Last Time

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

- Exploration vs 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

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

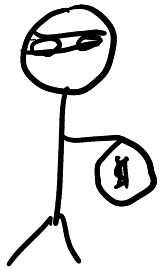
estimate T, R

solve MDP with T, R

Guiding Questions

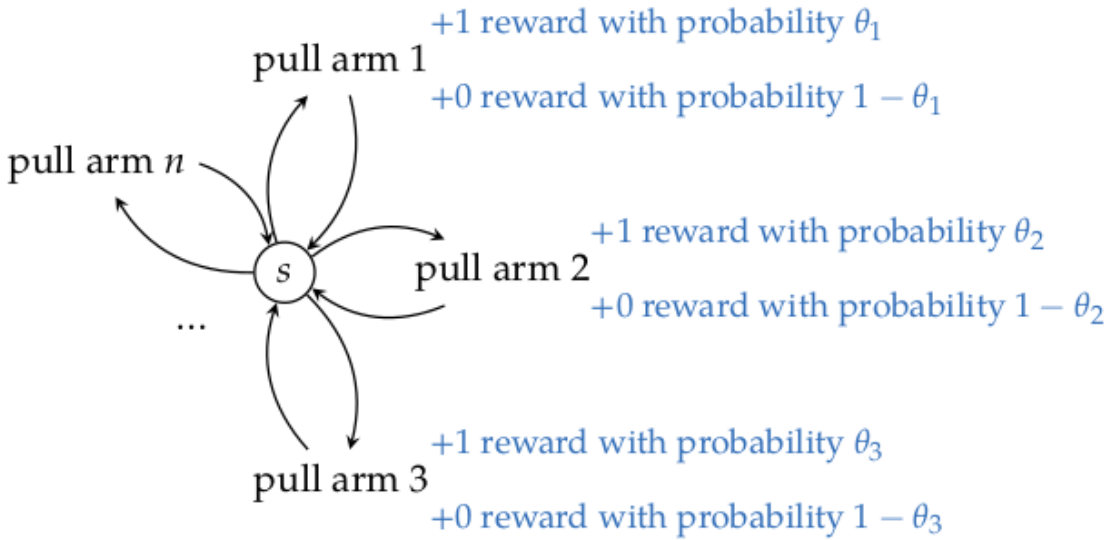
- What are the best ways to trade off Exploration and Exploitation?

Bandits

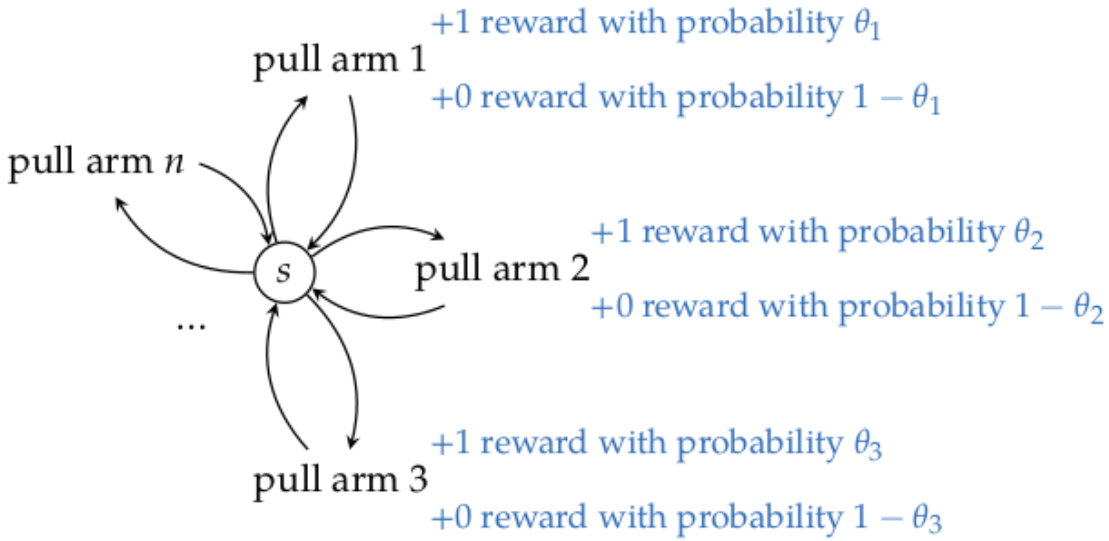


Bandits

Bandits

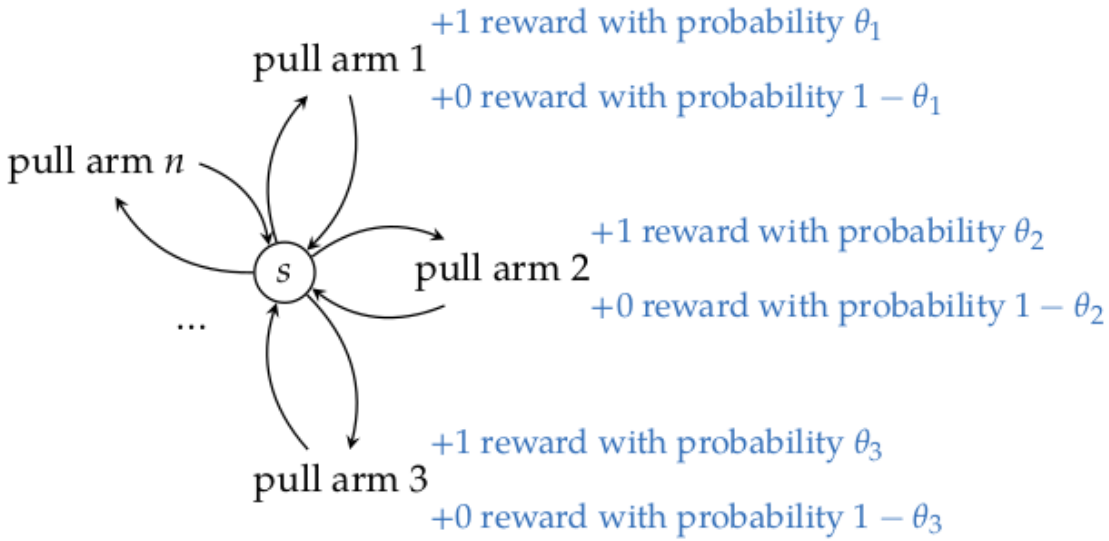

+1 reward with probability θ_1
pull arm 1
+0 reward with probability $1-\theta_1$
pull arm n
+1 reward with probability θ_2
s
pull arm 2
+0 reward with probability $1-\theta_2$
...
+1 reward with probability θ_3
pull arm 3
+0 reward with probability $1-\theta_3$

Bandits



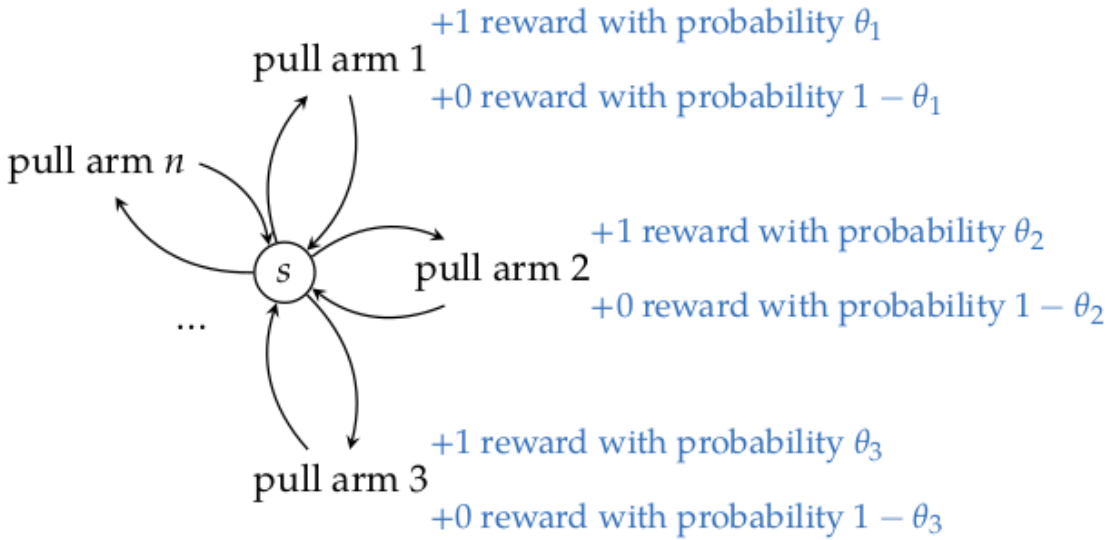
- Bernoulli Bandit with parameters θ

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

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

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

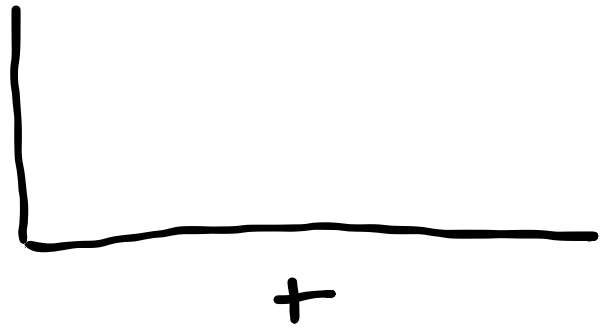
Undirected Strategies

Undirected Strategies

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

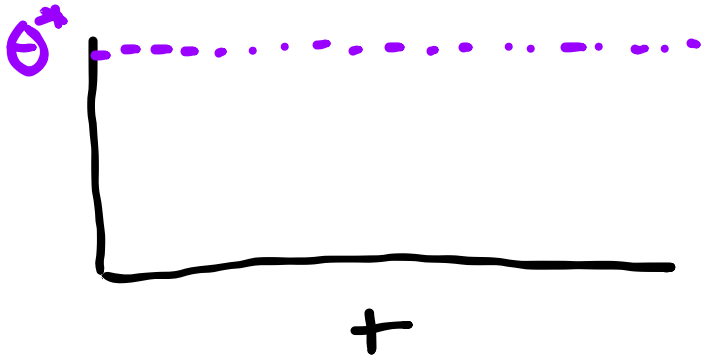
Undirected Strategies

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



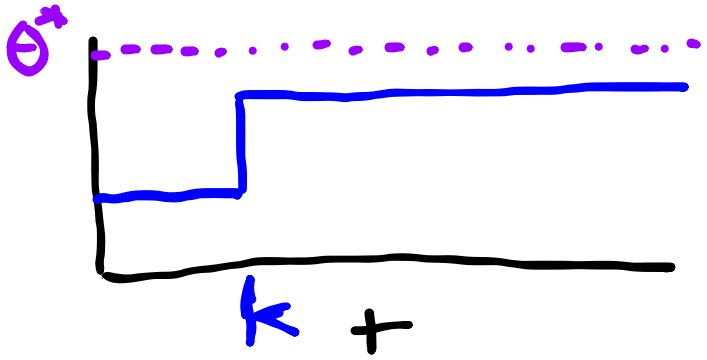
Undirected Strategies

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



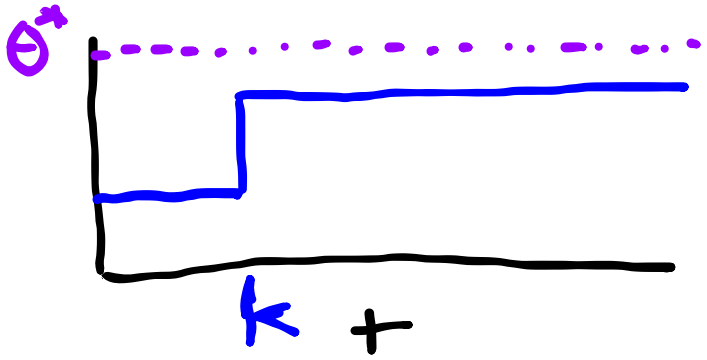
Undirected Strategies

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



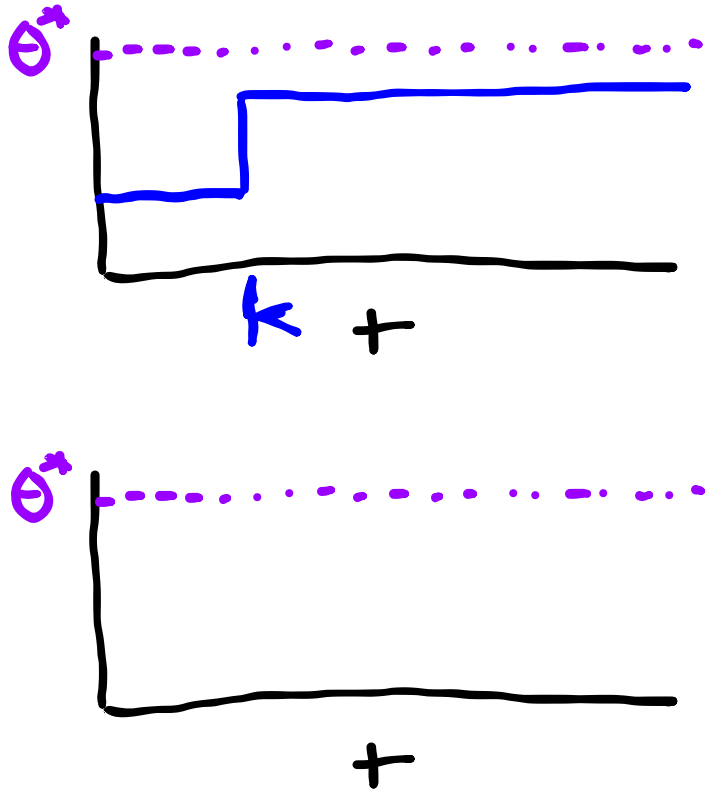
Undirected Strategies

- 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 $\operatorname{argmax}_a \rho_a$



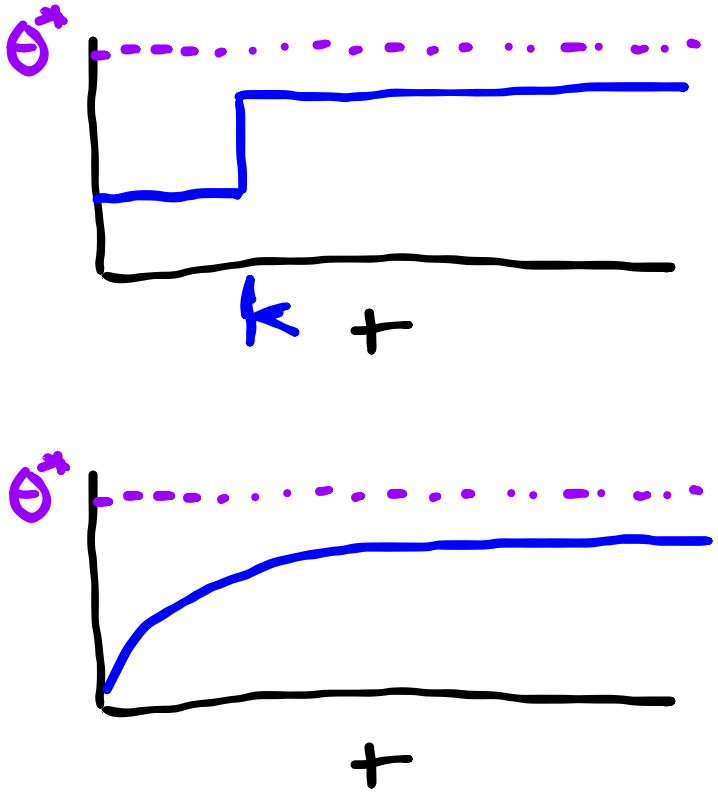
Undirected Strategies

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



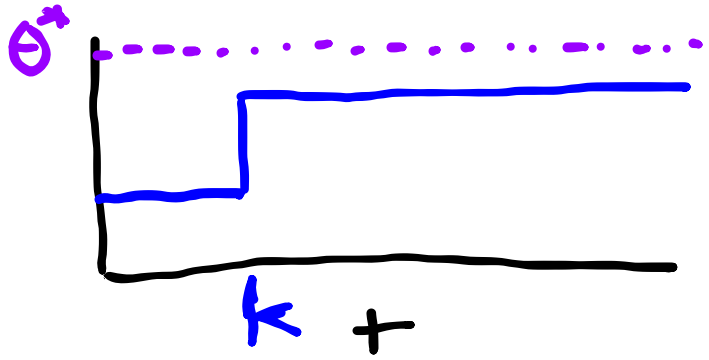
Undirected Strategies

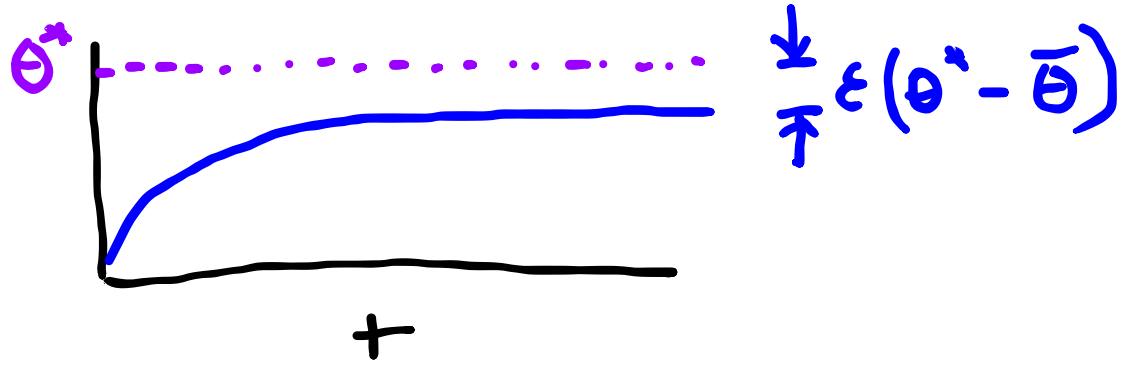
- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$
- ϵ - greedy
  With probability ϵ, choose randomly
  Otherwise choose $\underset{a}{\operatorname{argmax}}\, \rho_a$



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





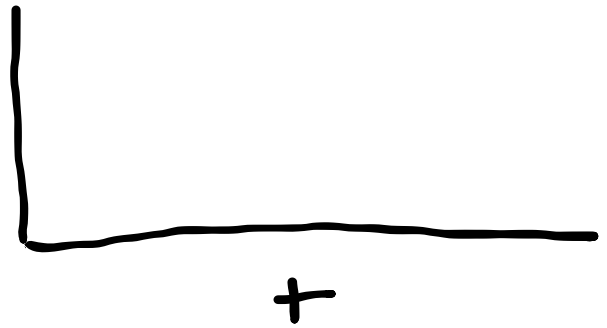
Directed Strategies (Non-Bayesian)

Directed Strategies

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

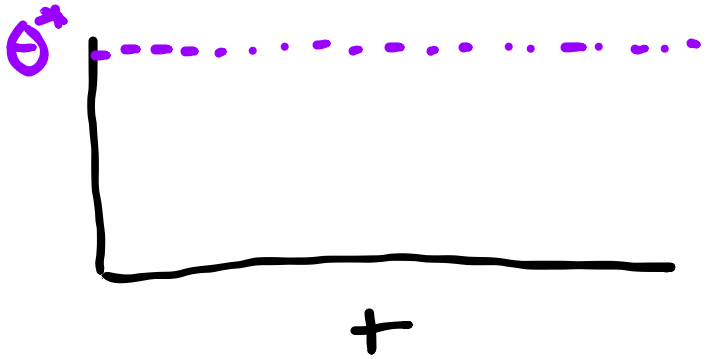
Directed Strategies

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



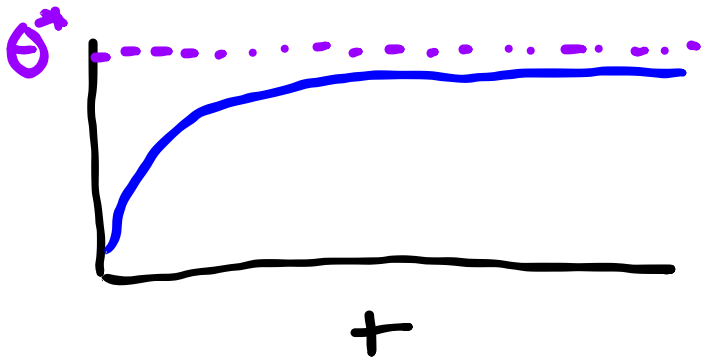
Directed Strategies

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



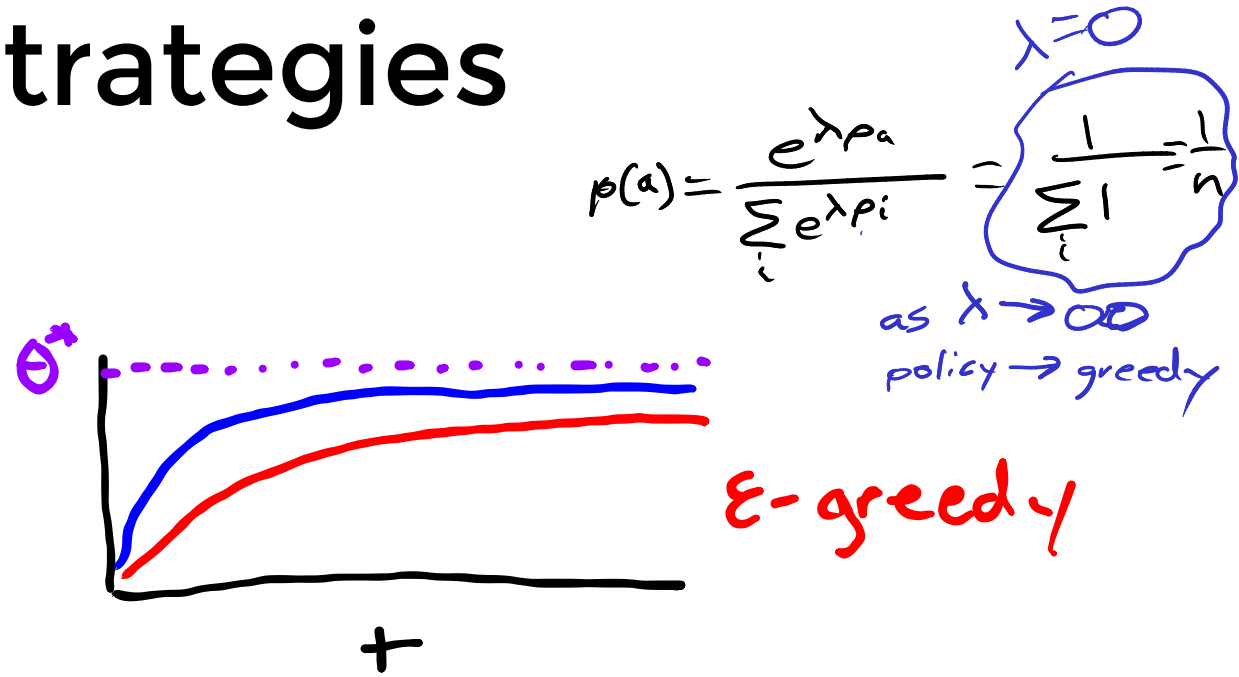
Directed Strategies

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



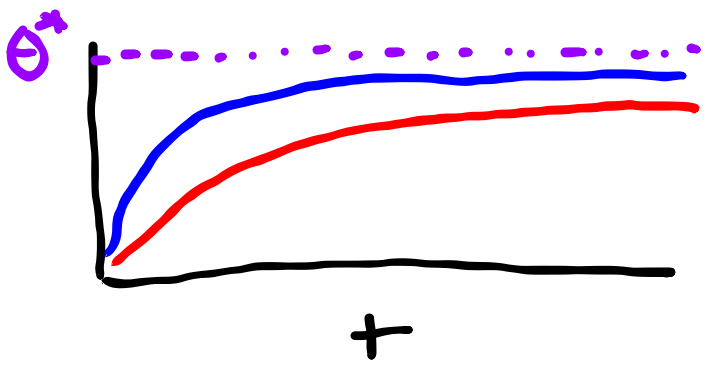
Directed Strategies

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



Directed Strategies

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



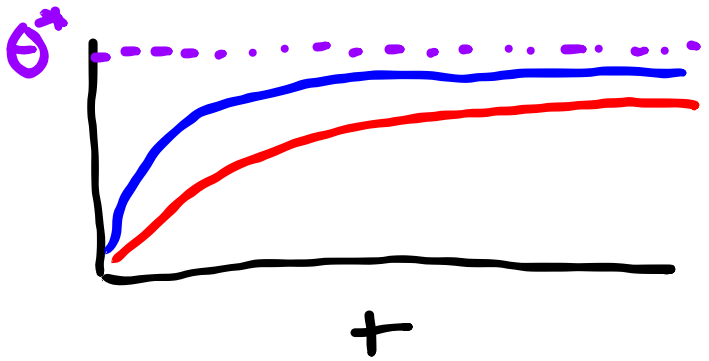
(remove gap with $\lambda \to \infty$)

Directed Strategies

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

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

Exploration



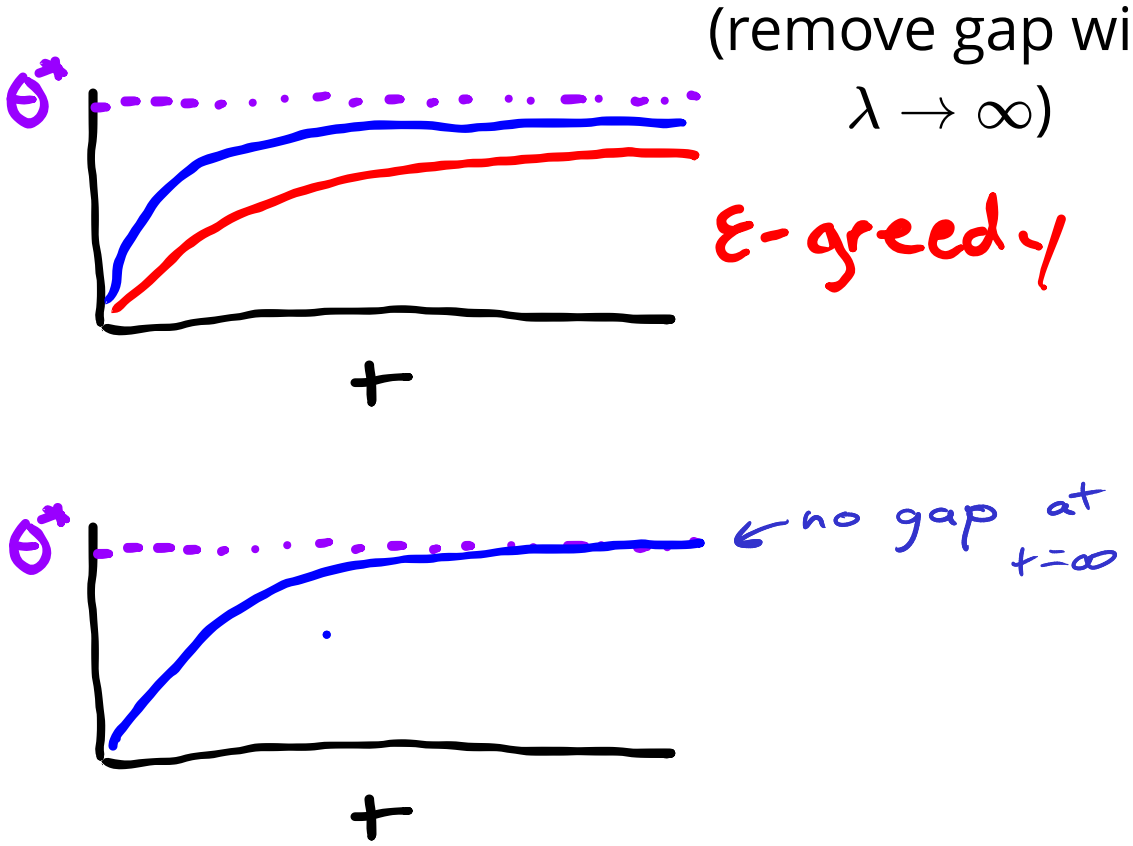
(remove gap with $\lambda \to \infty$)

ε-greedy

Directed Strategies

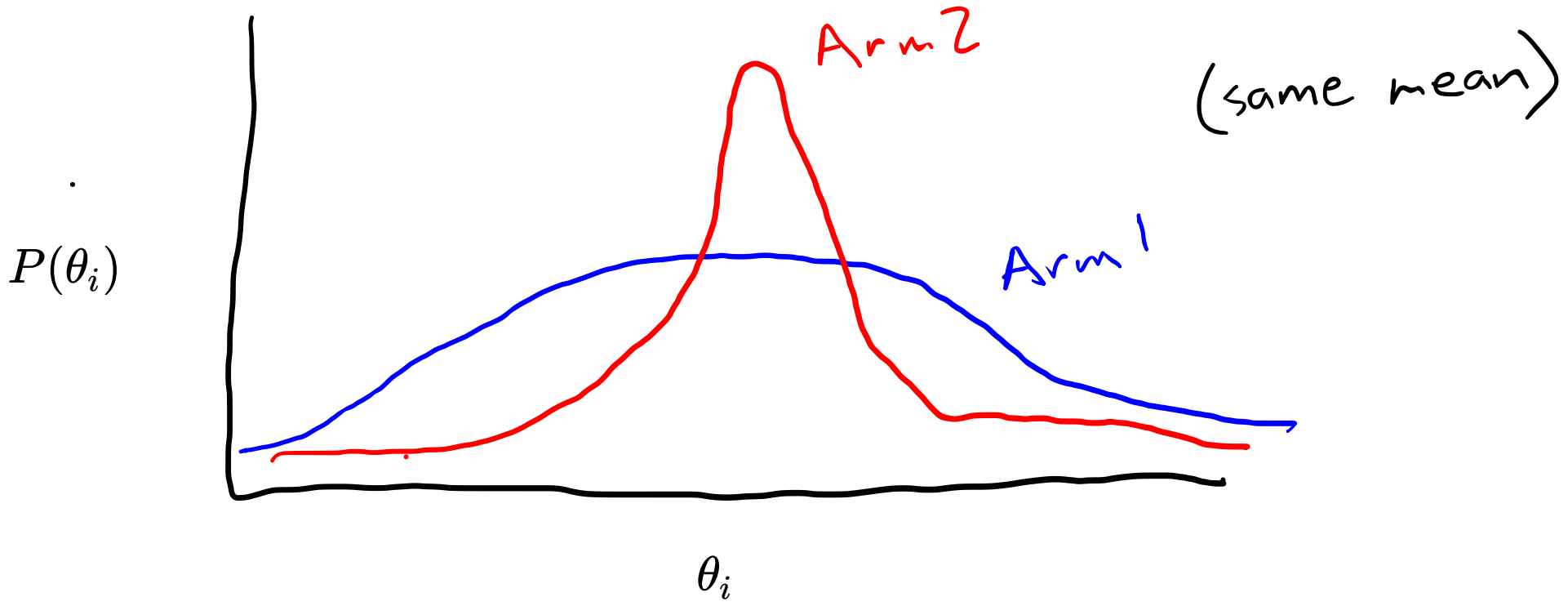
- Softmax
  Choose a with probability proportional to $e^{\lambda \rho_a}$

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



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

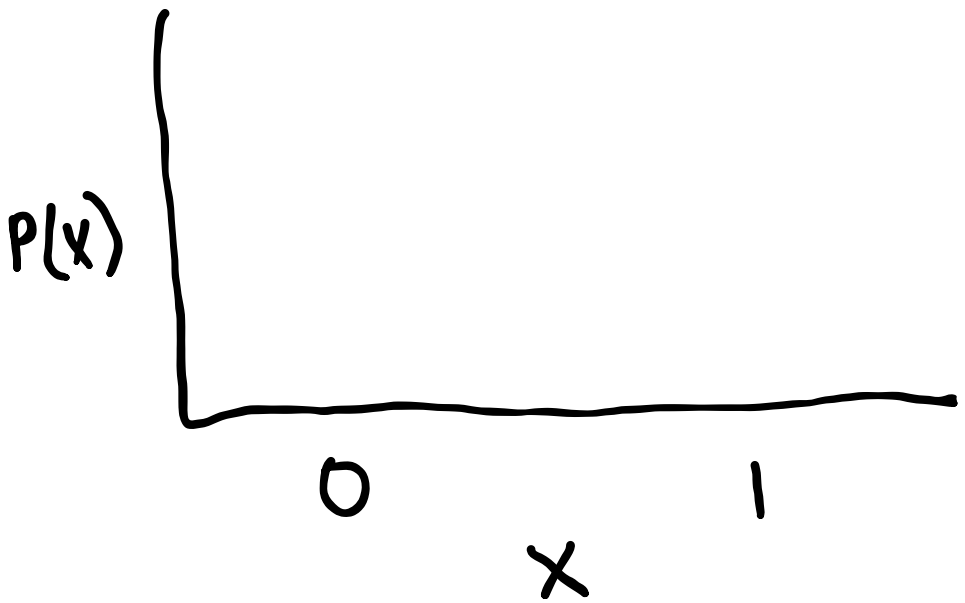
Bernoulli Distribution

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

Bayesian Estimation

Bernoulli Distribution

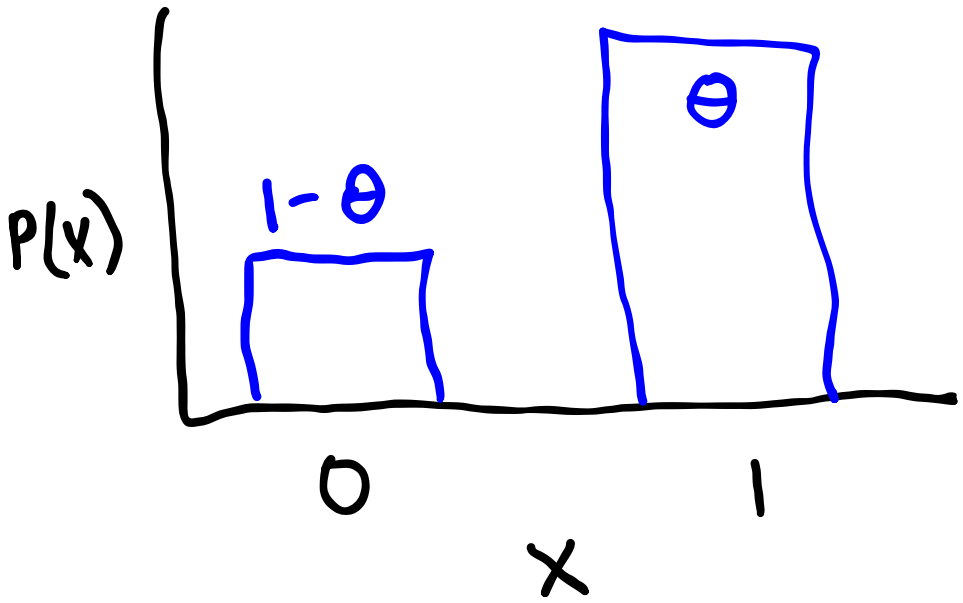
$$\text{Bernoulli}(\theta)$$



Bayesian Estimation

Bernoulli Distribution

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

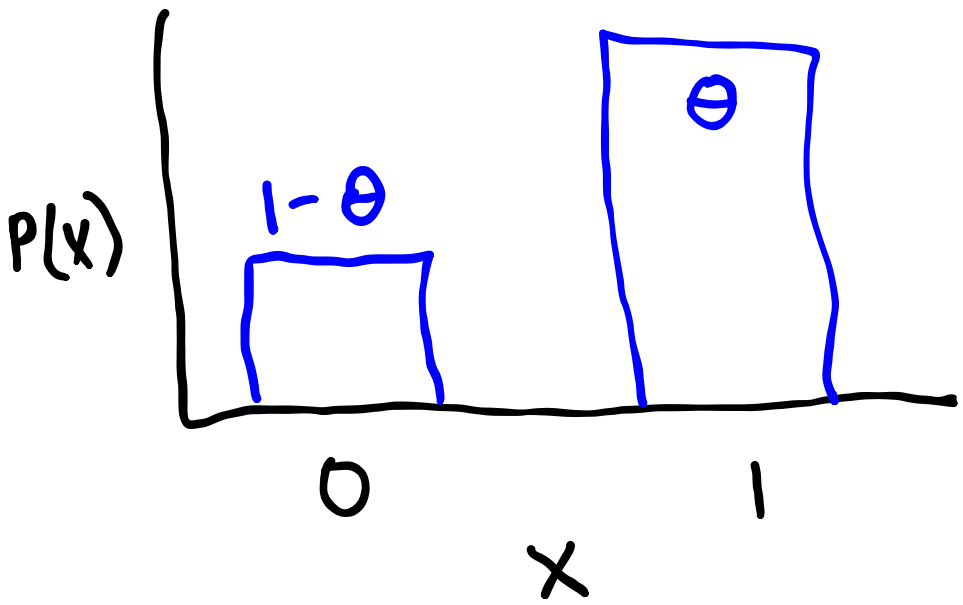


Bayesian Estimation

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?

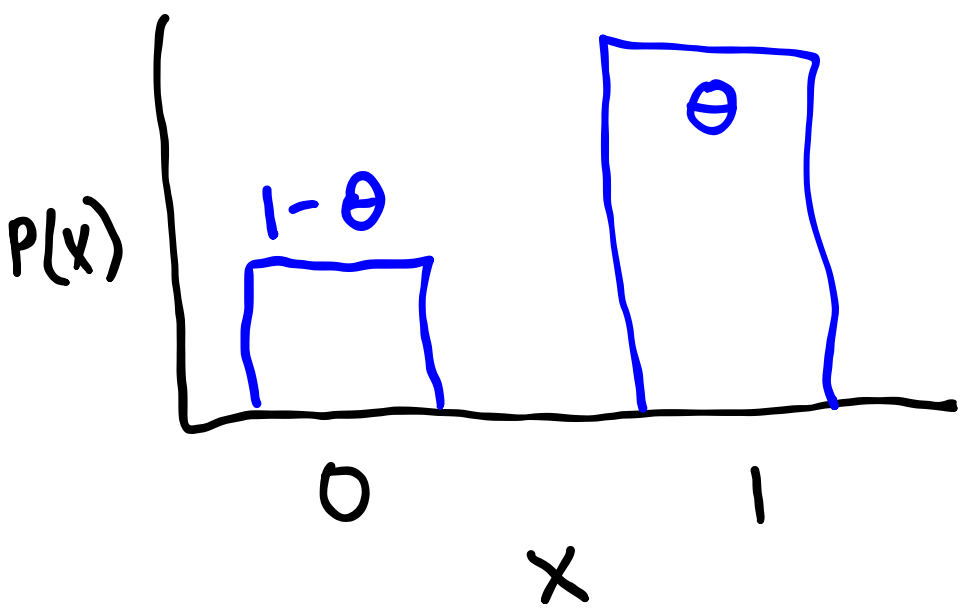


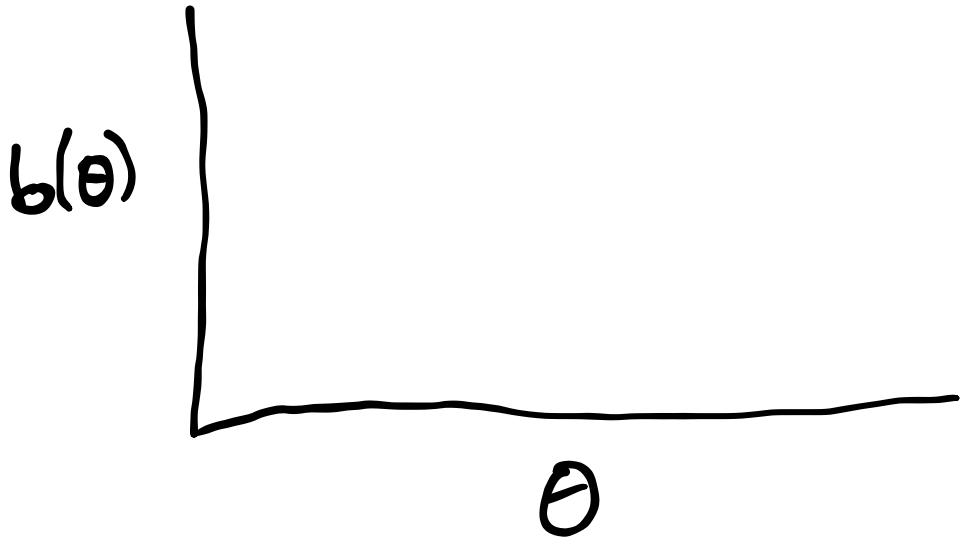
Bayesian Estimation

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?



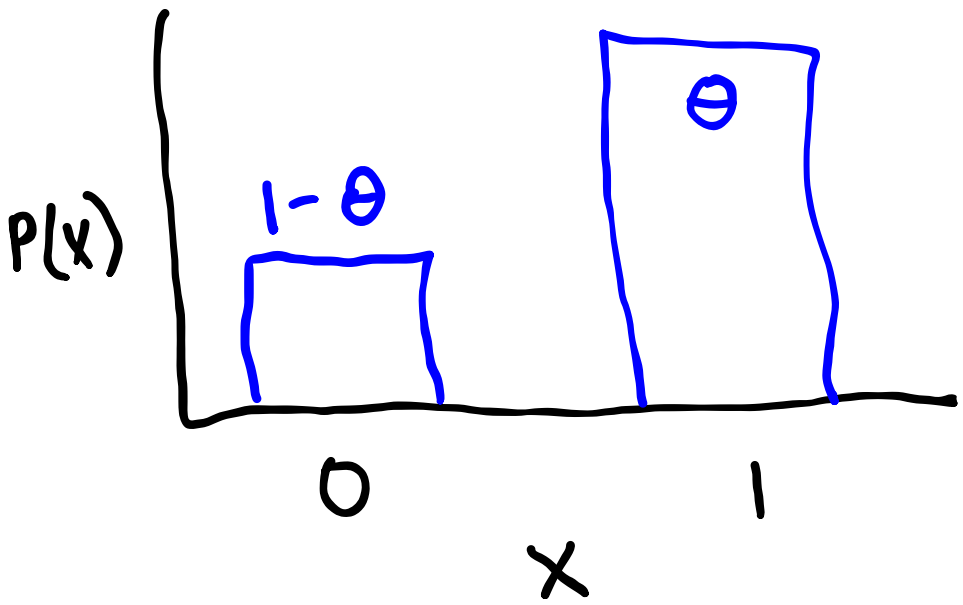


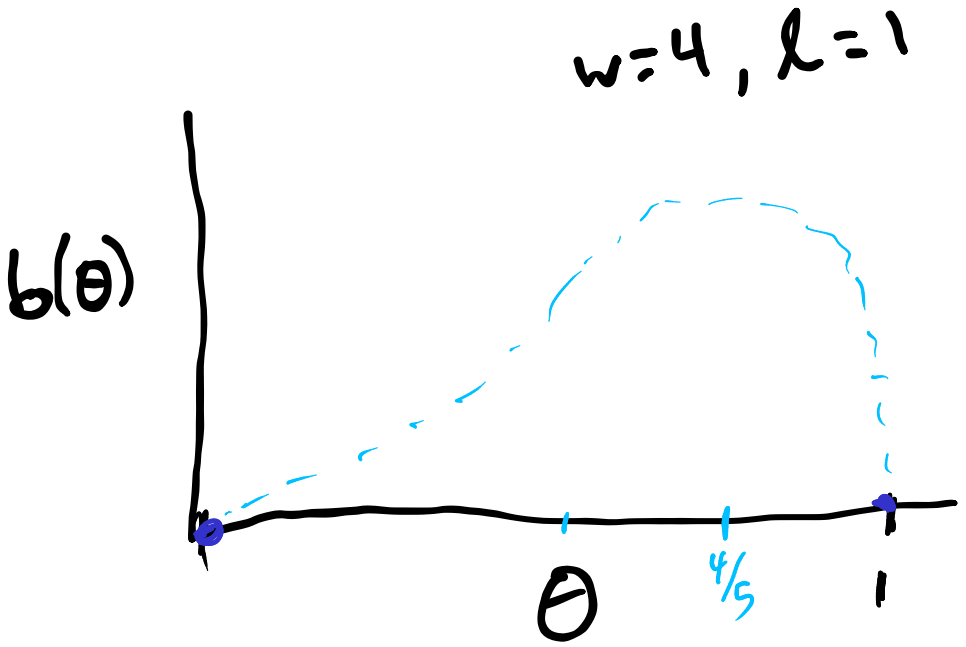
Bayesian Estimation

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?

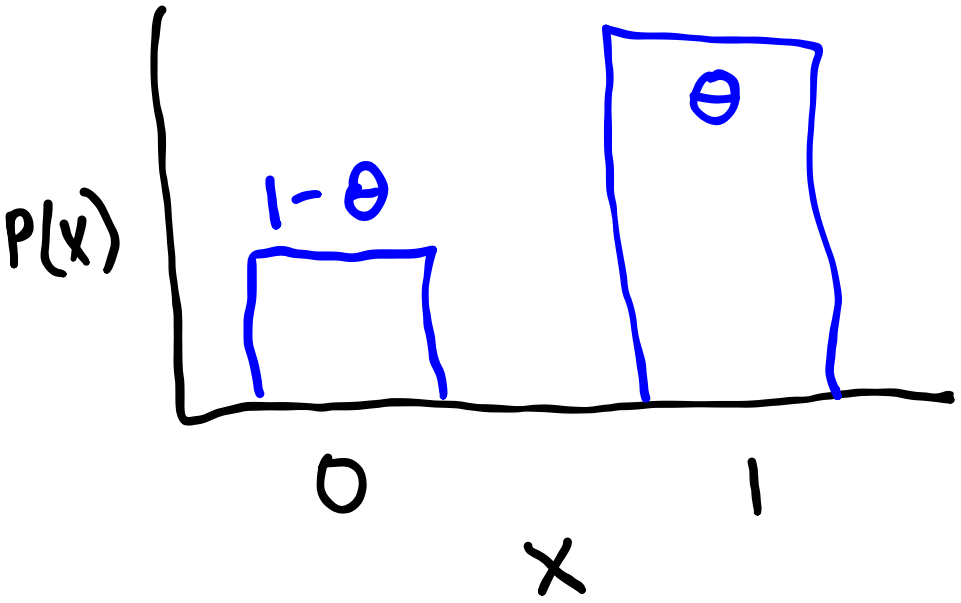




Bayesian Estimation

Bernoulli Distribution

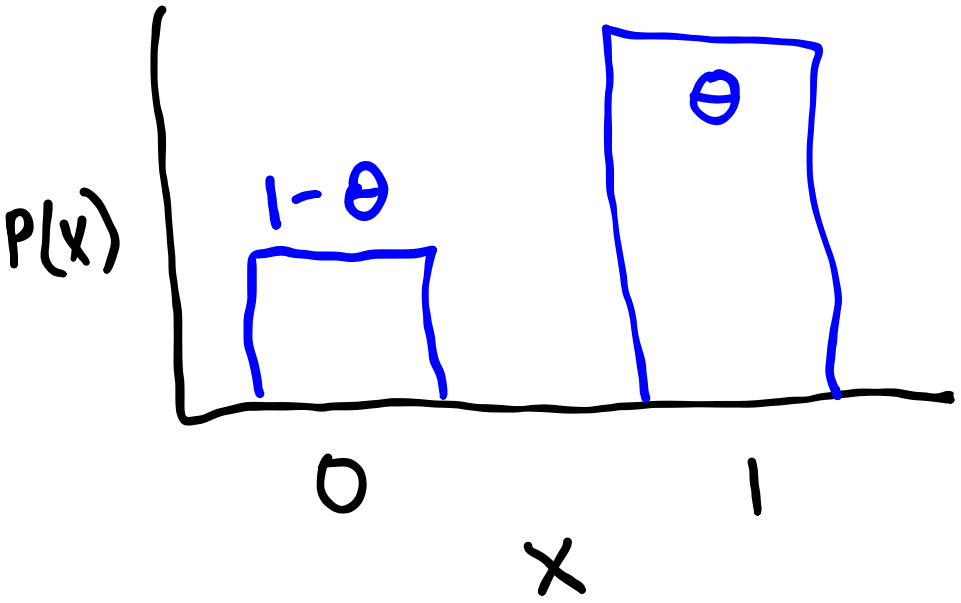
$$\text{Bernoulli}(\theta)$$



Bayesian Estimation

Bernoulli Distribution

$\text{Bernoulli}(\theta)$



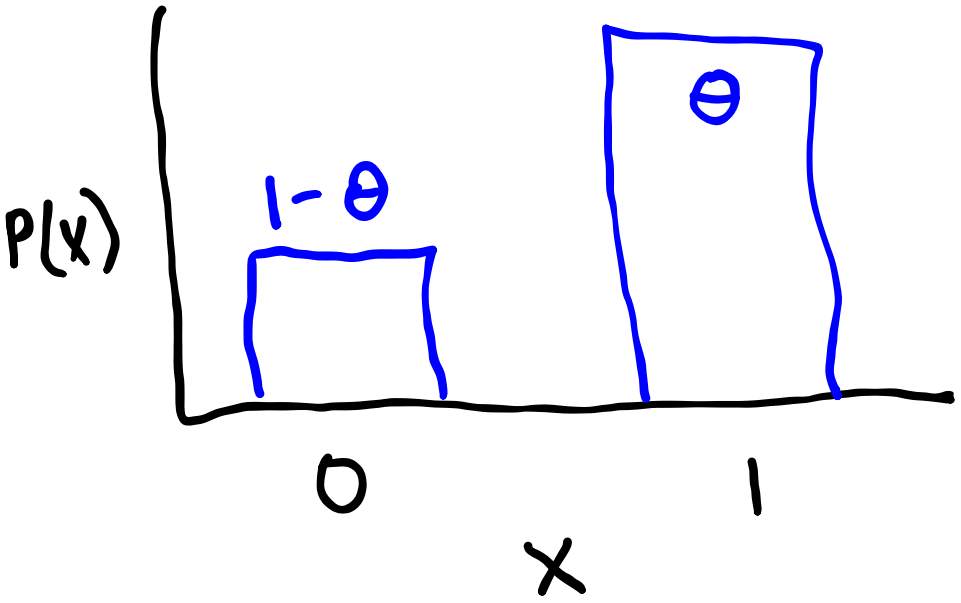
Beta Distribution

(distribution over Bernoulli distributions)

Bayesian Estimation

Bernoulli Distribution

$\text{Bernoulli}(\theta)$



Beta Distribution

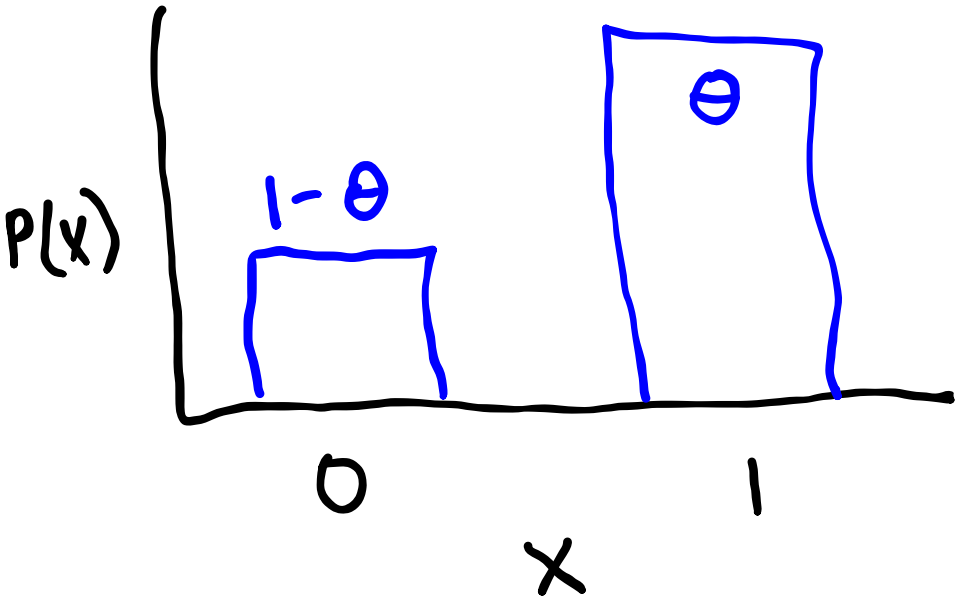
(distribution over Bernoulli distributions)

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

Bayesian Estimation

Bernoulli Distribution

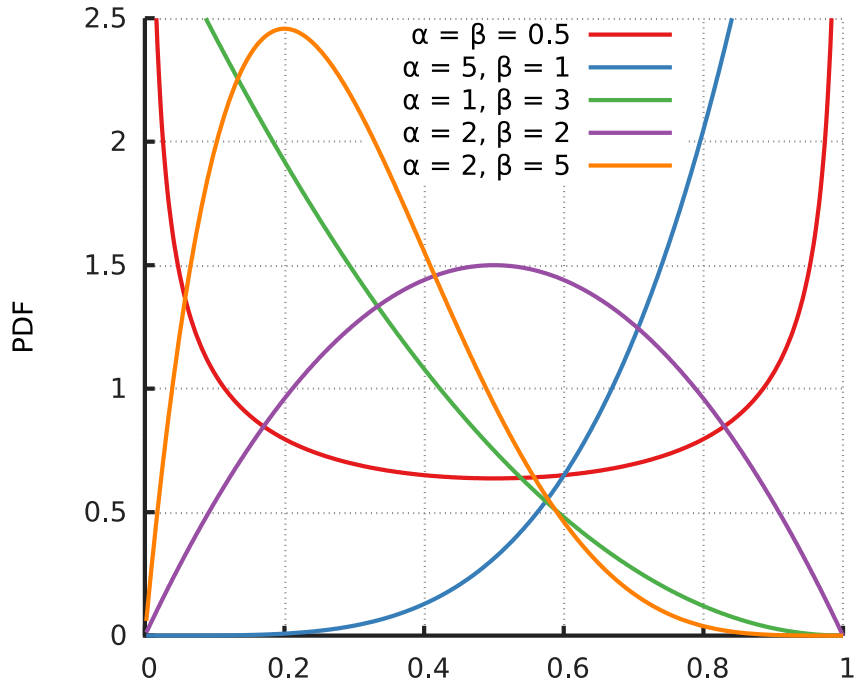
$$\text{Bernoulli}(\theta)$$



Beta Distribution

(distribution over Bernoulli distributions)

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



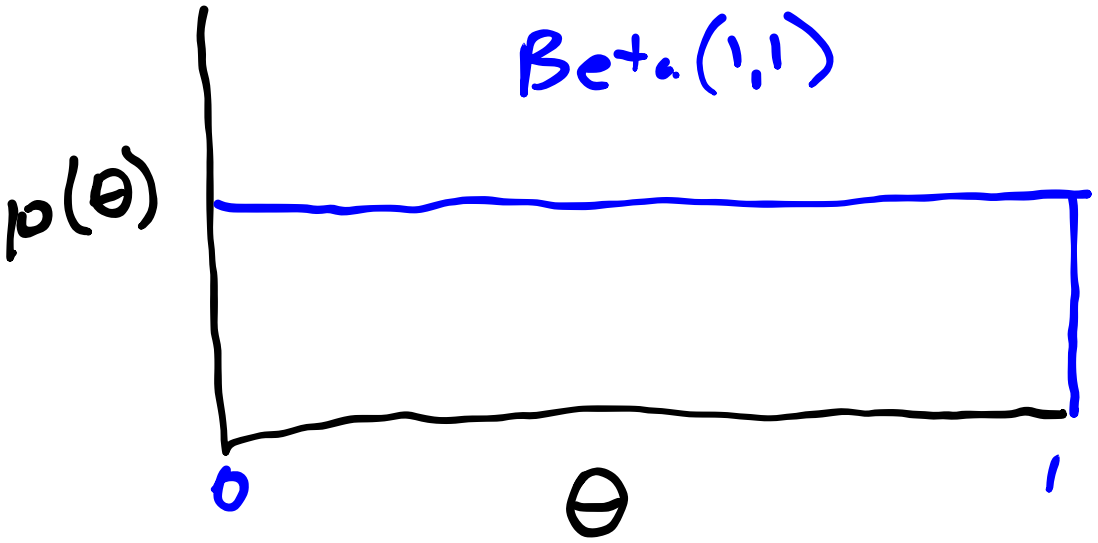
Bayesian Estimation

Bayesian Estimation

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

Bayesian Estimation

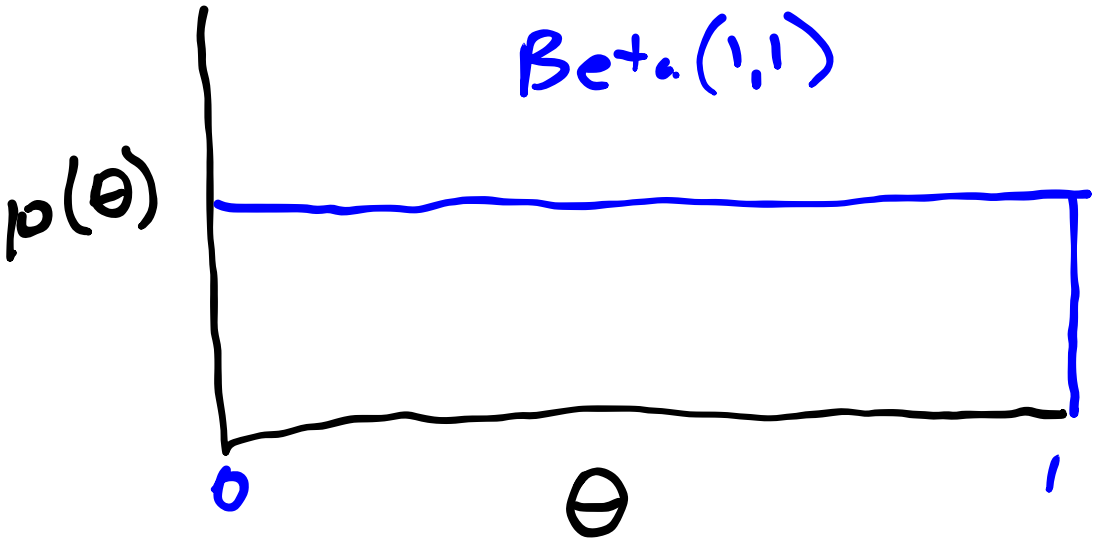
Given a $\mathrm{Beta}(1, 1)$ prior distribution



Bayesian Estimation

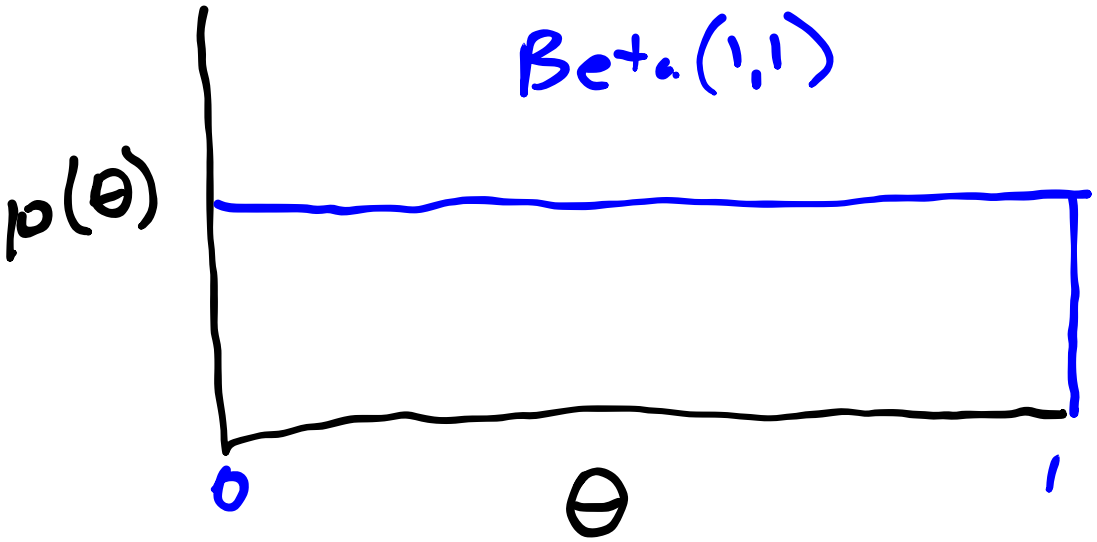
Given a $\mathrm{Beta}(1, 1)$ prior distribution

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

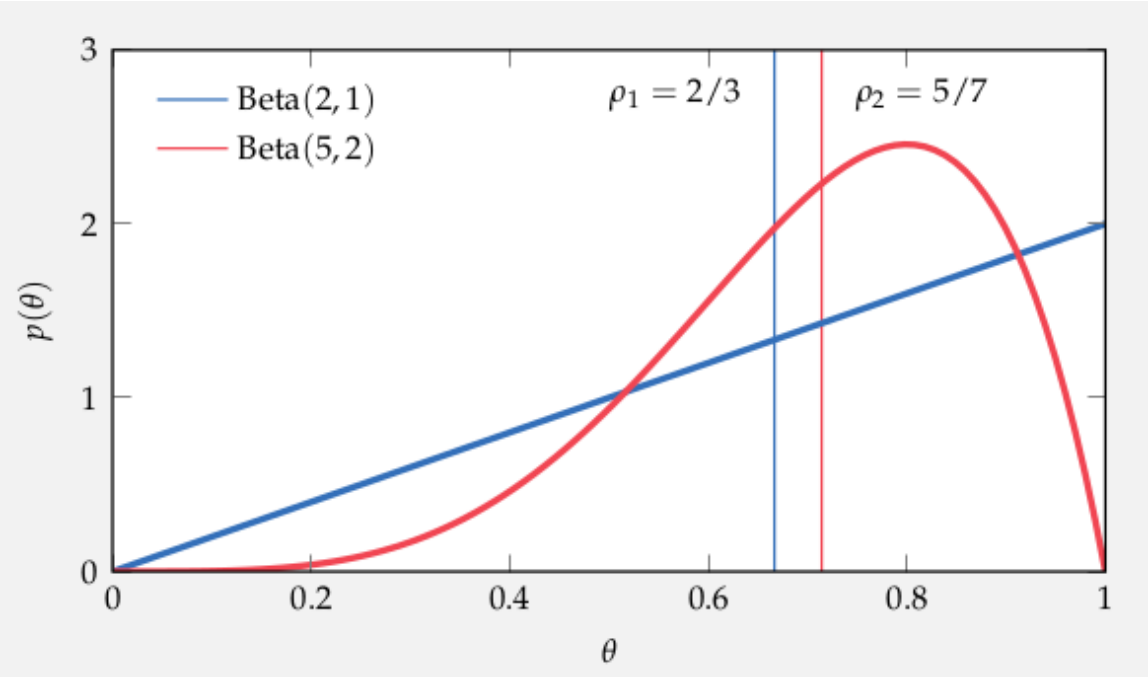


Bayesian Estimation

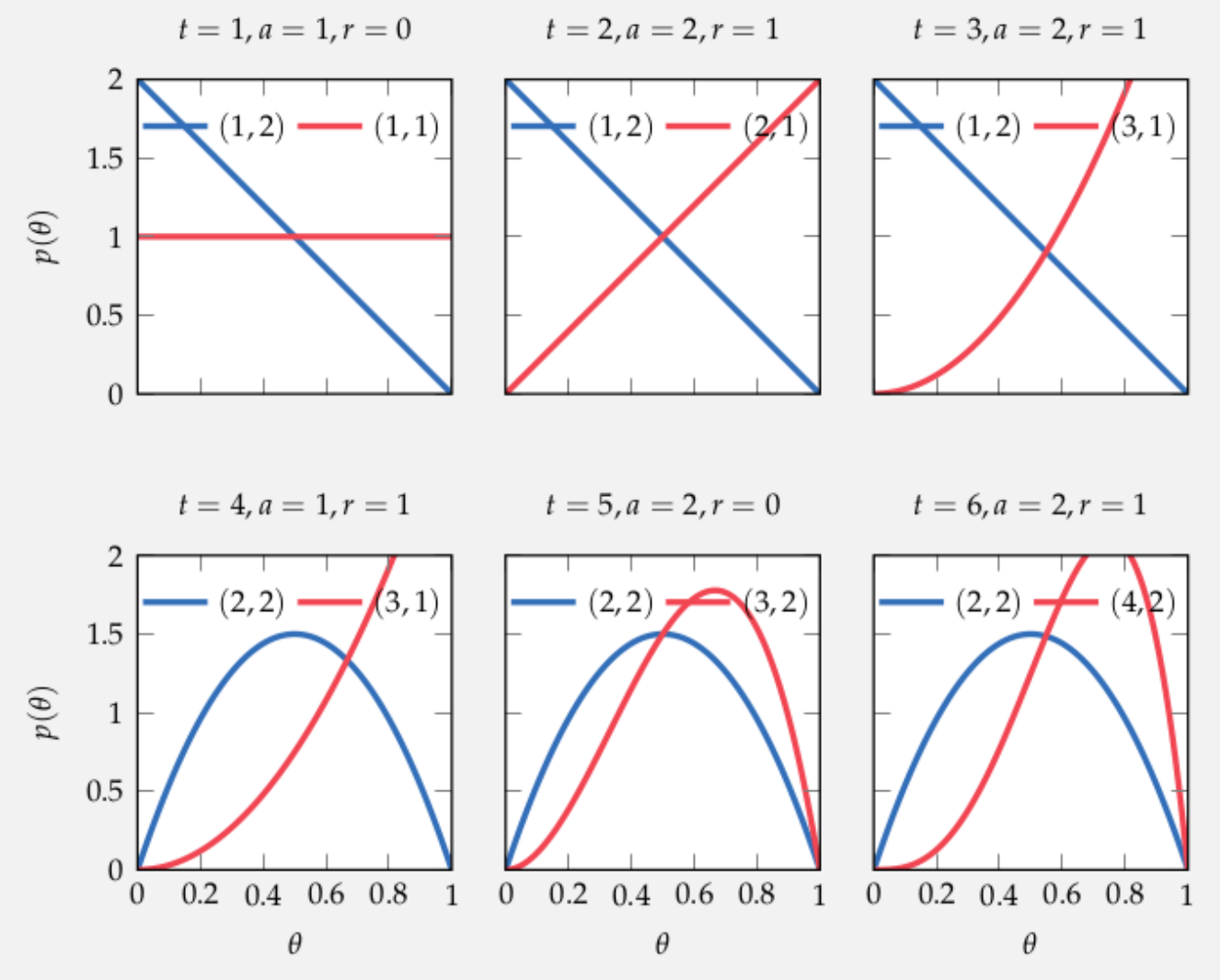
Given a $\mathrm{Beta}(1,1)$ prior distribution



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

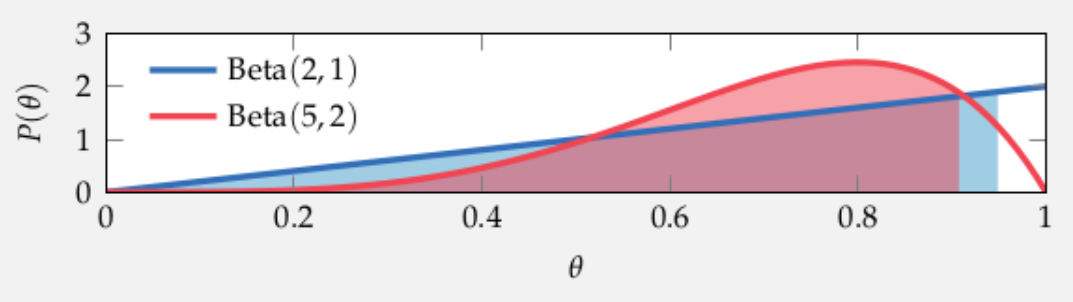
- Quantile Selection
  Choose a for which the α quantile of $b(\theta)$ is highest

Bayesian Bandit Algorithms

higher α
more optimistic

$\alpha = 0.9$

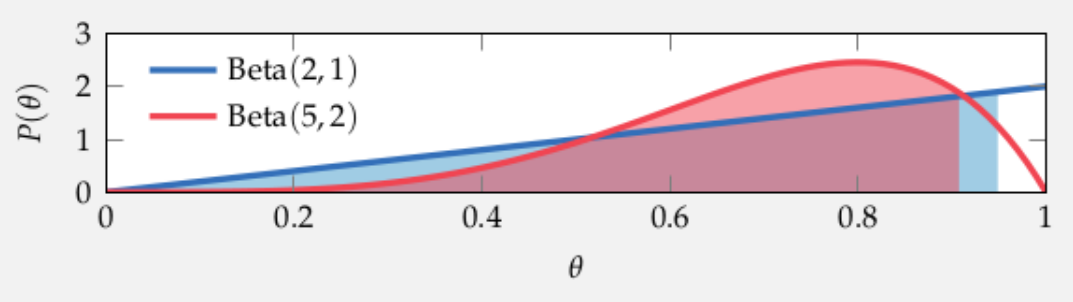
- Quantile Selection
  Choose a for which the α quantile of $b(\theta)$ is highest



Bayesian Bandit Algorithms

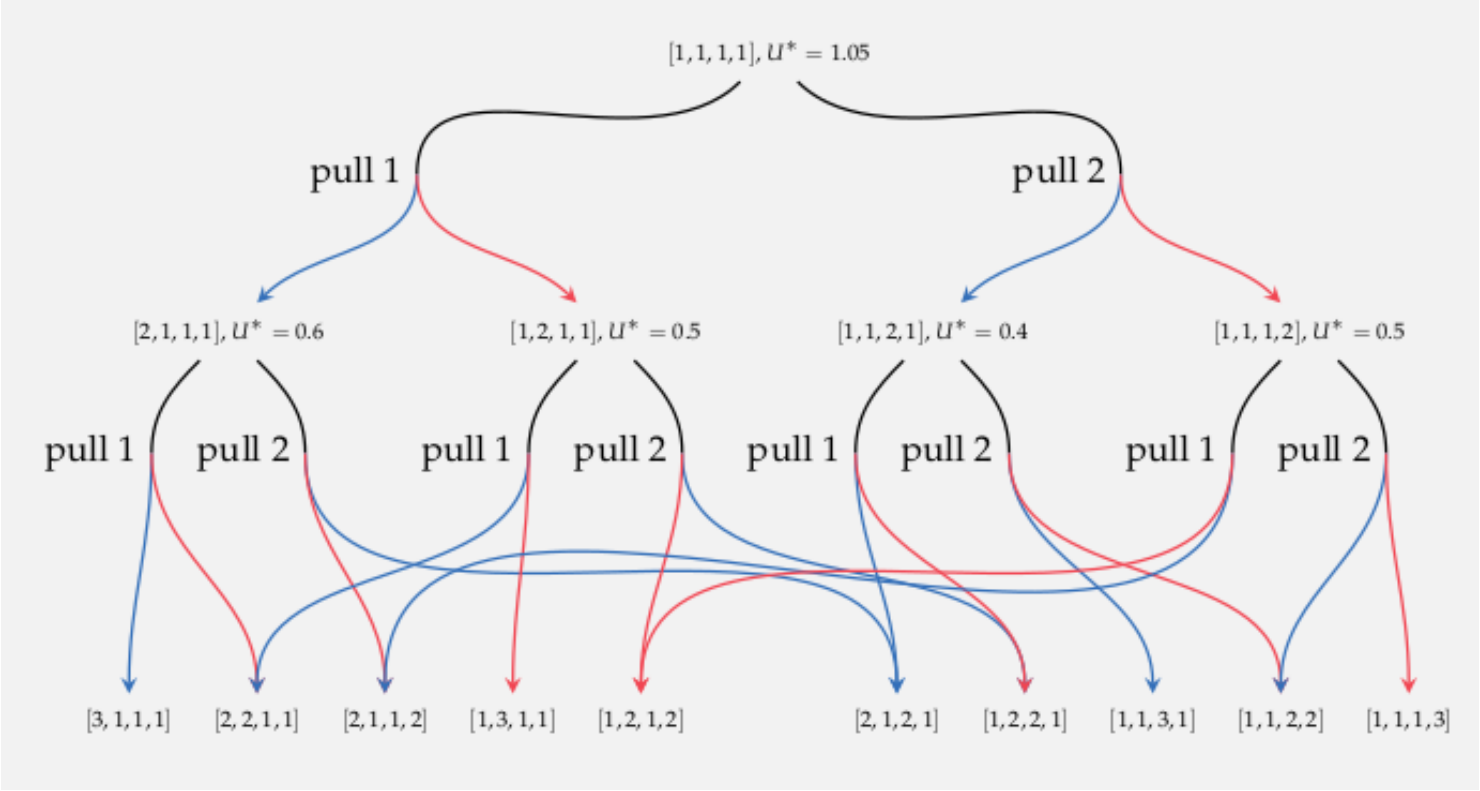
$$\alpha = 0.9$$

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



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

Optimal Algorithm - Dynamic Programming



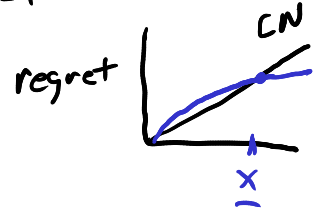
Review

Easier to Implement, Faster

Optimal in limit

for a Bernoulli Bandit

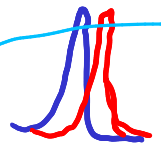
$$\text{Regret} \equiv \theta^* N - \sum_{t=1}^{N} r_t$$



		Optimal in limit	Regret
Undirected	-greedy	No	$O(N)$
	-εgreedy	$\varepsilon \to 0$	$O(N)$
	-explore-commit	$k \to \infty$	$O(N)$
	-softmax	$\lambda \to \infty$	$O(N)$
	-UCB	✓	$O(\log(N))$
Bayesian	-Quantile Selection	✓	"
	-Thompson Sampling	✓	"
	-Dynamic Programming	✓	

softmax: $\propto e^{\lambda p_a}$

UCB: $p_a + c\sqrt{\frac{\log N}{N(a)}}$



Less Regret

$g(x) = O(f(x))$

$\exists\, C$ s.t. $g(x) \le C f(x)$ $\forall x > \forall$

Guiding Questions

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

UCB