Last Time

Last Time

T? R?

- What is Reinforcement Learning?
- What are the main challenges in Reinforcement Learning?
  - Exploration + Explaitation
  - Credit Assignmen
  - Generalization

Last Time

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

$\pi_\theta \xrightarrow{\pi_\theta} \pi_{\theta'}$

On Policy
Off Policy
Batch

Model Based
Model Free

TR
π_θ Q

Deep
Tabular

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

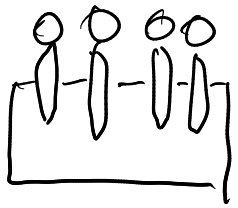
- choose action a
- gain experience
- estimate T, R
- solve MDP with T, R

Guiding Questions

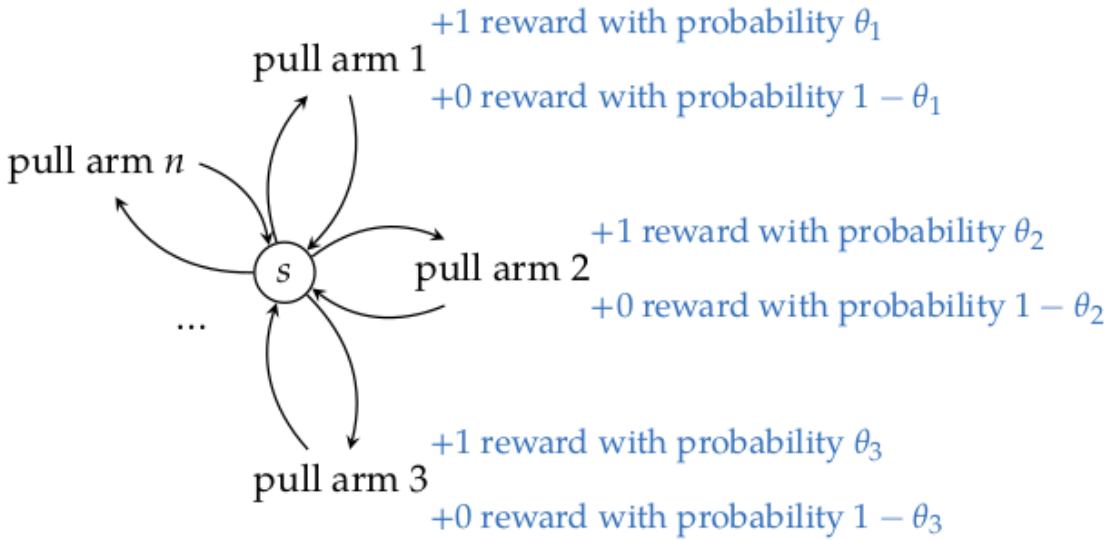
- What are the best ways to trade off Exploration and Exploitation

Bandits

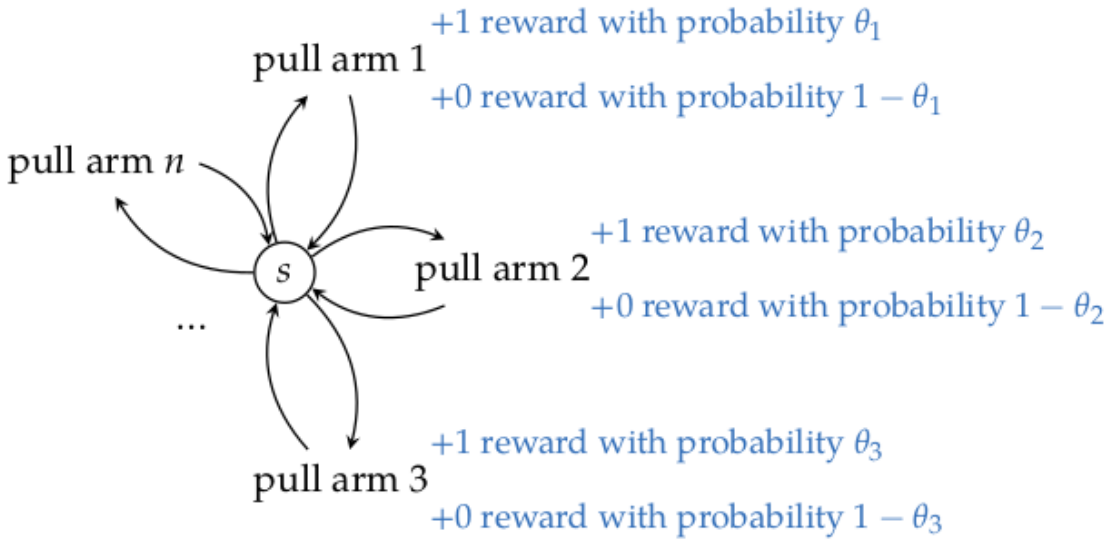
Bandits



Bandits

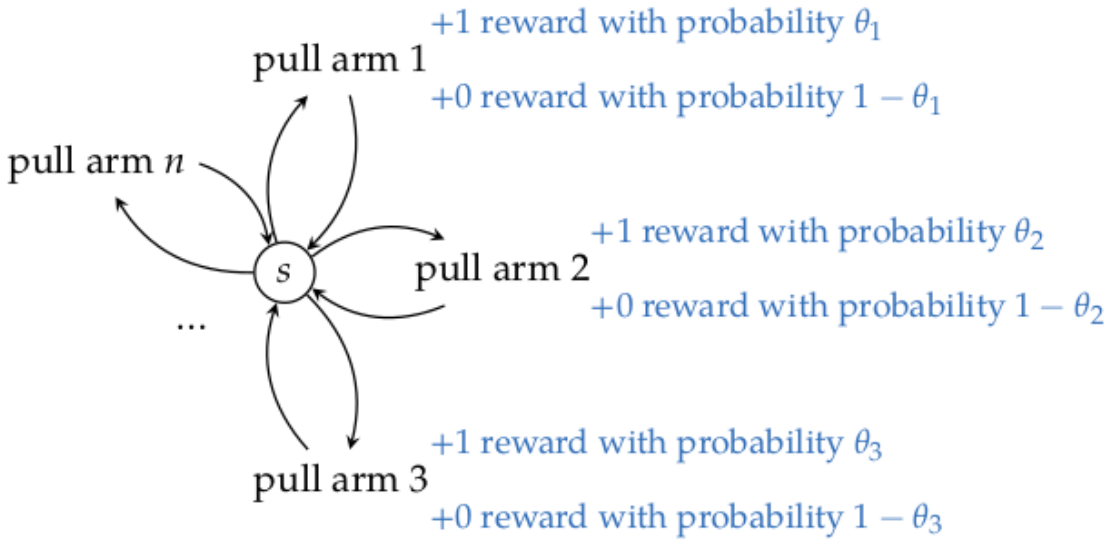

+1 reward with probability θ_1
pull arm 1
+0 reward with probability $1-\theta_1$
pull arm n
+1 reward with probability θ_2
s
pull arm 2
...
+0 reward with probability $1-\theta_2$
+1 reward with probability θ_3
pull arm 3
+0 reward with probability $1-\theta_3$

Bandits



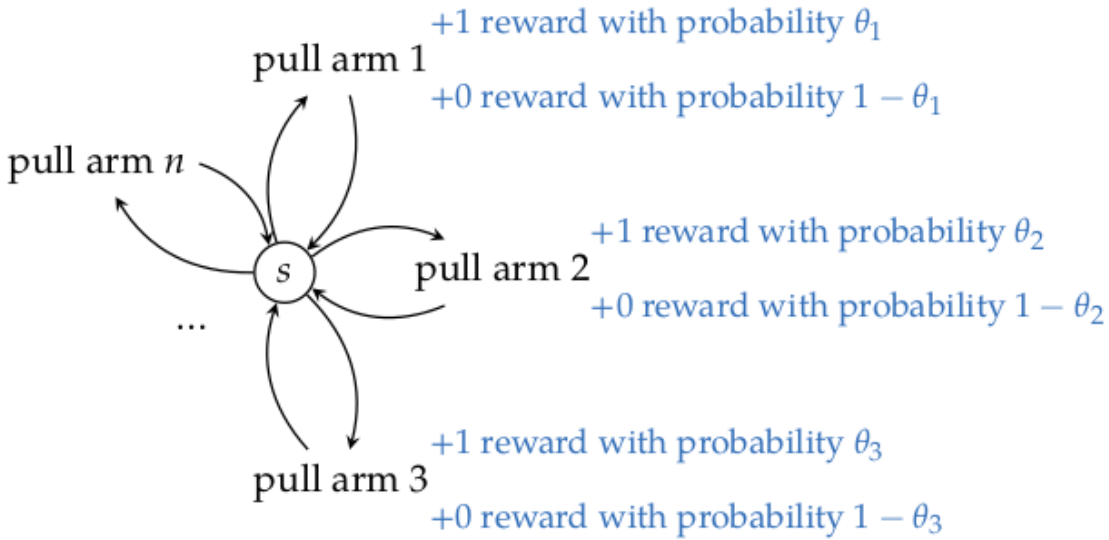
- Bernoulli Bandit with parameters θ

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

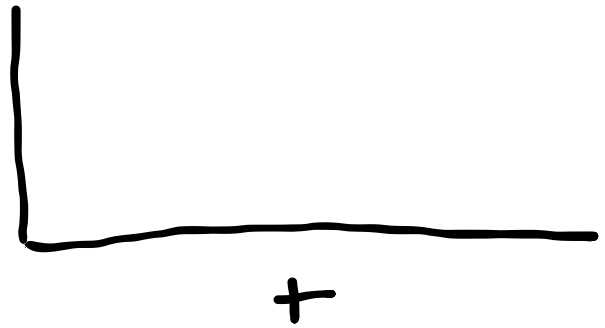
Undirected Strategies

Undirected Strategies

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

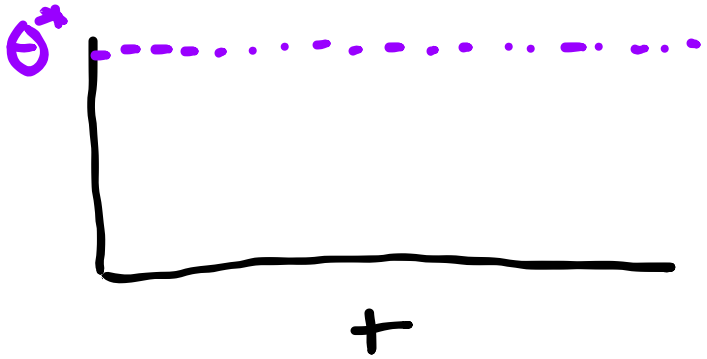
Undirected Strategies

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



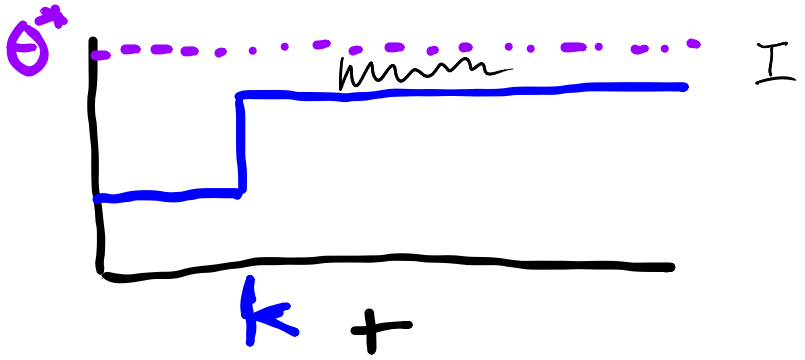
Undirected Strategies

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



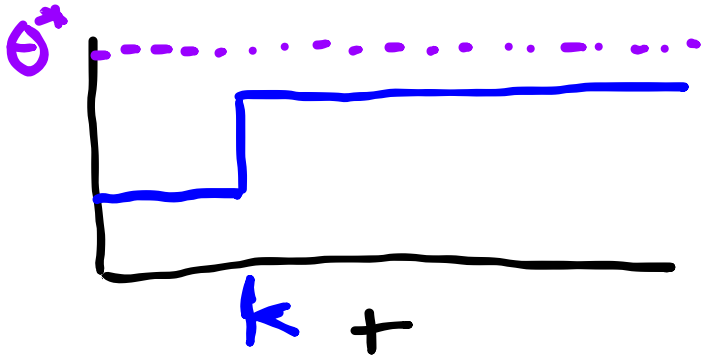
Undirected Strategies

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



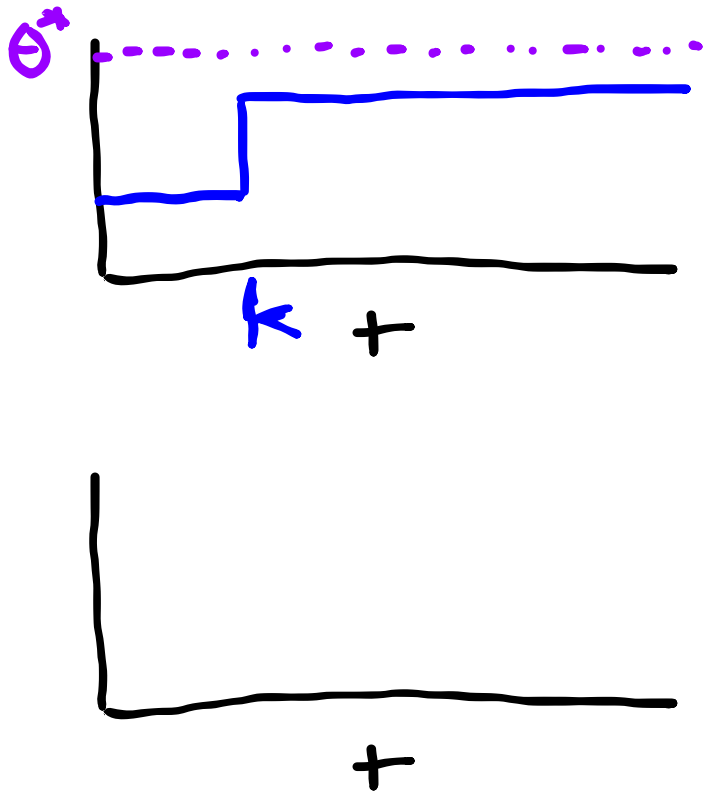
Undirected Strategies

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



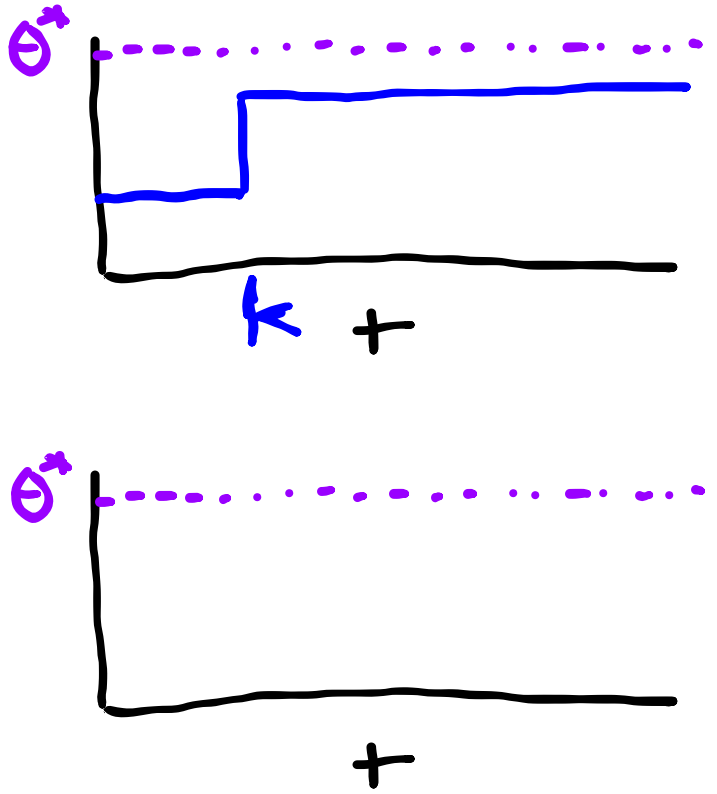
Undirected Strategies

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



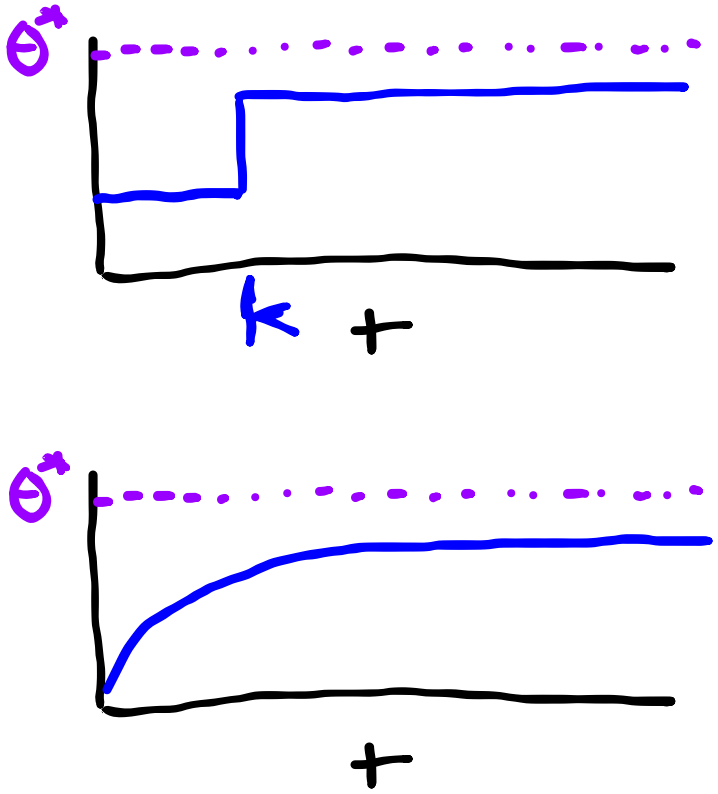
Undirected Strategies

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



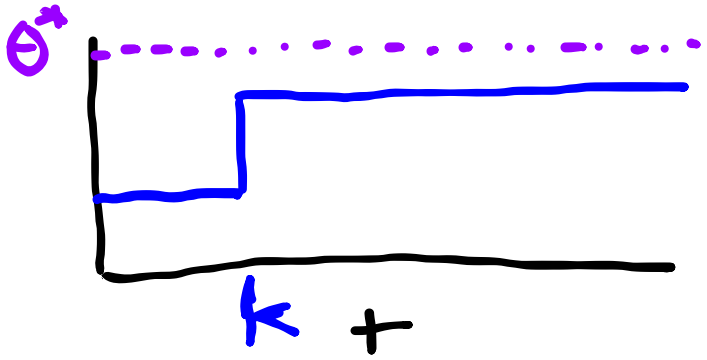
Undirected Strategies

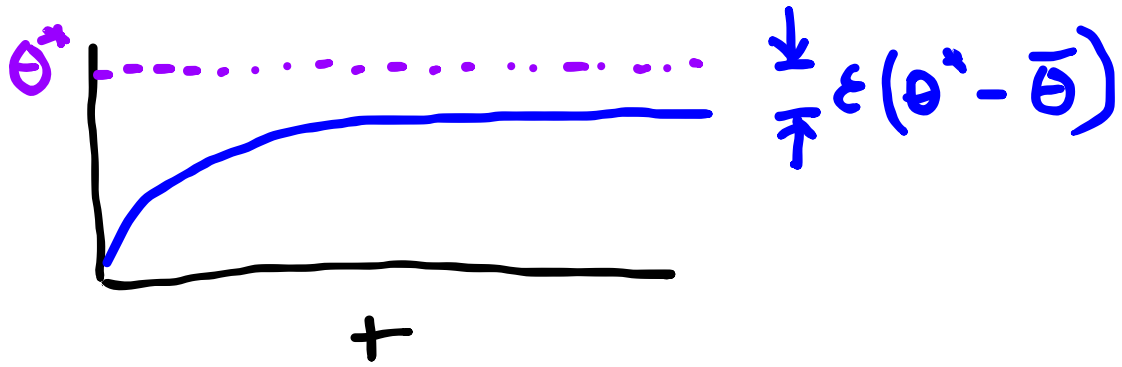
- Explore then Commit
  Choose a randomly for k steps
  Then choose $\underset{a}{\operatorname{argmax}}\, \rho_a$
- ϵ - greedy
  With probability ϵ, choose randomly
  Otherwise choose $\underset{a}{\operatorname{argmax}}\, \rho_a$

$\varepsilon \leftarrow \alpha\varepsilon$

$\alpha \in (0,1)$

close to 1





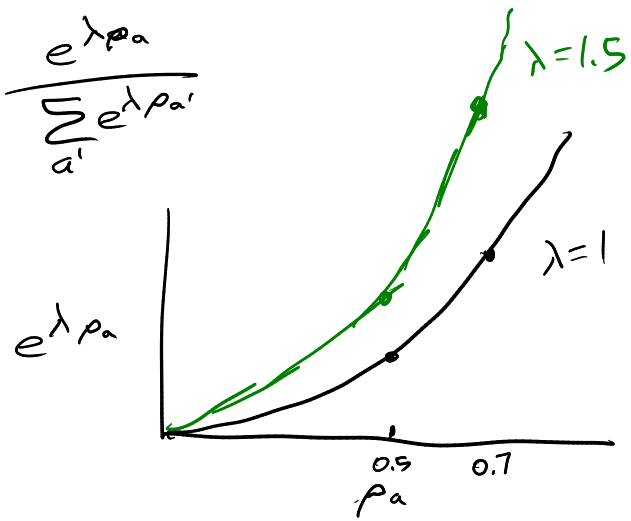
Directed Strategies

Directed Strategies

$$p(a) = \frac{e^{\lambda \rho_a}}{\sum_{a'} e^{\lambda \rho_{a'}}}$$

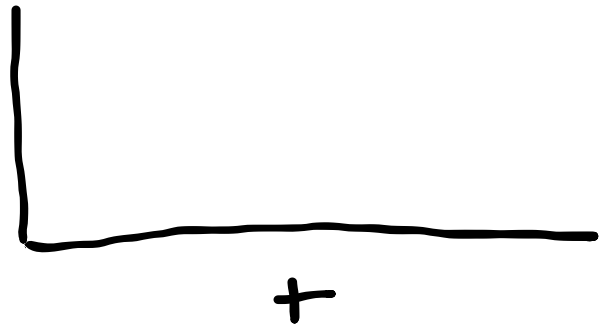
as $\lambda \to \infty$ looks like $\operatorname{argmax}_a \rho$

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



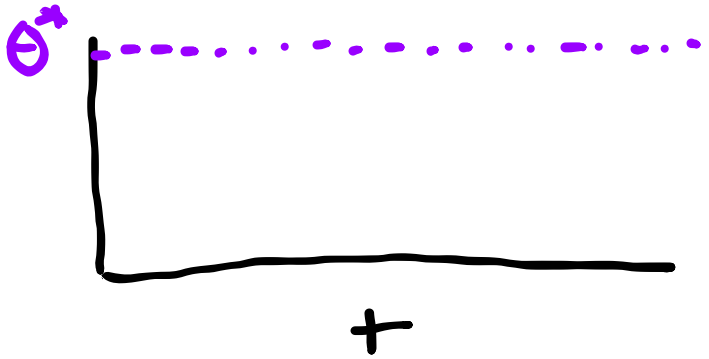
Directed Strategies

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



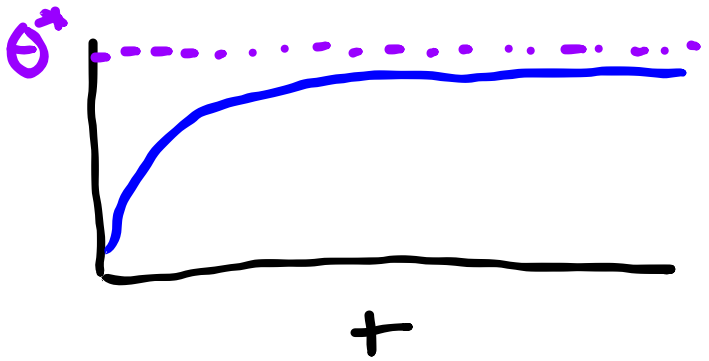
Directed Strategies

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



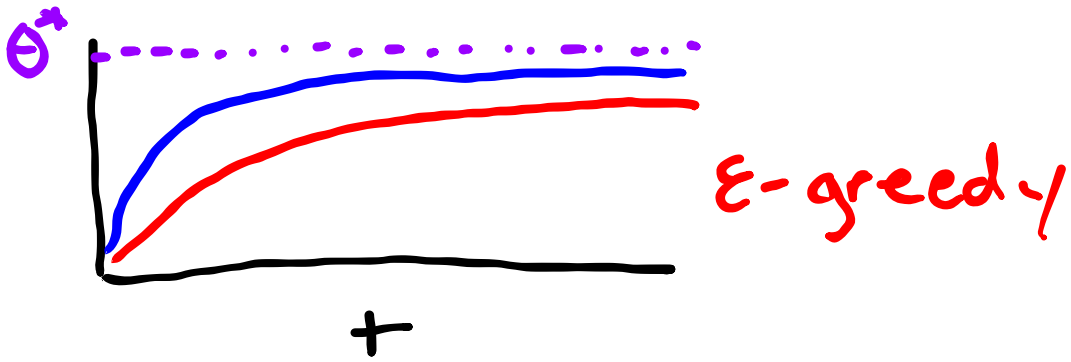
Directed Strategies

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



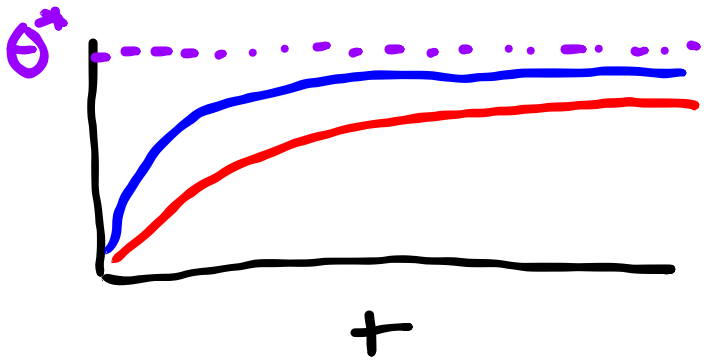
Directed Strategies

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



Directed Strategies

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

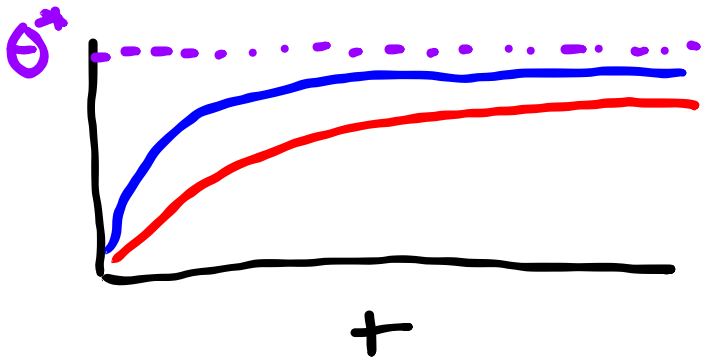


(remove gap with $\lambda \to \infty$)

Directed Strategies

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

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

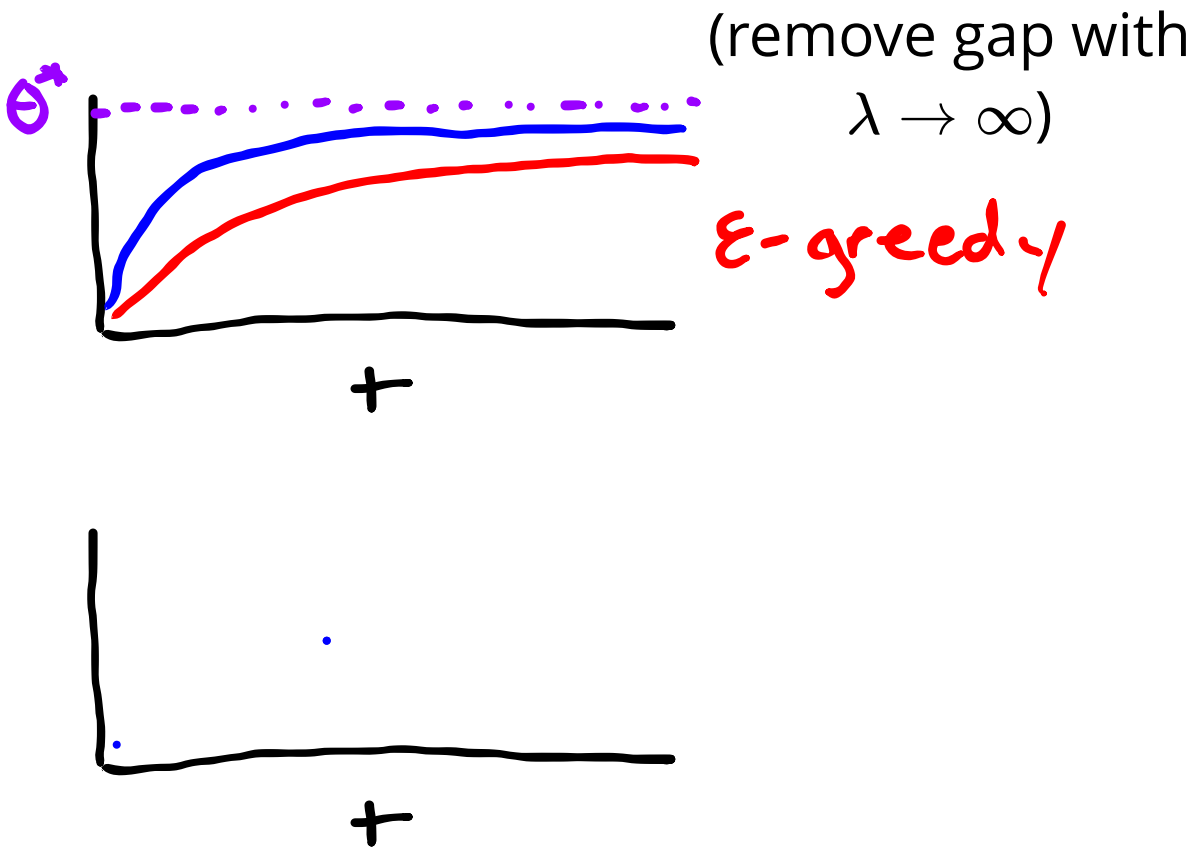


(remove gap with $\lambda \to \infty$)

Directed Strategies

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

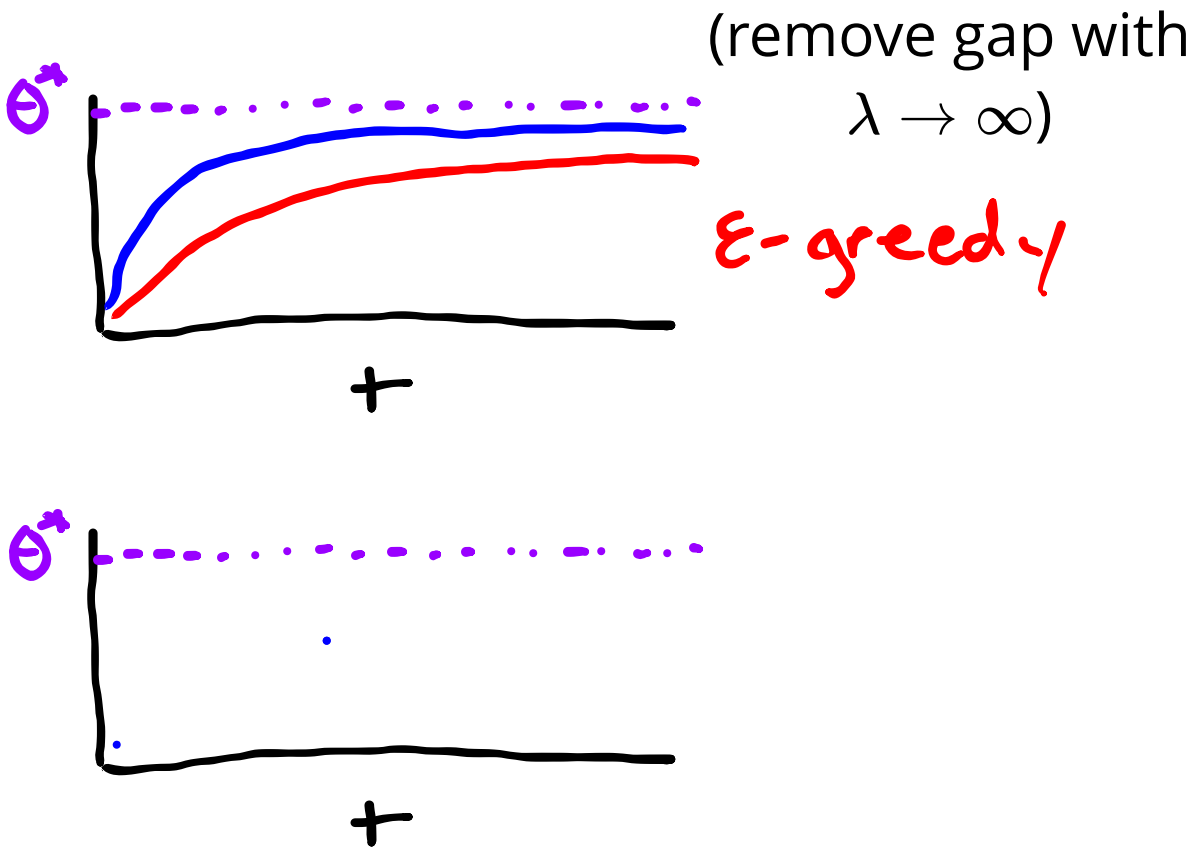
- Upper Confidence Bound (UCB)
  Choose $\underset{a}{\operatorname{argmax}}\, \rho_a + c\sqrt{\frac{\log N}{N(a)}}$



Directed Strategies

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

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

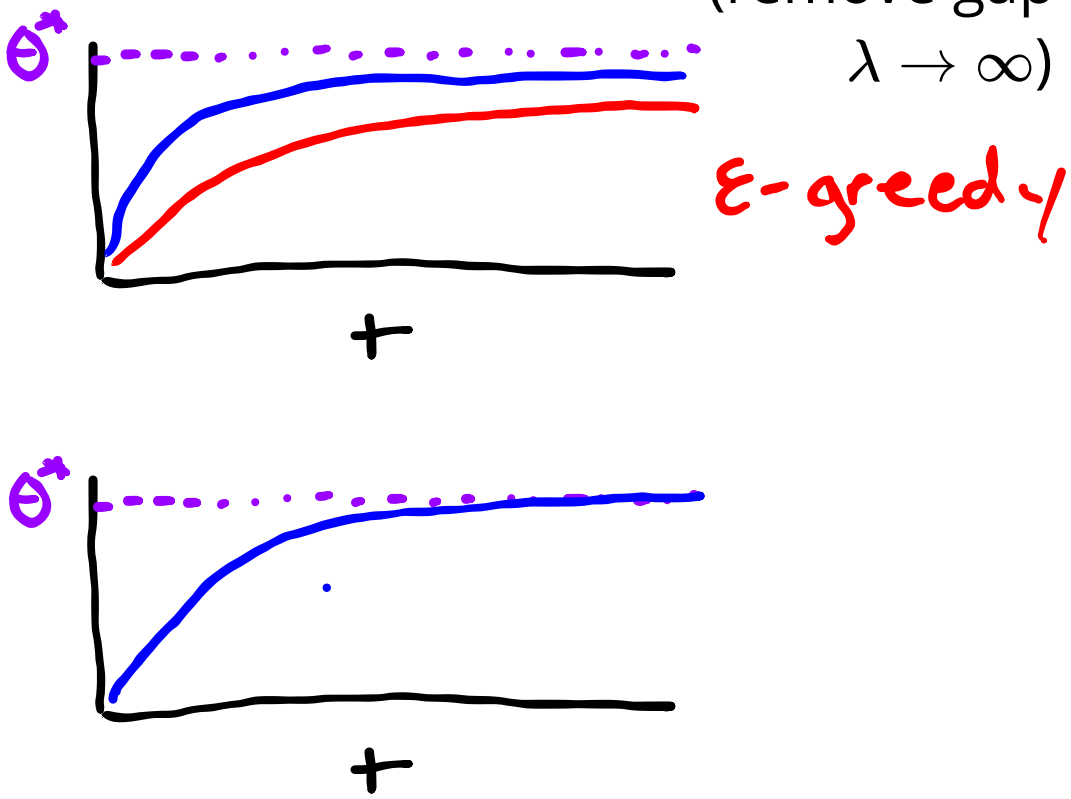


Directed Strategies

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

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

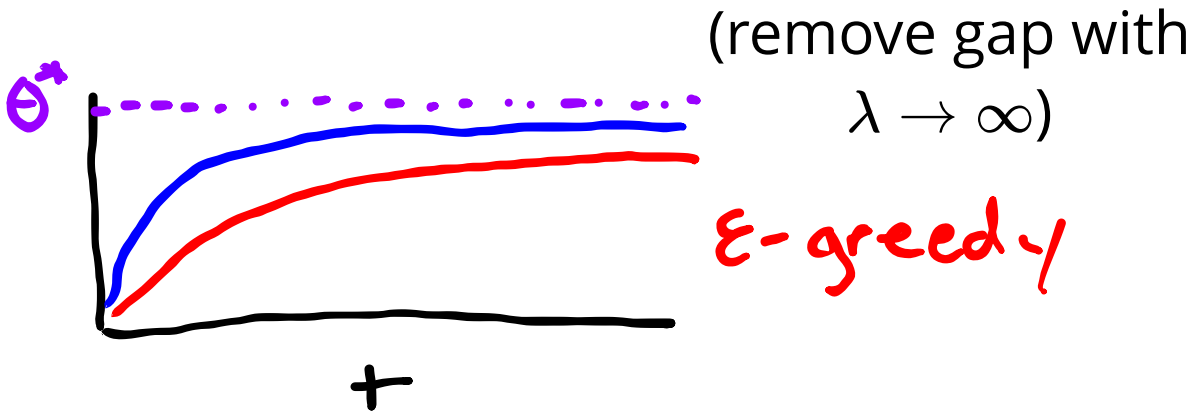
(remove gap with $\lambda \to \infty$)

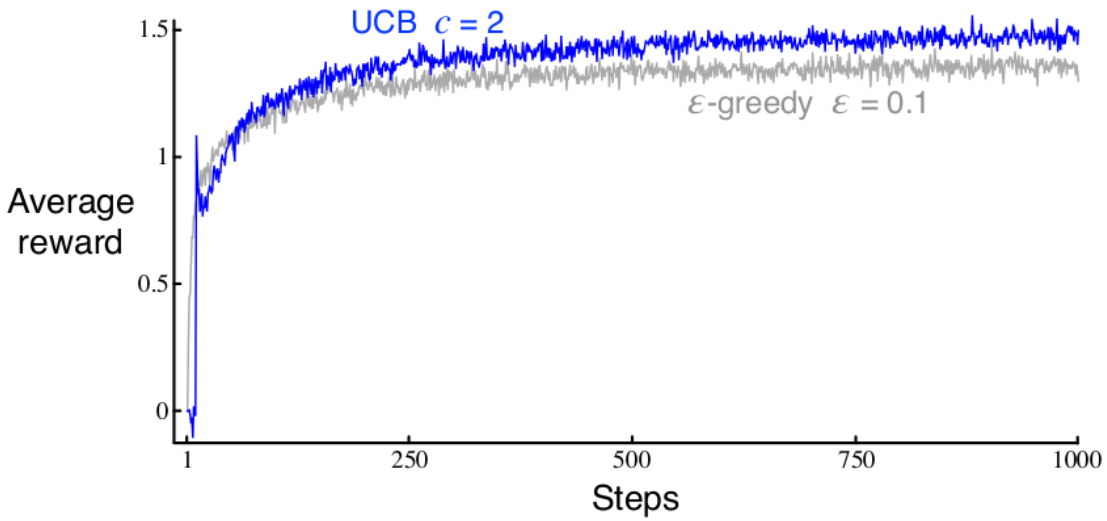


Directed Strategies

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

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





Bayesian Estimation

Bayesian Estimation

Bernoulli Distribution

Bayesian Estimation

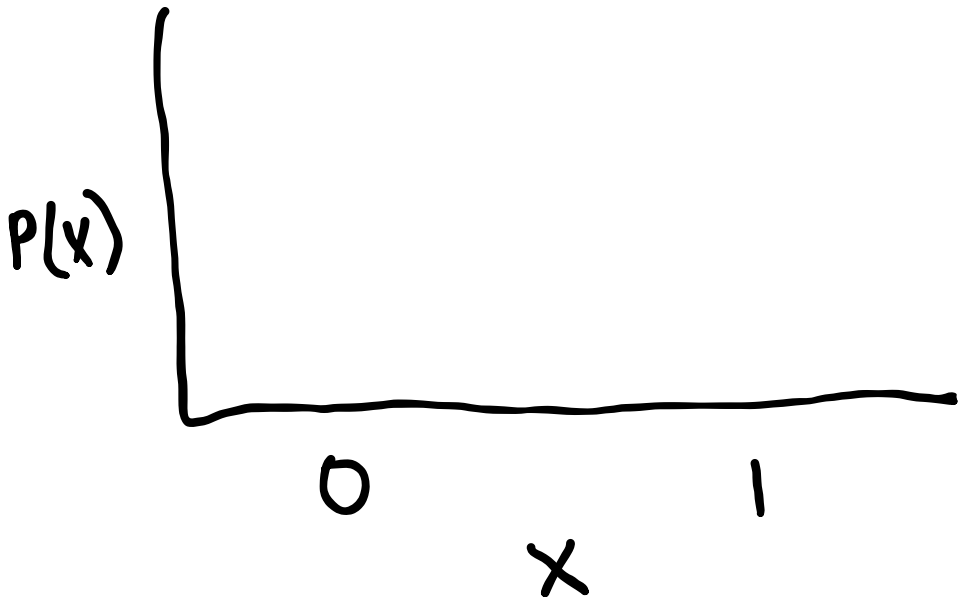
Bernoulli Distribution

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

Bayesian Estimation

Bernoulli Distribution

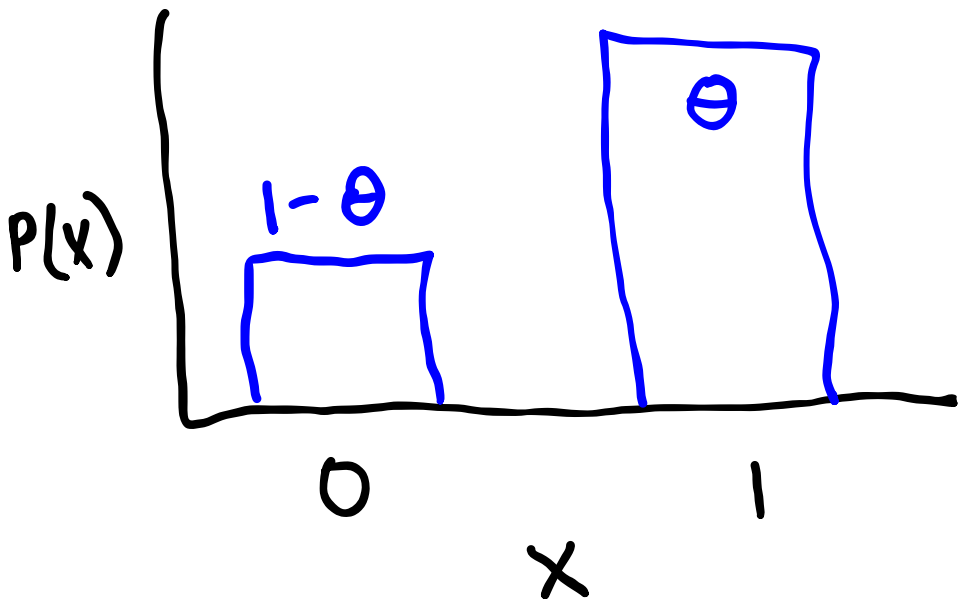
$X \sim \text{Bernoulli}(\theta)$



Bayesian Estimation

Bernoulli Distribution

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

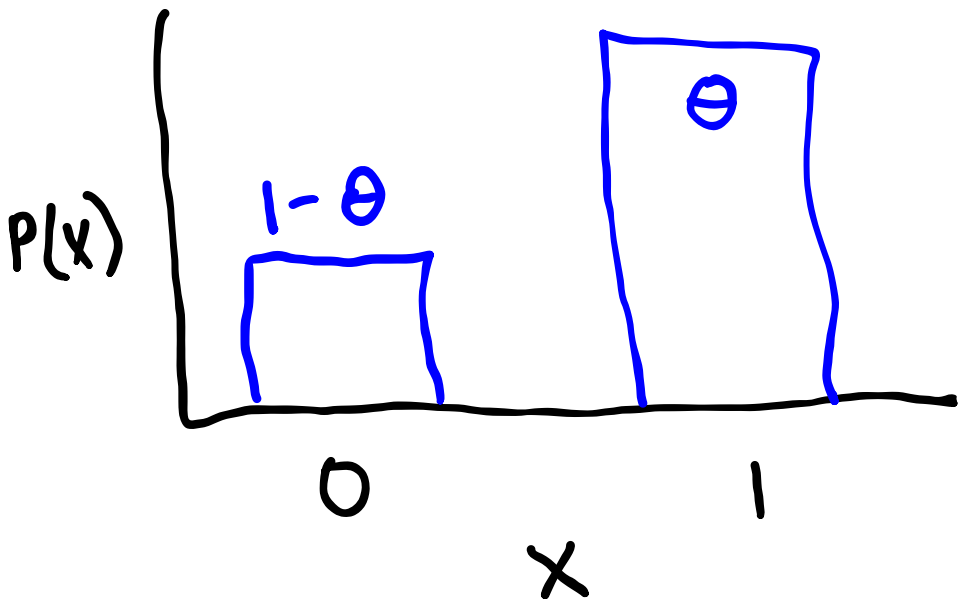


Bayesian Estimation

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?

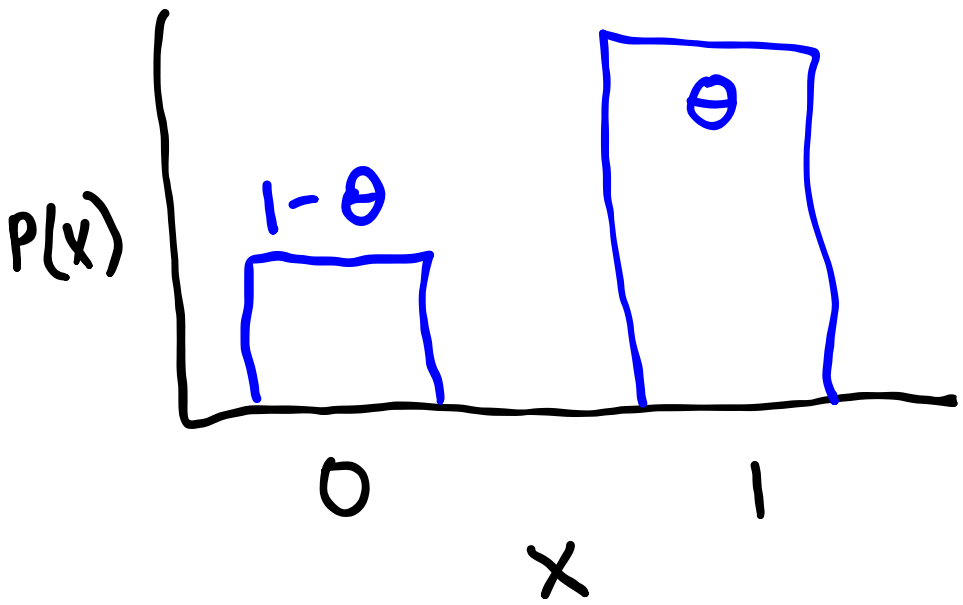


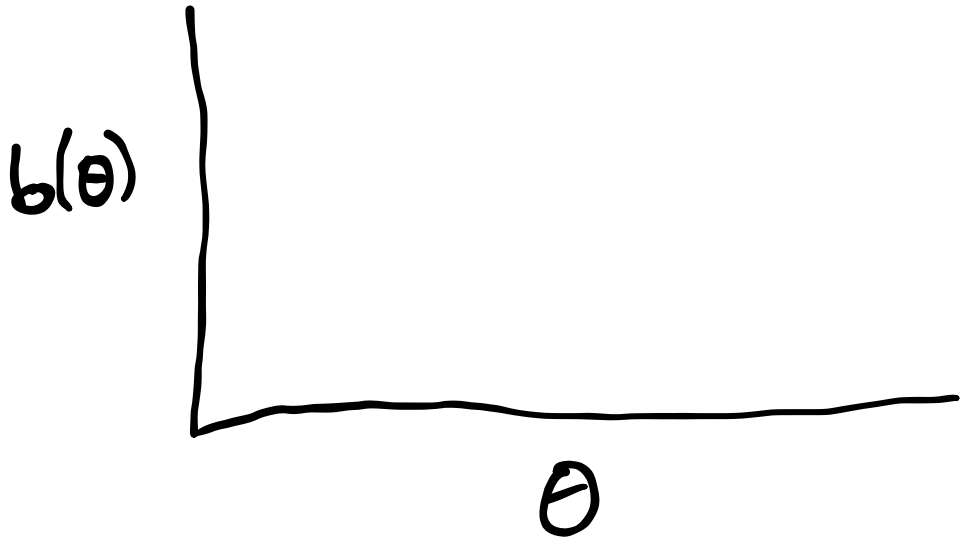
Bayesian Estimation

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?



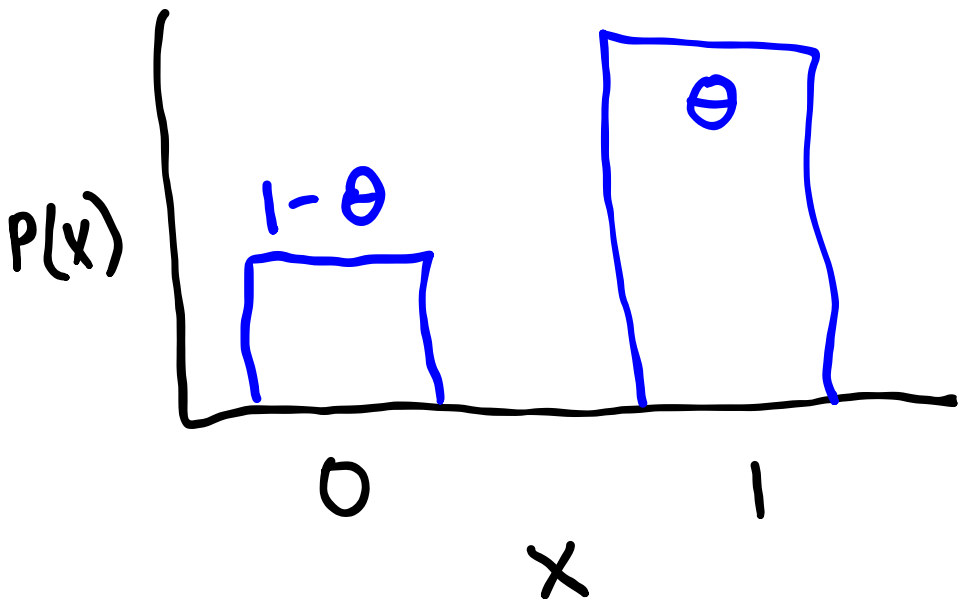


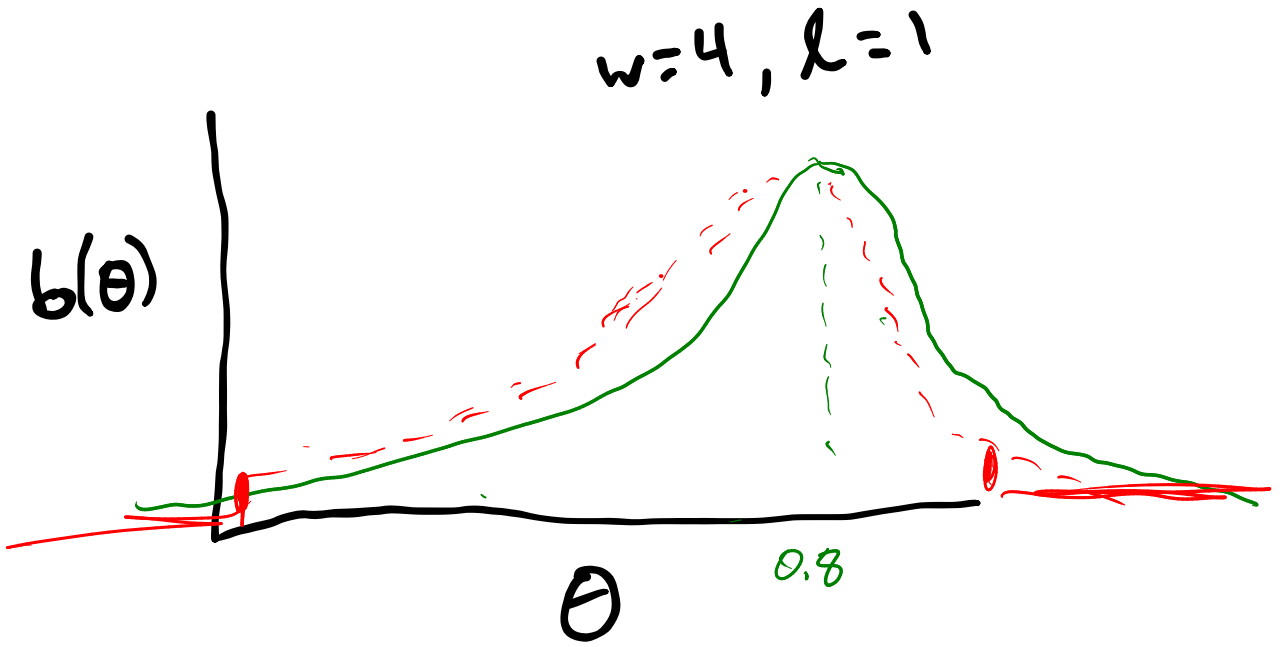
Bayesian Estimation

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?

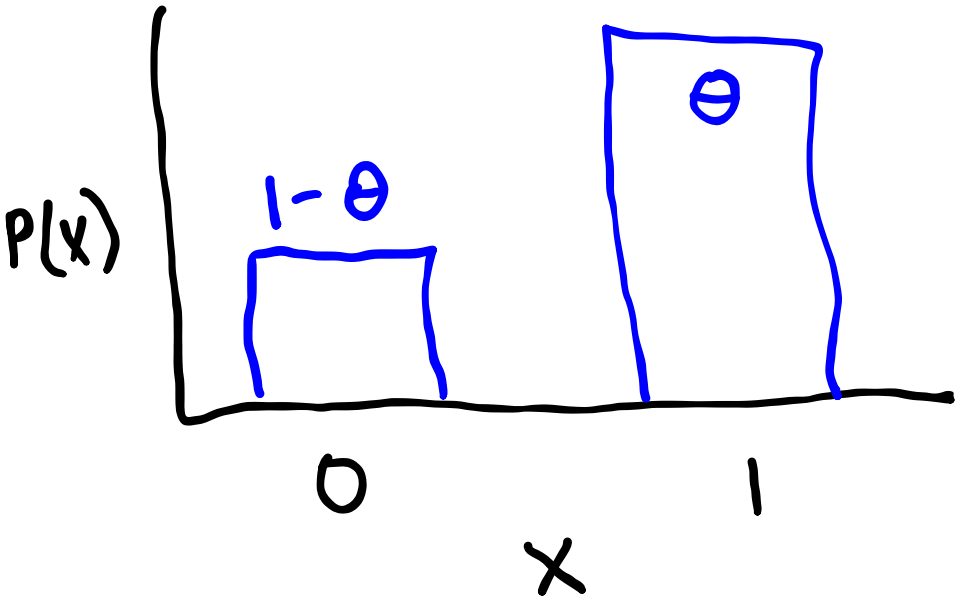




Bayesian Estimation

Bernoulli Distribution

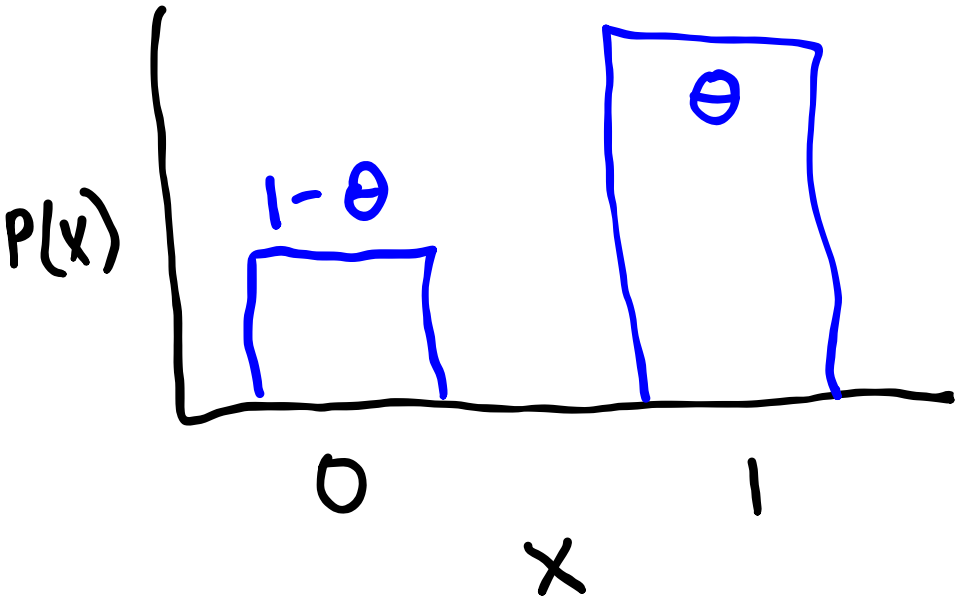
$$\text{Bernoulli}(\theta)$$



Bayesian Estimation

Bernoulli Distribution

$\text{Bernoulli}(\theta)$



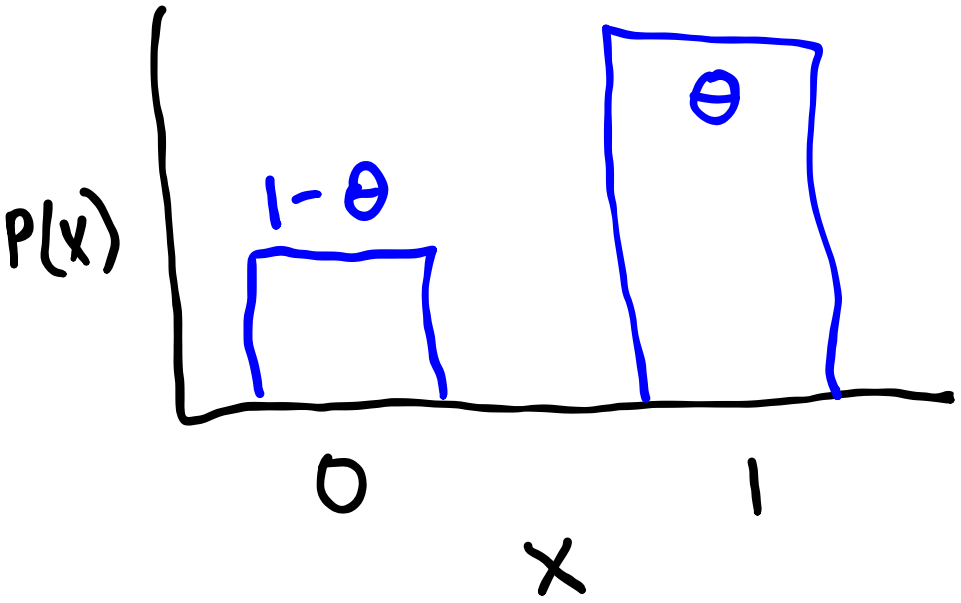
Beta Distribution

(distribution over Bernoulli distributions)

Bayesian Estimation

Bernoulli Distribution

$\text{Bernoulli}(\theta)$



Beta Distribution

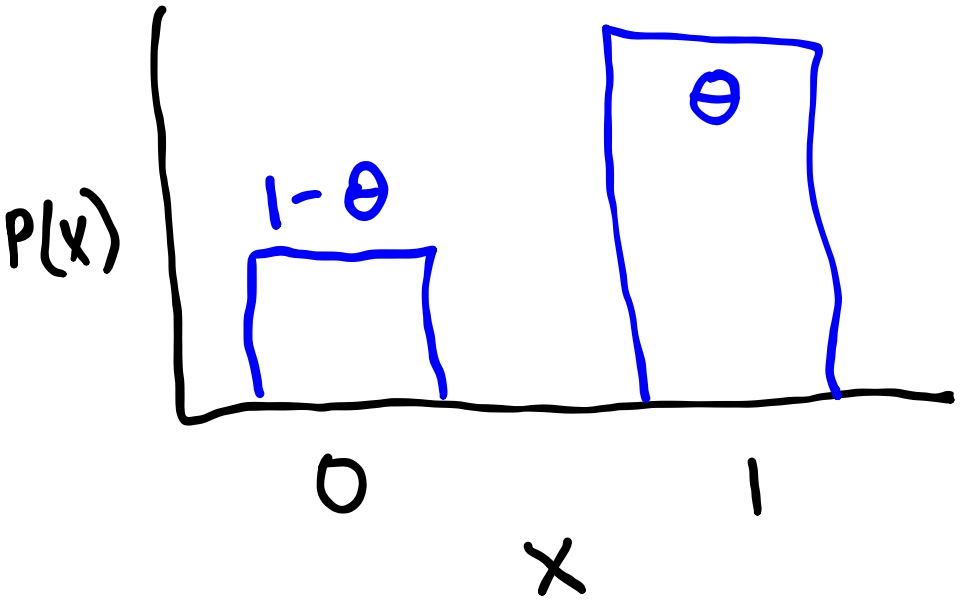
(distribution over Bernoulli distributions)

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

Bayesian Estimation

Bernoulli Distribution

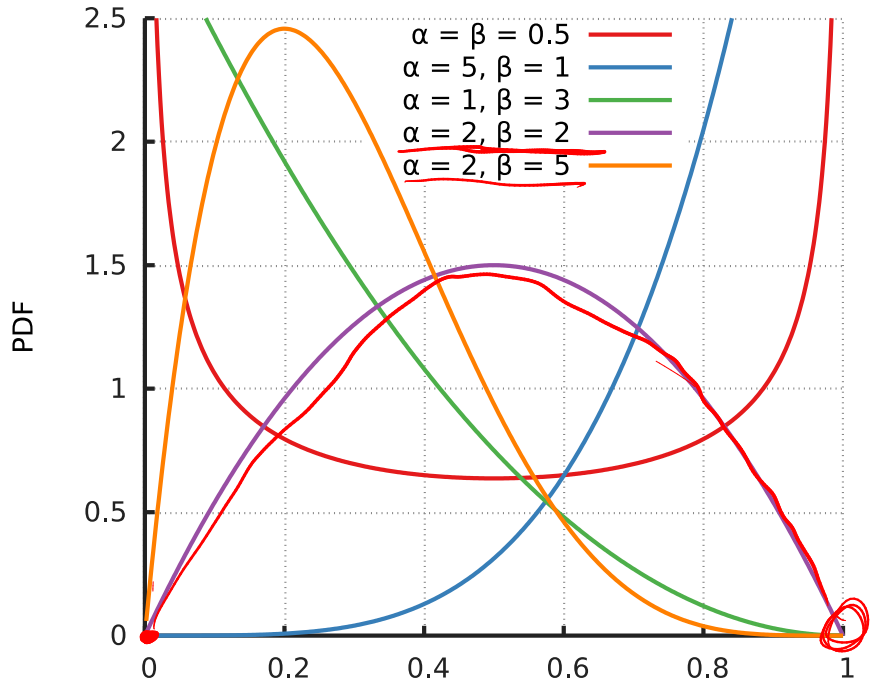
$$\text{Bernoulli}(\theta)$$



Beta Distribution

(distribution over Bernoulli distributions)

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



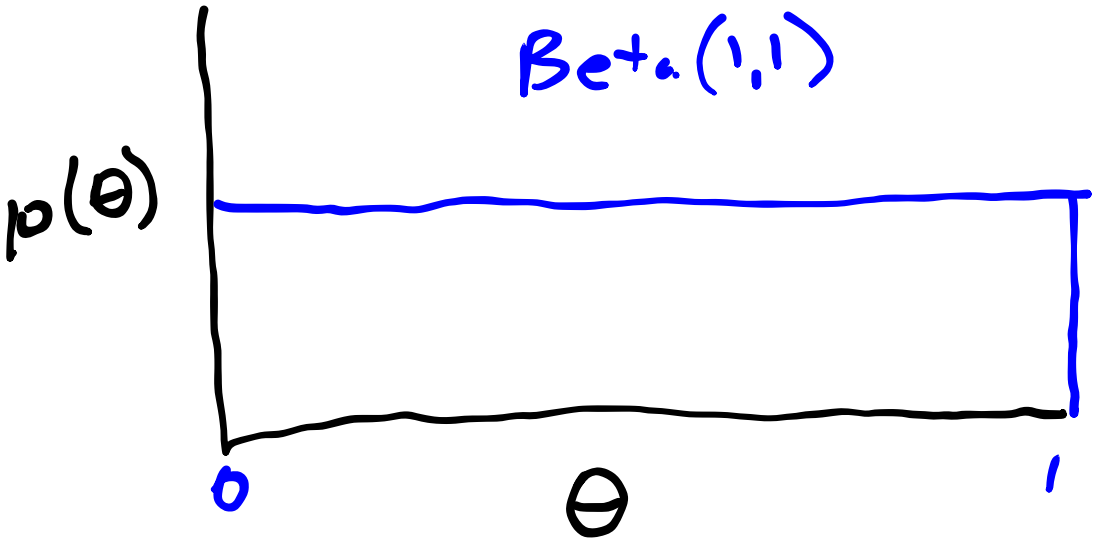
Bayesian Estimation

Bayesian Estimation

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

Bayesian Estimation

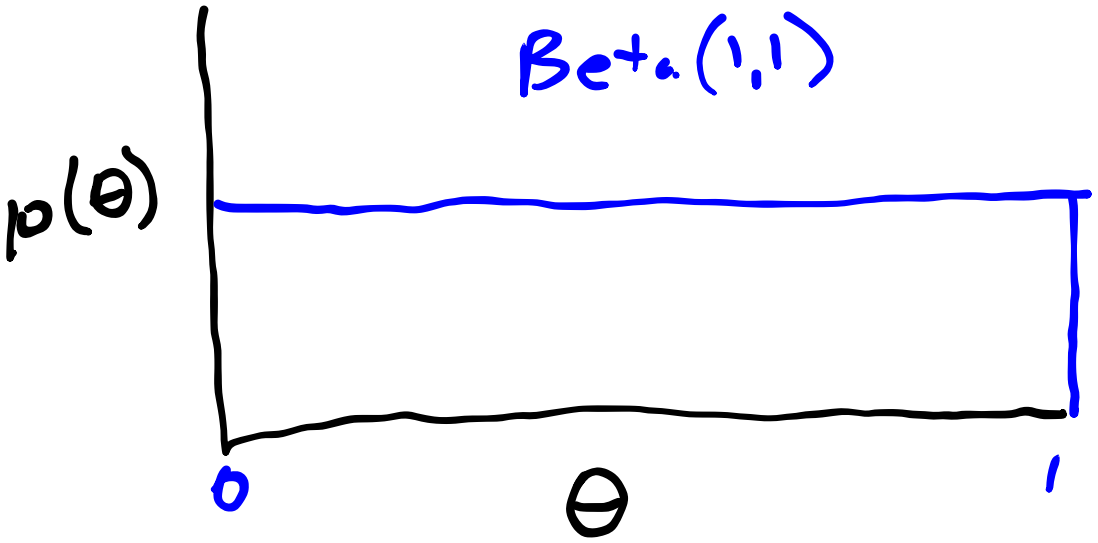
Given a $\mathrm{Beta}(1, 1)$ prior distribution



Bayesian Estimation

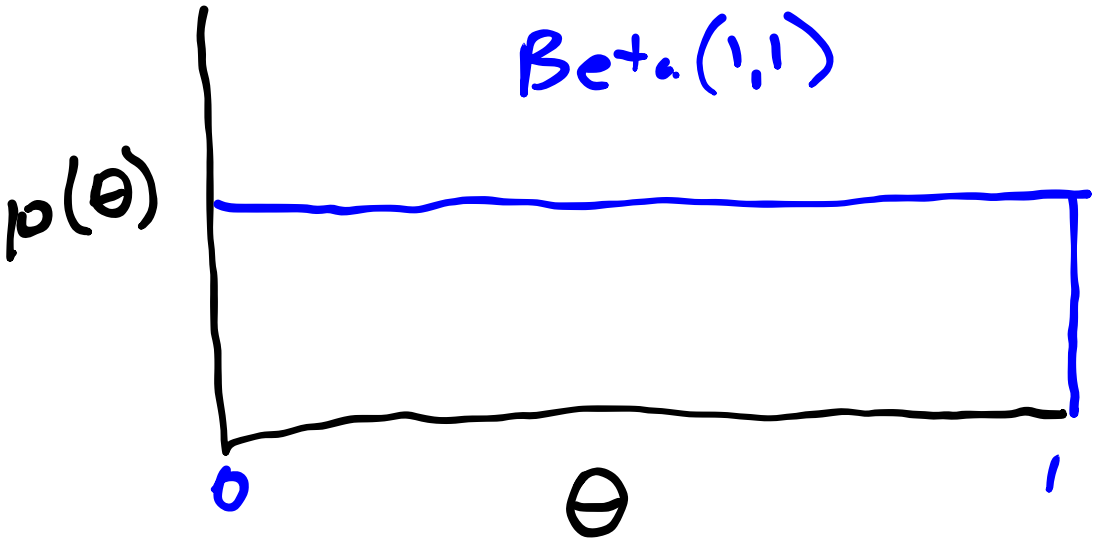
Given a $\mathrm{Beta}(1,1)$ prior distribution

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

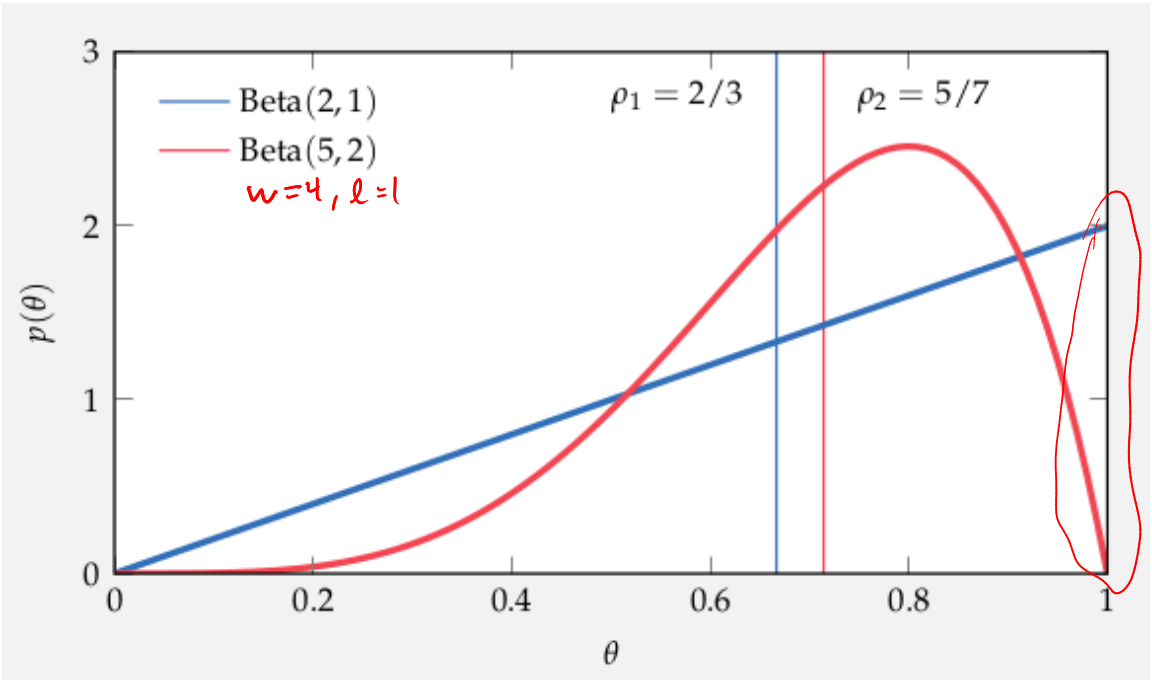


Bayesian Estimation

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

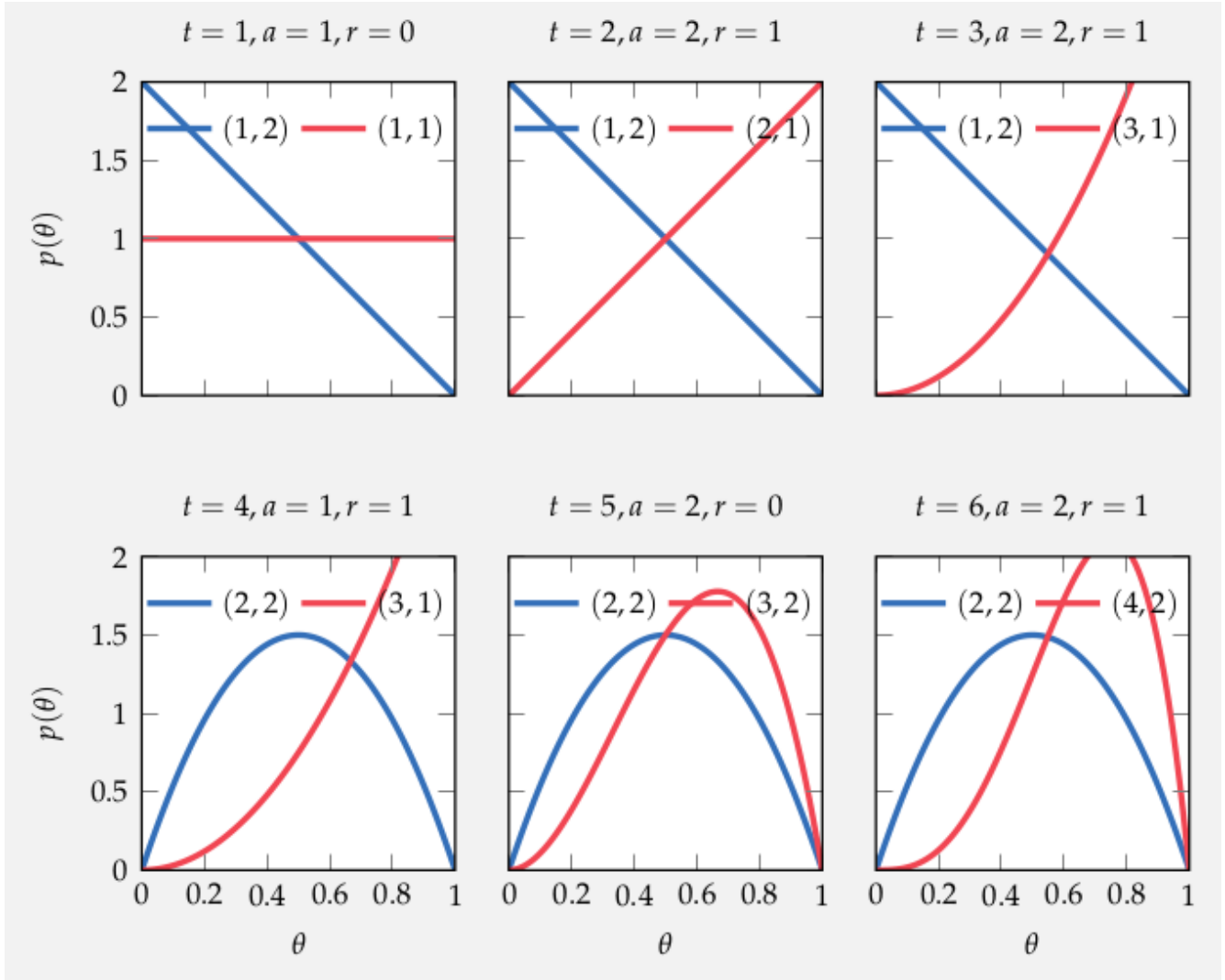


The posterior distribution of θ is $\mathrm{Beta}(w+1, l+1)$



Bayesian Estimation

arm 1
arm 2



Bayesian Bandit Algorithms

Bayesian Bandit Algorithms

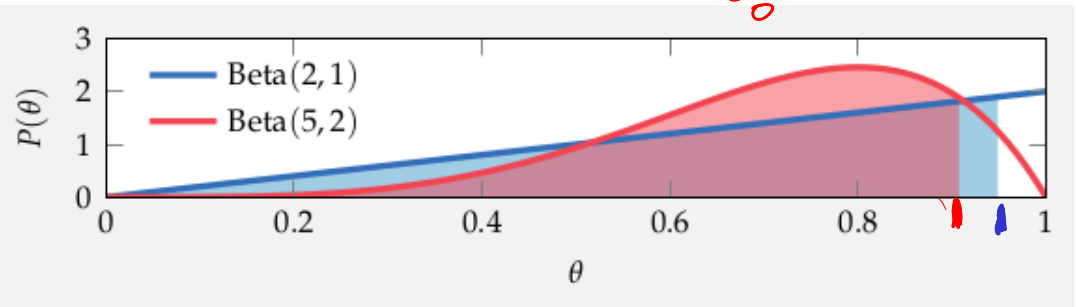
- Quantile Selection
  Choose a for which the α quantile of $b(\theta)$ is highest

Bayesian Bandit Algorithms

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

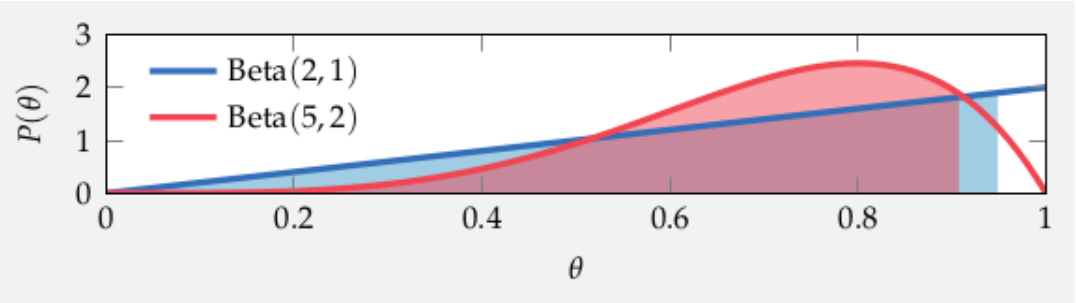
α quantile

θ s.t. $\int_0^{\theta} p(\theta')\, d\theta' = \alpha$

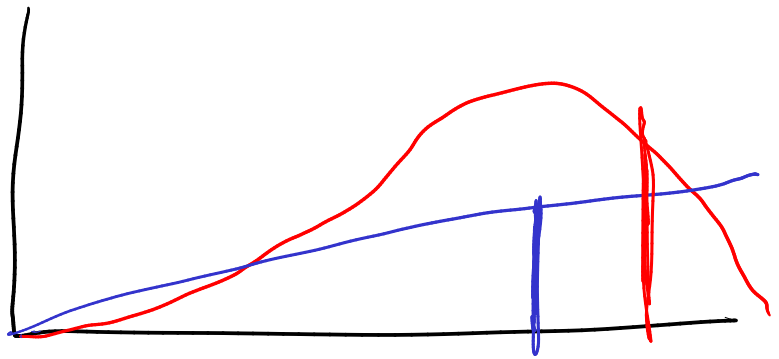


Bayesian Bandit Algorithms

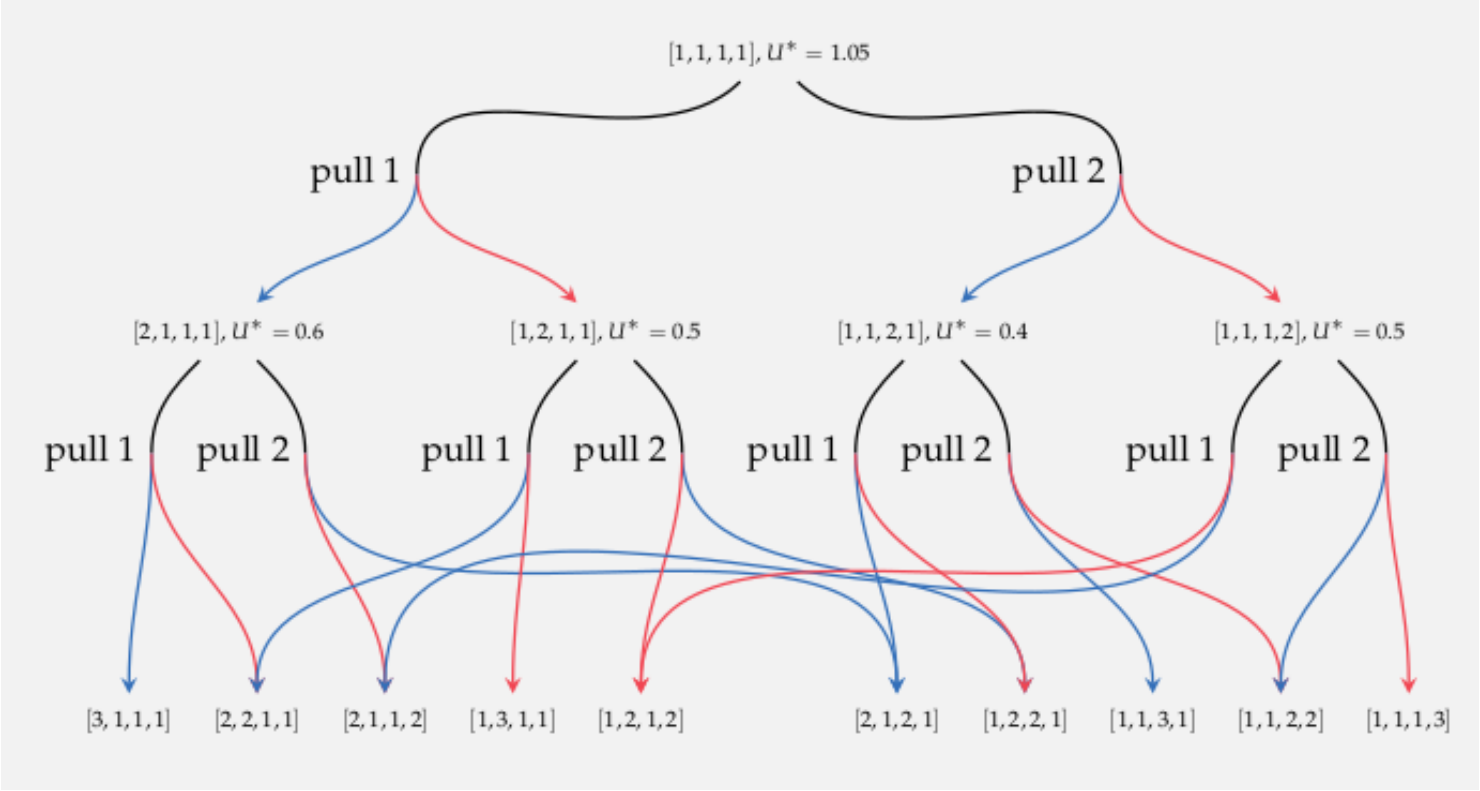
- Quantile Selection
  Choose a for which the α quantile of $b(\theta)$ is highest



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



Optimal Algorithm - Dynamic Programming



Review

Optimal in Limit

Easier to Implement
Faster

- greedy
- ε-greedy — $\varepsilon \to 0$
- explore-commit — $k \to \infty$
- softmax — $\propto e^{\lambda p_a}$ — $\lambda \to \infty$
- UCB — $p_a + c\sqrt{\frac{\log N}{N(a)}}$ ✓
- Quantile Selection ✓
- Thompson Sampling ✓
- Dynamic Programming ✓

Bayesian

Less Regret

for a Bernoulli Bandit

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

$O(N)$

$O(N)$

$O(N)$

$O(N)$

$O(\log(N))$

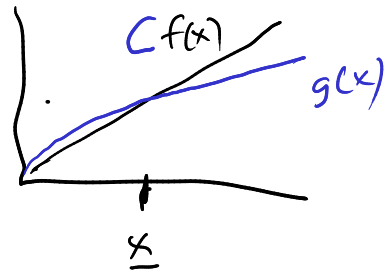
$=$

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

$\exists\ C$ s.t.

$g(x) \leq C f(x)$

$\forall x > \underline{x}$



Guiding Questions

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