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?

Online
Offline

On Policy
Off Policy
Batch

Tabular
Deep

MDP

Model-free ← $\hat{Q}, \hat{\pi}$

Model-Based ← $\hat{T}, \hat{R}$

Last Time

Last Time

First RL Algorithm:

Last Time

First RL Algorithm:

Tabular Maximum Likelihood Model-Based Reinforcement Learning

Last Time

First RL Algorithm:

Tabular Maximum Likelihood Model-Based Reinforcement Learning

loop

  choose action a ← E/E

  gain experience

  estimate T, R

  solve MDP with T, R

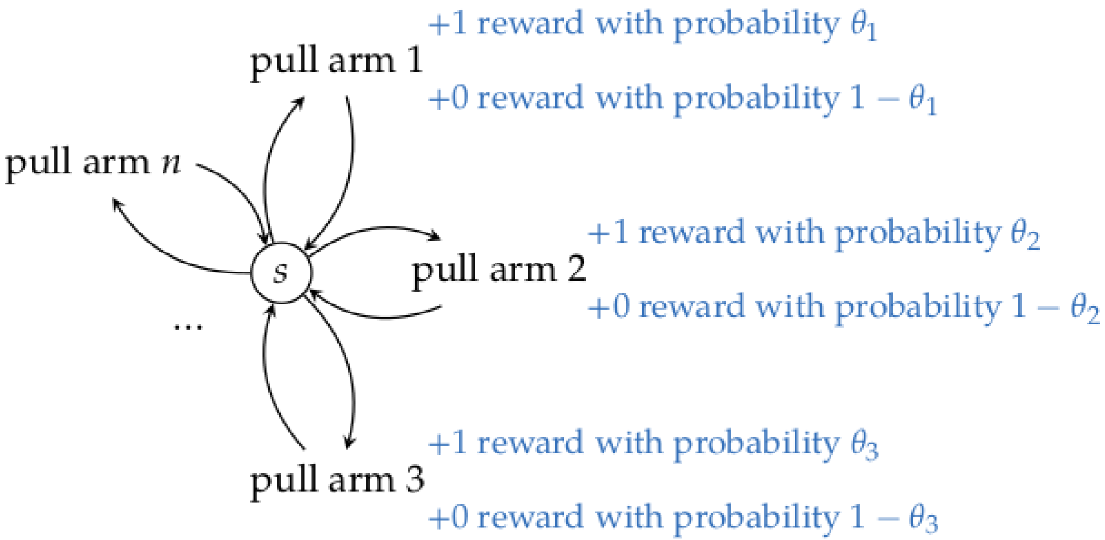
Guiding Questions

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

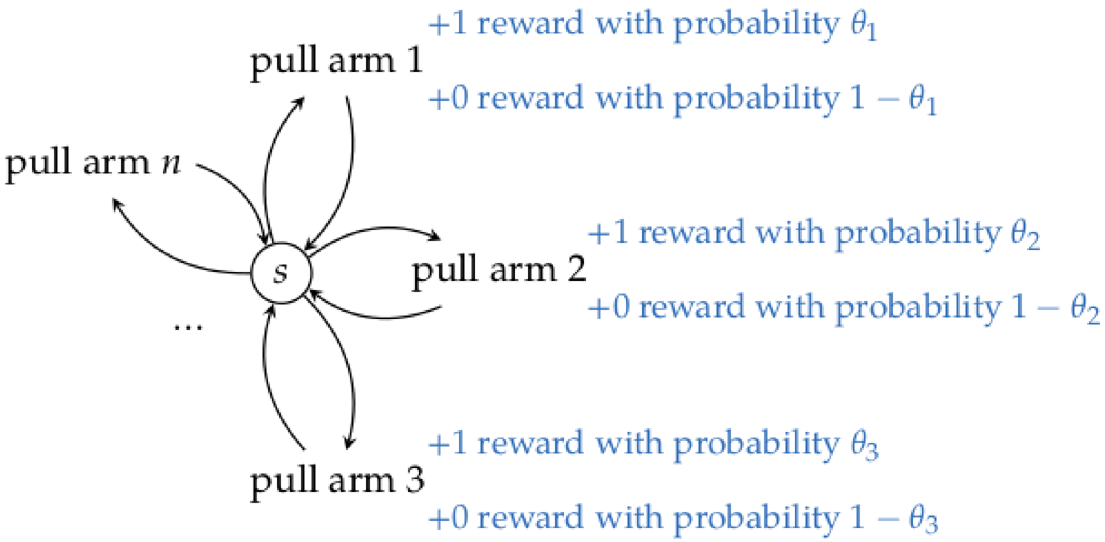
Bandits

Bandits

Bandits

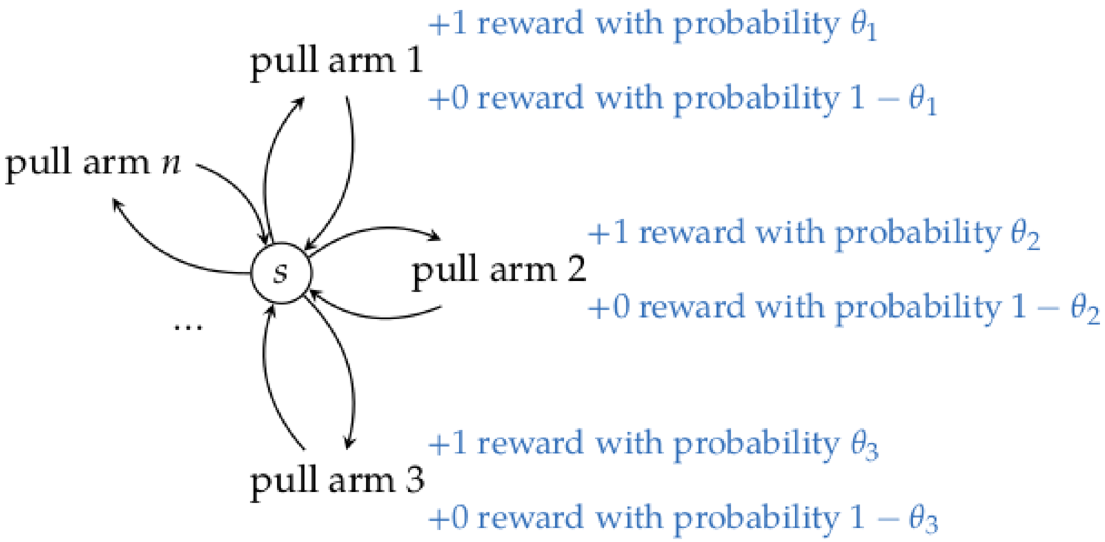

pull arm 1
+1 reward with probability θ_1
+0 reward with probability $1 - \theta_1$
pull arm n
s
pull arm 2
+1 reward with probability θ_2
+0 reward with probability $1 - \theta_2$
...
pull arm 3
+1 reward with probability θ_3
+0 reward with probability $1 - \theta_3$

Bandits



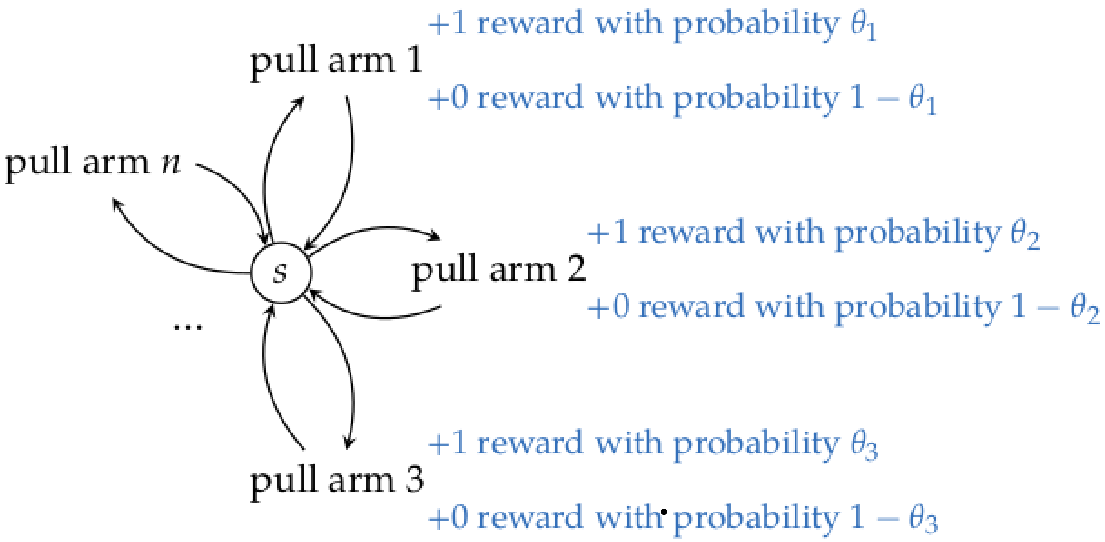
- Bernoulli Bandit with parameters θ

Bandits



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

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."*

Greedy Strategy

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

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

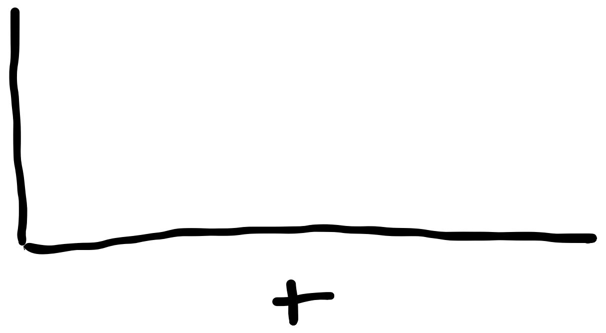
Undirected Strategies

Undirected Strategies

- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$

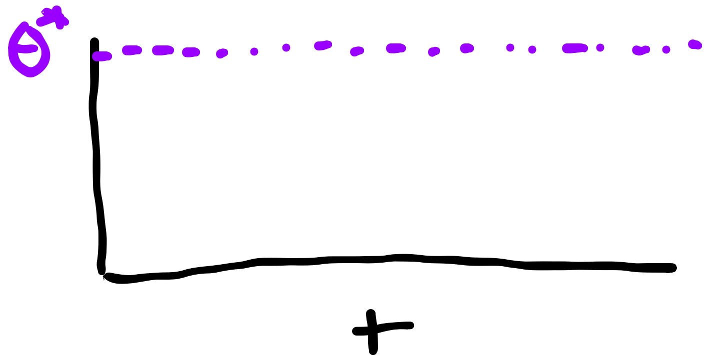
Undirected Strategies

- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}} \rho_a$



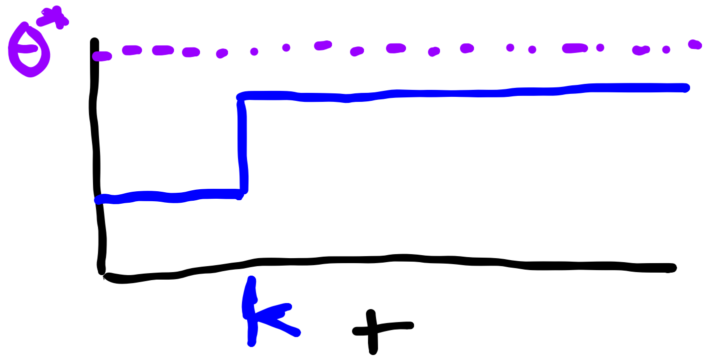
Undirected Strategies

- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$



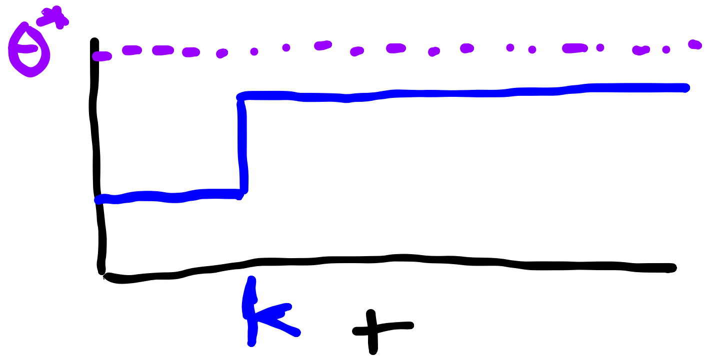
Undirected Strategies

- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$



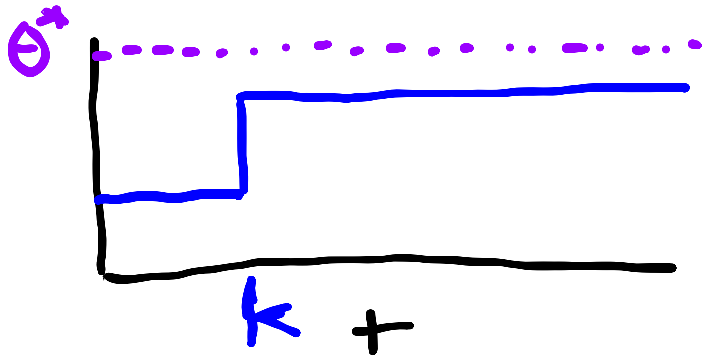
Undirected Strategies

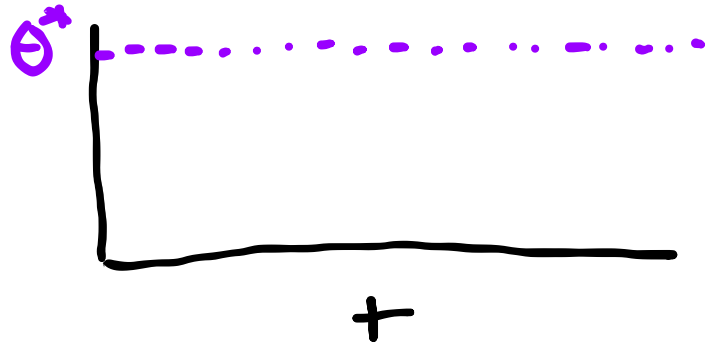
- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$
- ϵ - greedy
  With probability ϵ, choose randomly
  Otherwise choose $\underset{a}{\operatorname{argmax}}\, \rho_a$



Undirected Strategies

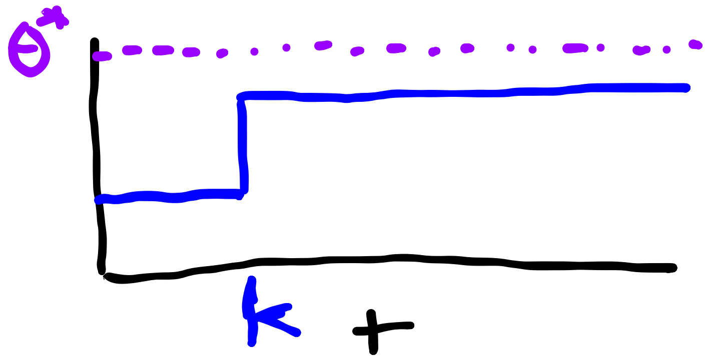
- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$
- ϵ - greedy
  With probability ϵ, choose randomly
  Otherwise choose $\underset{a}{\operatorname{argmax}}\, \rho_a$

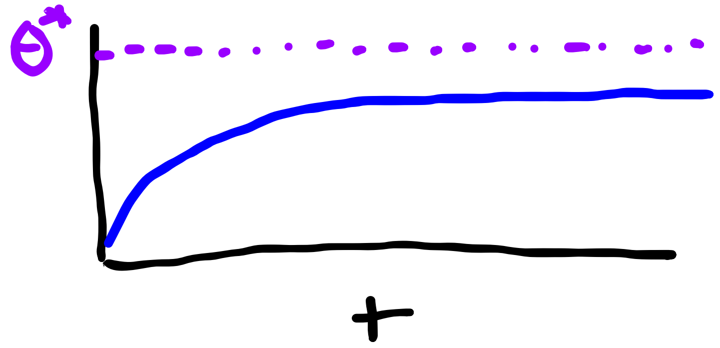




Undirected Strategies

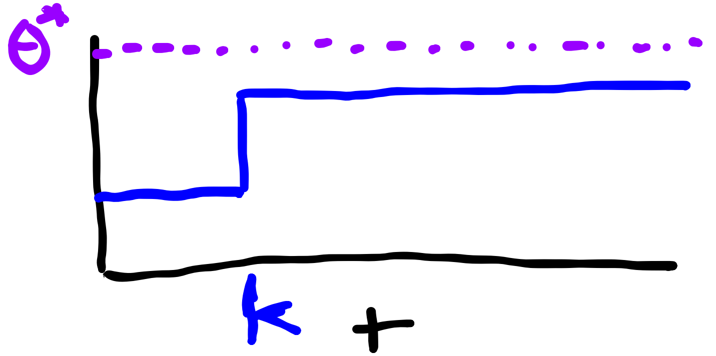
- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$
- ϵ - greedy
  With probability ϵ, choose randomly
  Otherwise choose $\underset{a}{\operatorname{argmax}}\, \rho_a$

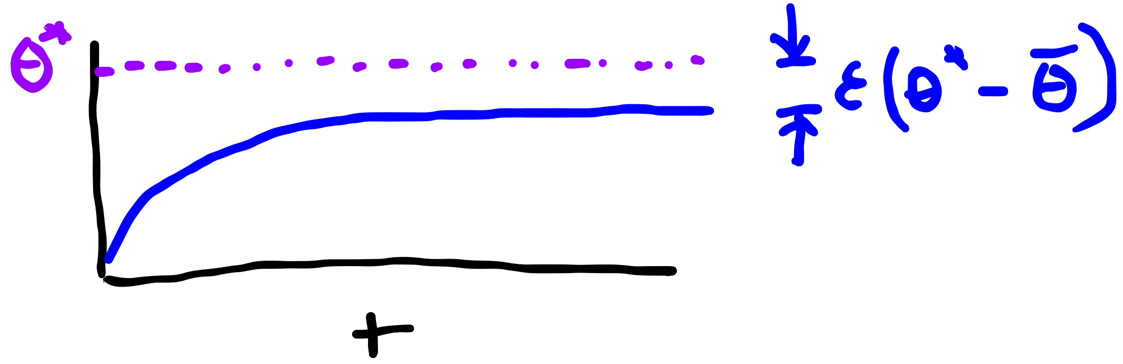




Undirected Strategies

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





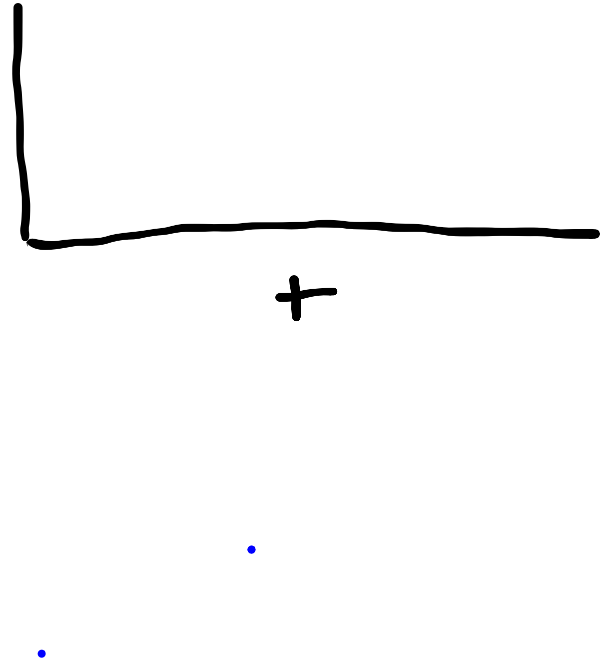
Directed Strategies

Directed Strategies

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

Directed Strategies

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



Directed Strategies

$$P(a) = \frac{e^{\lambda \rho_a}}{\sum_i e^{\lambda \rho_i}}$$

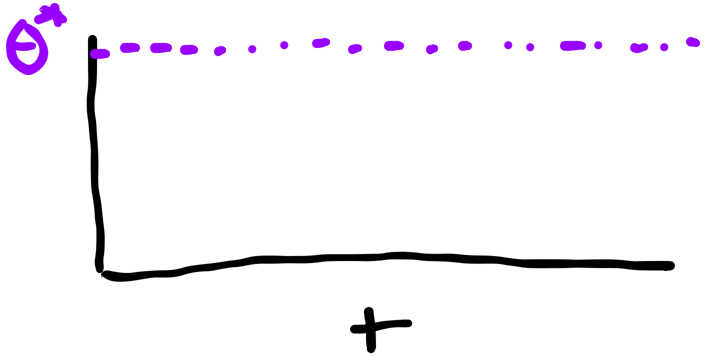
$\lambda = 0$

$\frac{1}{\sum_i 1} = \frac{1}{n}$

$\lambda \to \infty$

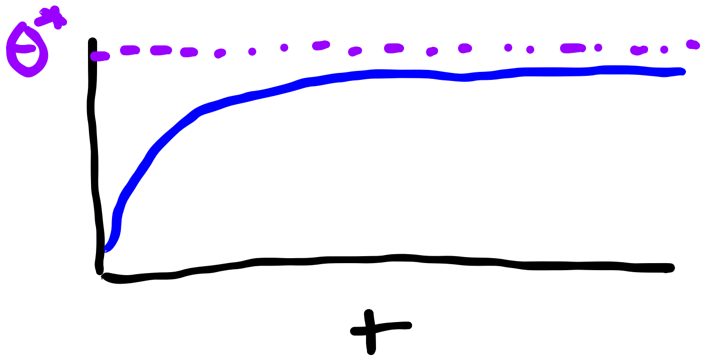
greedy

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



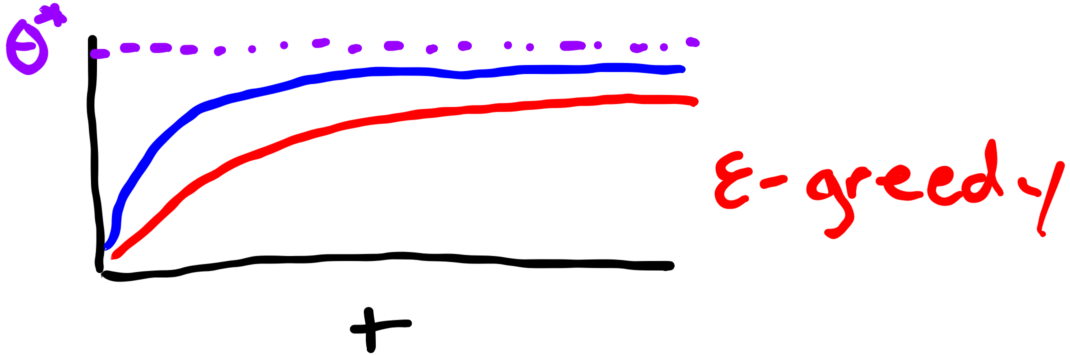
Directed Strategies

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



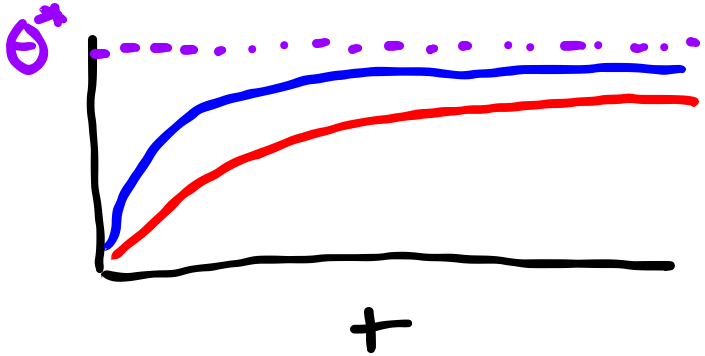
Directed Strategies

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



Directed Strategies

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



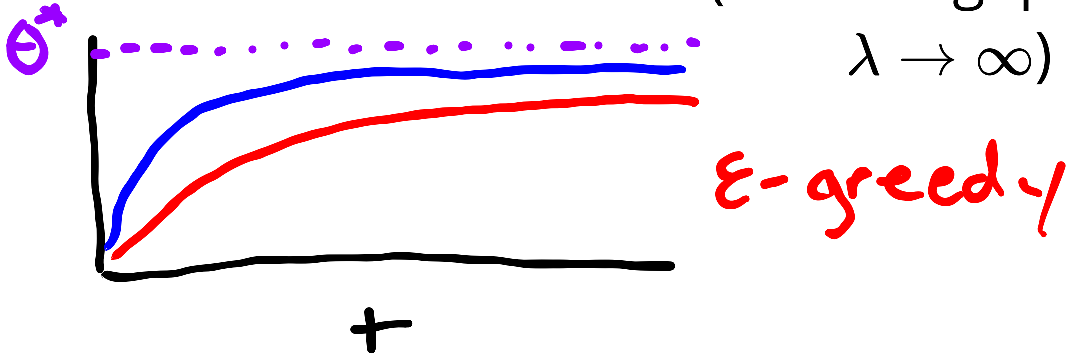
(remove gap with $\lambda \to \infty$)

ε-greedy

Directed Strategies

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

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



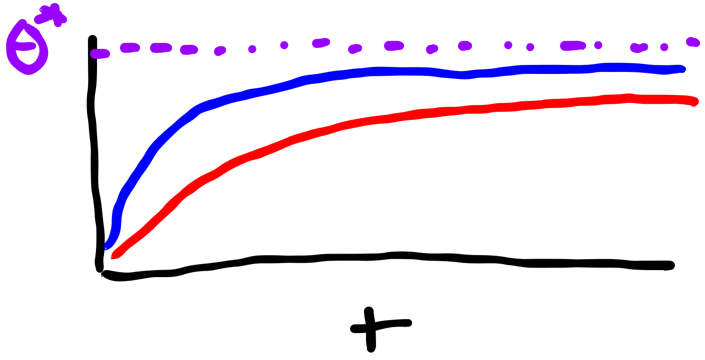
(remove gap with $\lambda \to \infty$)

Directed Strategies

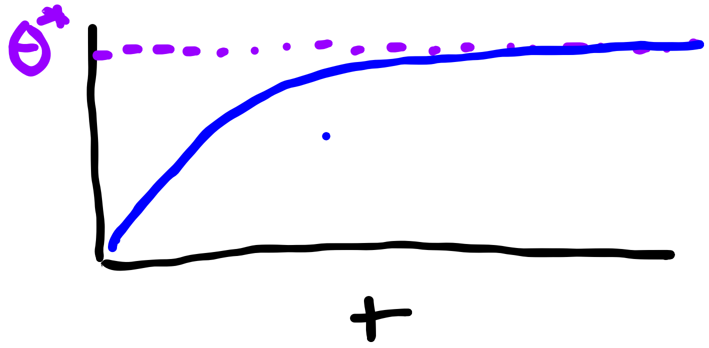
- Softmax
  Choose a with probability proportional to $e^{\lambda \rho_a}$

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

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

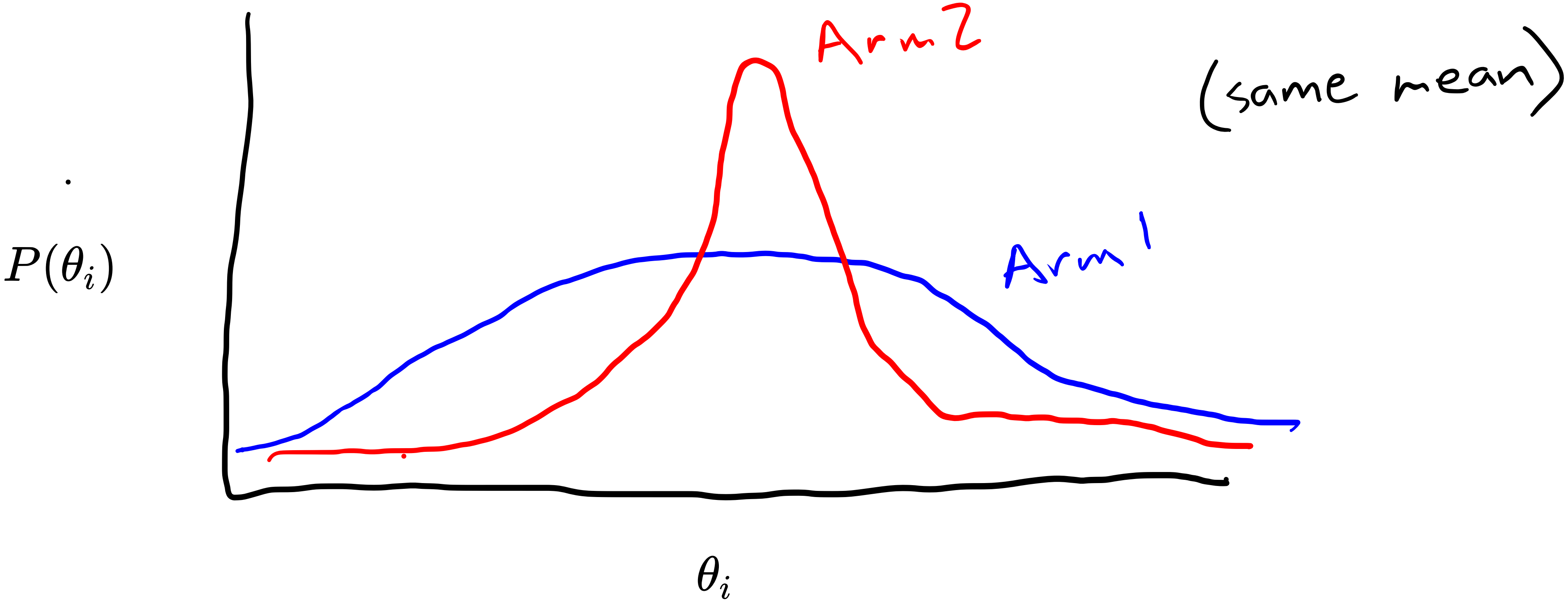


ε-greedy



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

Bayesian Estimation

Bernoulli Distribution

Bayesian Estimation

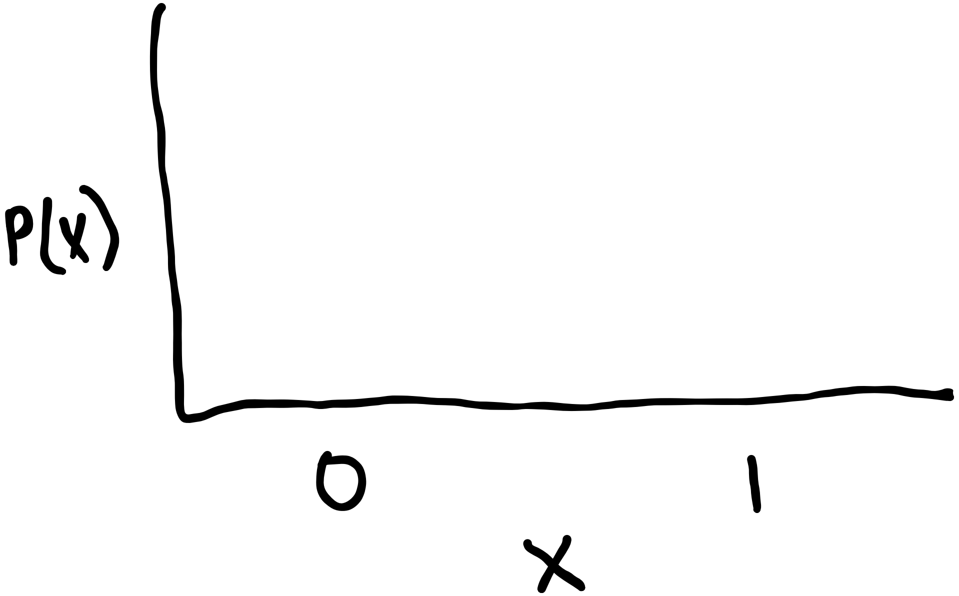
Bernoulli Distribution

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

Bayesian Estimation

Bernoulli Distribution

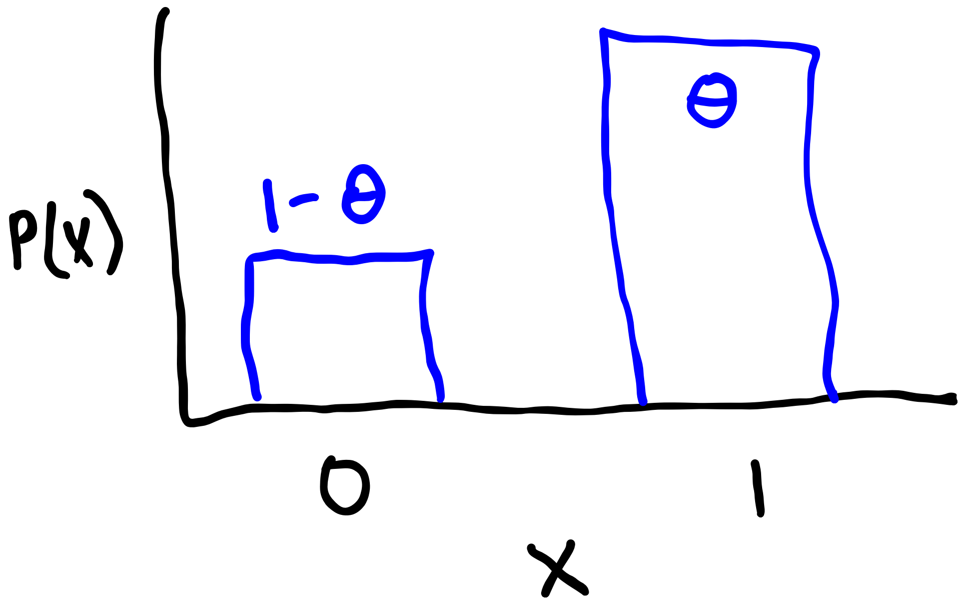
$$\text{Bernoulli}(\theta)$$



Bayesian Estimation

Bernoulli Distribution

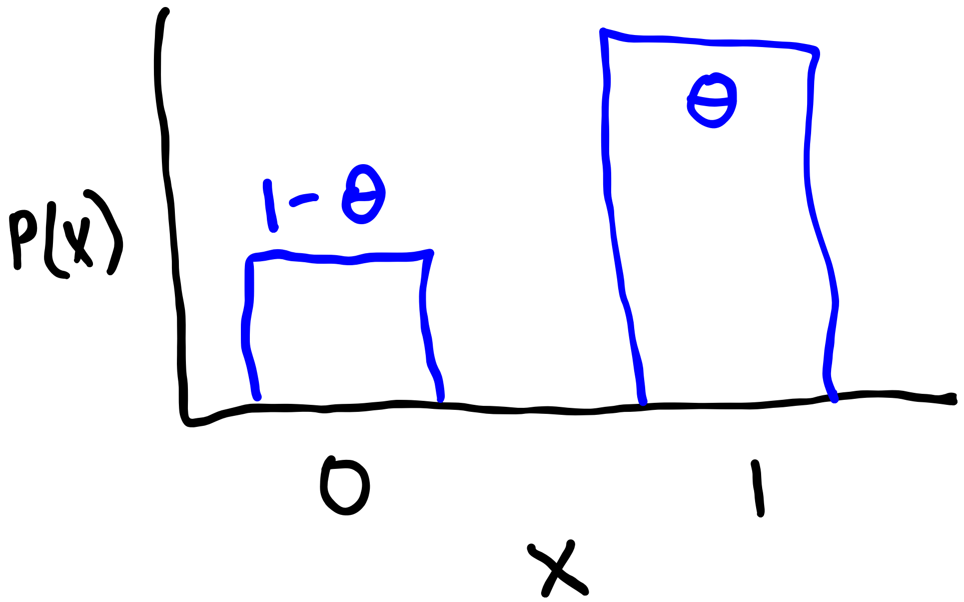
$$\text{Bernoulli}(\theta)$$



Bayesian Estimation

Bernoulli Distribution

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



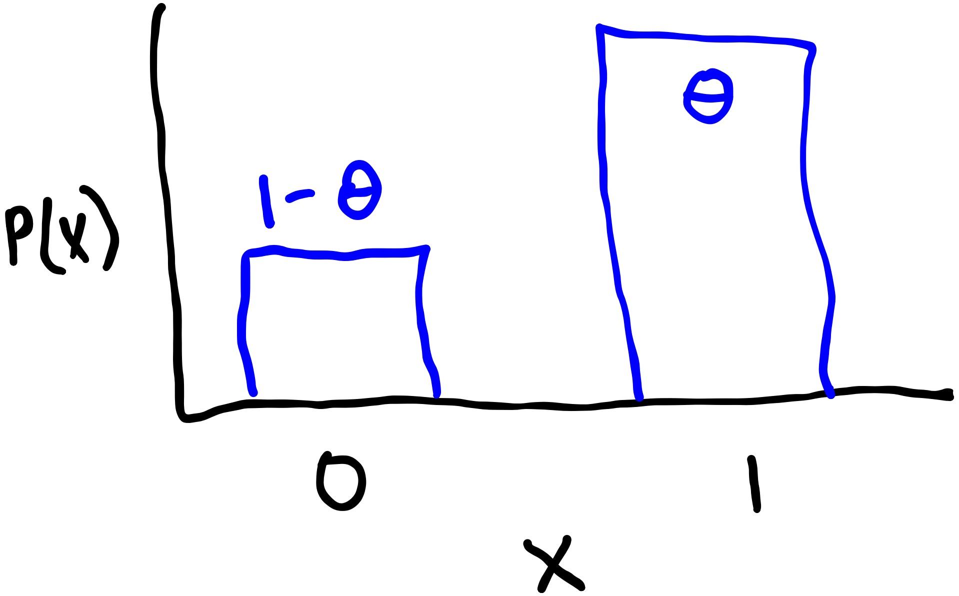
Discussion: Given that I have received w wins and l losses, what should my belief (probability distribution) about θ look like?

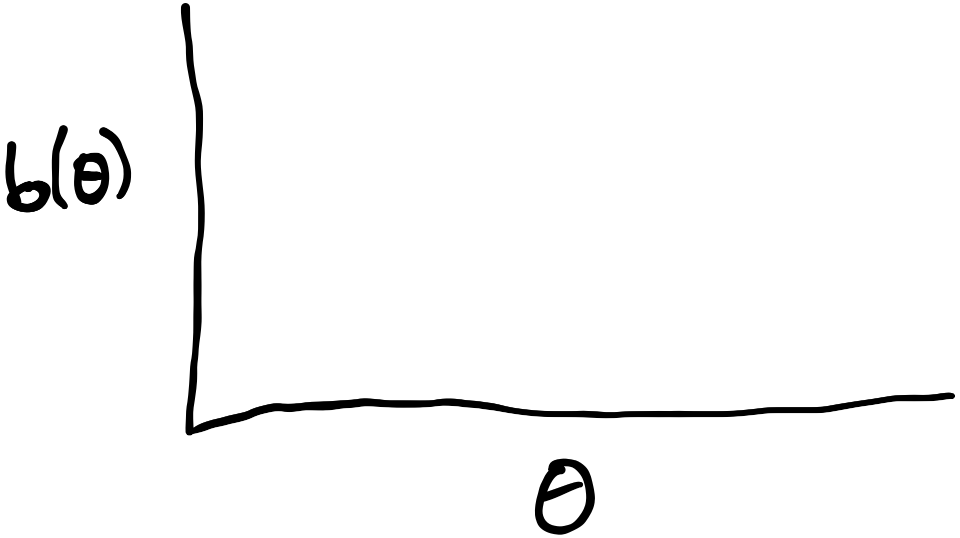
Bayesian Estimation

Bernoulli Distribution

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

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

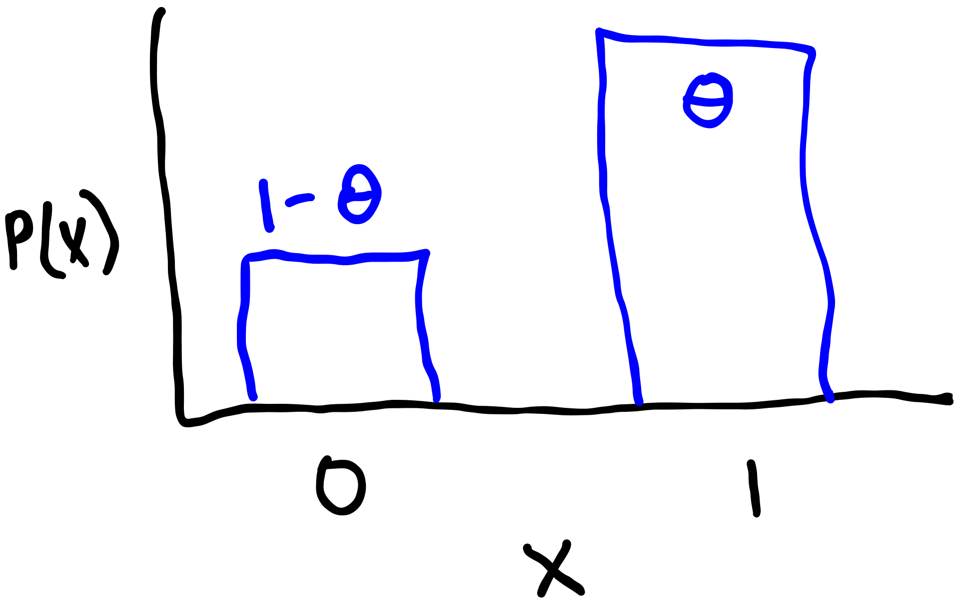




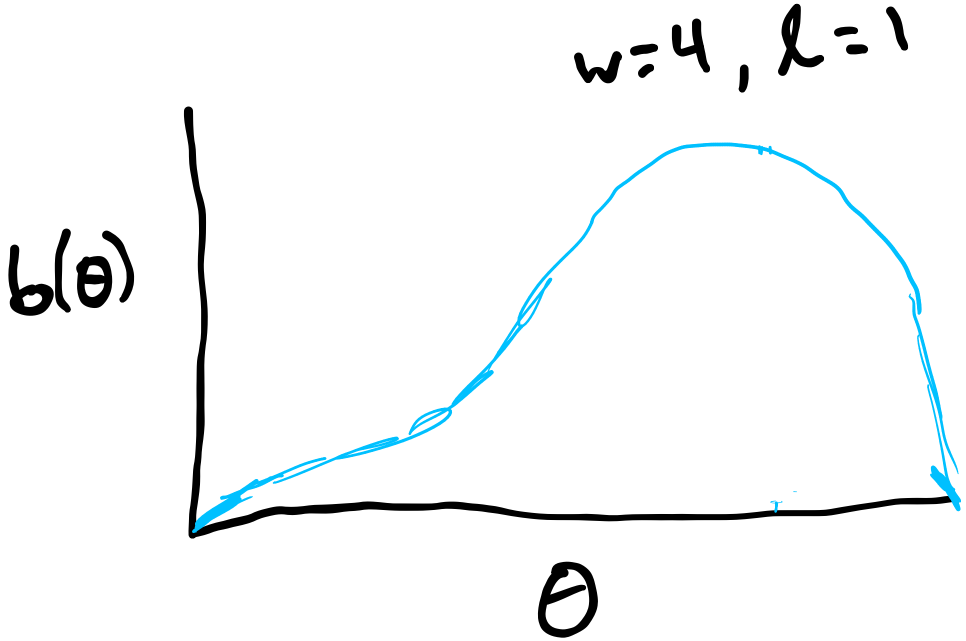
Bayesian Estimation

Bernoulli Distribution

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



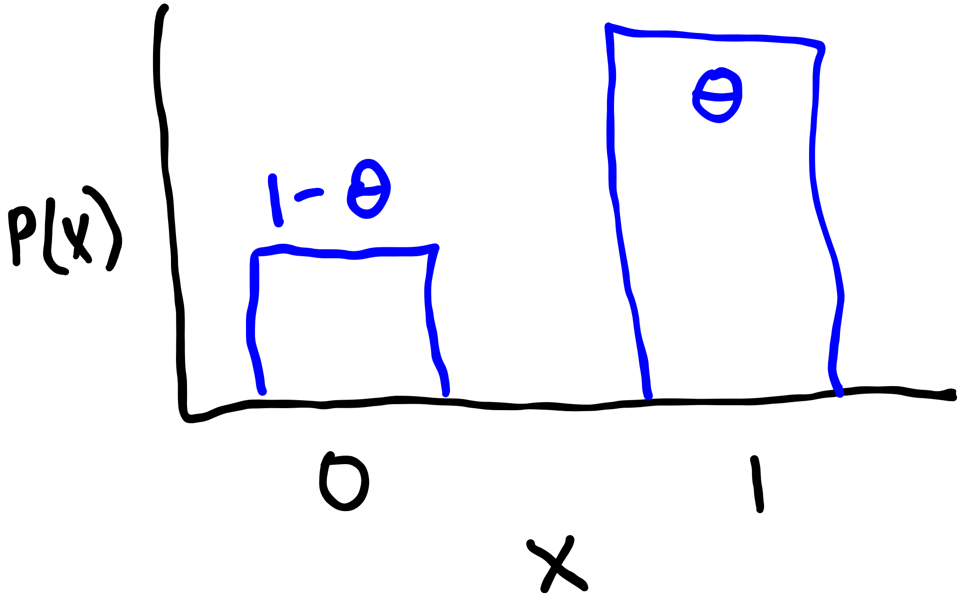
Discussion: Given that I have received w wins and l losses, what should my belief (probability distribution) about θ look like?



Bayesian Estimation

Bernoulli Distribution

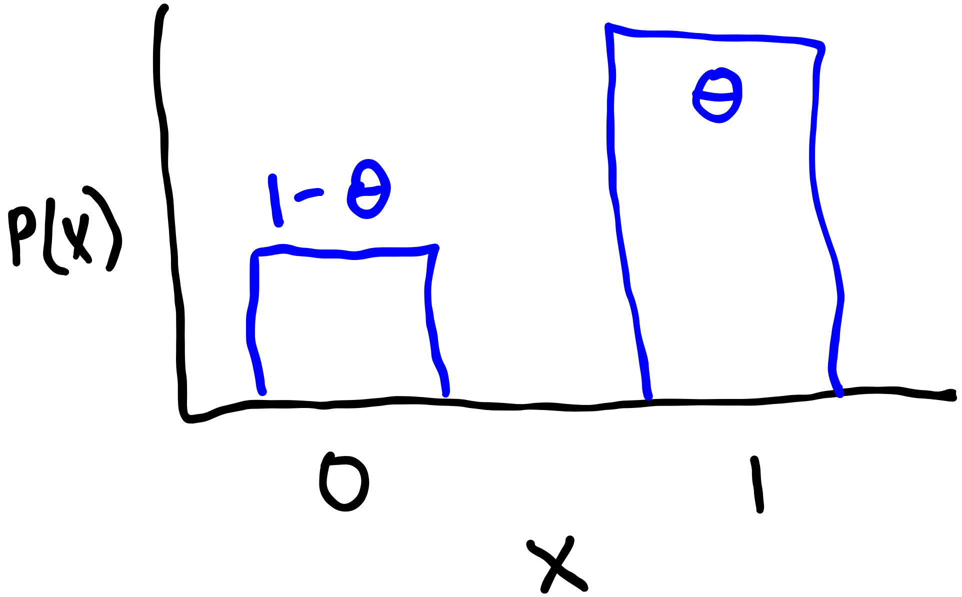
$$\text{Bernoulli}(\theta)$$



Bayesian Estimation

Bernoulli Distribution

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



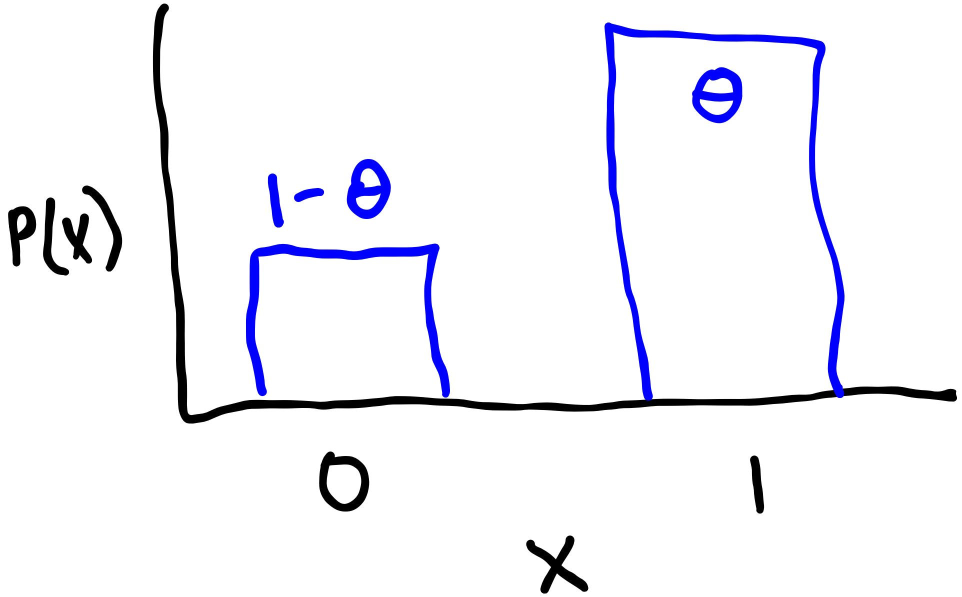
Beta Distribution

(distribution over Bernoulli distributions)

Bayesian Estimation

Bernoulli Distribution

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



Beta Distribution

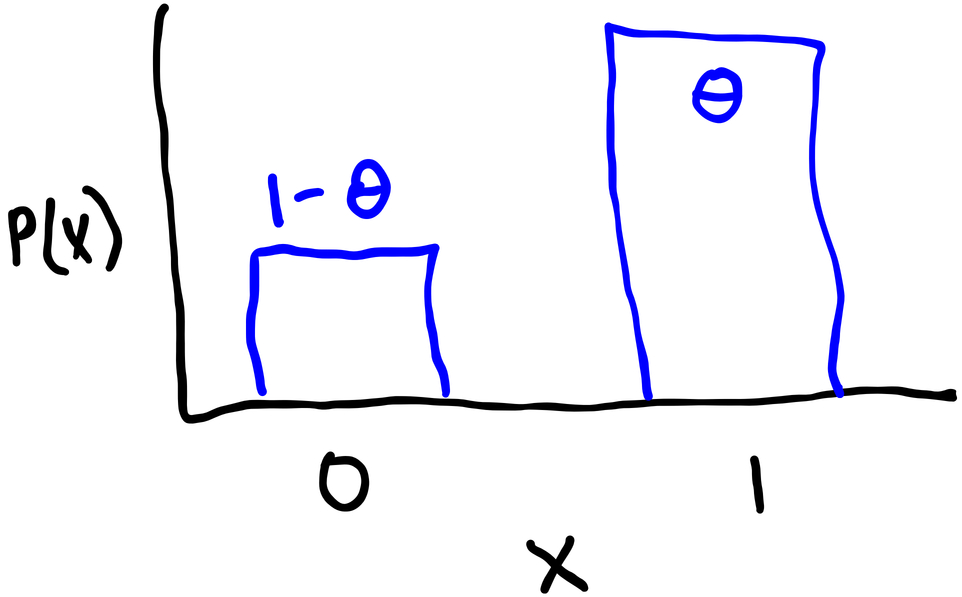
(distribution over Bernoulli distributions)

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

Bayesian Estimation

Bernoulli Distribution

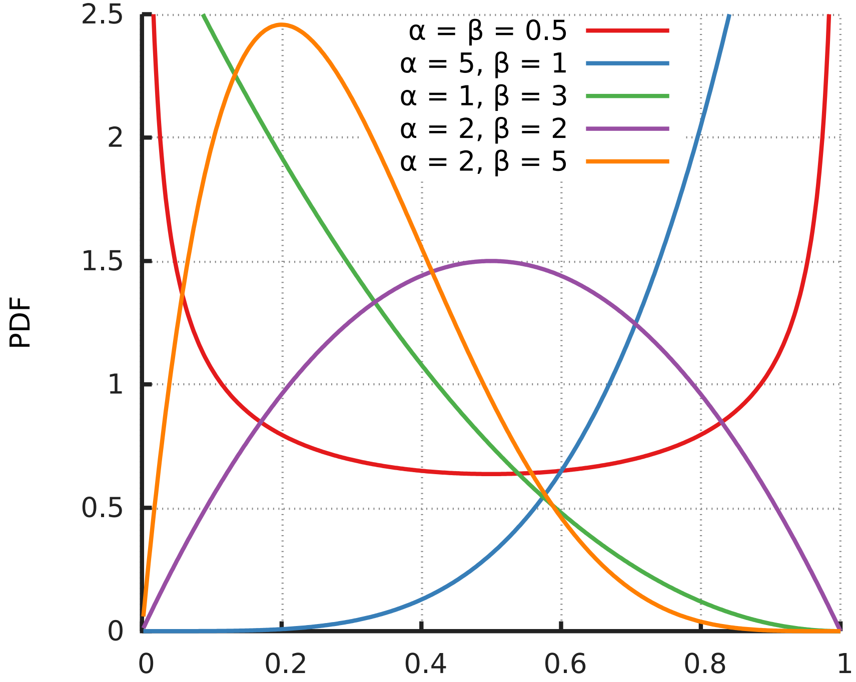
$$\mathrm{Bernoulli}(\theta)$$



Beta Distribution

(distribution over Bernoulli distributions)

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



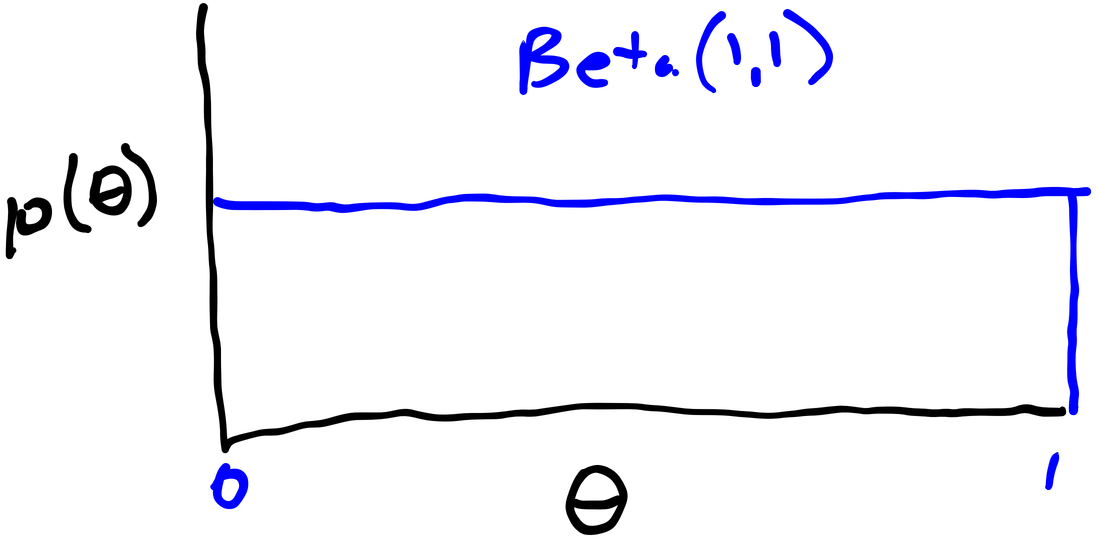
Bayesian Estimation

Bayesian Estimation

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

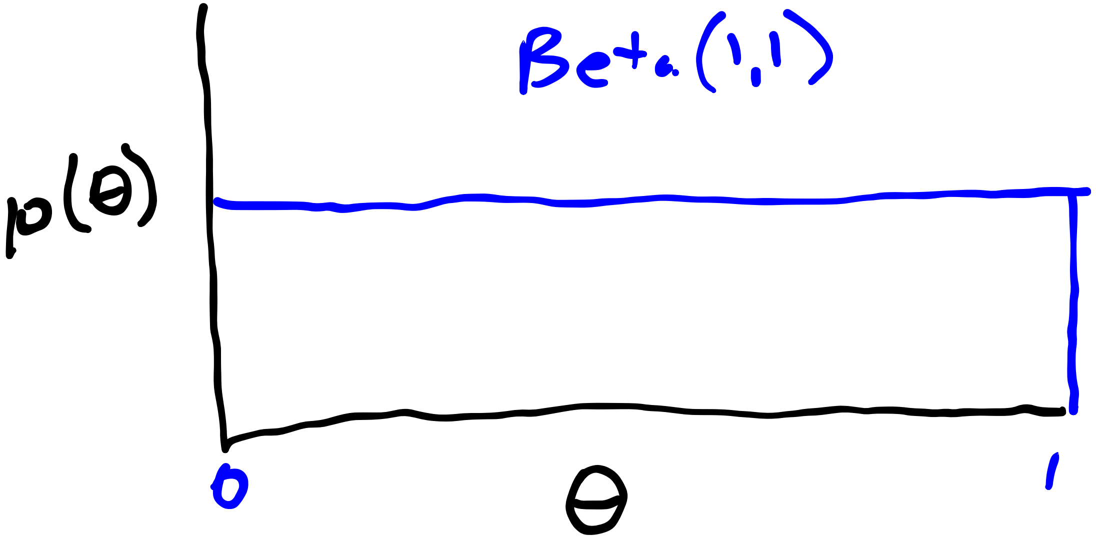
Bayesian Estimation

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



Bayesian Estimation

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

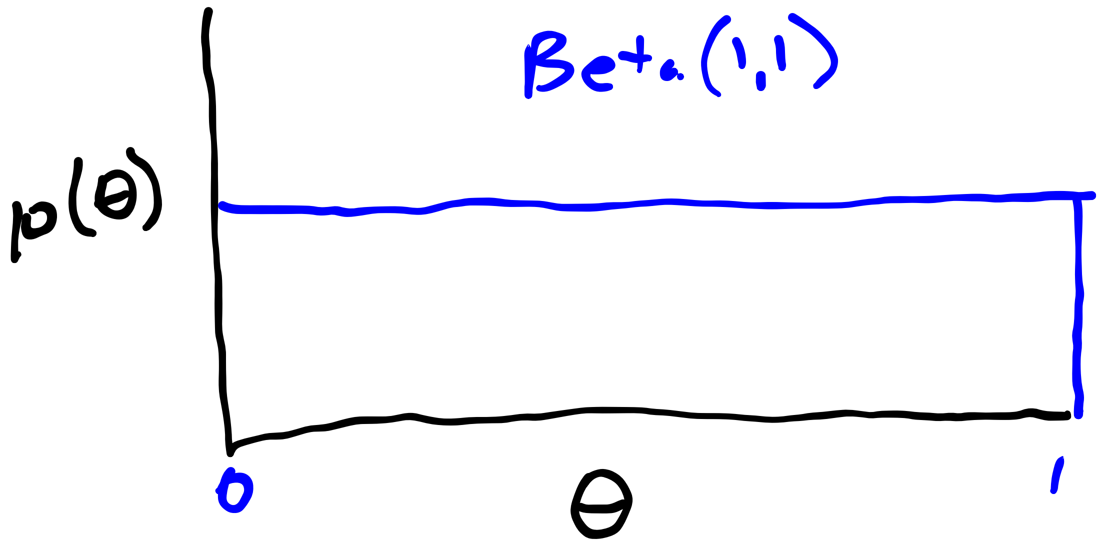


The posterior distribution of θ is

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

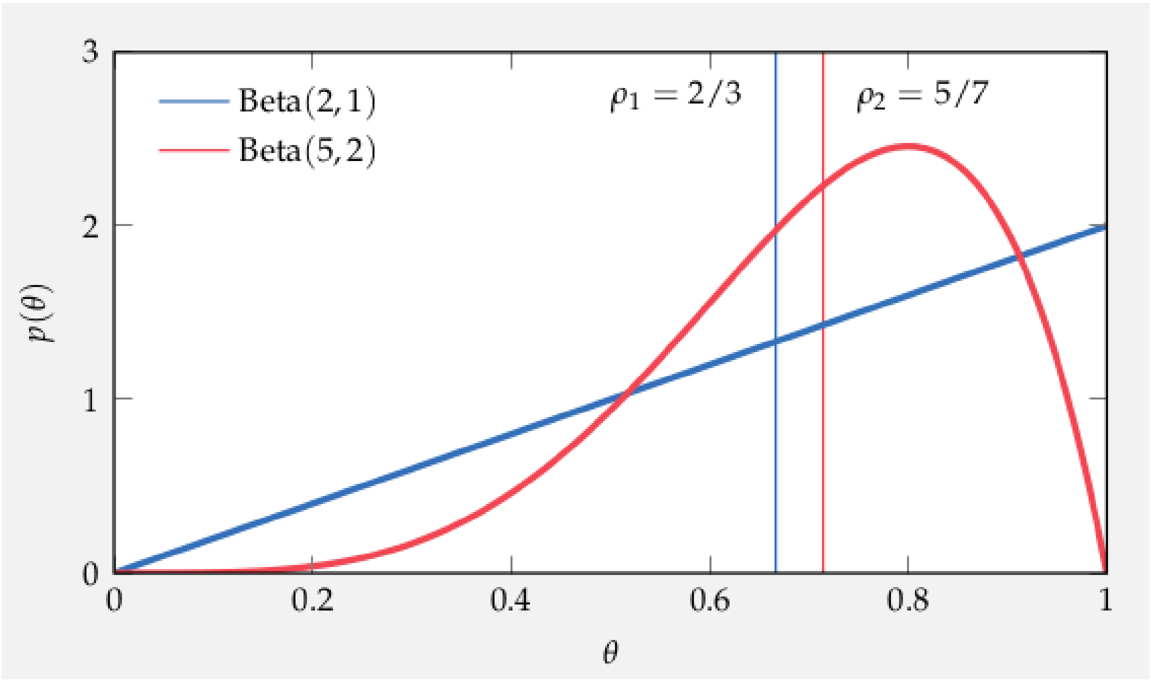
Bayesian Estimation

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

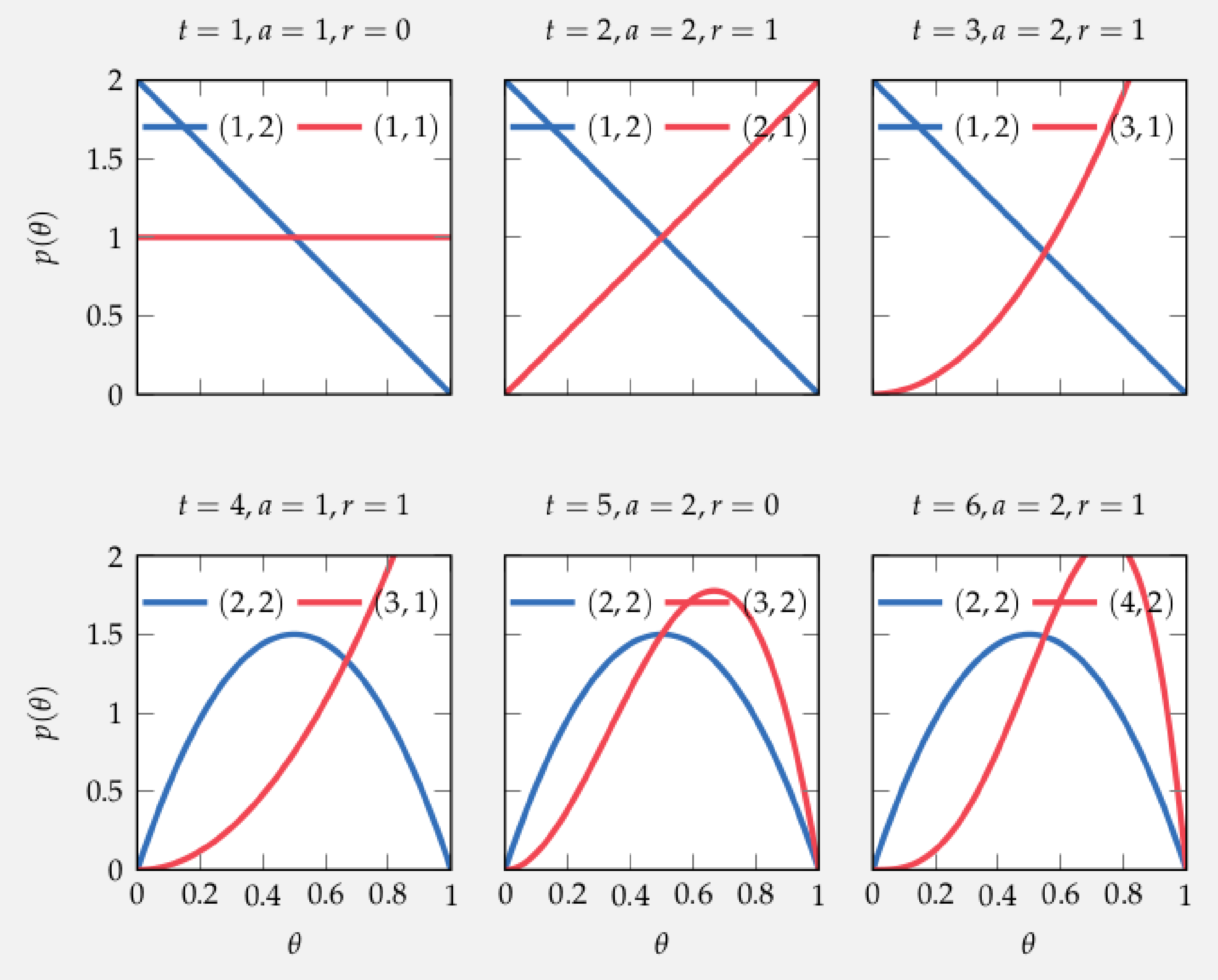


The posterior distribution of θ is

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



Bayesian Estimation



t = time

a = arm pulled

r = reward

Bayesian Bandit Algorithms

Bayesian Bandit Algorithms

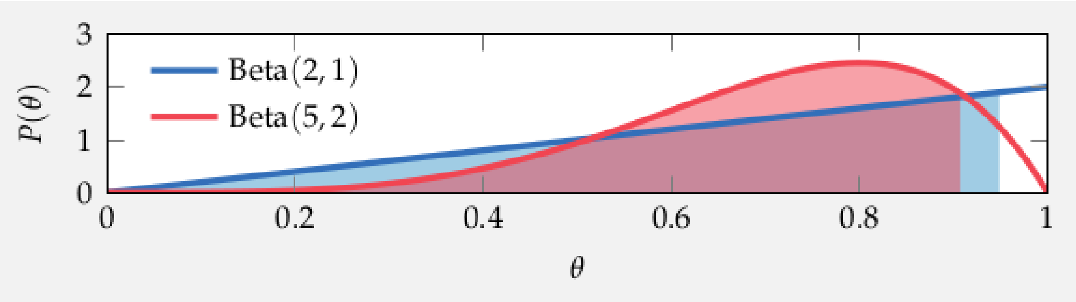
- Quantile Selection
  Choose a for which the α quantile of $b(\theta)$ is highest

Bayesian Bandit Algorithms

$\alpha = 0.9$

higher α = more optimistic

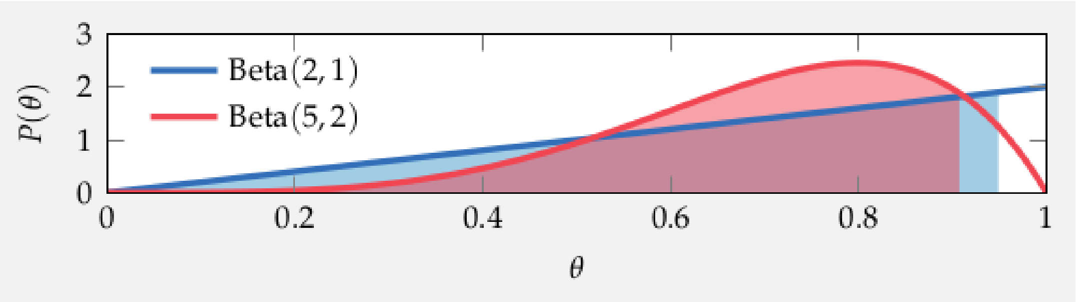
- Quantile Selection
  Choose a for which the α quantile of $b(\theta)$ is highest



Bayesian Bandit Algorithms

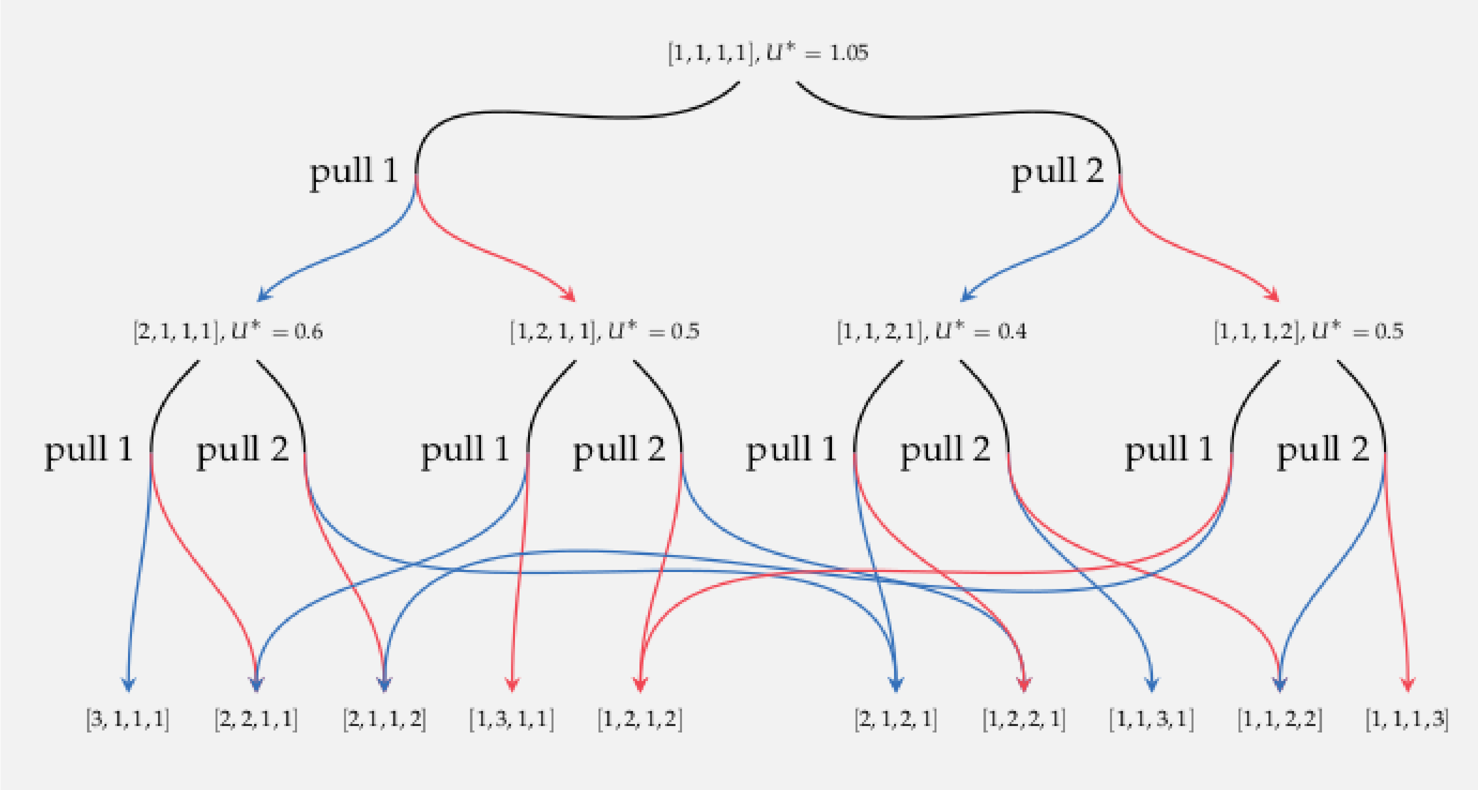
$\alpha = 0.9$

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



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

Optimal Algorithm - Dynamic Programming



Review

Easier to Implement
Faster

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

		Optimal in limit	Regret
undirected	- Greedy	No	$O(N)$
undirected	- ε-Greedy	$\varepsilon \to 0$	$O(N)$
undirected	- explore-commit	$k \to \infty$	$O(N)$
	- softmax $\propto e^{\lambda p_a}$	$\lambda \to \infty$	$O(N)$
	- UCB $\text{argmax}\ p_a + c\sqrt{\frac{\log N}{N(a)}}$	✓	$O(\log(N))$
Bayesian	- Quantile Selection	✓	"
Bayesian	- Thompson Sampling	✓	"
	- Dynamic Programming	✓	

Less Regret

Guiding Questions

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