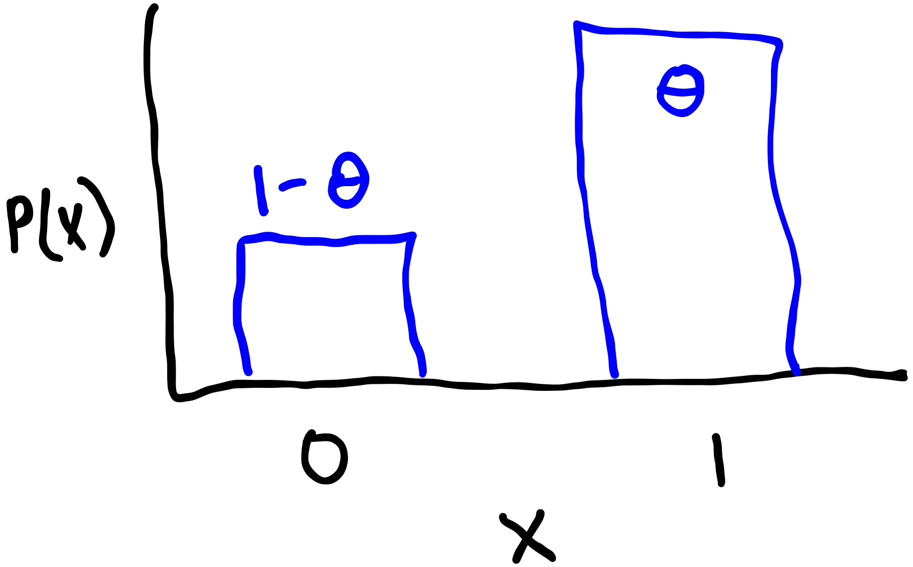




Bayesian Estimation

Bernoulli Distribution

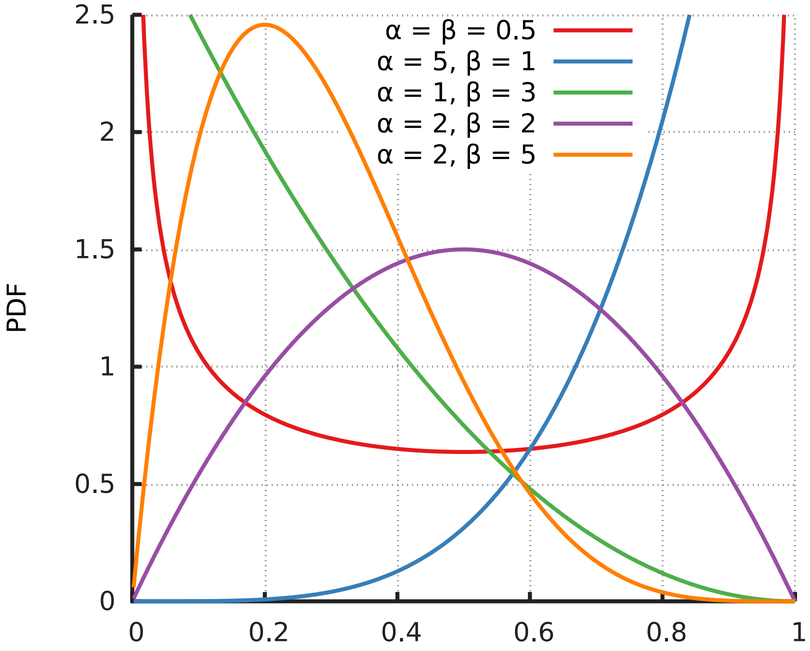
$$\text{Bernoulli}(\theta)$$



Beta Distribution

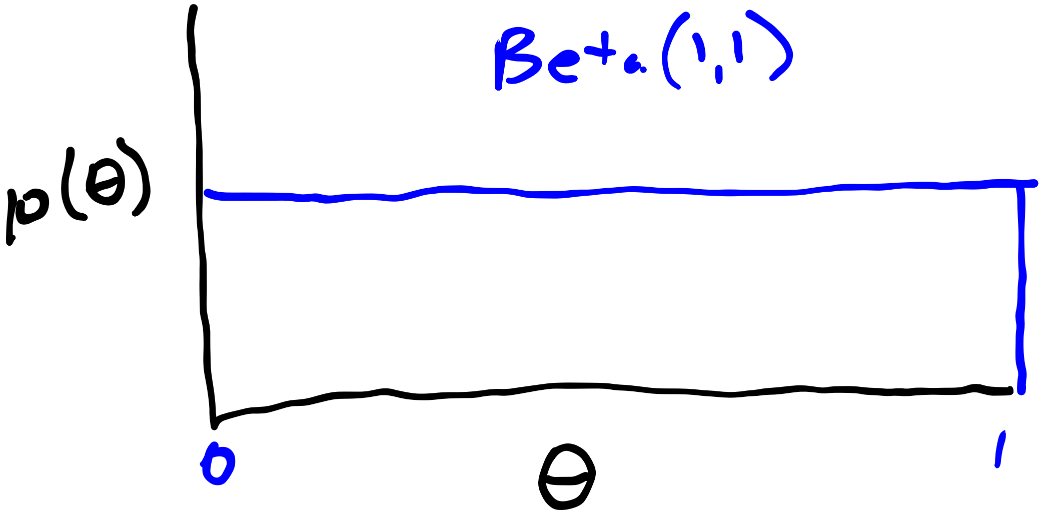
(distribution over Bernoulli distributions)

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

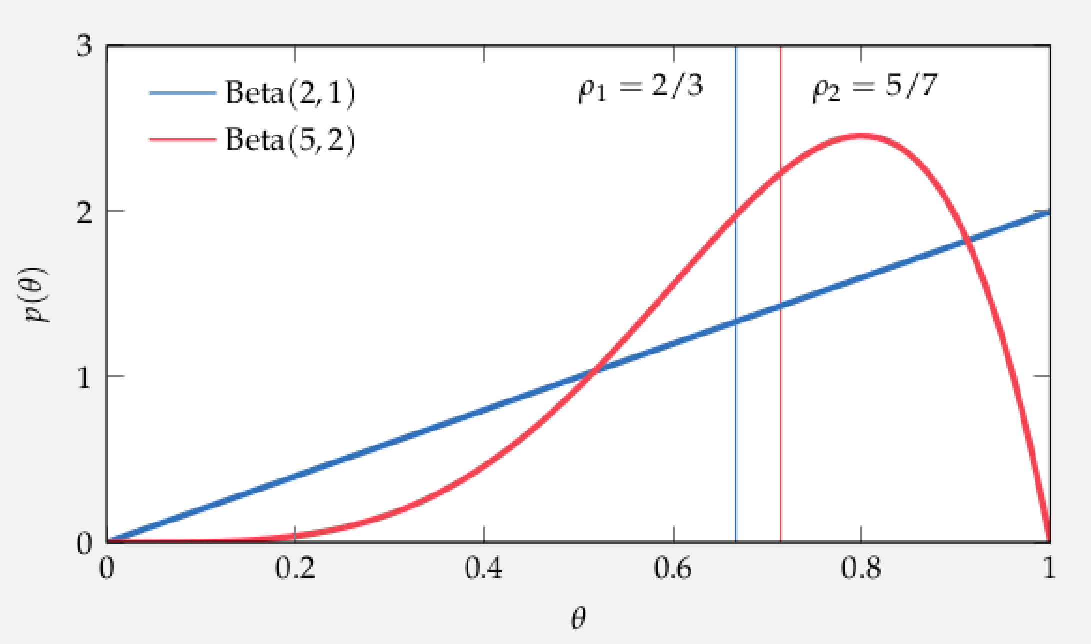


Bayesian Estimation

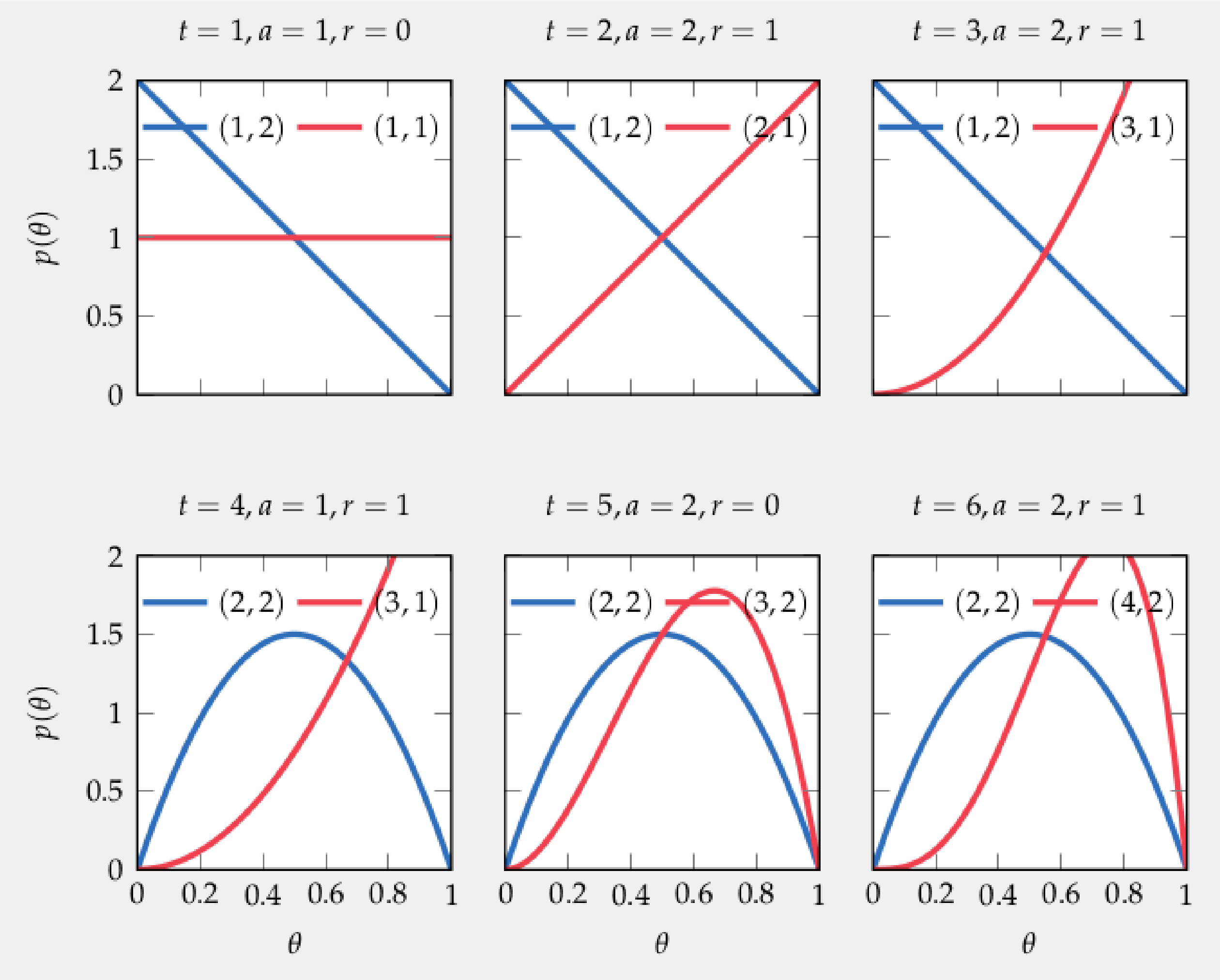
Given a $\mathrm{Beta}(1,1)$ prior distribution



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



Bayesian Estimation



t = time

a = arm pulled

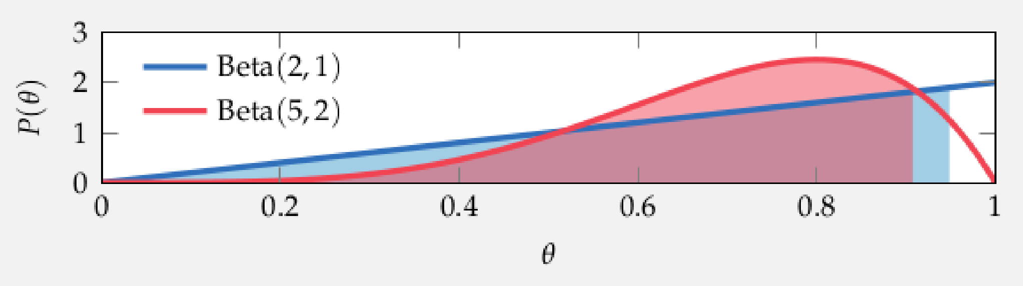
r = reward

Bayesian Bandit Algorithms

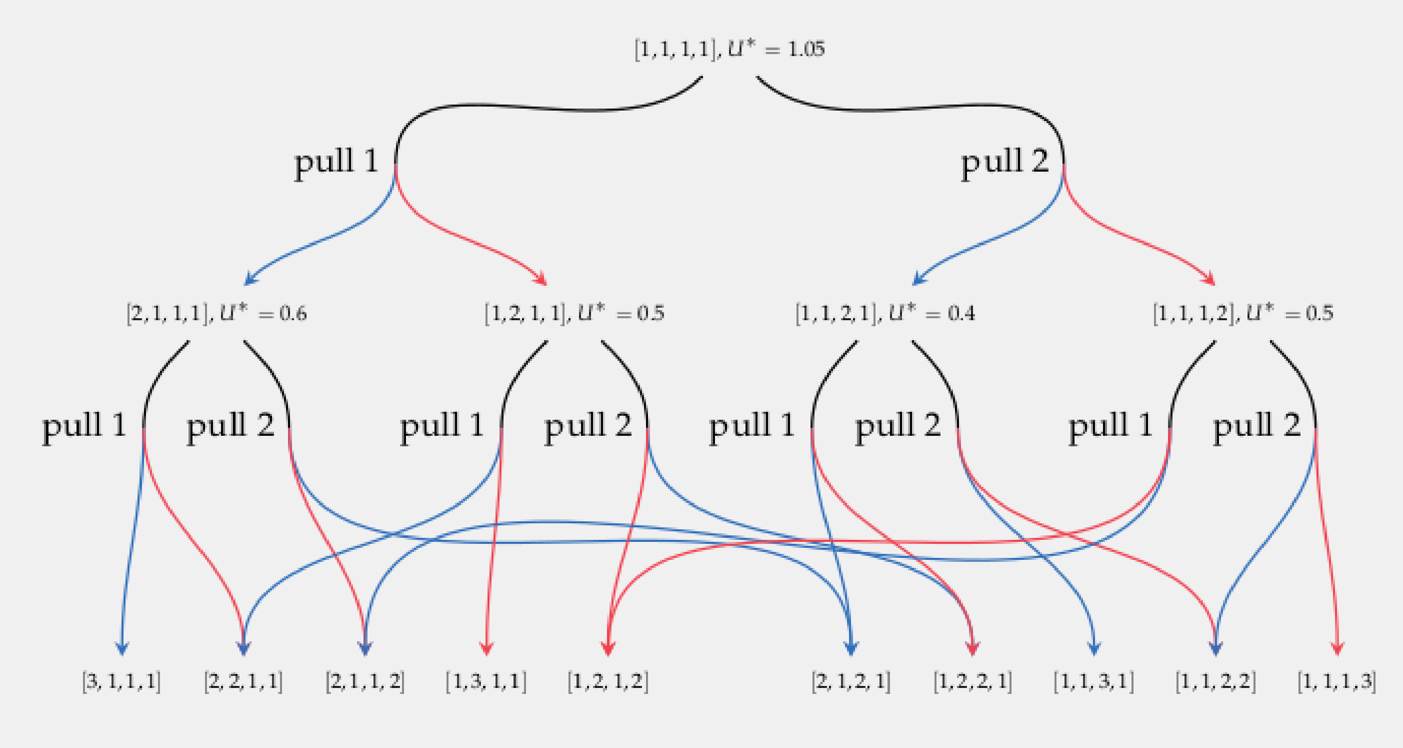
- Quantile Selection
  Choose a for which the α quantile of $b(\theta)$ is highest

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

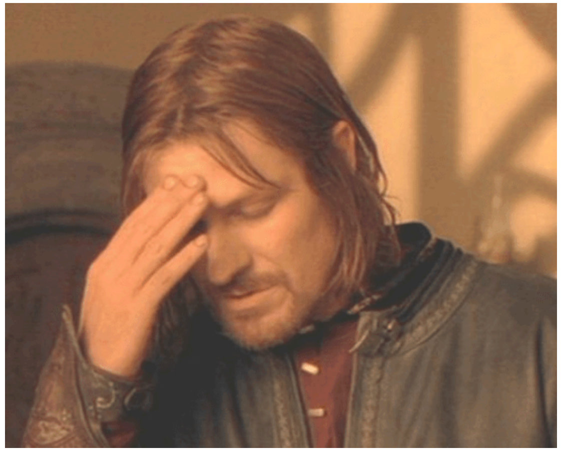
$\alpha = 0.9$



Optimal Algorithm - Dynamic Programming



Regret Analysis



$$\text{Regret}(n) \equiv \theta^* n - \sum_{t=1}^{n} r_t$$

Recall: $f(n) = O(g(n))$ means that there exists a $C > 0$ and $N > 0$ such that $f(n) < C\, g(n)$ for all $n > N$.

Roughly:

- $O(n)$ regret means you might keep picking the wrong arm forever
- $O(\log(n))$ regret means that you keep learning

Review

Guiding Questions

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