

Restless Bandits with Average Reward: Breaking the Uniform Global Attractor Assumption

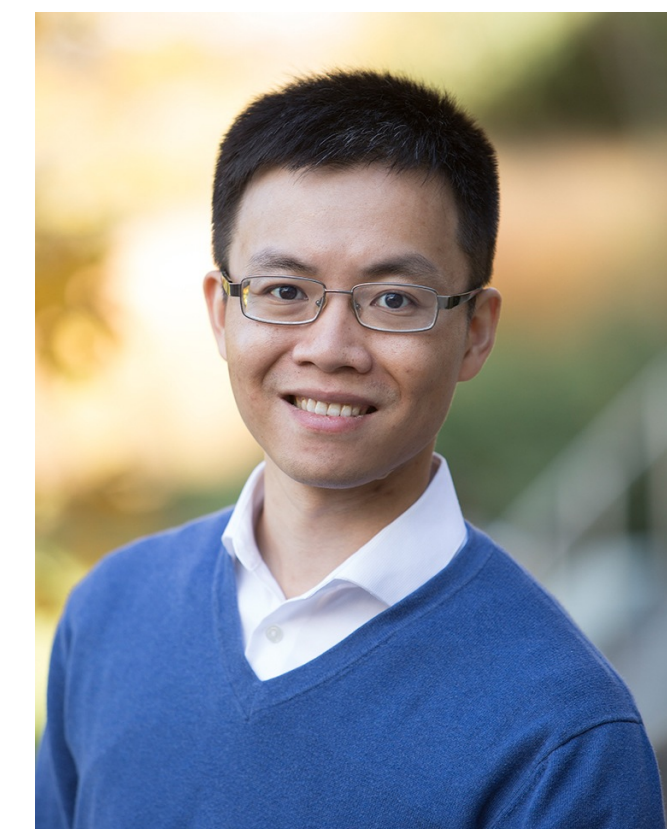
Weina Wang
Carnegie Mellon University



Yige Hong
Carnegie Mellon

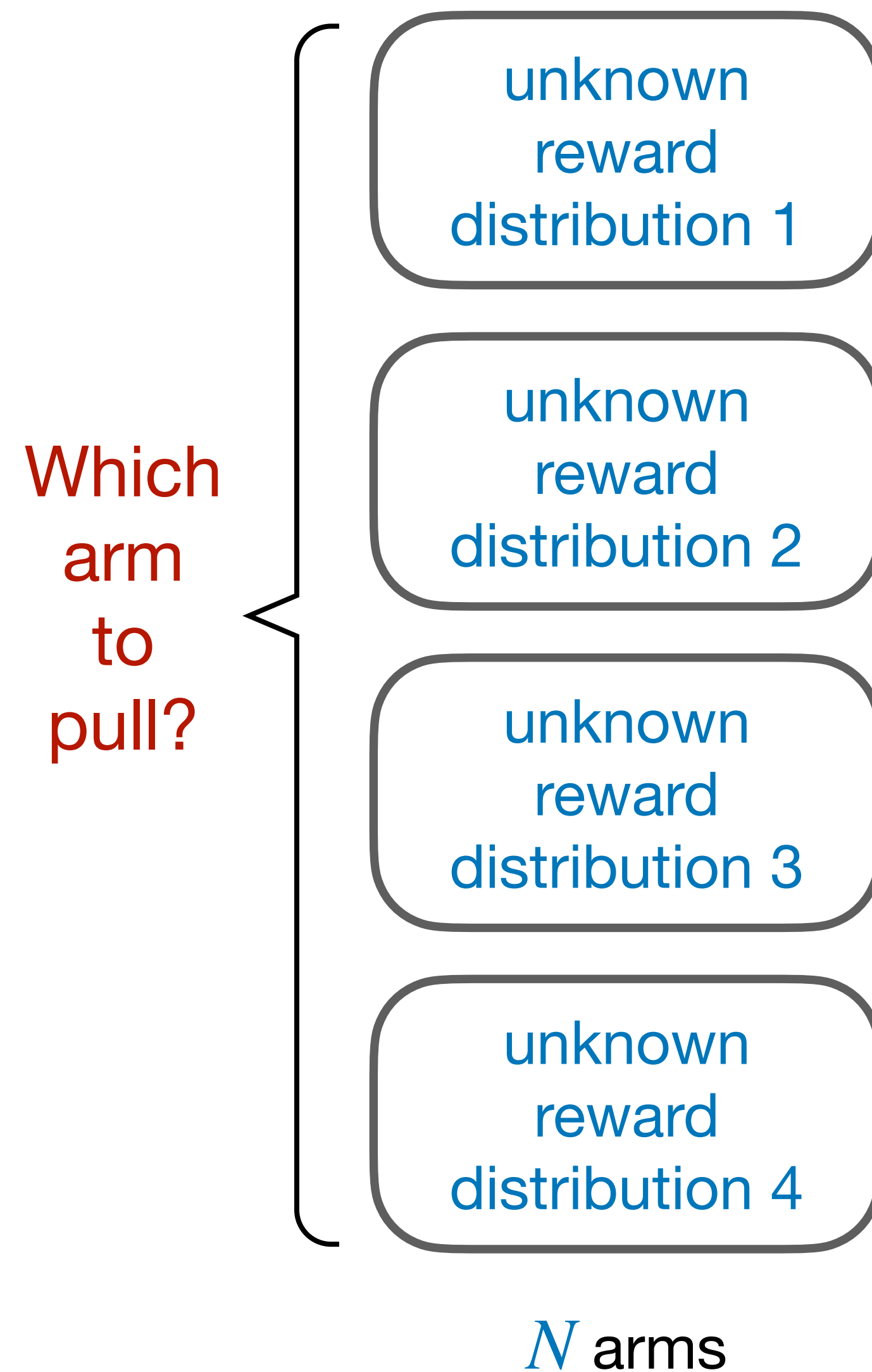


Qiaomin Xie
UW—Madison

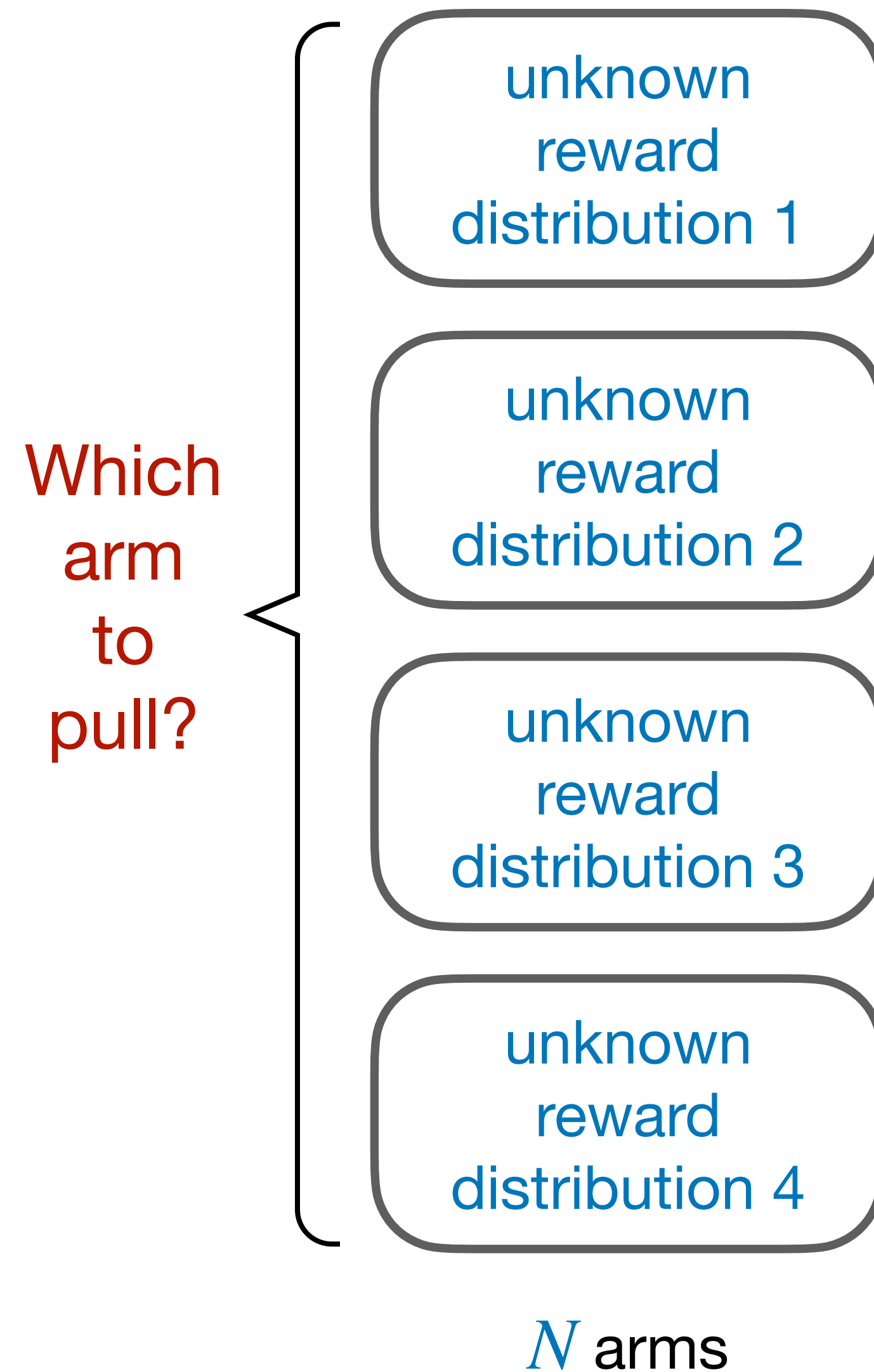


Yudong Chen
UW—Madison

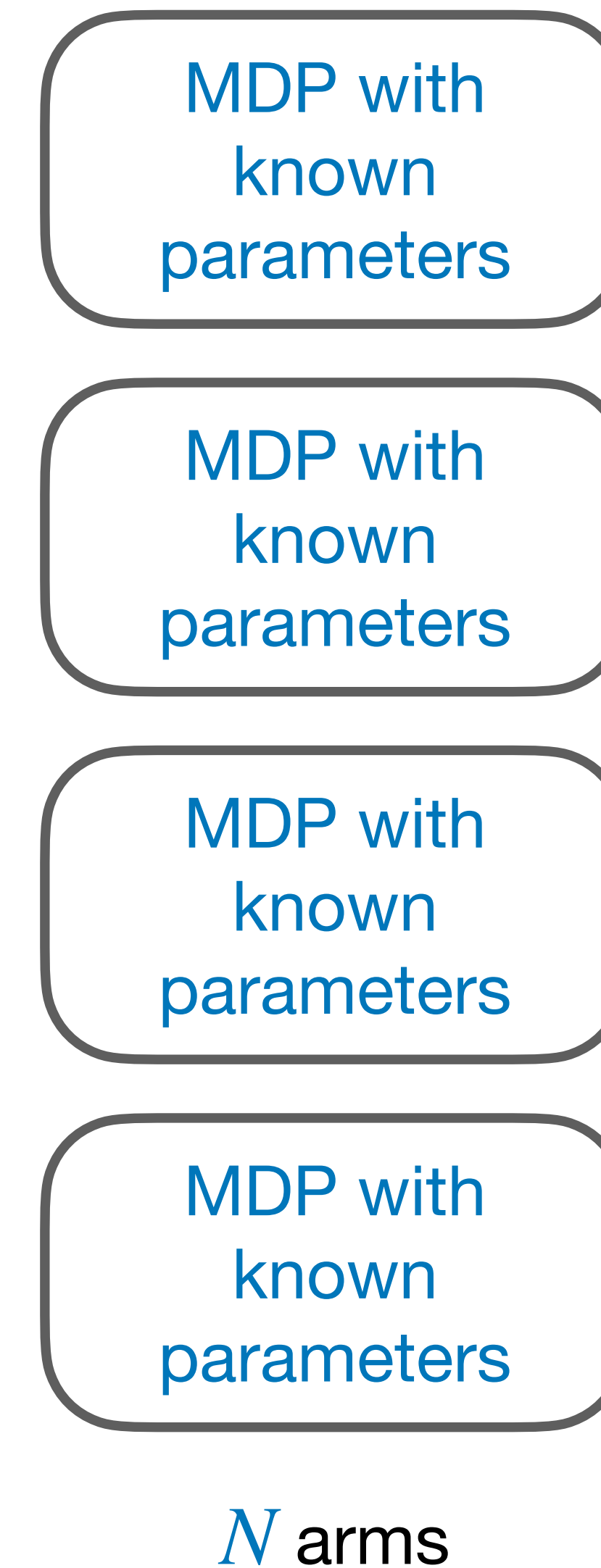
Stochastic multi-armed bandits



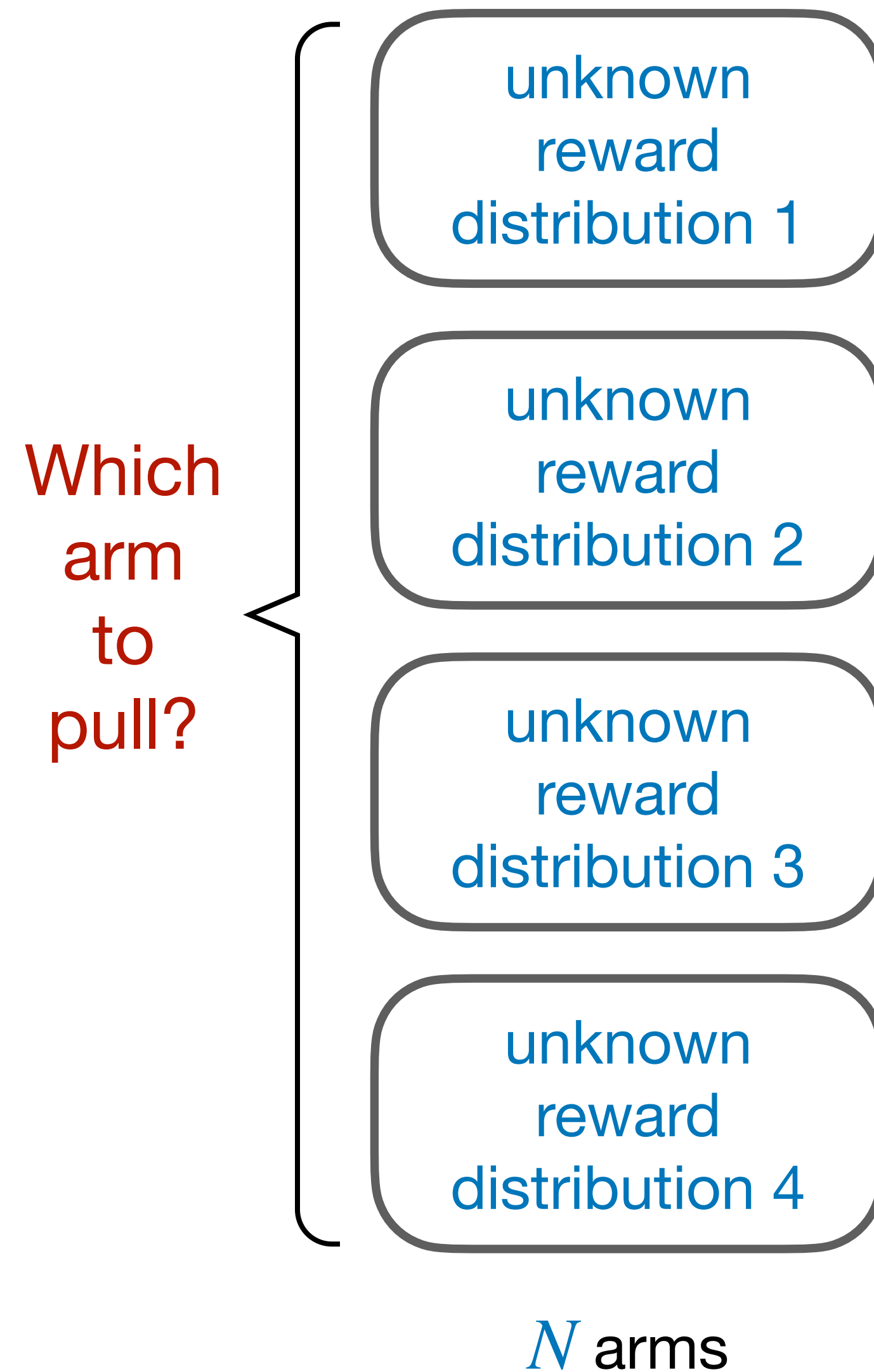
Stochastic multi-armed bandits



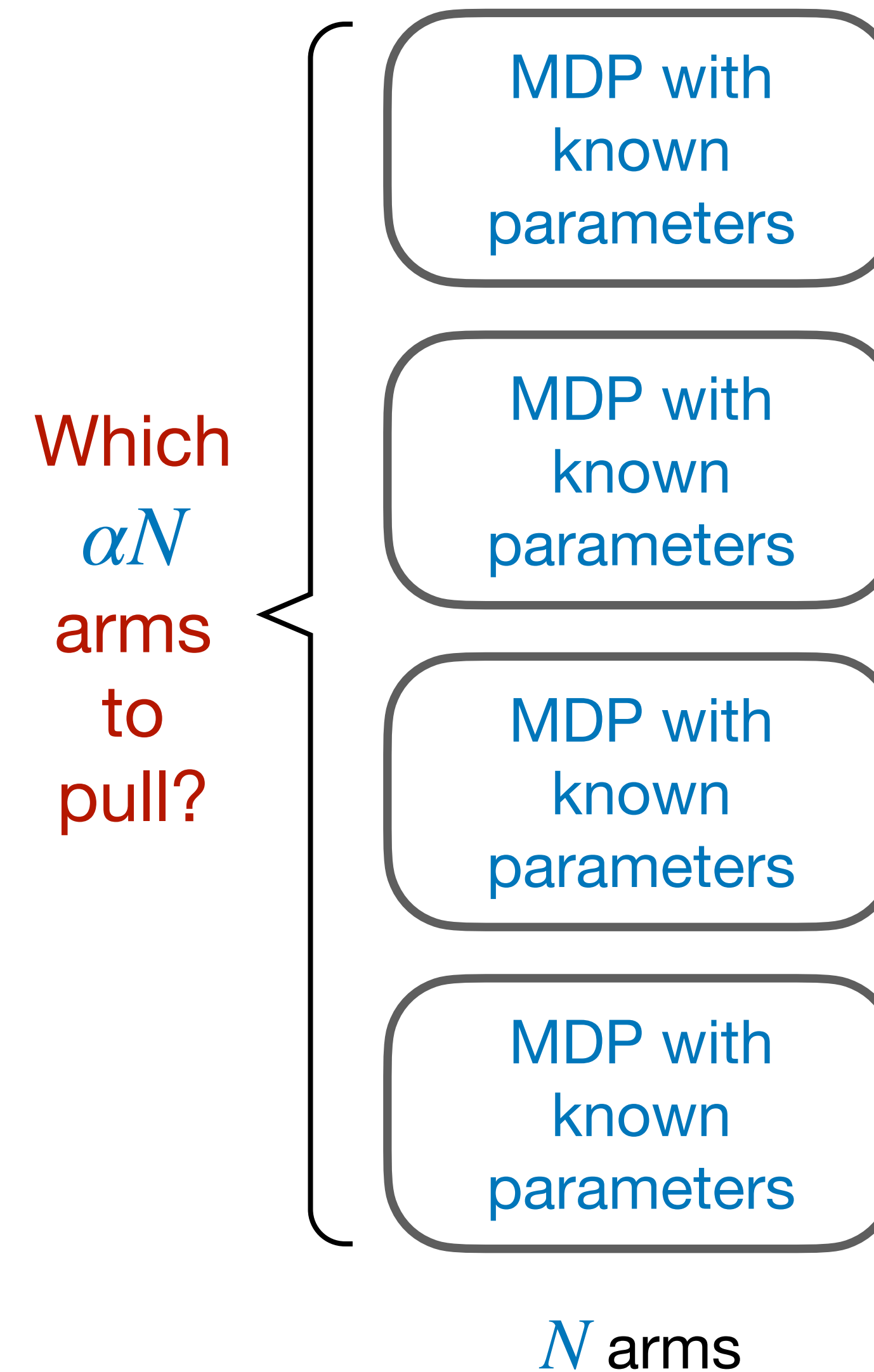
Restless bandits [Whittle 88]



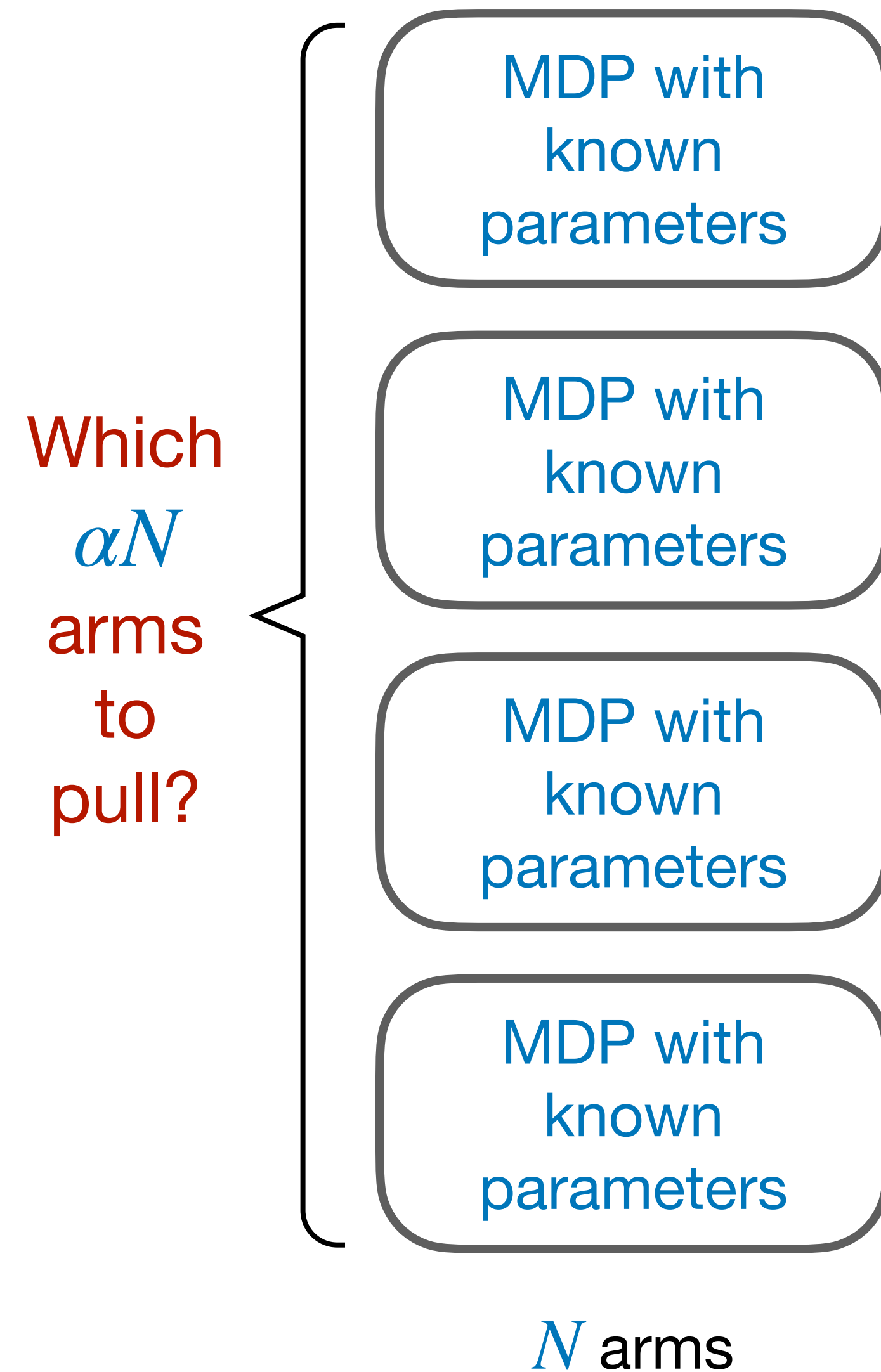
Stochastic multi-armed bandits



Restless bandits [Whittle 88]

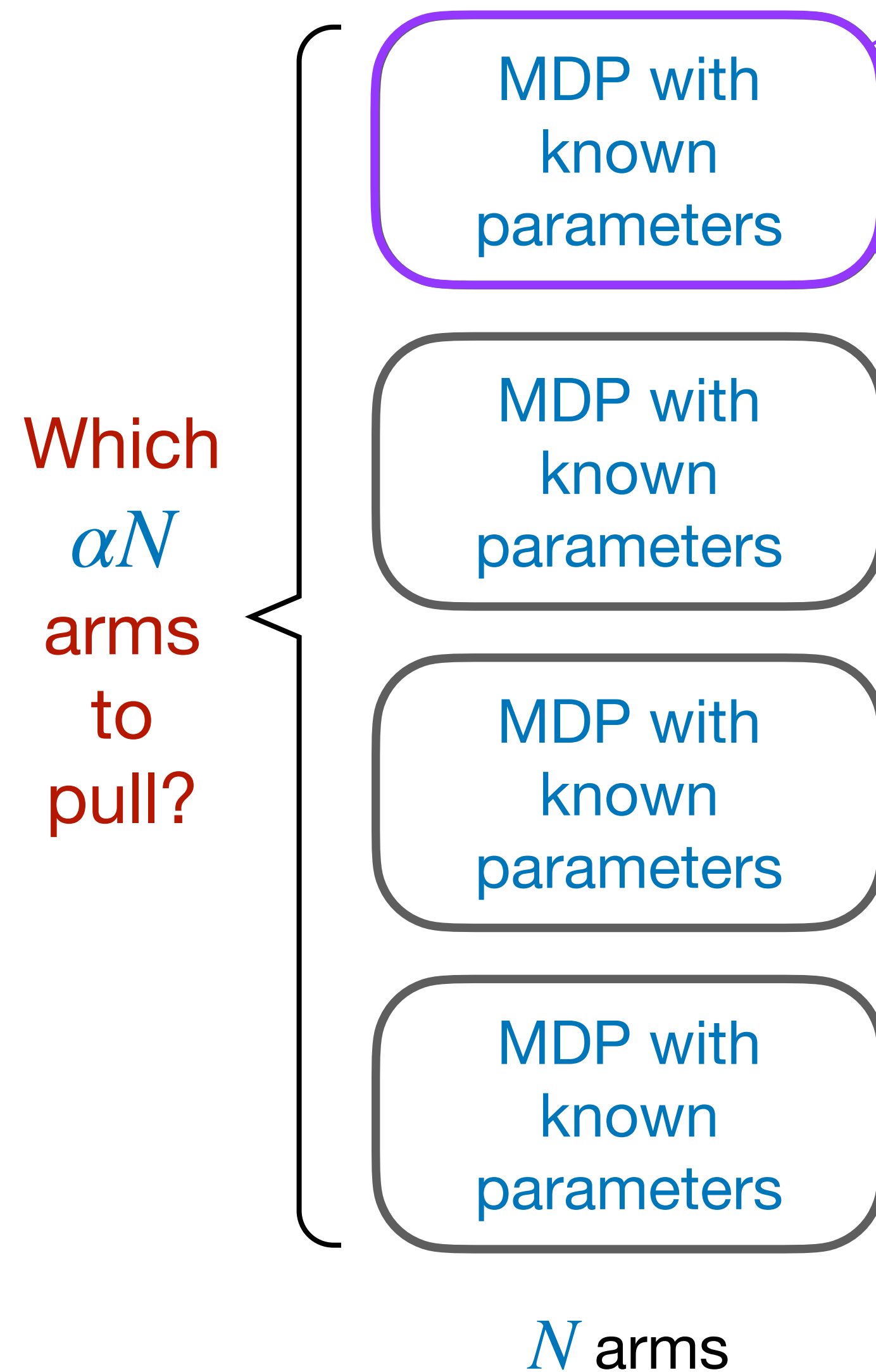


Restless bandits



Restless bandits

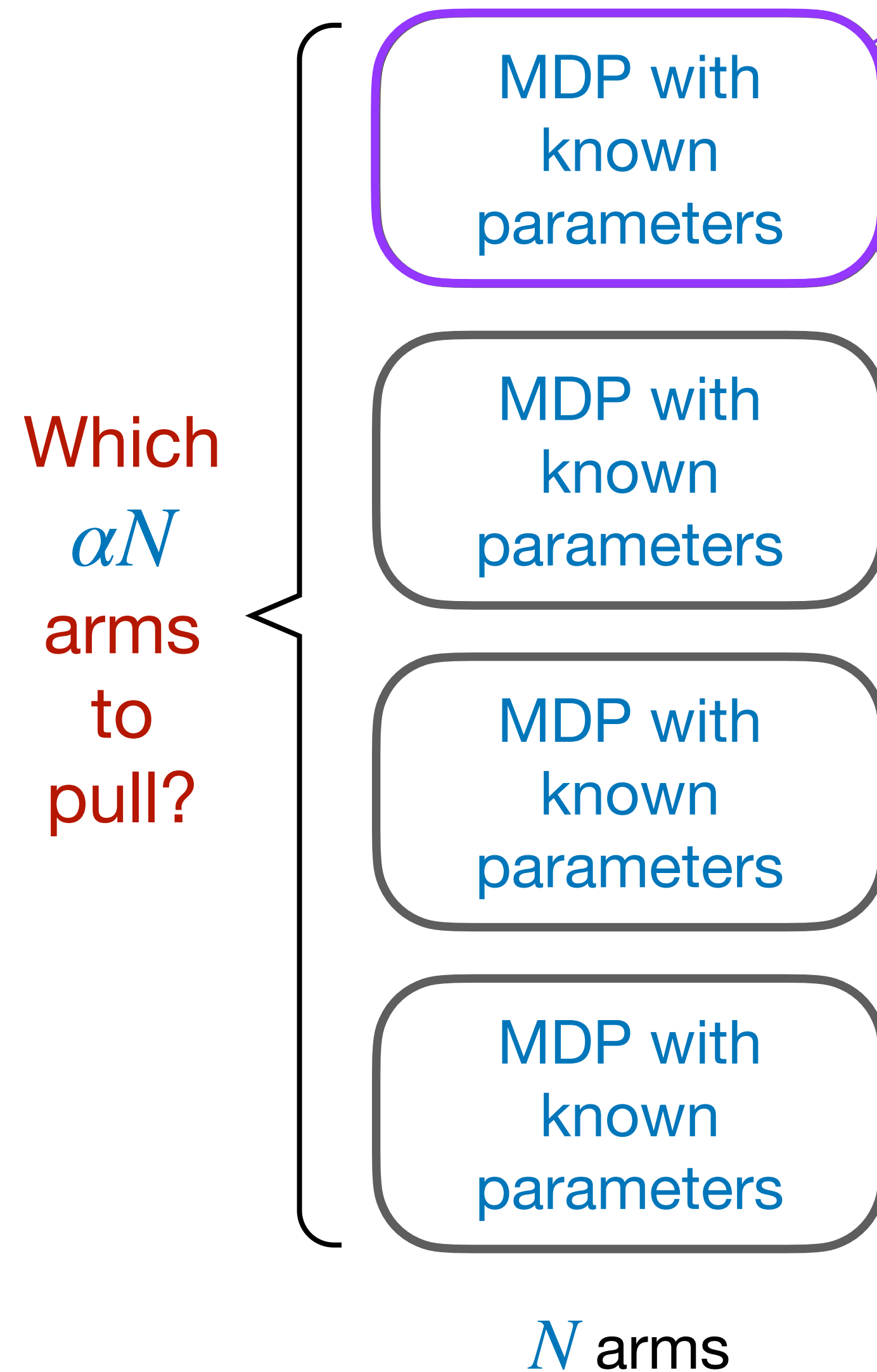
Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:



Restless bandits

Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

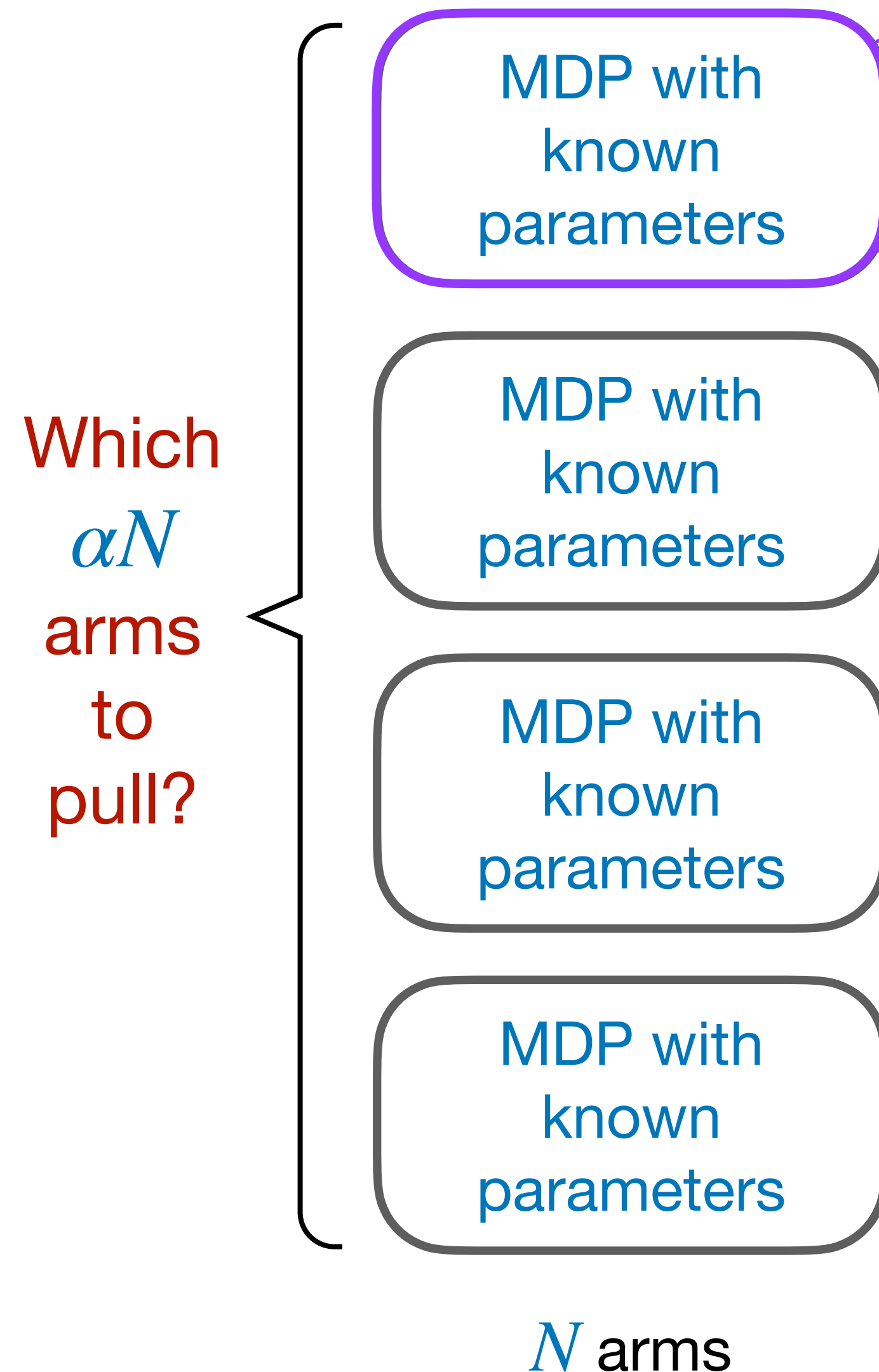
- State space: \mathcal{S} , finite



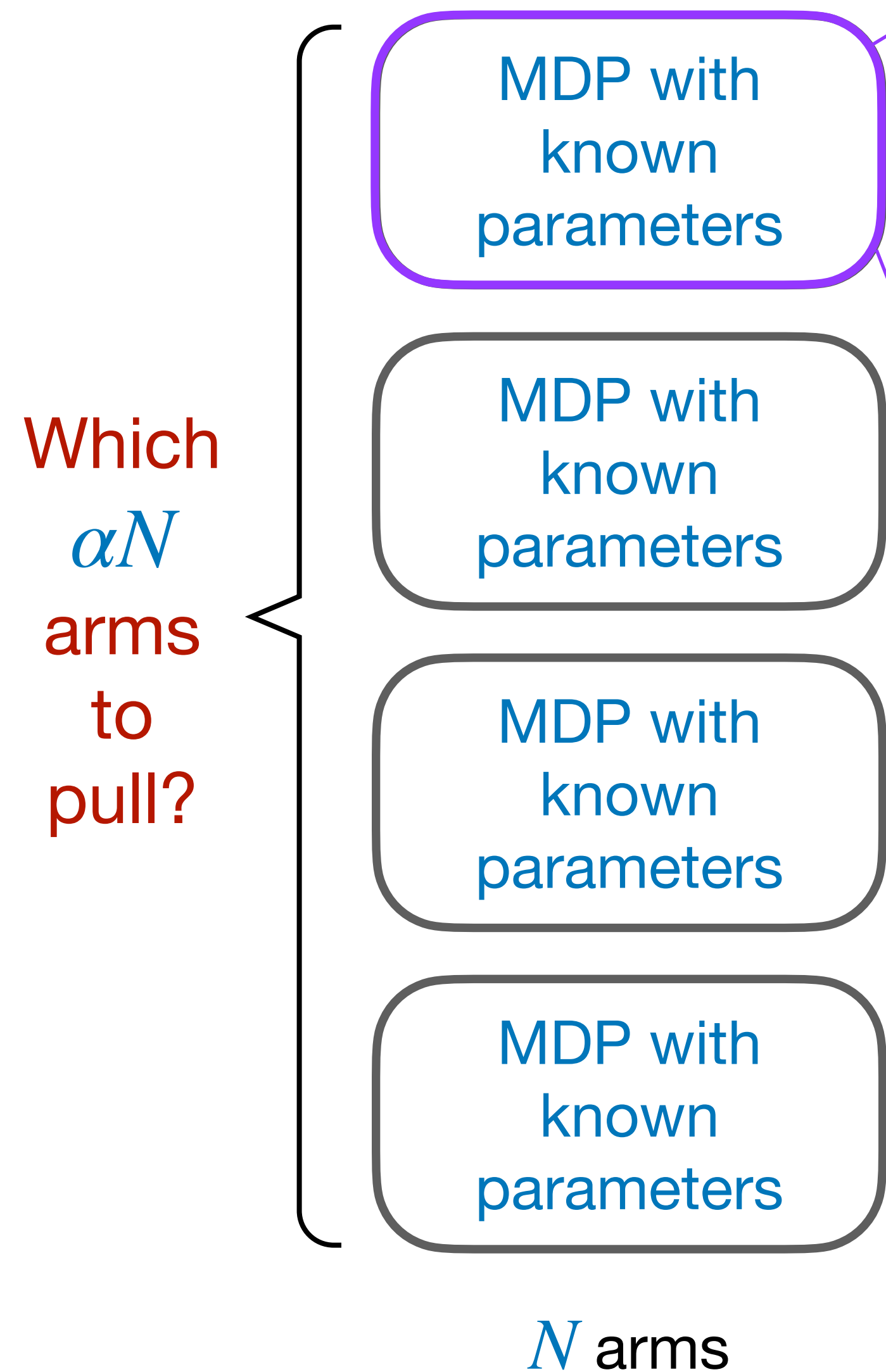
Restless bandits

Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$



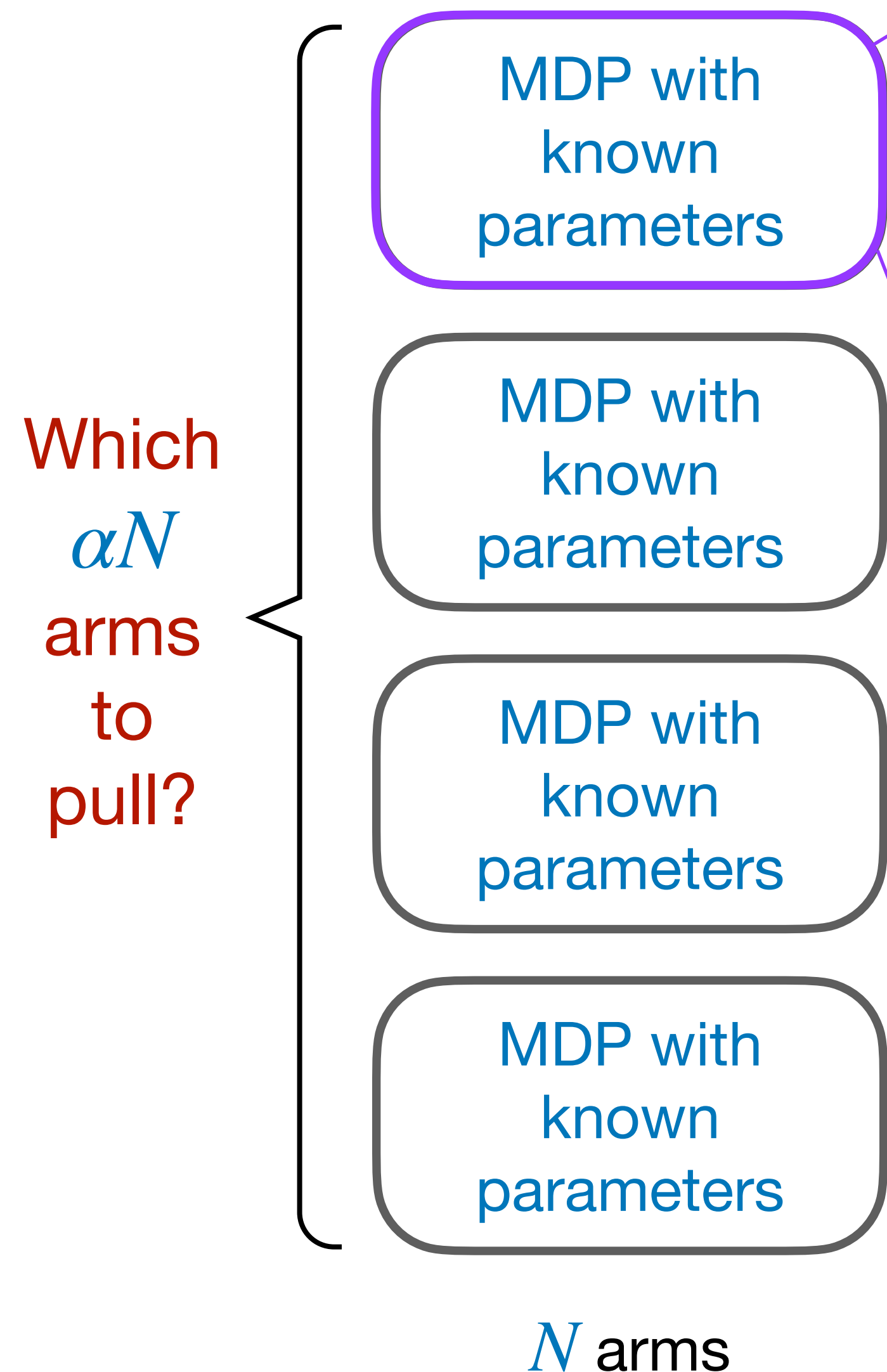
Restless bandits



Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”

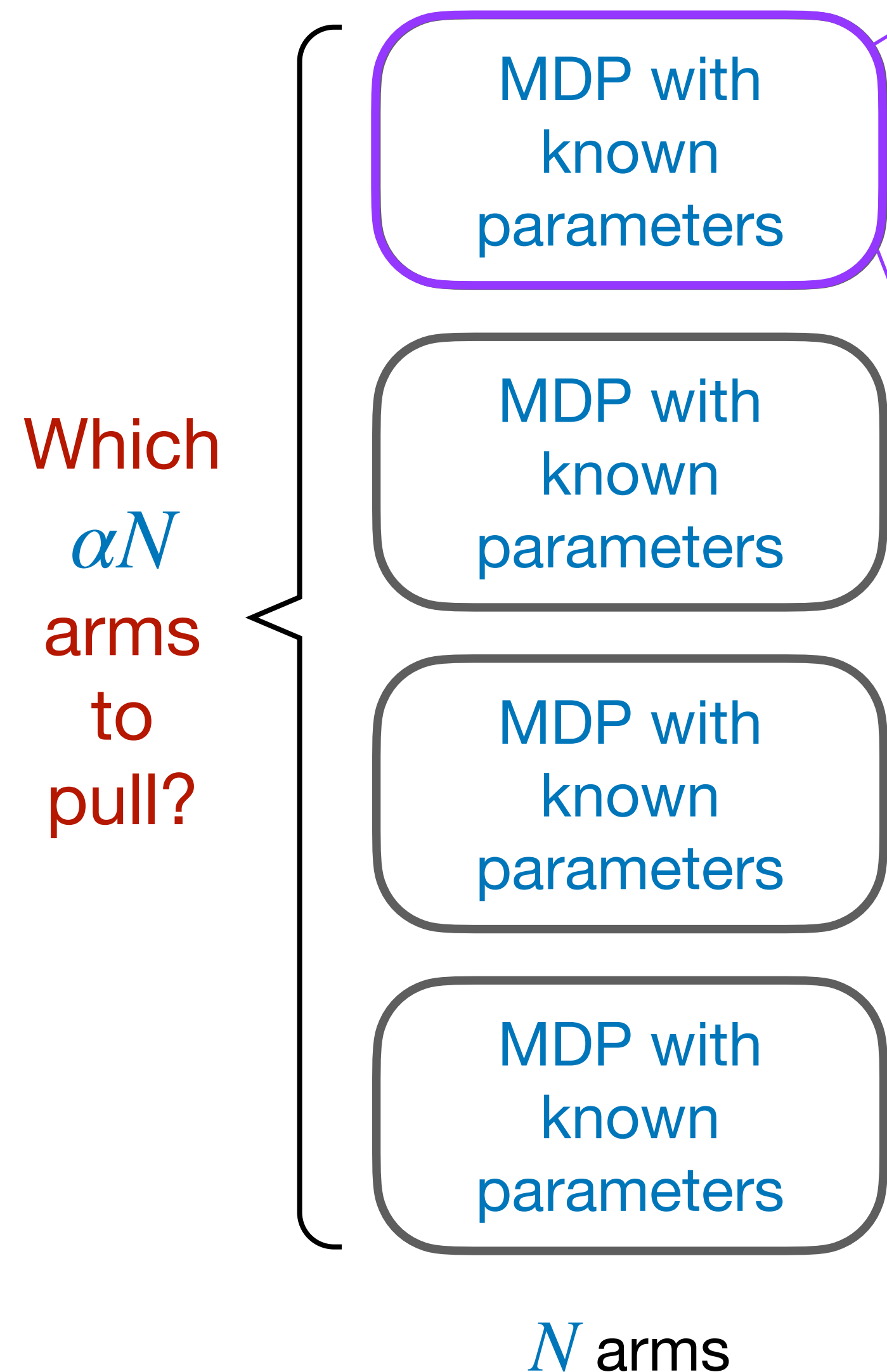
Restless bandits



Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”
 - Passive = “not pulling”

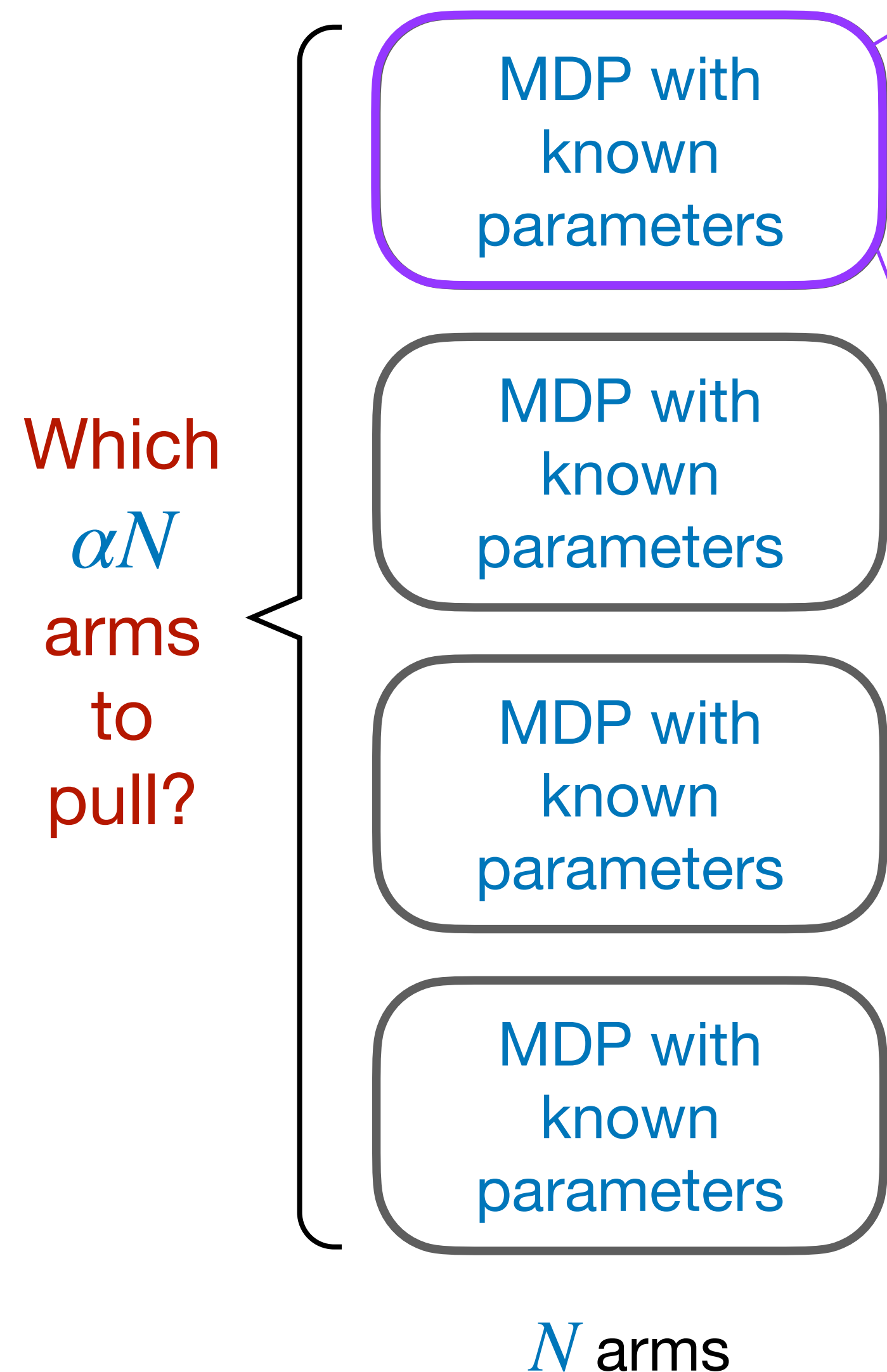
Restless bandits



Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”
 - Passive = “not pulling”
- Transition probabilities:

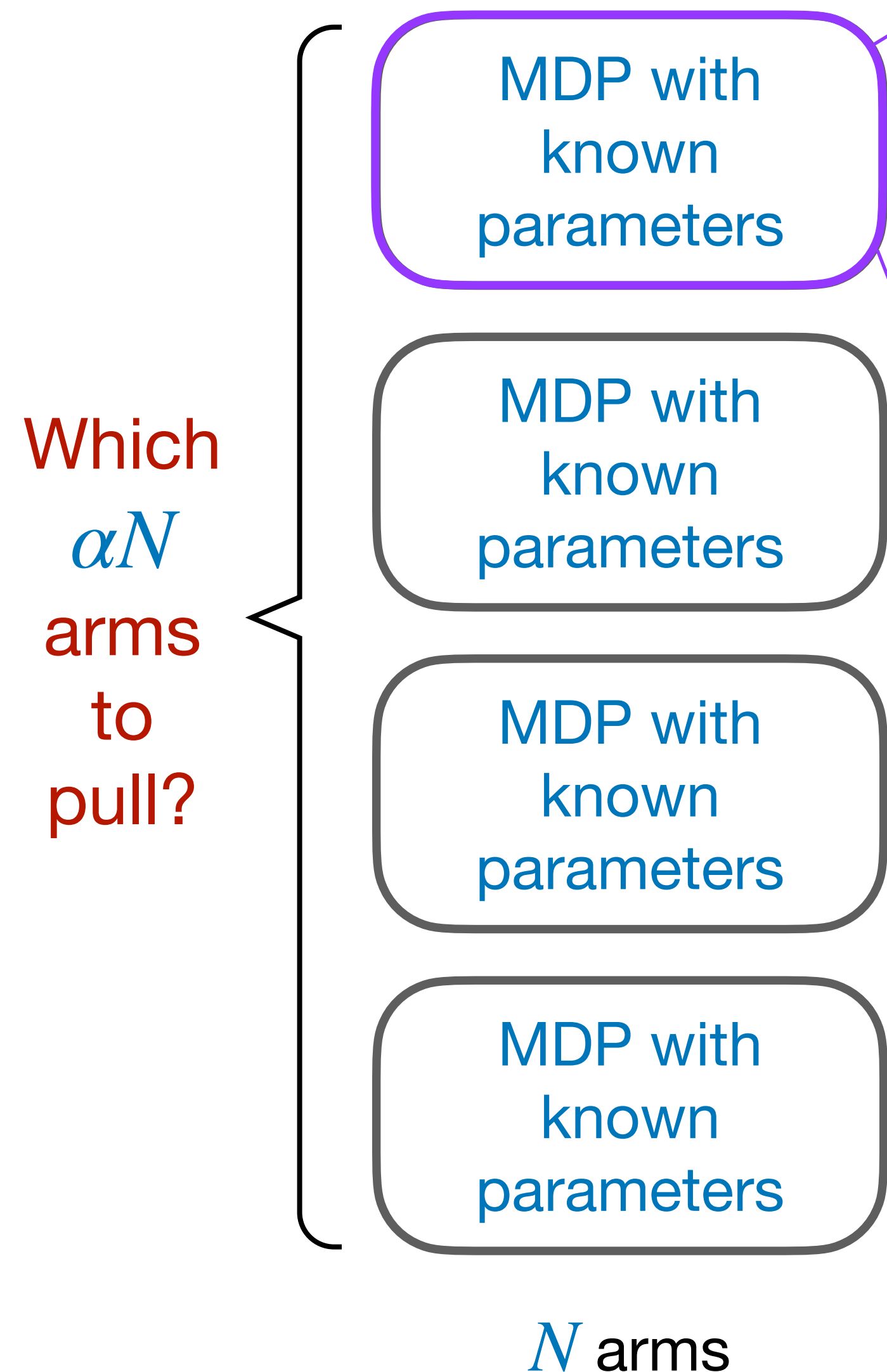
Restless bandits



Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”
 - Passive = “not pulling”
- Transition probabilities:
 - $P(s' \mid s, \text{active})$

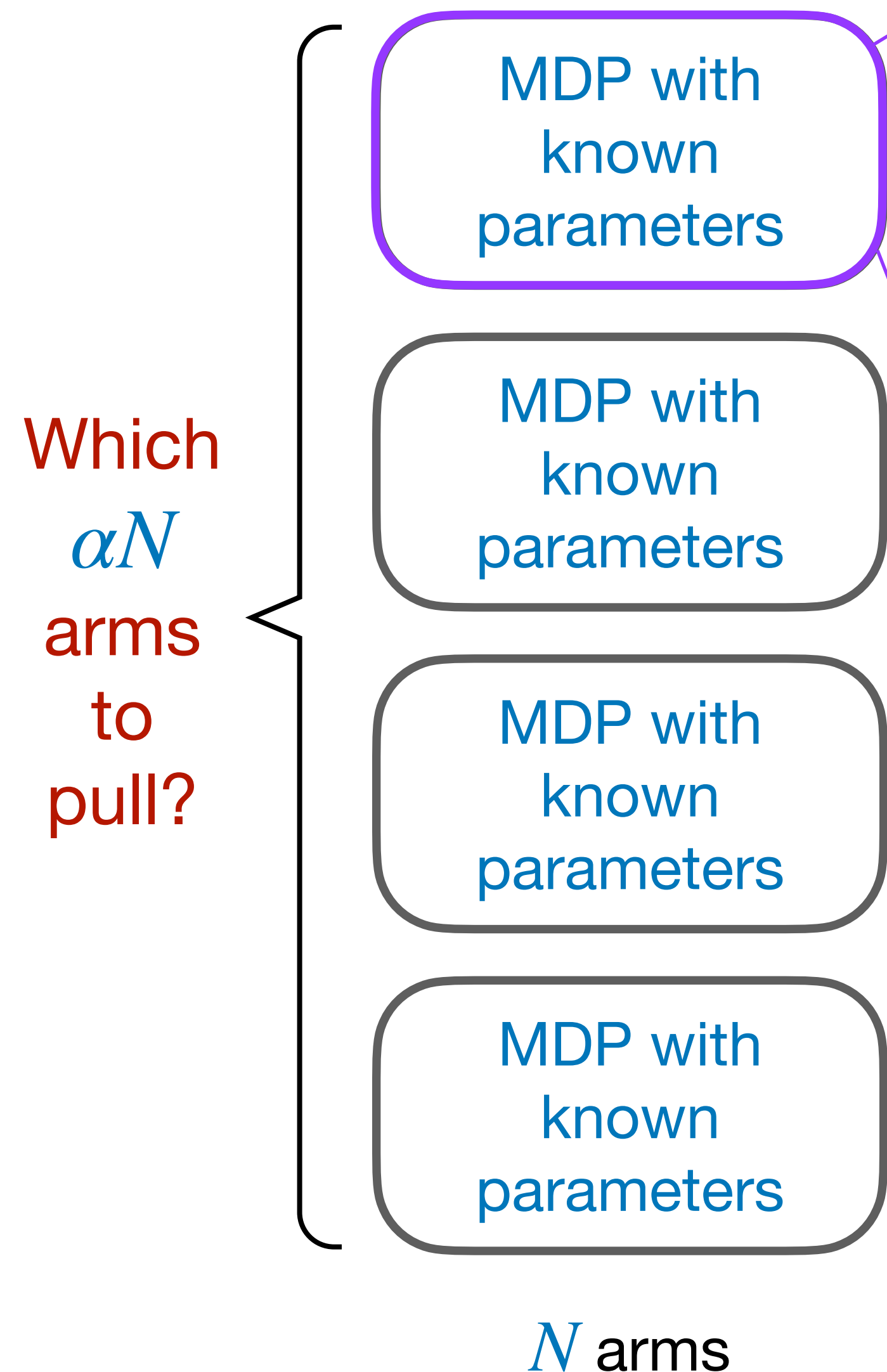
Restless bandits



Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”
 - Passive = “not pulling”
- Transition probabilities:
 - $P(s' \mid s, \text{active})$
 - $P(s' \mid s, \text{passive})$ “restless”

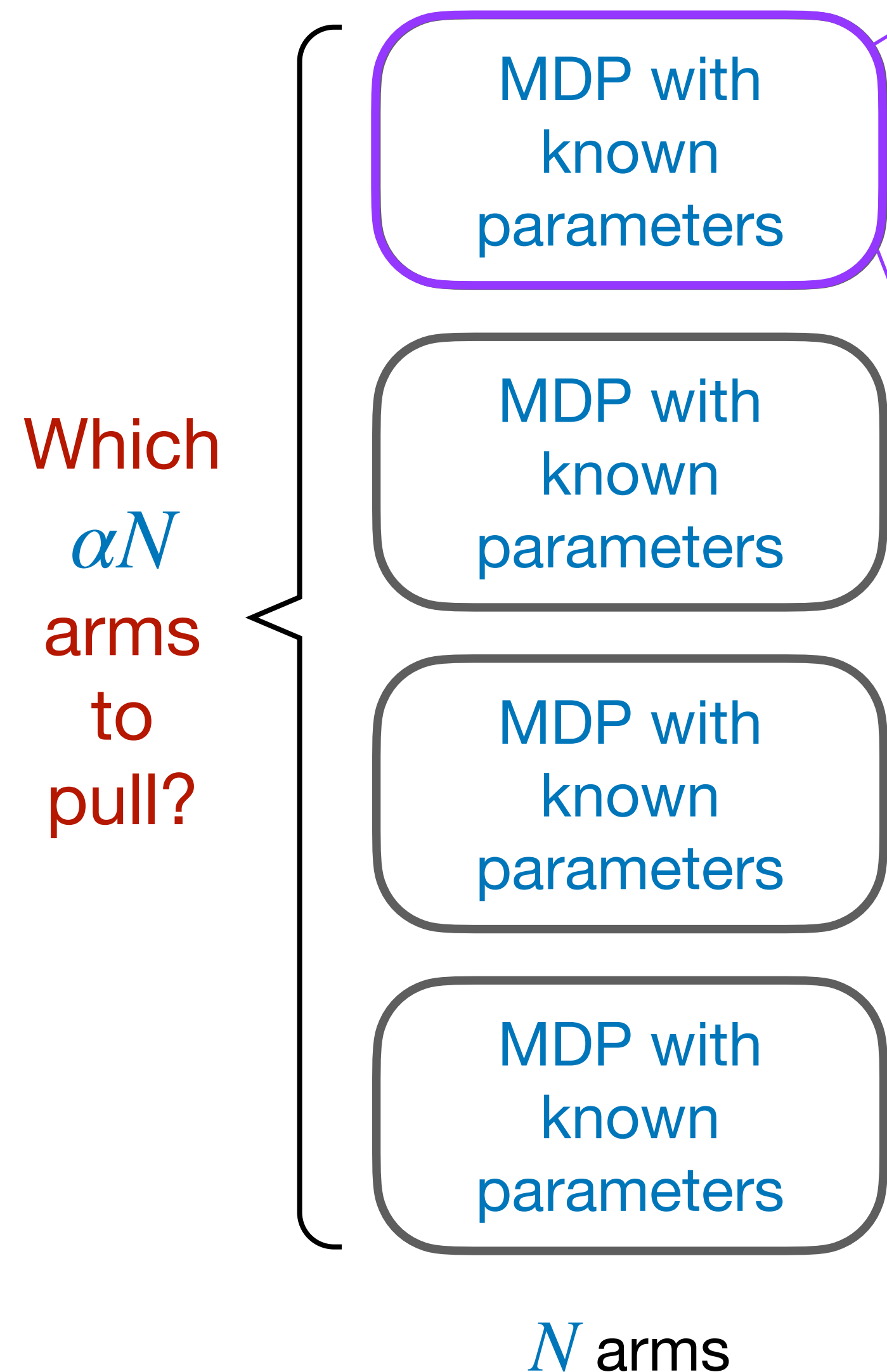
Restless bandits



Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”
 - Passive = “not pulling”
- Transition probabilities:
 - $P(s' \mid s, \text{active})$
 - $P(s' \mid s, \text{passive})$ “restless”
- Reward: $r(s, a)$

Restless bandits



Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”
 - Passive = “not pulling”
- Transition probabilities:
 - $P(s' | s, \text{active})$
 - $P(s' | s, \text{passive})$ “restless”
- Reward: $r(s, a)$

Goal: Design a policy to maximize long-term average of total reward

Restless bandits

Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r)$:

- State space: \mathcal{S} , finite
- Action space: $\mathcal{A} = \{\text{active}, \text{passive}\}$
 - Active = “pulling”
 - Passive = “not pulling”
- Transition probabilities:
 - $P(s' | s, \text{active})$
 - $P(s' | s, \text{passive})$ “restless”
- Reward: $r(s, a)$

Which
 αN
arms
to
pull?

αN : budget

MDP with
known
parameters

MDP with
known
parameters

MDP with
known
parameters

MDP with
known
parameters

N arms

Goal: Design a policy to maximize long-term average of total reward

Restless bandits

Which
 αN
arms
to
pull?

MDP with
known
parameters

MDP with
known
parameters

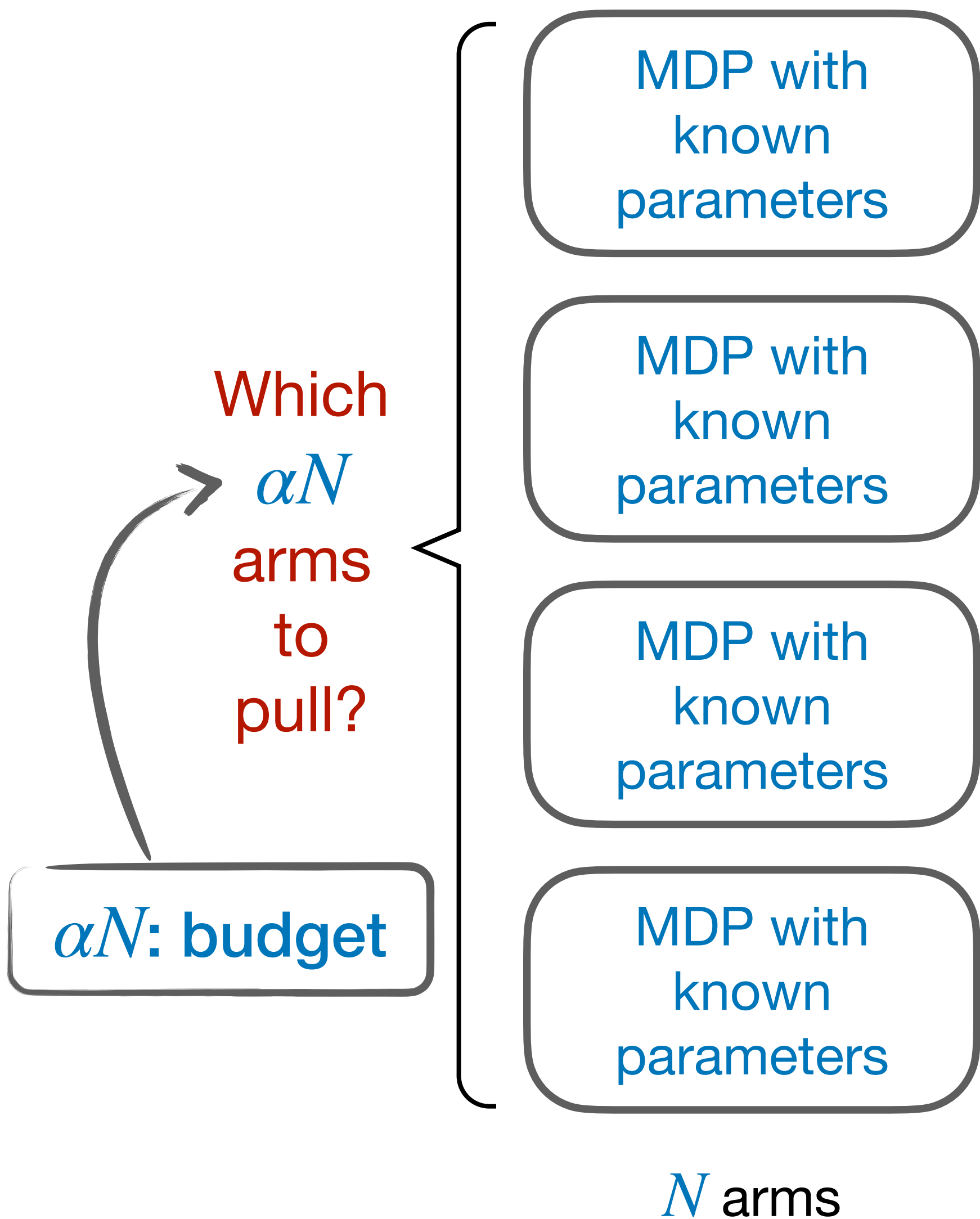
MDP with
known
parameters

MDP with
known
parameters

αN : budget

N arms

Restless bandits



maximize
policy π

subject to

N -armed MDP

$$V_N^\pi \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^\pi(t), A_i^\pi(t))]$$

$$\sum_{i=1}^N \mathbf{1}_{\{A_i^\pi(t)=\text{active}\}} = \alpha N, \quad \forall t \geq 0$$

Restless bandits

Which αN arms to pull?

MDP with known parameters

MDP with known parameters

MDP with known parameters

MDP with known parameters

N arms

αN : budget

N -armed MDP

maximize
policy π

$$V_N^\pi \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^\pi(t), A_i^\pi(t))]$$

subject to

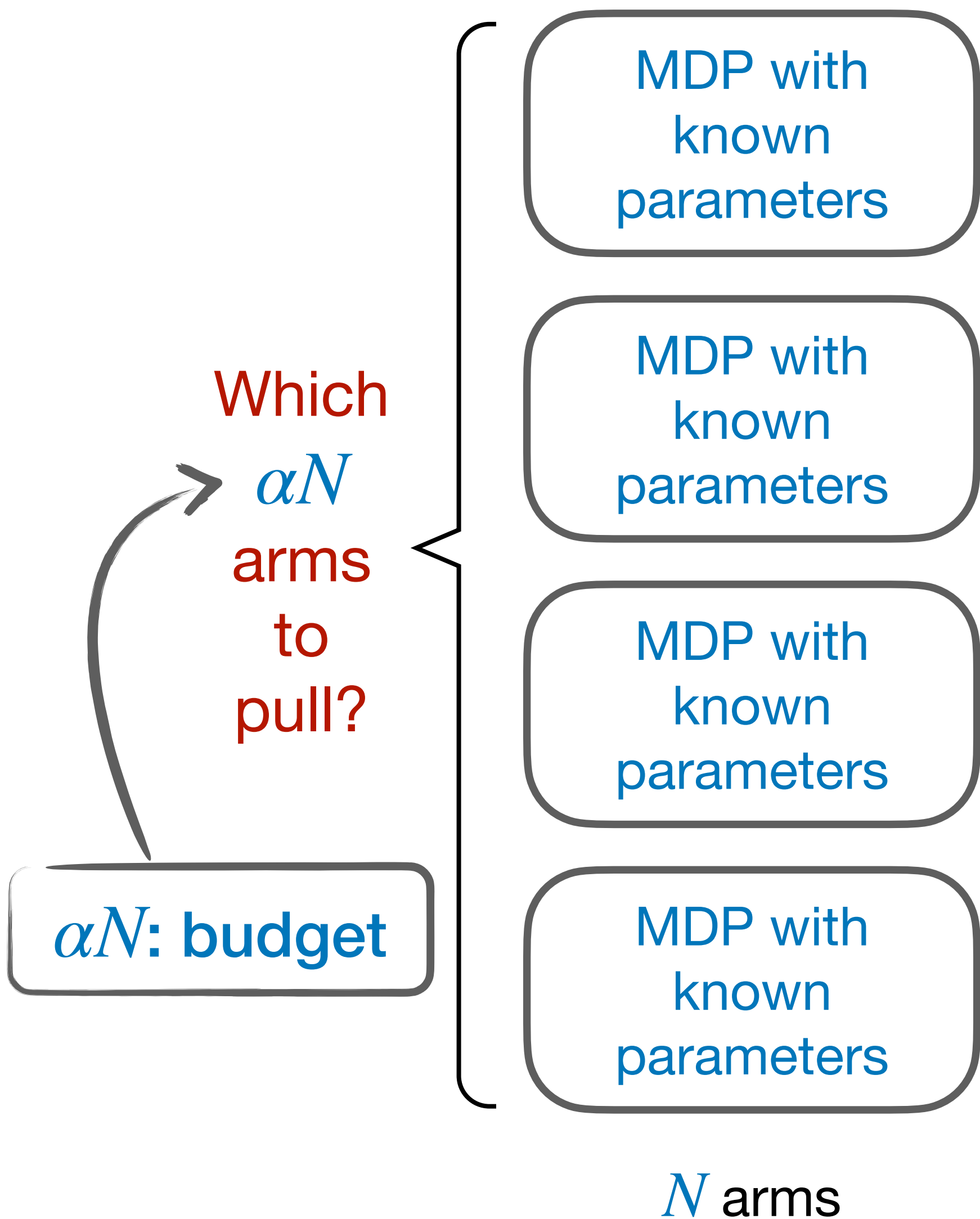
$$\sum_{i=1}^N \mathbf{1}_{\{A_i^\pi(t)=\text{active}\}} = \alpha N, \quad \forall t \geq 0$$

Goal (refined):

Design a computationally efficient policy π such that

$$V_N^* - V_N^\pi \rightarrow 0 \text{ as } N \rightarrow \infty$$

Restless bandits



N -armed MDP

maximize
policy π

$$V_N^\pi \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^\pi(t), A_i^\pi(t))]$$

subject to

$$\sum_{i=1}^N \mathbf{1}_{\{A_i^\pi(t)=\text{active}\}} = \alpha N, \quad \forall t \geq 0$$

Goal (refined):

Design a computationally efficient policy π such that

$$V_N^* - V_N^\pi \rightarrow 0 \text{ as } N \rightarrow \infty$$

↓
optimality gap

Prior work

Discrete
time
setting

Continuous
time
setting

Prior work

	Paper	Policy	Optimality Gap	Conditions*
Discrete time setting				
Continuous time setting				

Prior work

	Paper	Policy	Optimality Gap	Conditions*
Discrete time setting				
Continuous time setting	Weber and Weiss 90	Whittle Index	$o(1)$	Indexability & UGAP
	Verloop 16	LP-Prioriy	$o(1)$	UGAP

Prior work

	Paper	Policy	Optimality Gap	Conditions*
Discrete time setting				
Continuous time setting	Weber and Weiss 90	Whittle Index	$o(1)$	Indexability & UGAP
	Verloop 16	LP-Prioriy	$o(1)$	UGAP
	Gast, Gaujal, and Yan 20	Whittle Index	$O\left(e^{-cN}\right)$	UGAP & Non-singular
	Gast, Gaujal, and Yan 22	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate

Prior work

	Paper	Policy	Optimality Gap	Conditions*
Discrete time setting	Gast, Gaujal, and Yan 20	Whittle Index	$O\left(e^{-cN}\right)$	UGAP & Non-singular
	Gast, Gaujal, and Yan 22	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
Continuous time setting	Weber and Weiss 90	Whittle Index	$o(1)$	Indexability & UGAP
	Verloop 16	LP-Prioriy	$o(1)$	UGAP
	Gast, Gaujal, and Yan 20	Whittle Index	$O\left(e^{-cN}\right)$	UGAP & Non-singular
	Gast, Gaujal, and Yan 22	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate

Prior work

Discrete
time
setting

Paper	Policy	Optimality Gap	Conditions*
Gast, Gaujal, and Yan 20	Whittle Index	$O(e^{-cN})$	UGAP & Non-singular
Gast, Gaujal, and Yan 22	LP-Priority	$O(e^{-cN})$	UGAP & Non-degenerate

Can we achieve asymptotic optimality without UGAP?

Continuous
time
setting

Weber and Weiss 90	Whittle Index	$o(1)$	Indexability & UGAP
Verloop 16	LP-Priority	$o(1)$	UGAP
Gast, Gaujal, and Yan 20	Whittle Index	$O(e^{-cN})$	UGAP & Non-singular
Gast, Gaujal, and Yan 22	LP-Priority	$O(e^{-cN})$	UGAP & Non-degenerate

Prior work

Discrete
time
setting

Paper	Policy	Optimality Gap	Conditions*
Gast, Gaujal, and Yan 20	Whittle Index	$O(e^{-cN})$	UGAP & Non-singular
Gast, Gaujal, and Yan 22	LP-Priority	$O(e^{-cN})$	UGAP & Non-degenerate

Can we achieve asymptotic optimality without UGAP?

Continuous
time
setting

Gast, Gaujal, and Yan 20	Whittle Index	$O(e^{-cN})$	UGAP & Non-singular
Gast, Gaujal, and Yan 22	LP-Priority	$O(e^{-cN})$	UGAP & Non-degenerate

Can we establish a non-trivial optimality gap without non-singular/non-degenerate assumption (and UGAP)?

Our results

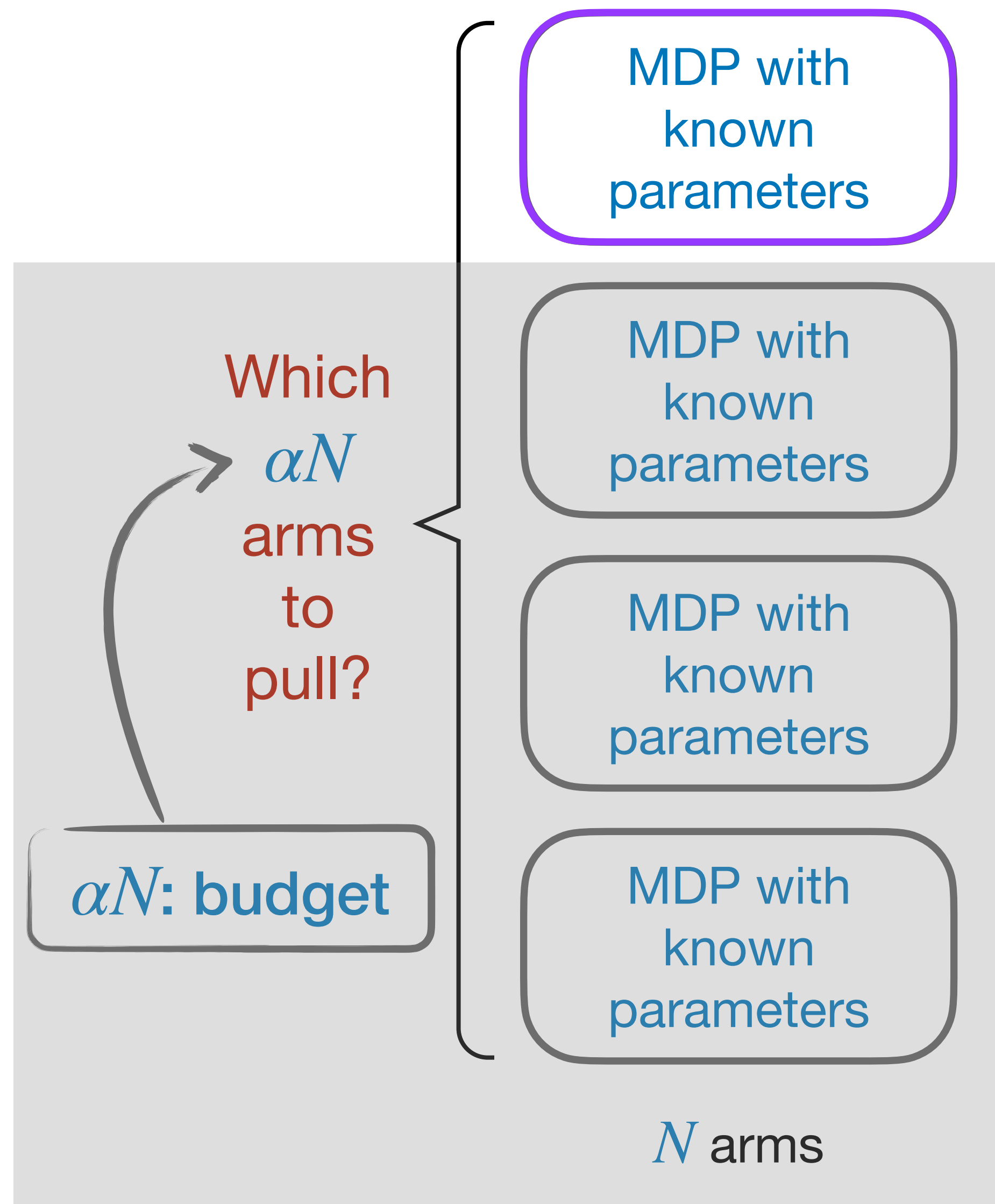
**Discrete
time
setting**

Paper	Policy	Optimality Gap	Conditions*
Gast, Gaujal, and Yan 20	Whittle Index	$O(e^{-cN})$	UGAP & Non-singular
Gast, Gaujal, and Yan 22	LP-Priority	$O(e^{-cN})$	UGAP & Non-degenerate
Our paper	FTVA	$O(1/\sqrt{N})$	SA

Our results

	Paper	Policy	Optimality Gap	Conditions*
Discrete time setting	Gast, Gaujal, and Yan 20	Whittle Index	$O\left(e^{-cN}\right)$	UGAP & Non-singular
	Gast, Gaujal, and Yan 22	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
	Our paper	FTVA	$O(1/\sqrt{N})$	SA
Continuous time setting	Weber and Weiss 90	Whittle Index	$o(1)$	UGAP
	Verloop 16	LP-Prioriy	$o(1)$	UGAP
	Gast, Gaujal, and Yan 20	Whittle Index	$O\left(e^{-cN}\right)$	UGAP & Non-singular
	Gast, Gaujal, and Yan 22	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
	Our paper	FTVA-CT	$O(1/\sqrt{N})$	—

Restless bandits



Single-armed MDP

N -armed MDP

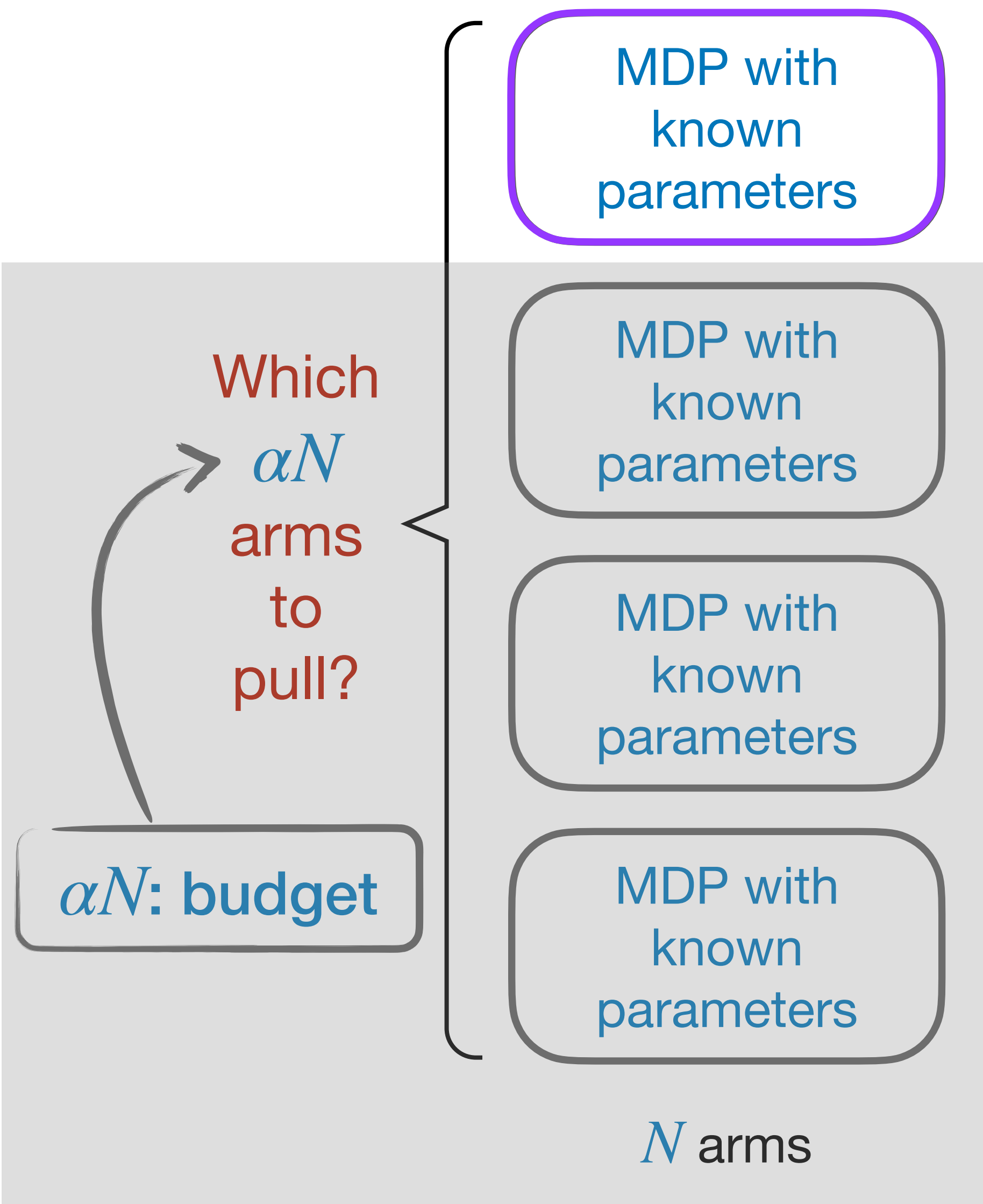
maximize
policy π

subject to

$$V_N^\pi \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^\pi(t), A_i^\pi(t))]$$

$$\sum_{i=1}^N \mathbf{1}_{\{A_i^\pi(t)=\text{active}\}} = \alpha N, \quad \forall t \geq 0$$

Restless bandits



Single-armed MDP

maximize
policy $\bar{\pi}$

$$V_1^{\bar{\pi}} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[r(S_1^{\bar{\pi}}(t), A_1^{\bar{\pi}}(t))]$$

N-armed MDP

maximize
policy π

$$V_N^{\pi} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^{\pi}(t), A_i^{\pi}(t))]$$

subject to

$$\sum_{i=1}^N \mathbf{1}_{\{A_i^{\pi}(t)=\text{active}\}} = \alpha N, \quad \forall t \geq 0$$

Restless bandits

Which αN arms to pull?

MDP with known parameters

MDP with known parameters

MDP with known parameters

MDP with known parameters

N arms

αN : budget

Single-armed MDP

maximize
policy $\bar{\pi}$

$$V_1^{\bar{\pi}} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[r(S_1^{\bar{\pi}}(t), A_1^{\bar{\pi}}(t))]$$

subject to

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\mathbf{1}_{\{A_1^{\bar{\pi}}(t)=\text{active}\}} \right] = \alpha$$

Relaxed budget constraint

N -armed MDP

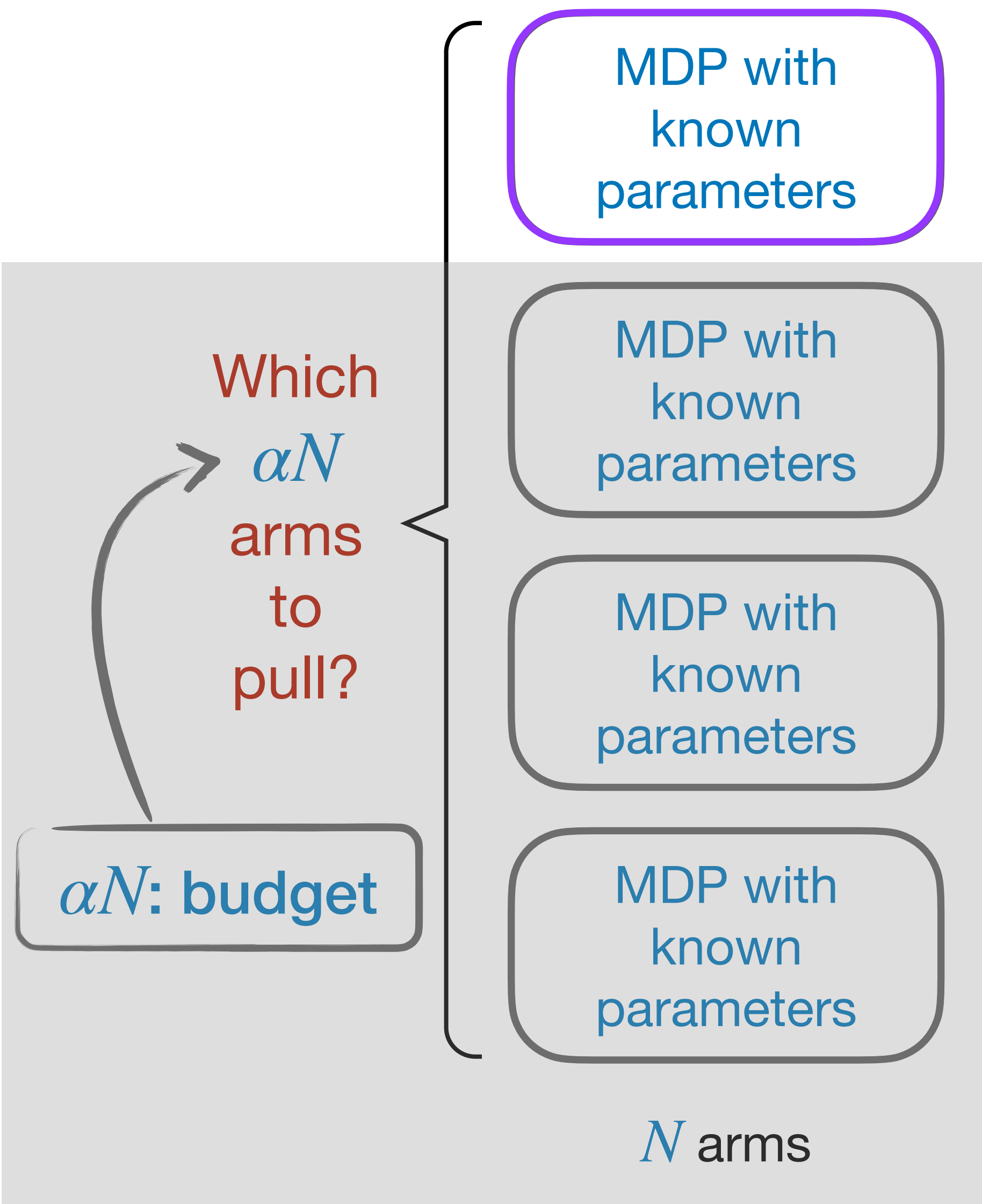
maximize
policy π

$$V_N^{\pi} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^{\pi}(t), A_i^{\pi}(t))]$$

subject to

$$\sum_{i=1}^N \mathbf{1}_{\{A_i^{\pi}(t)=\text{active}\}} = \alpha N, \quad \forall t \geq 0$$

Restless bandits



Single-armed MDP

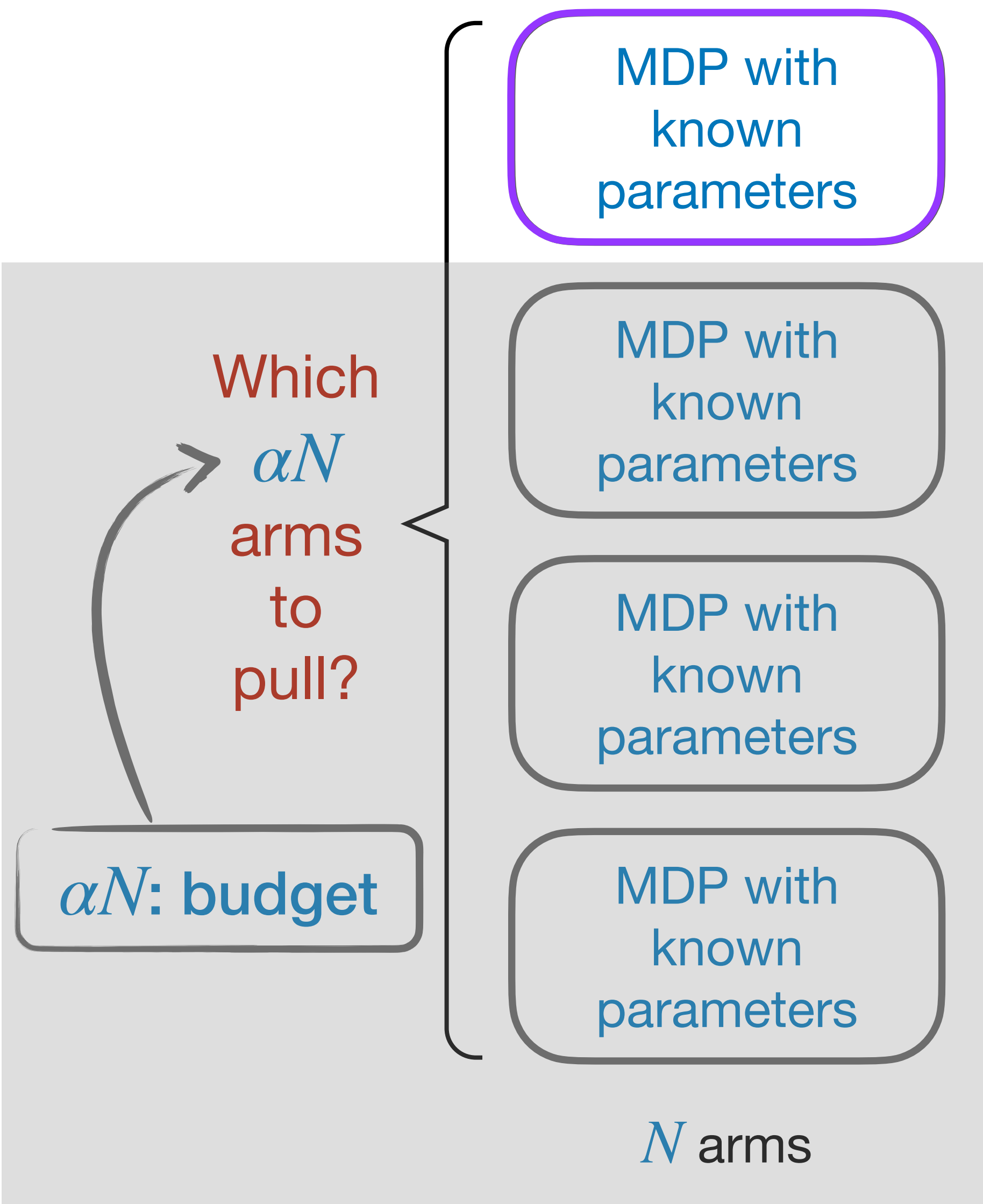
maximize
policy $\bar{\pi}$

$$V_1^{\bar{\pi}} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[r(S_1^{\bar{\pi}}(t), A_1^{\bar{\pi}}(t))]$$

subject to

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\mathbf{1}_{\{A_1^{\bar{\pi}}(t)=\text{active}\}} \right] = \alpha$$

Restless bandits



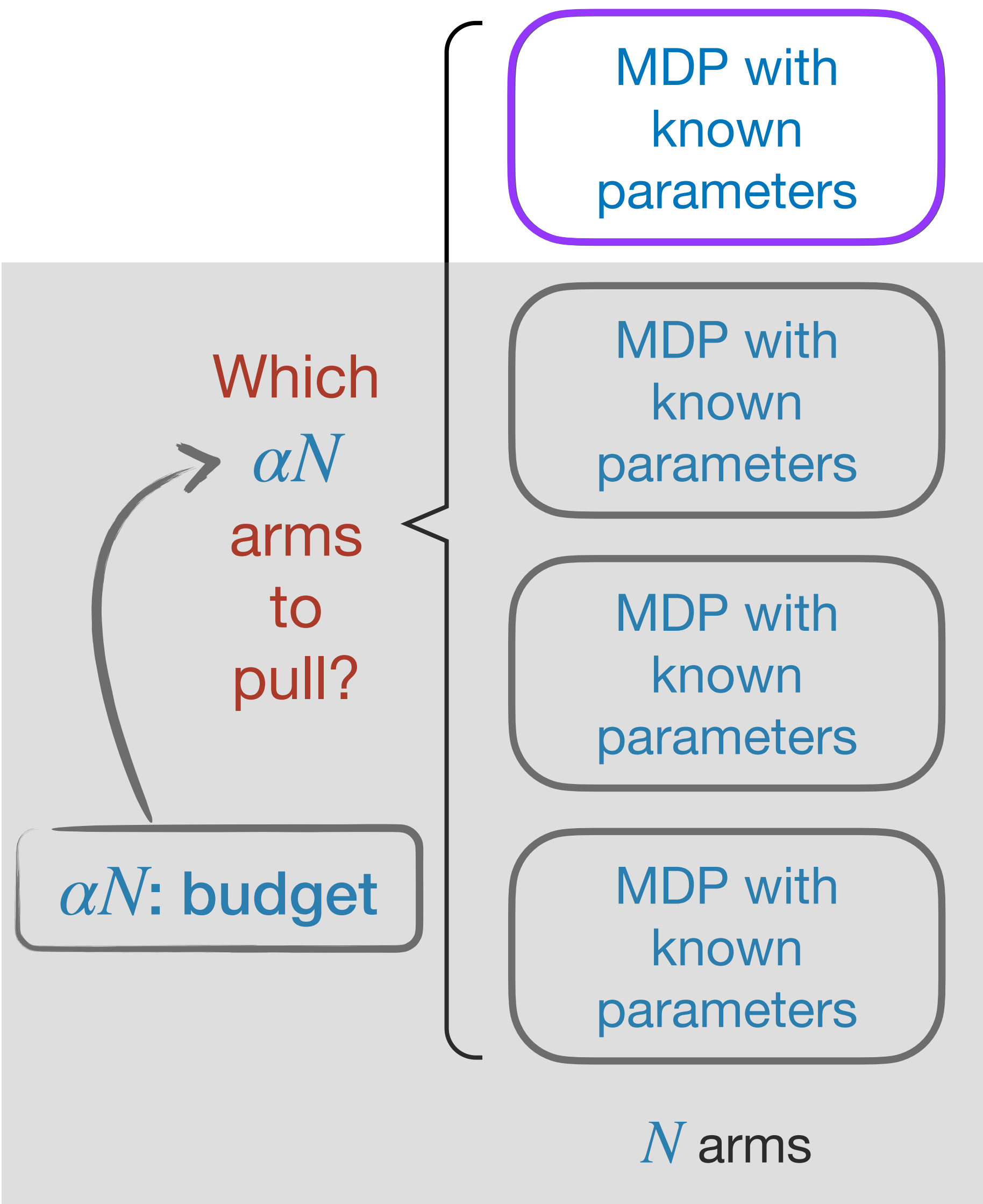
Single-armed MDP

maximize policy $\bar{\pi}$ $V_1^{\bar{\pi}} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[r(S_1^{\bar{\pi}}(t), A_1^{\bar{\pi}}(t))]$

subject to $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\mathbf{1}_{\{A_1^{\bar{\pi}}(t)=\text{active}\}} \right] = \alpha$

- Upper bound: $V_N^* \leq V_1^{\bar{\pi}^*}$, for all N

Restless bandits



Single-armed MDP

maximize policy $\bar{\pi}$ $V_1^{\bar{\pi}} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[r(S_1^{\bar{\pi}}(t), A_1^{\bar{\pi}}(t))]$

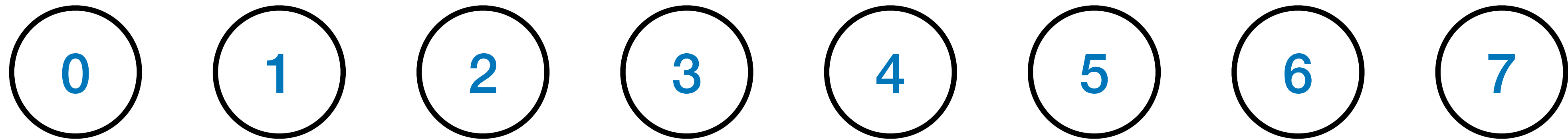
subject to $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\mathbf{1}_{\{A_1^{\bar{\pi}}(t)=\text{active}\}} \right] = \alpha$

- Upper bound: $V_N^* \leq V_1^{\bar{\pi}^*}$, for all N
- The single-armed MDP can be solved as a linear program

Example

MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

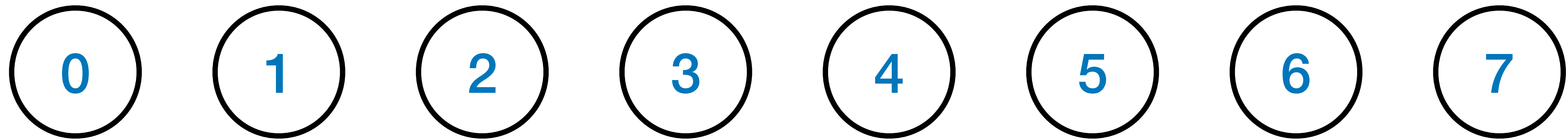


Example

MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

- State space of each arm: $\mathcal{S} = \{0, 1, 2, 3, 4, 5, 6, 7\}$

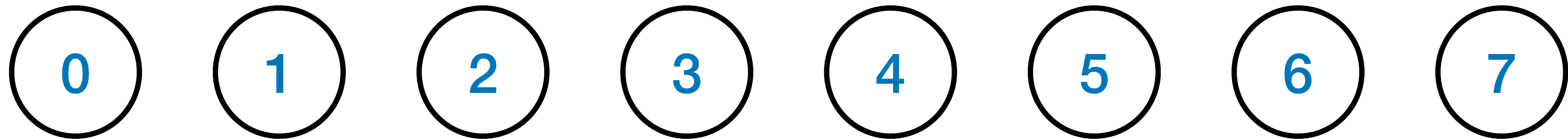


Example

MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

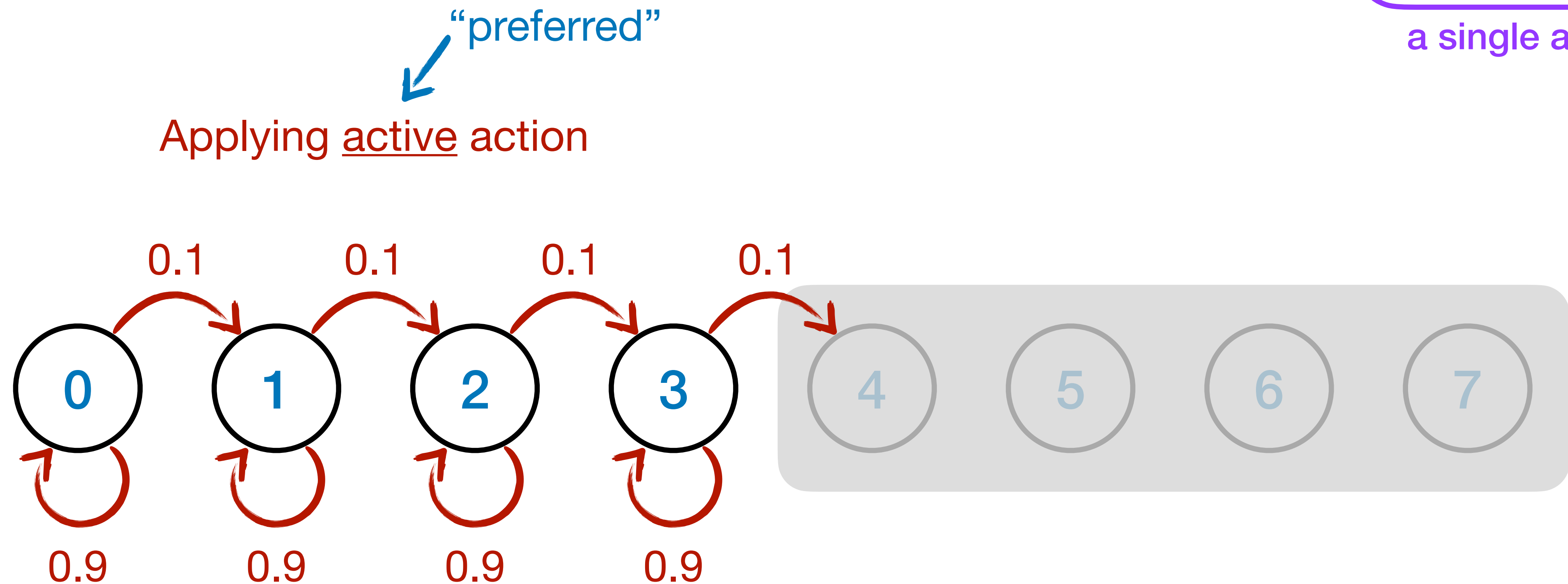
- State space of each arm: $\mathcal{S} = \{0, 1, 2, 3, 4, 5, 6, 7\}$
- Get a unit of reward only when going from State 7 to State 0



Example

MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

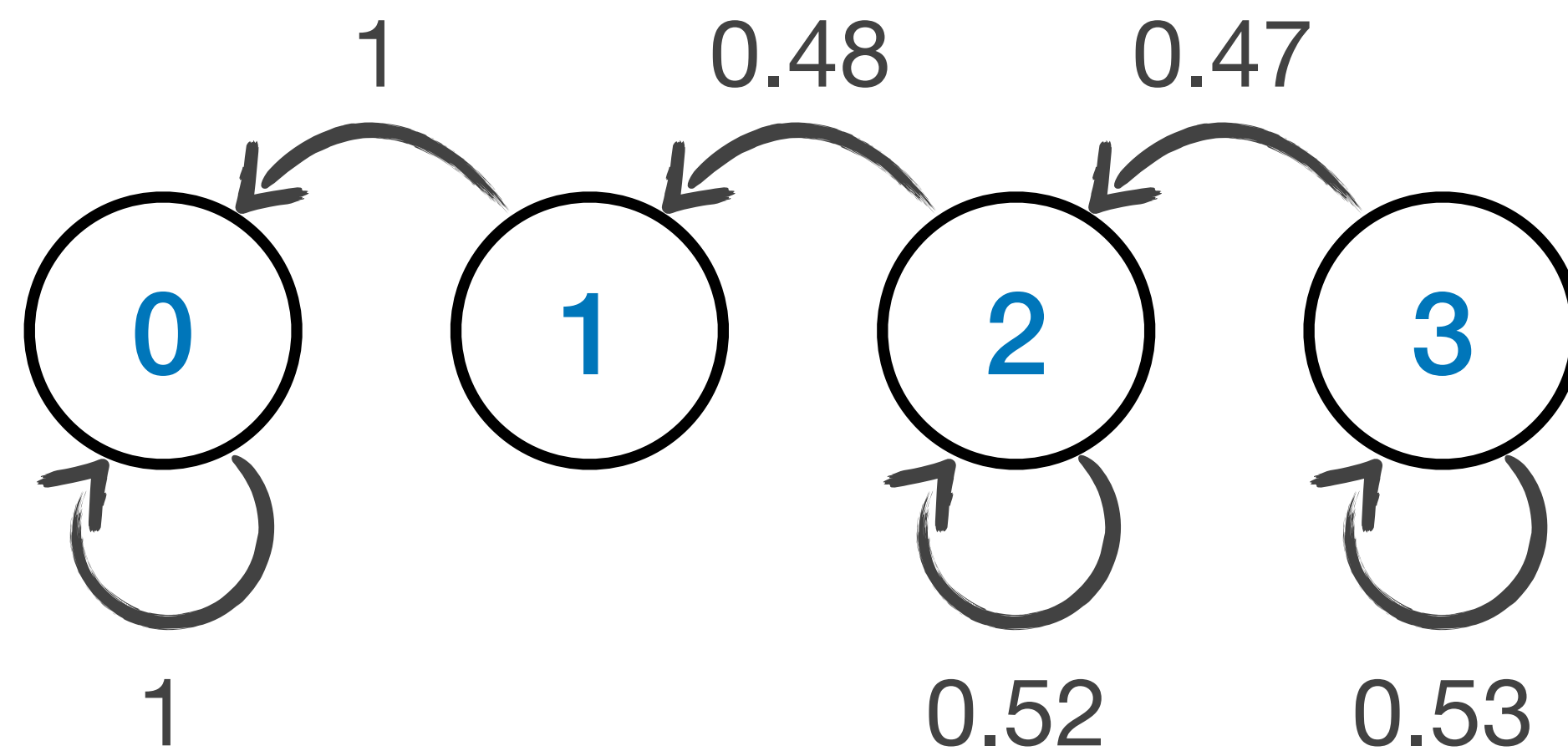


Example

MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

Applying passive action



Example

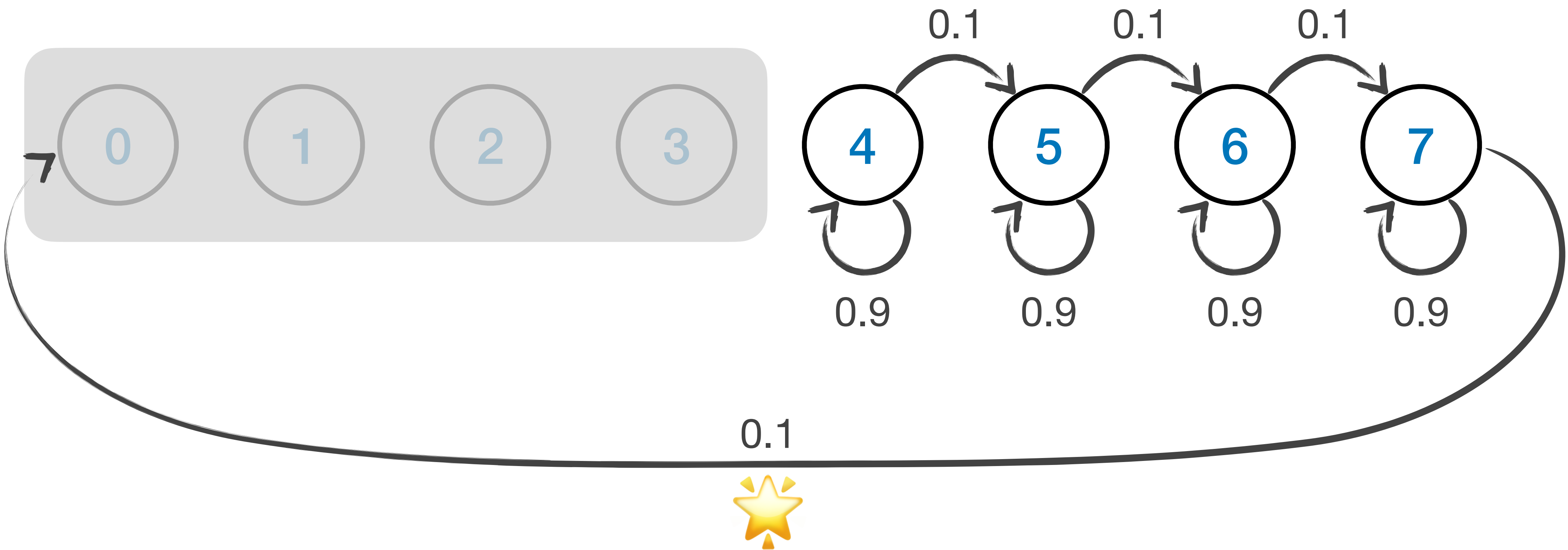
MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

“preferred”



Applying passive action

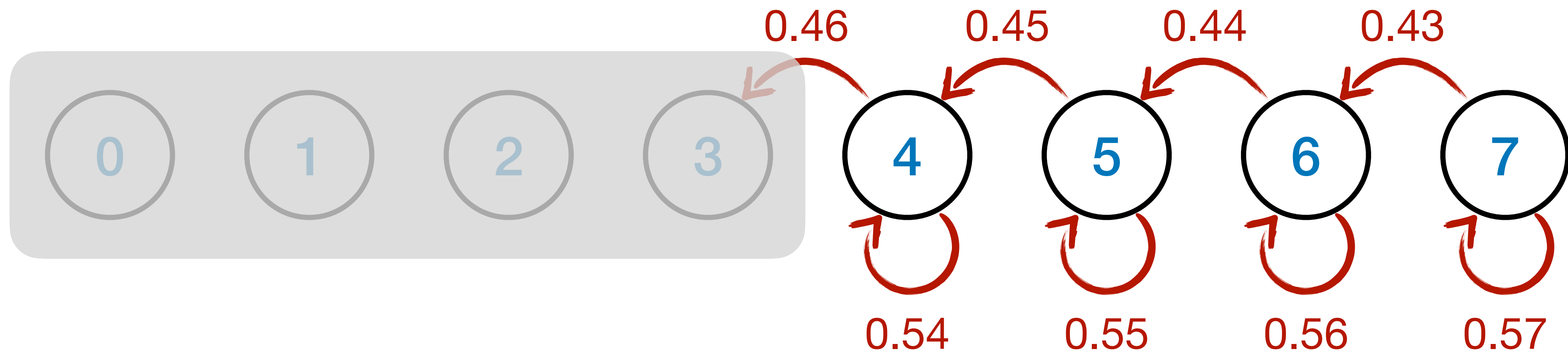


Example

MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

Applying active action



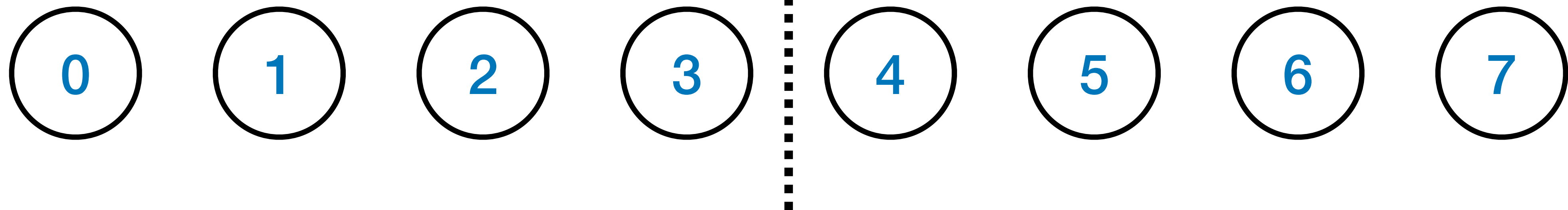
Example

MDP
 $(\mathcal{S}, \mathcal{A}, P, r)$

a single arm

Preferred action: active

Preferred action: passive



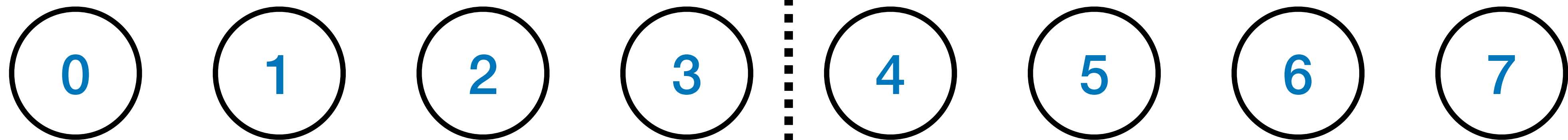
Example

MDP
($\mathcal{S}, \mathcal{A}, P, r$)

a single arm

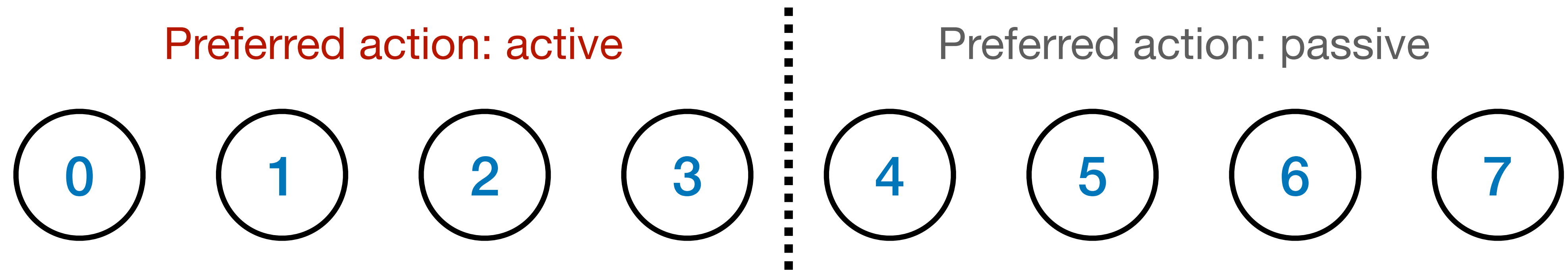
Preferred action: active

Preferred action: passive

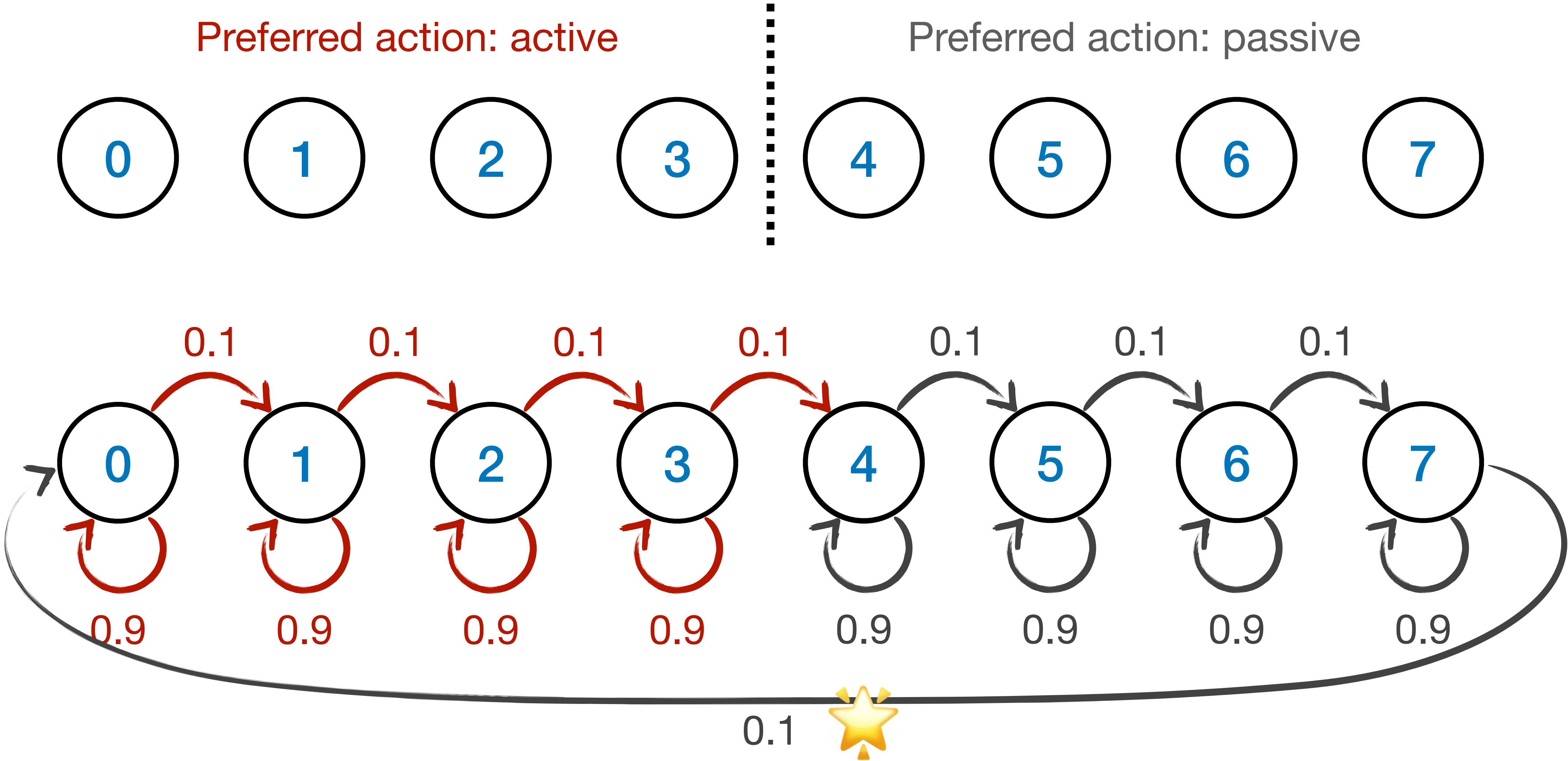


Budget: $\alpha = \frac{1}{2}$

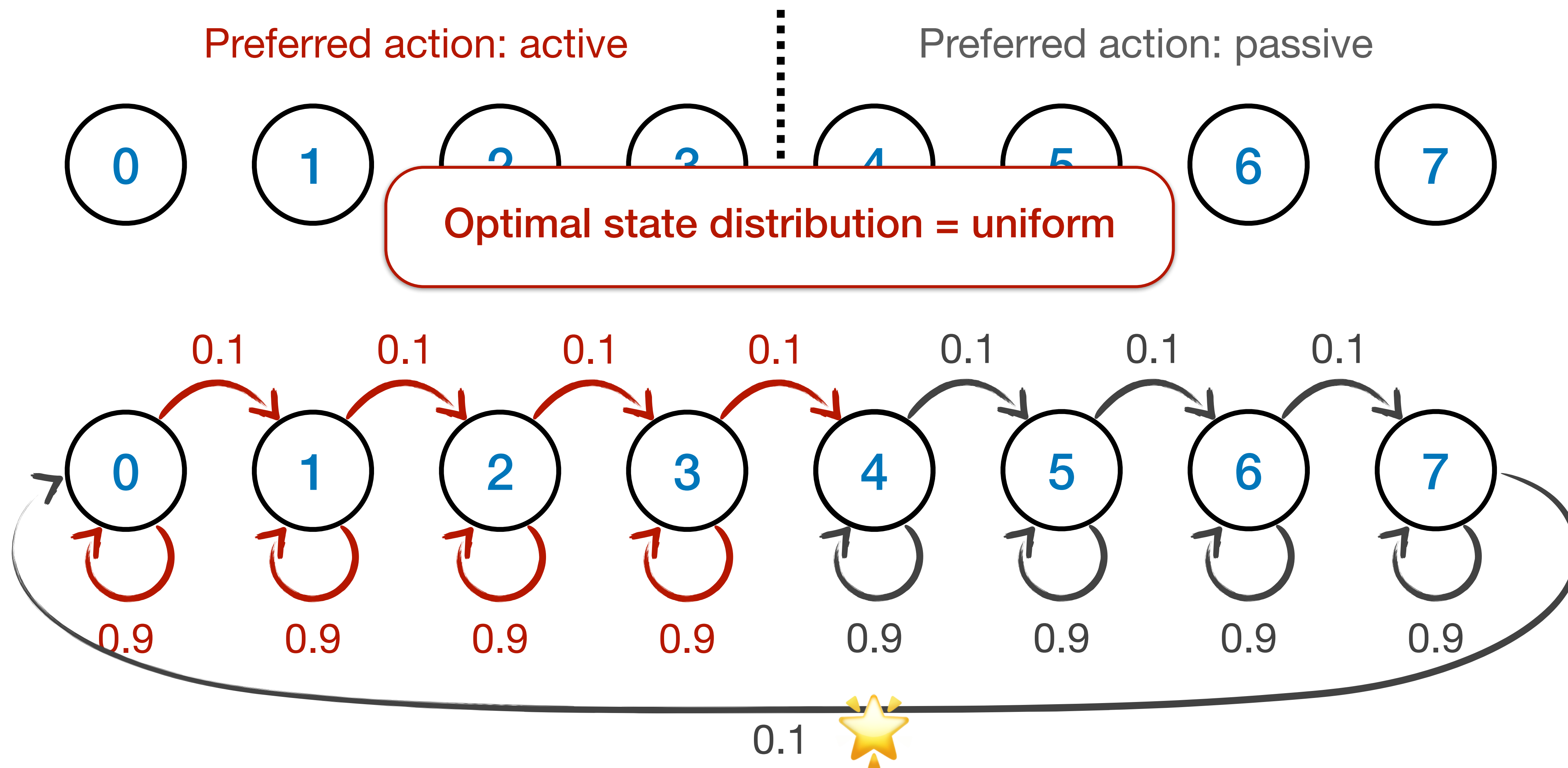
Optimal single-armed policy



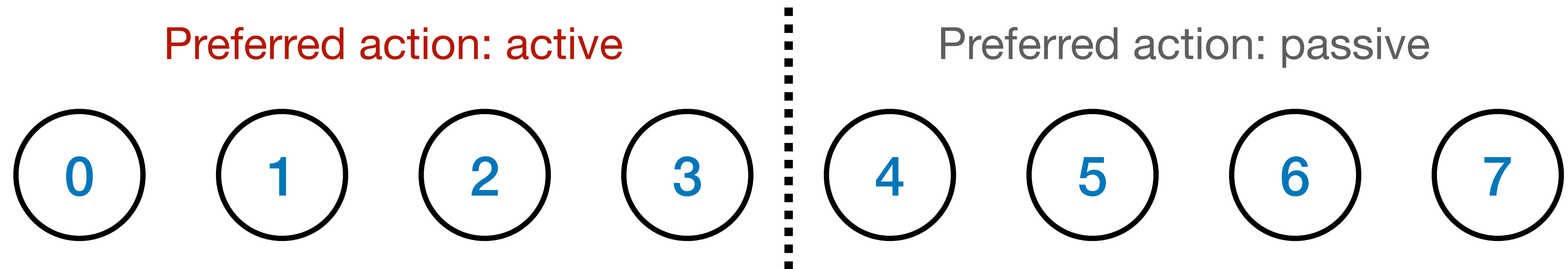
Optimal single-armed policy



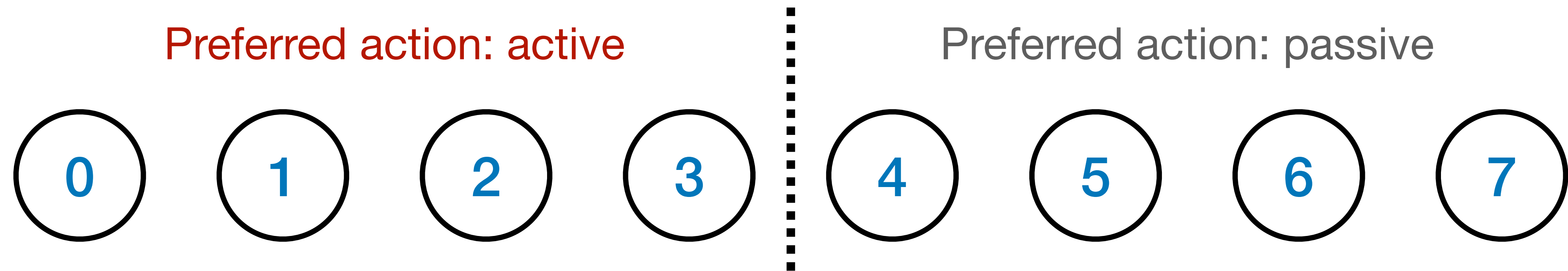
Optimal single-armed policy



LP-priority policy

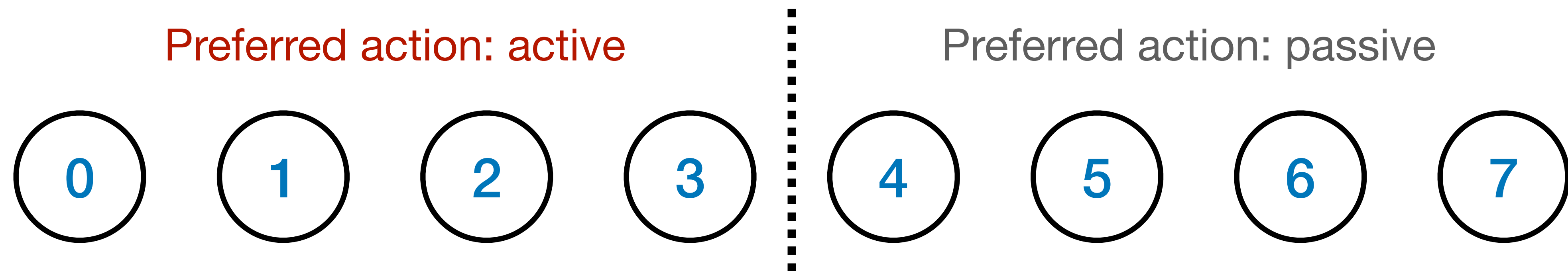


LP-priority policy



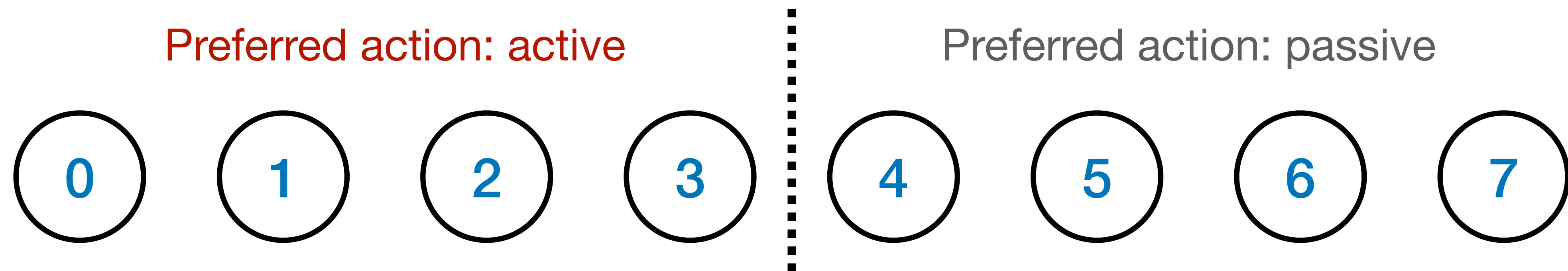
- LP-priority policy gives the states a priority order

LP-priority policy



- LP-priority policy gives the states a priority order
- In the N -armed system, the policy starts pulling arms in the state with the highest priority, and then goes down the order until it has pulled αN arms

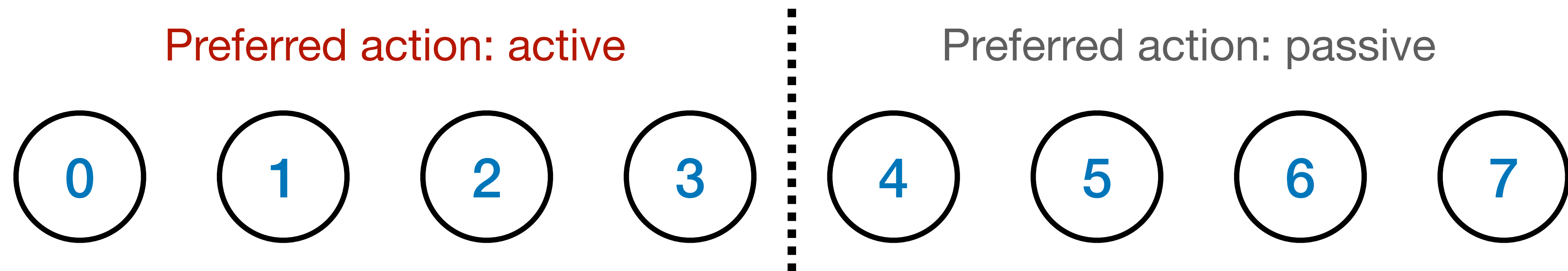
LP-priority policy



- LP-priority policy gives the states a priority order
- In the N -armed system, the policy starts pulling arms in the state with the highest priority, and then goes down the order until it has pulled αN arms

What priority order would you assign to the states in this example?

LP-priority policy

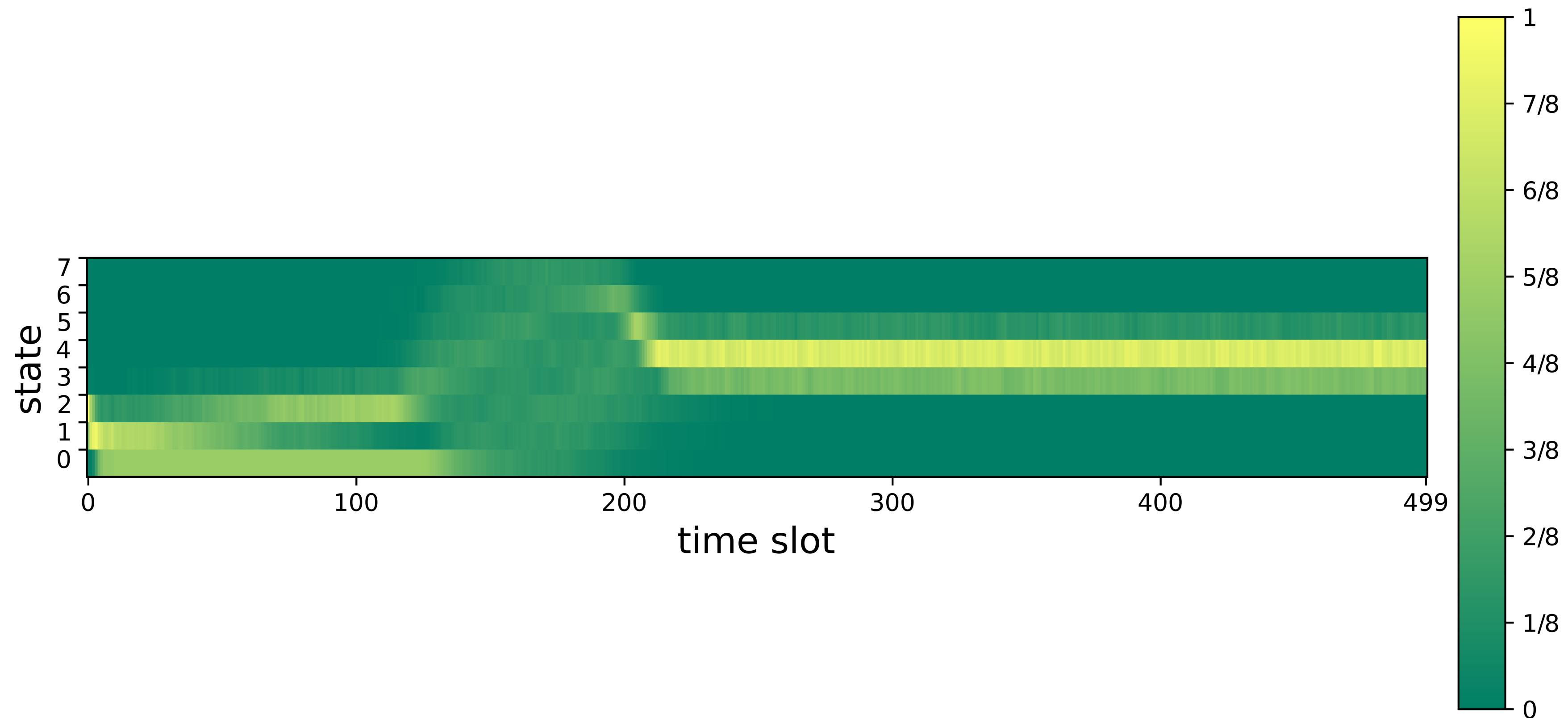


- LP-priority policy gives the states a priority order
- In the N -armed system, the policy starts pulling arms in the state with the highest priority, and then goes down the order until it has pulled αN arms

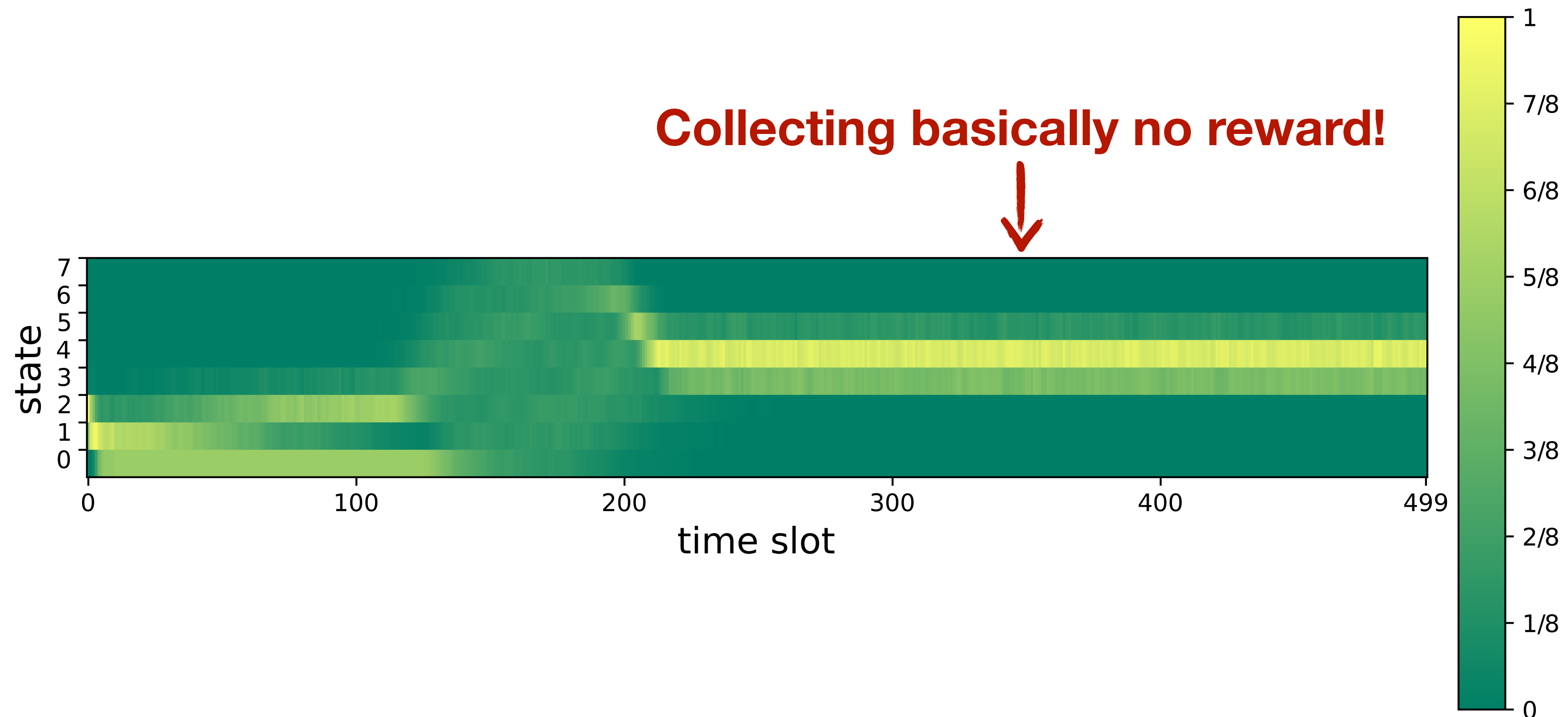
What priority order would you assign to the states in this example?

LP-index: $1 > 2 > 3 > 0 > 7 > 6 > 5 > 4$

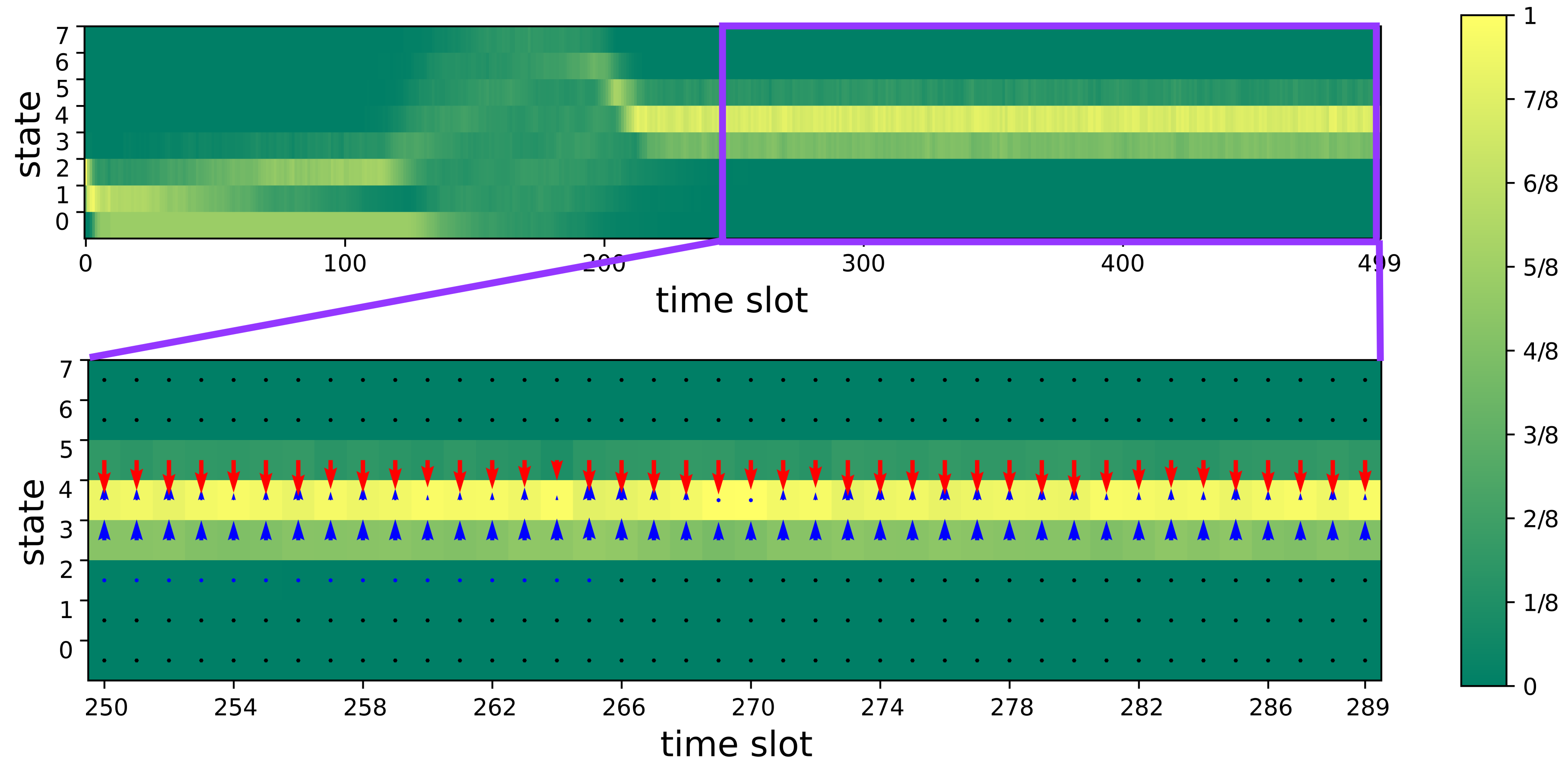
LP-priority: Empirical state distribution over time



LP-priority: Empirical state distribution over time

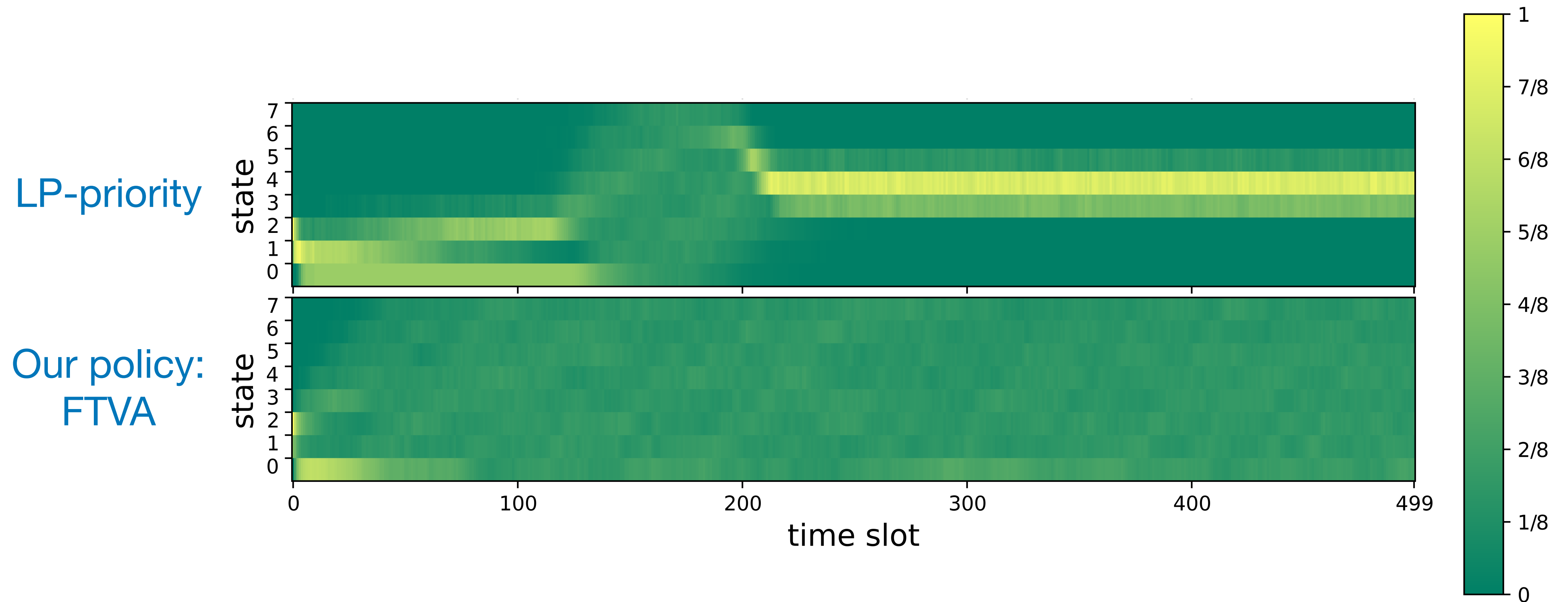


LP-priority: Why suboptimal?

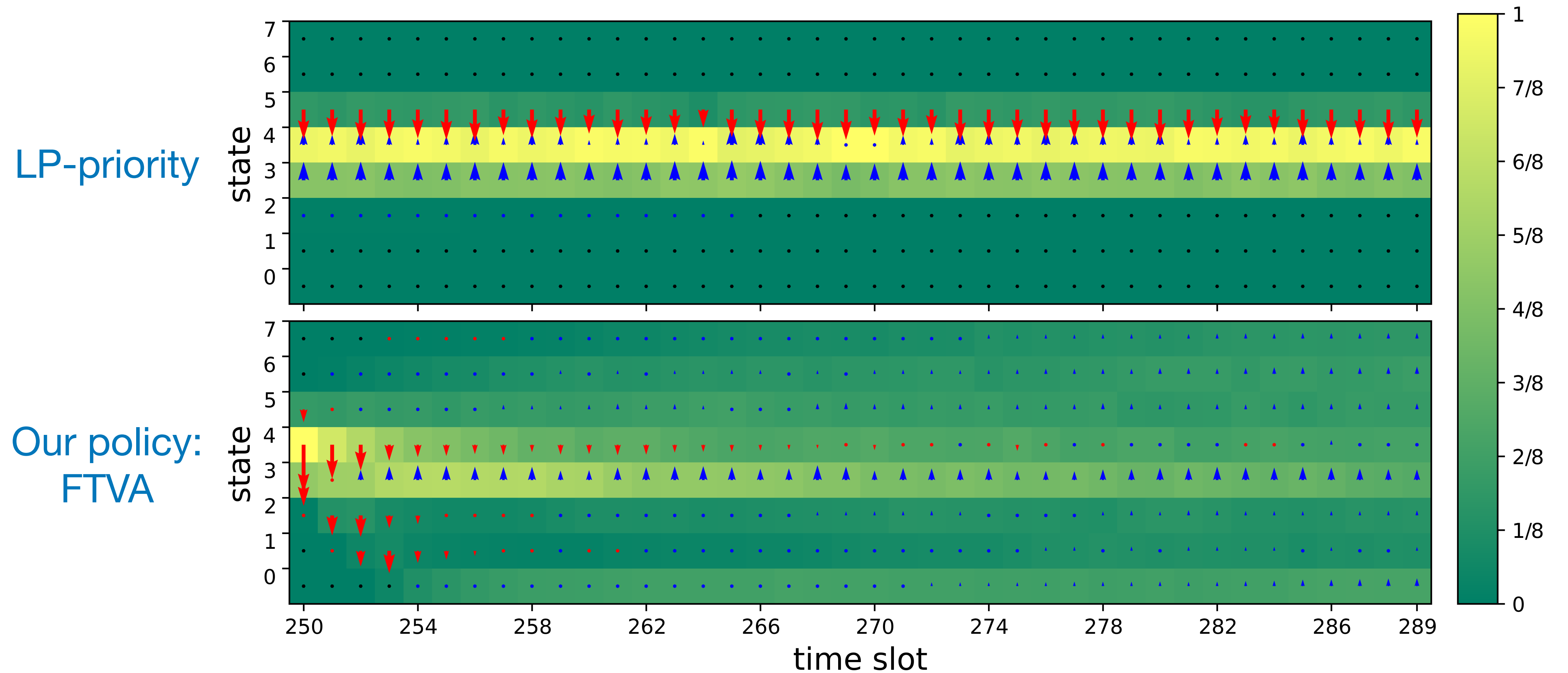


Priority order: 1 > 2 > 3 > 0 > 7 > 6 > 5 > 4

LP-priority vs our policy



LP-priority vs our policy



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$

Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

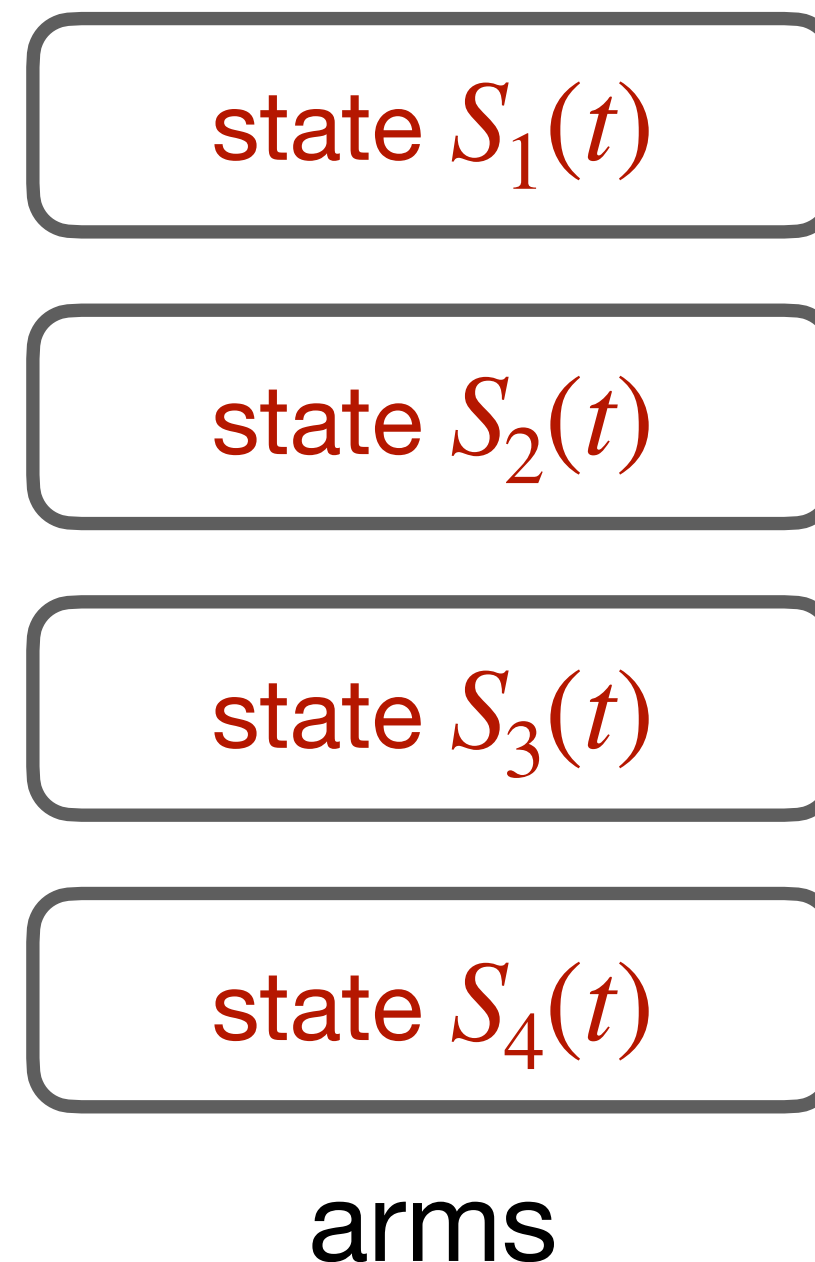
- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$

Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

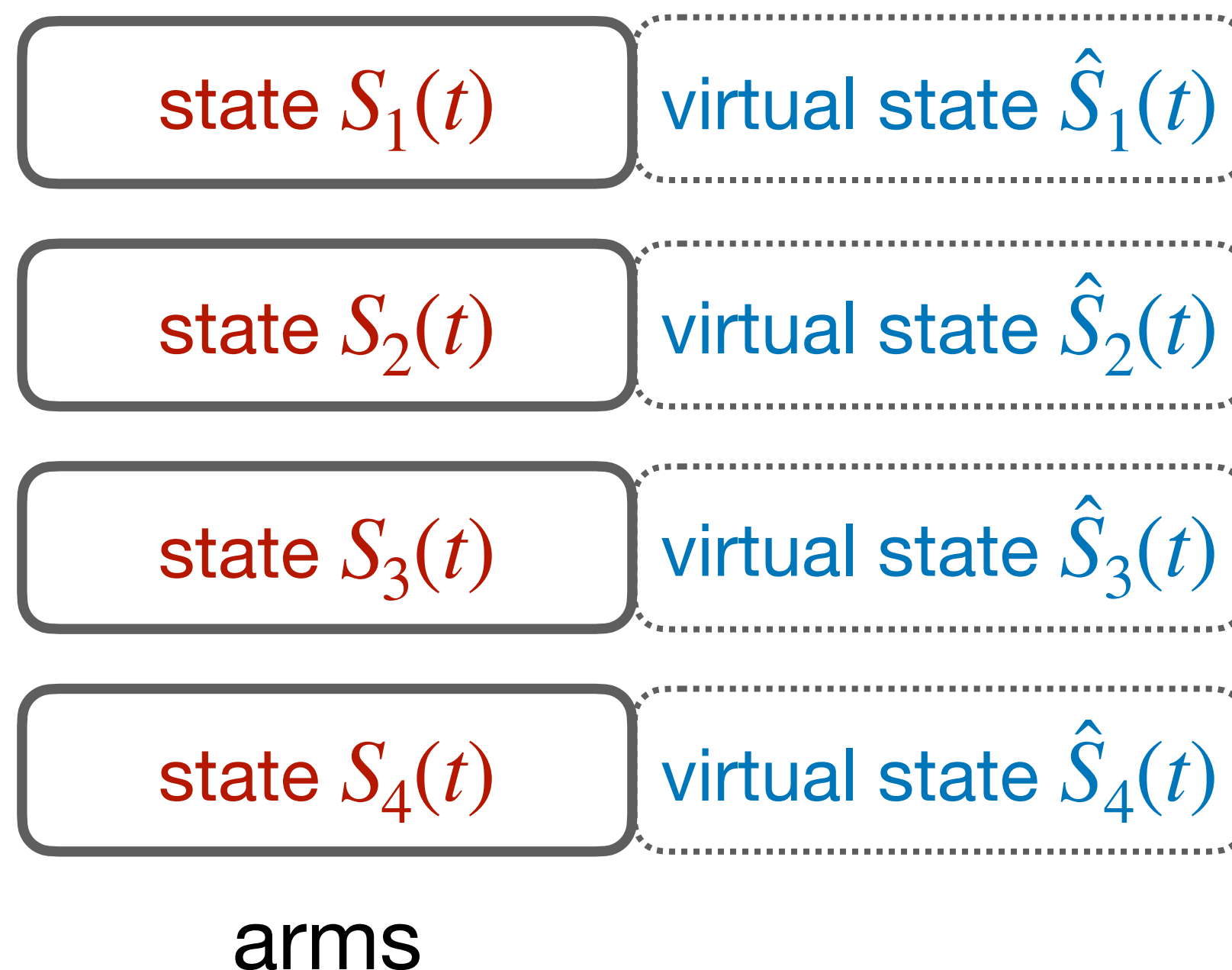
Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



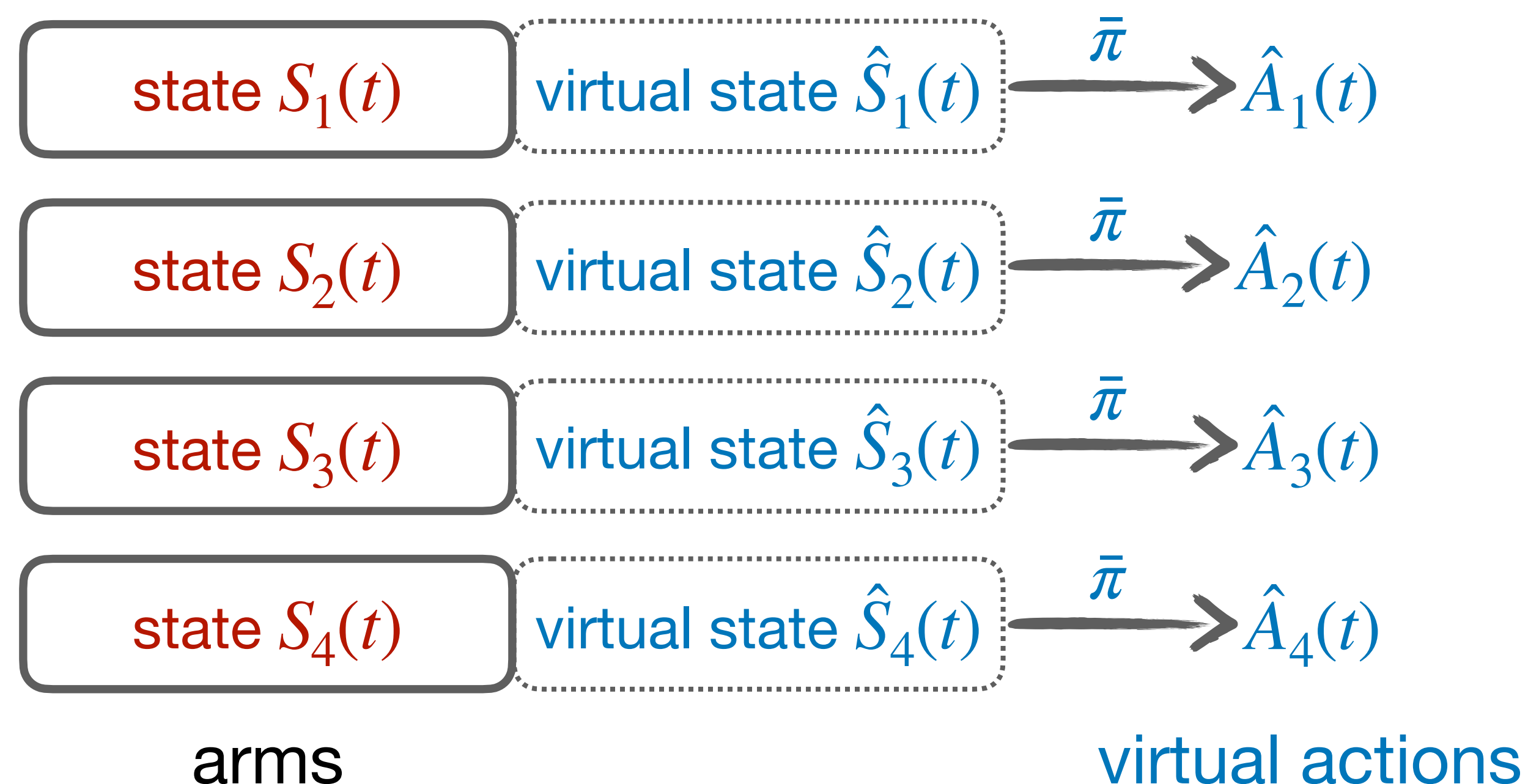
Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



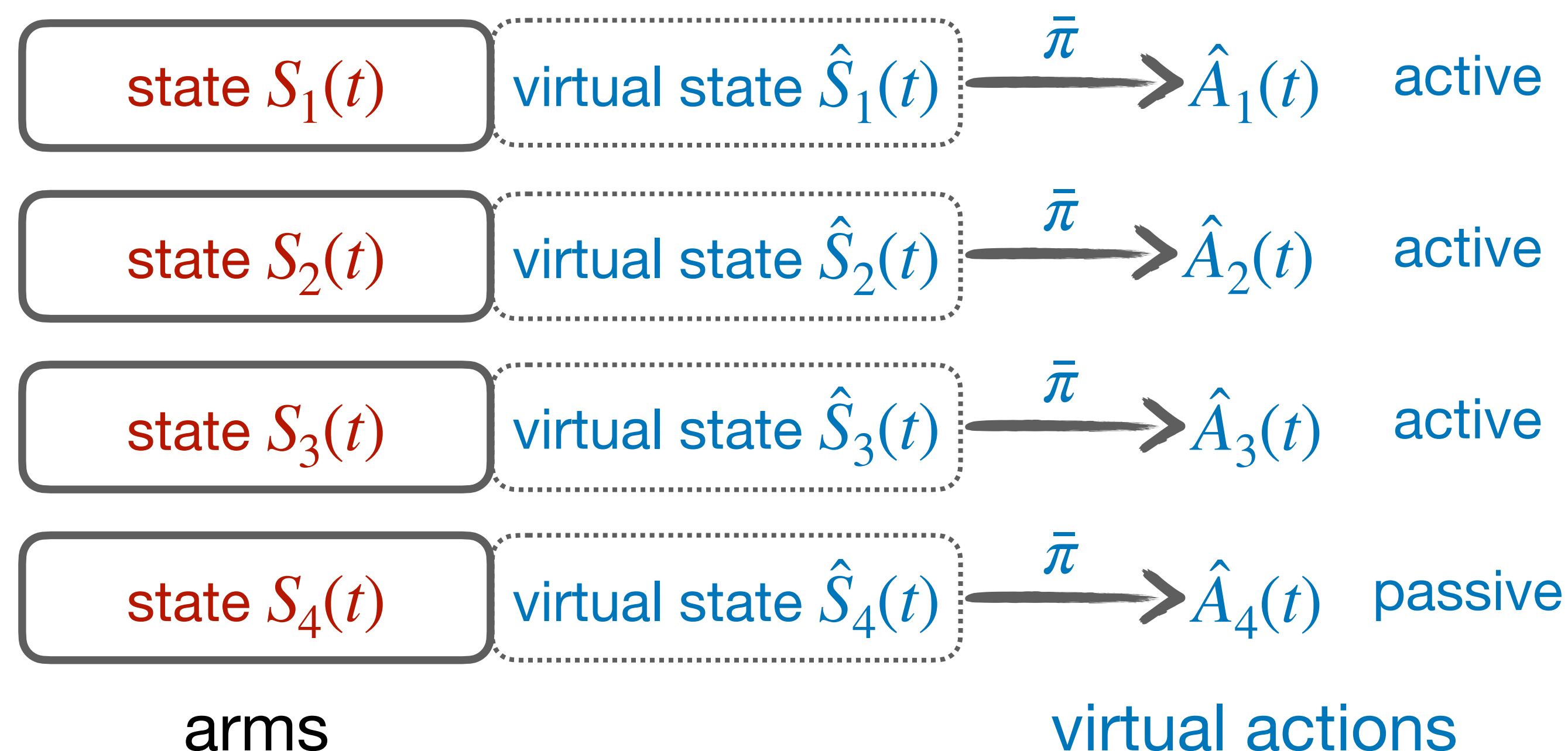
Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



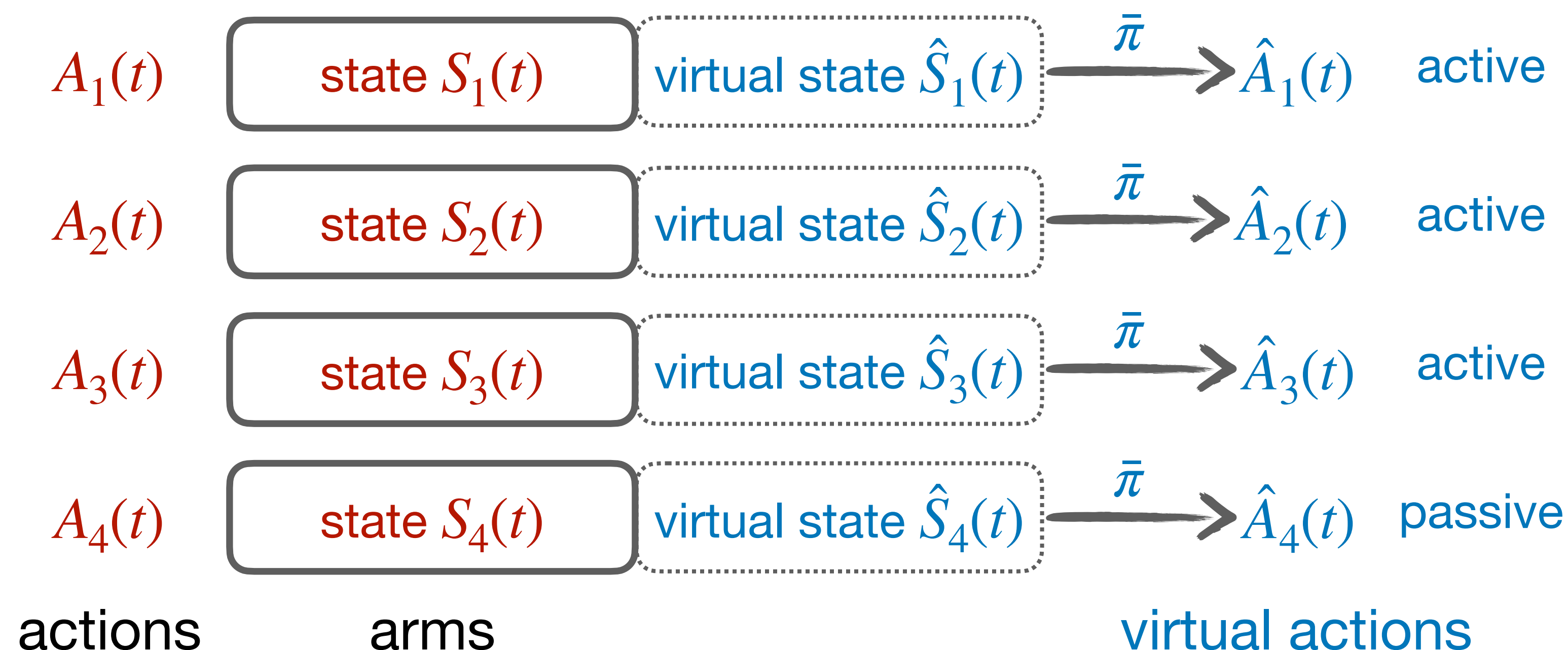
Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



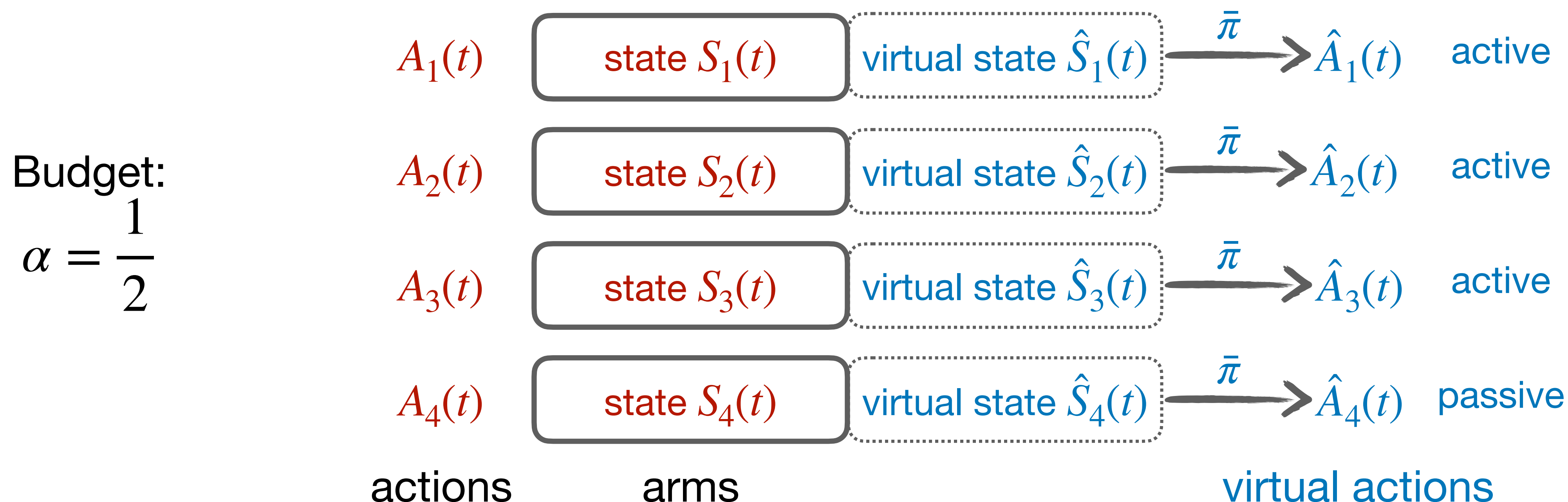
Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



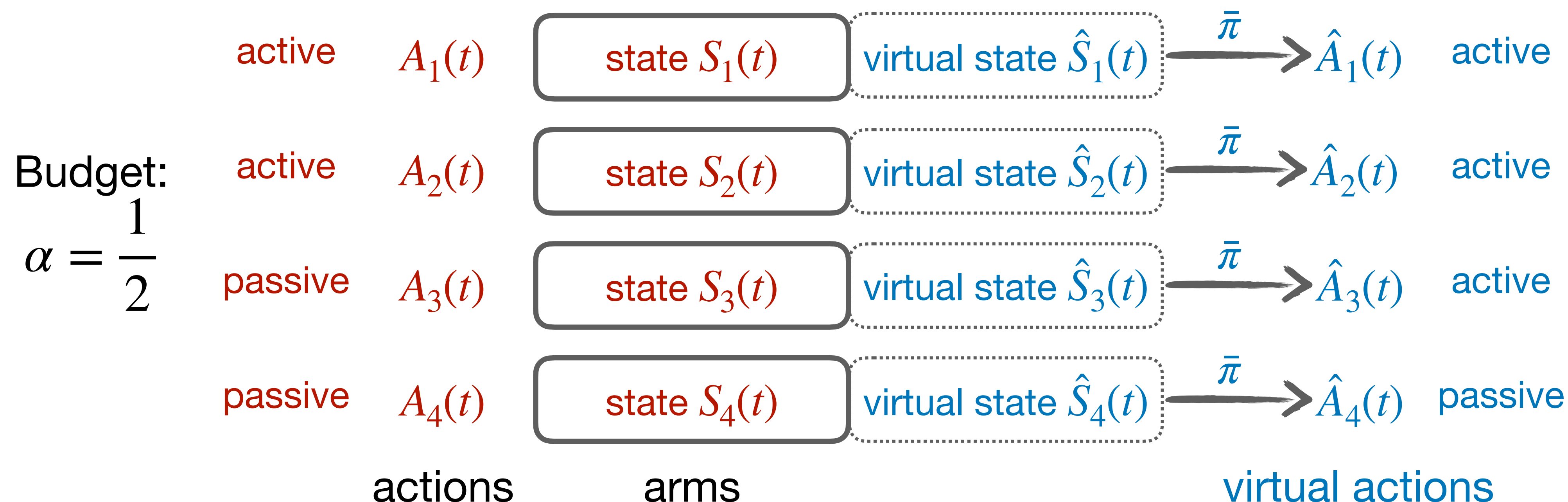
Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



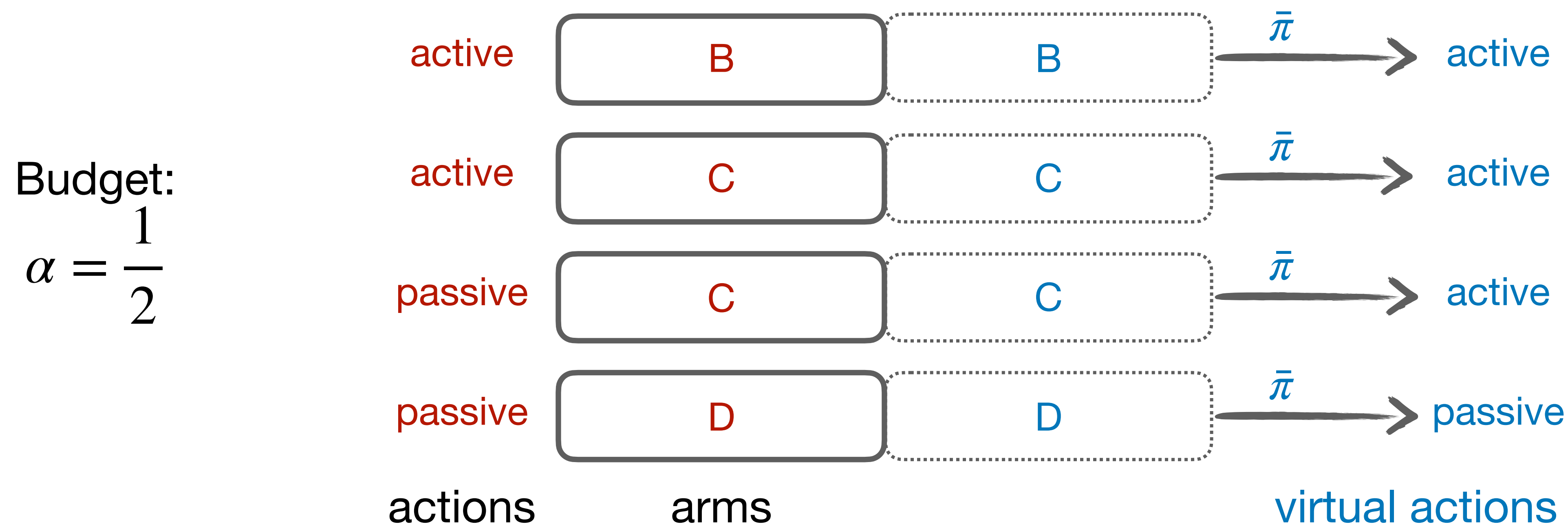
Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

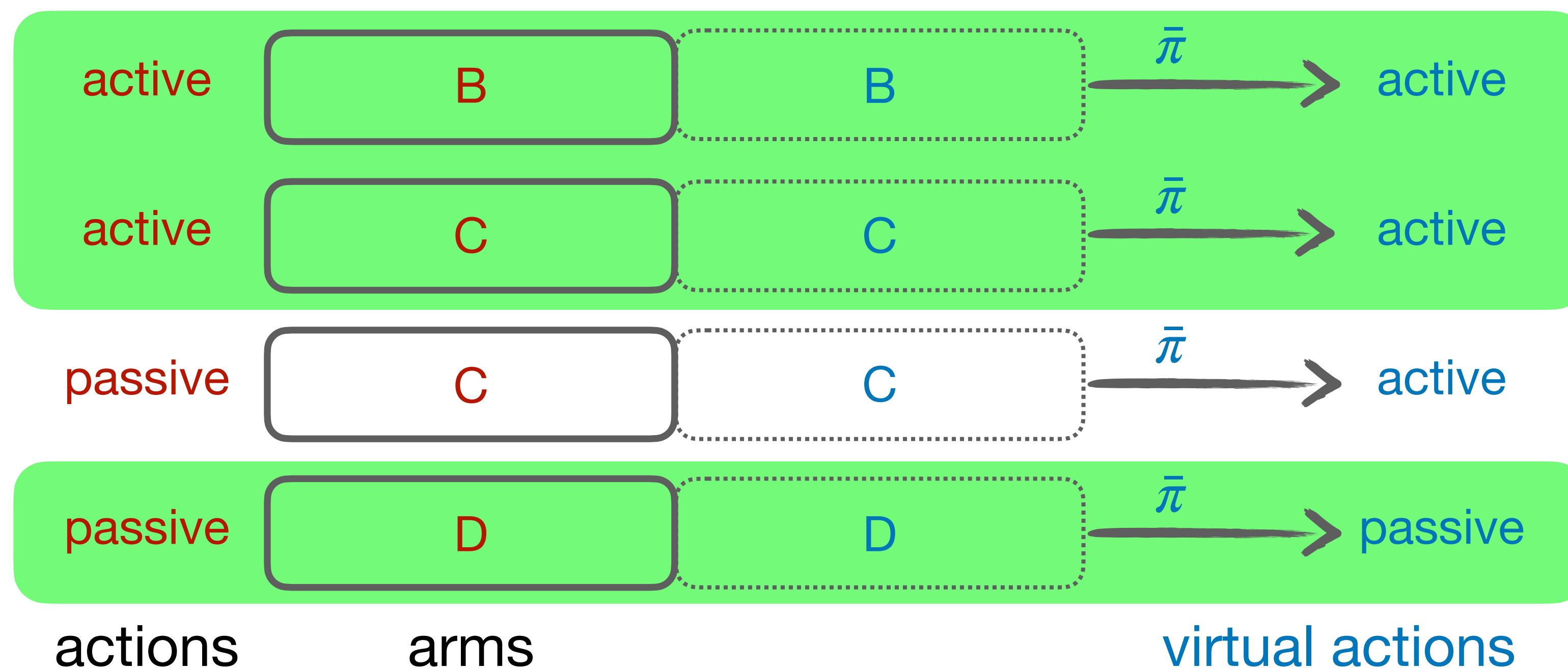
- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

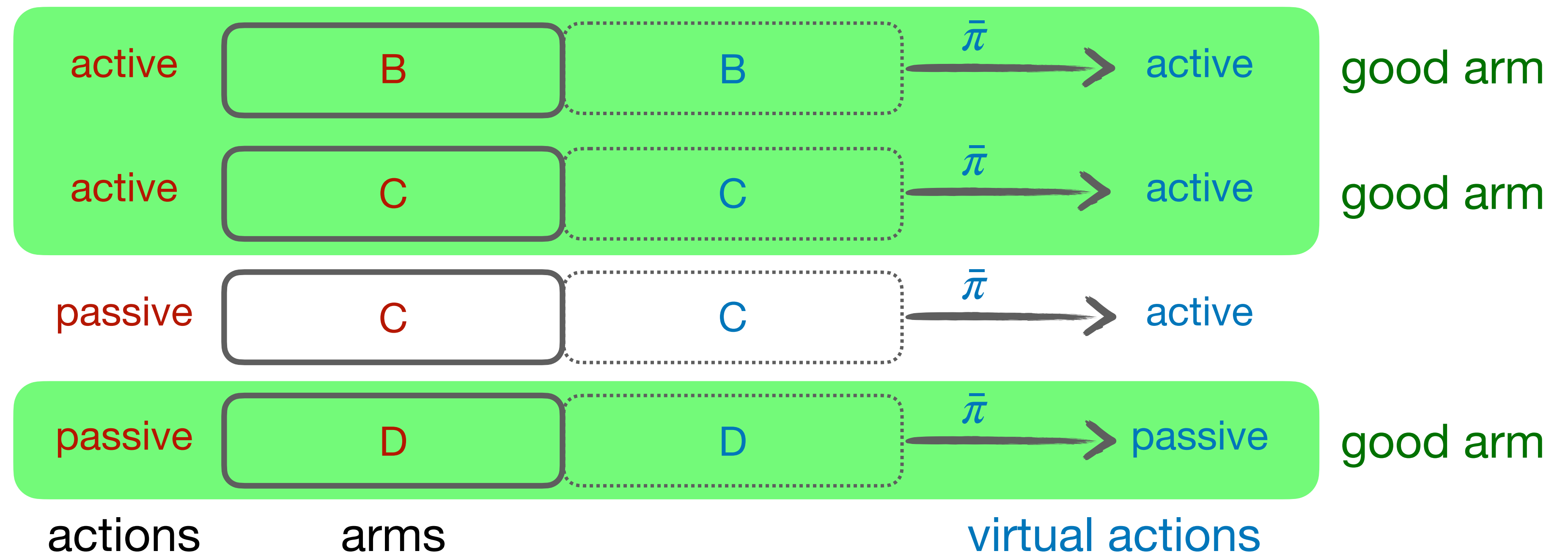
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

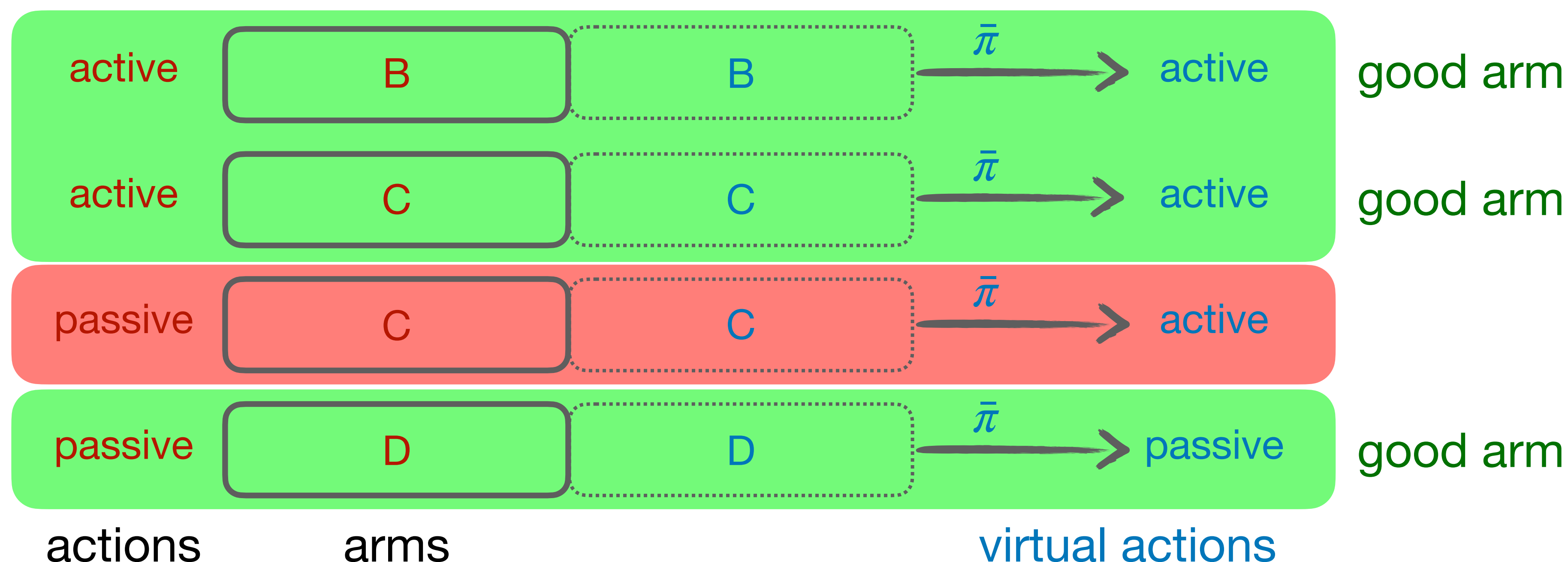
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

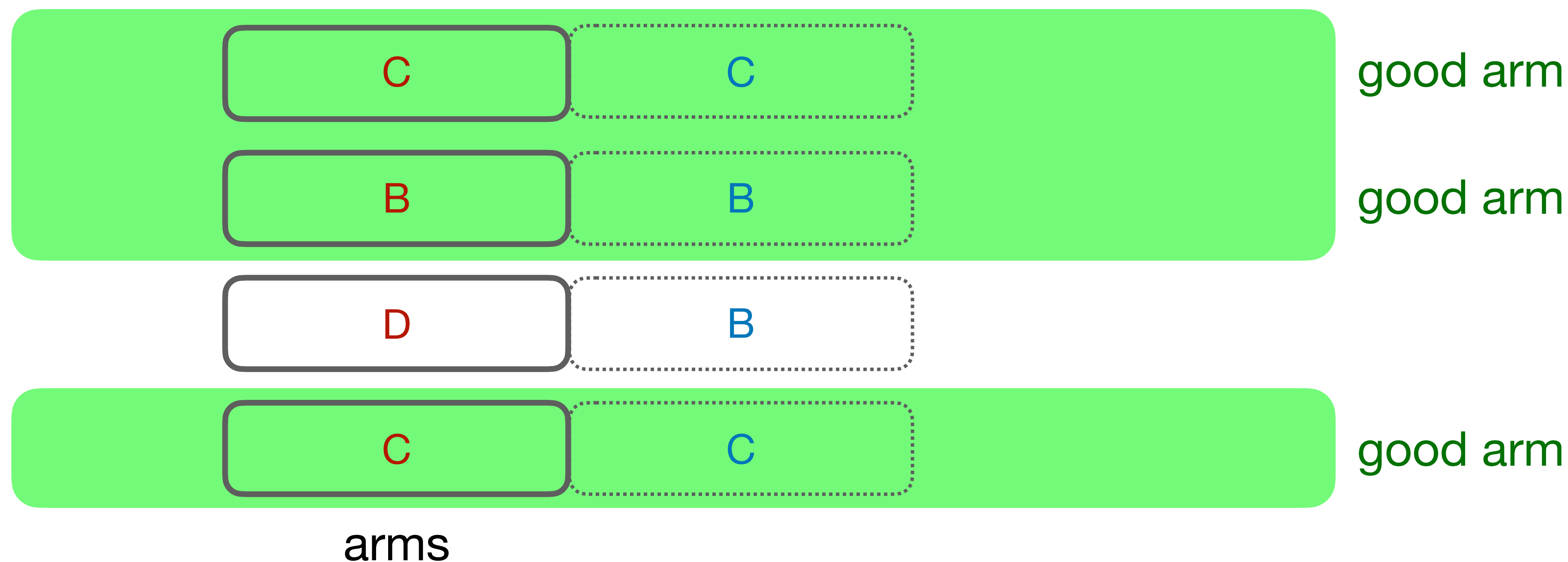
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

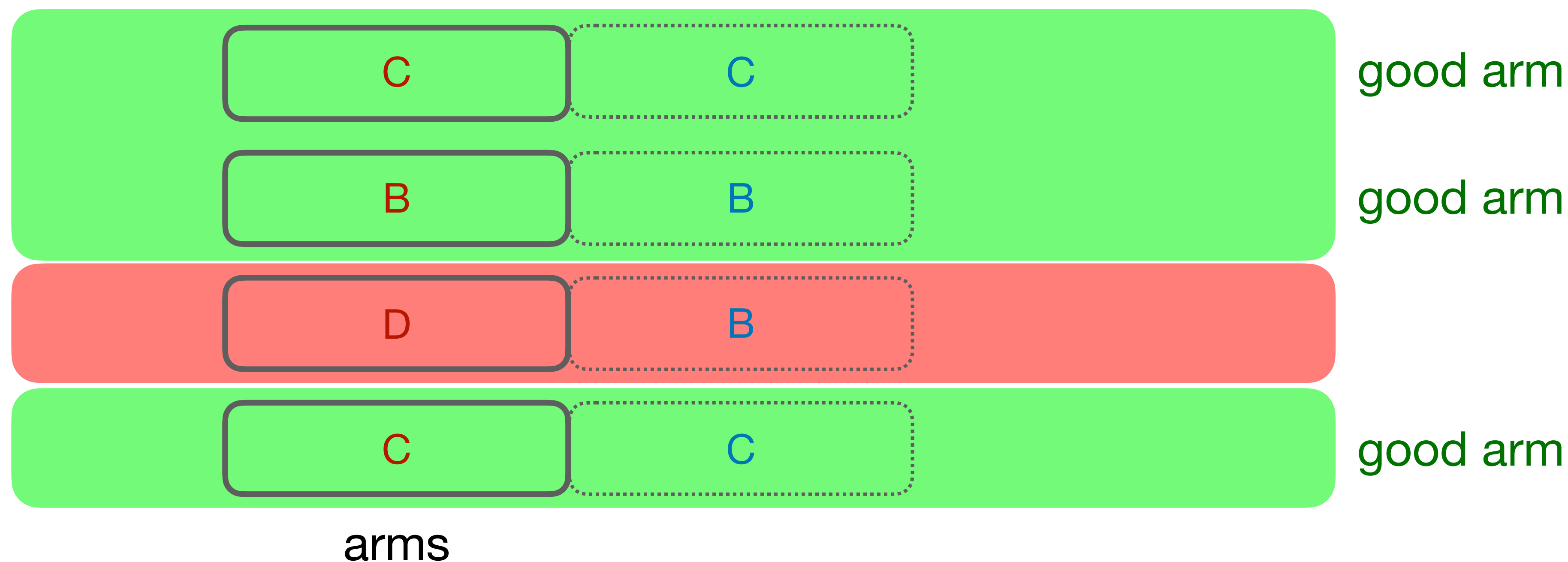
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

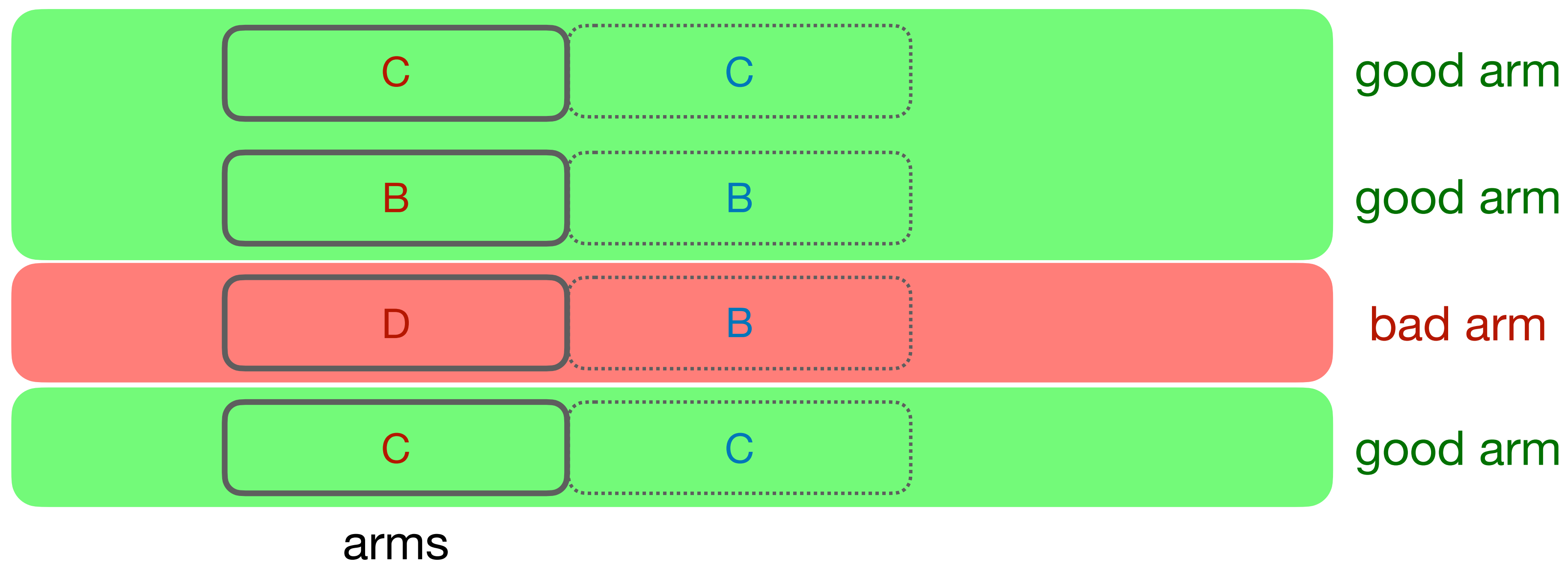
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

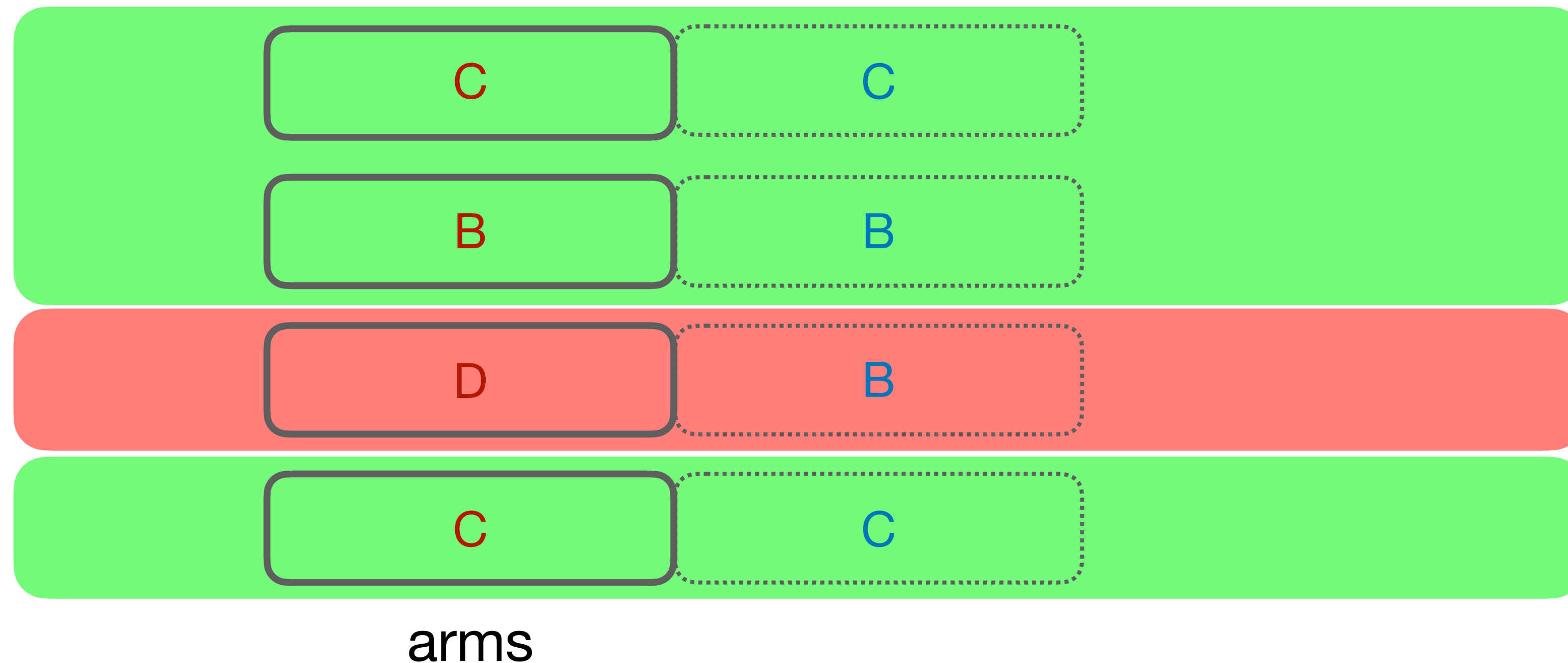
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

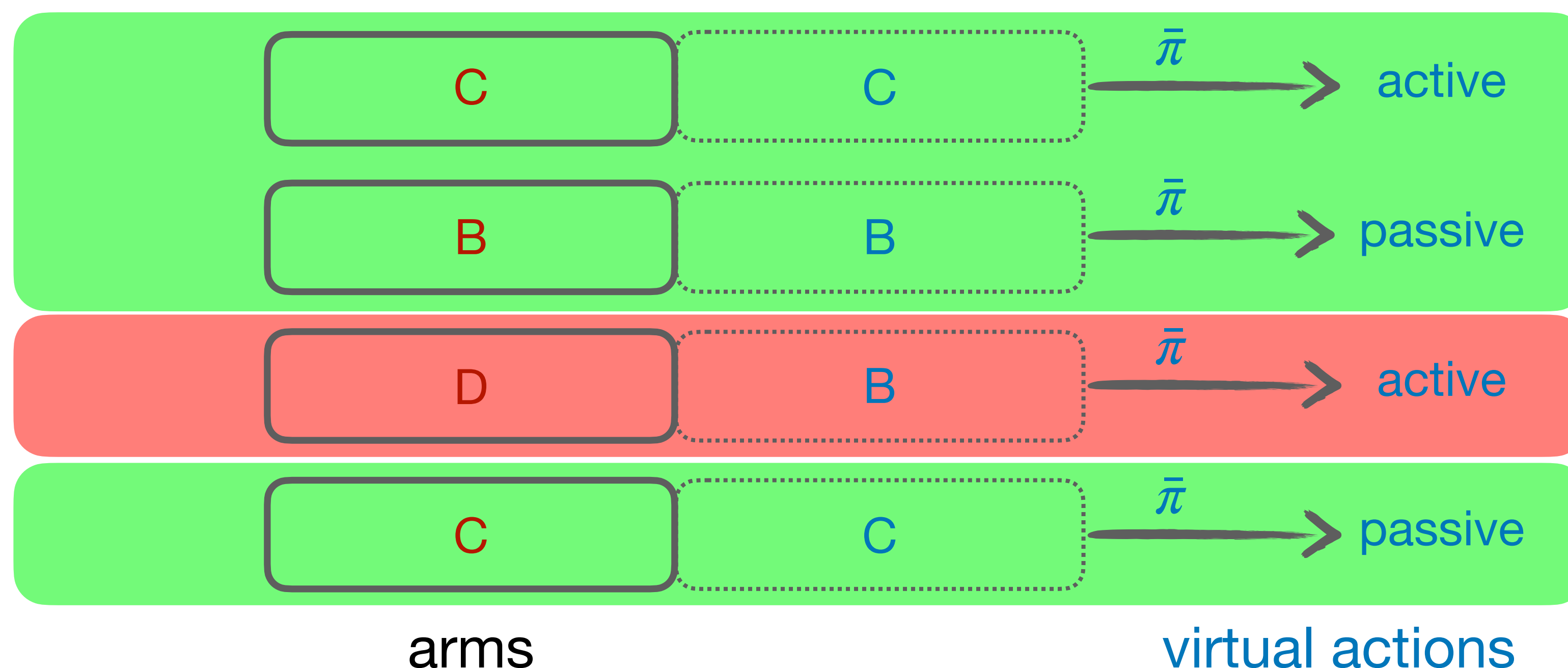
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

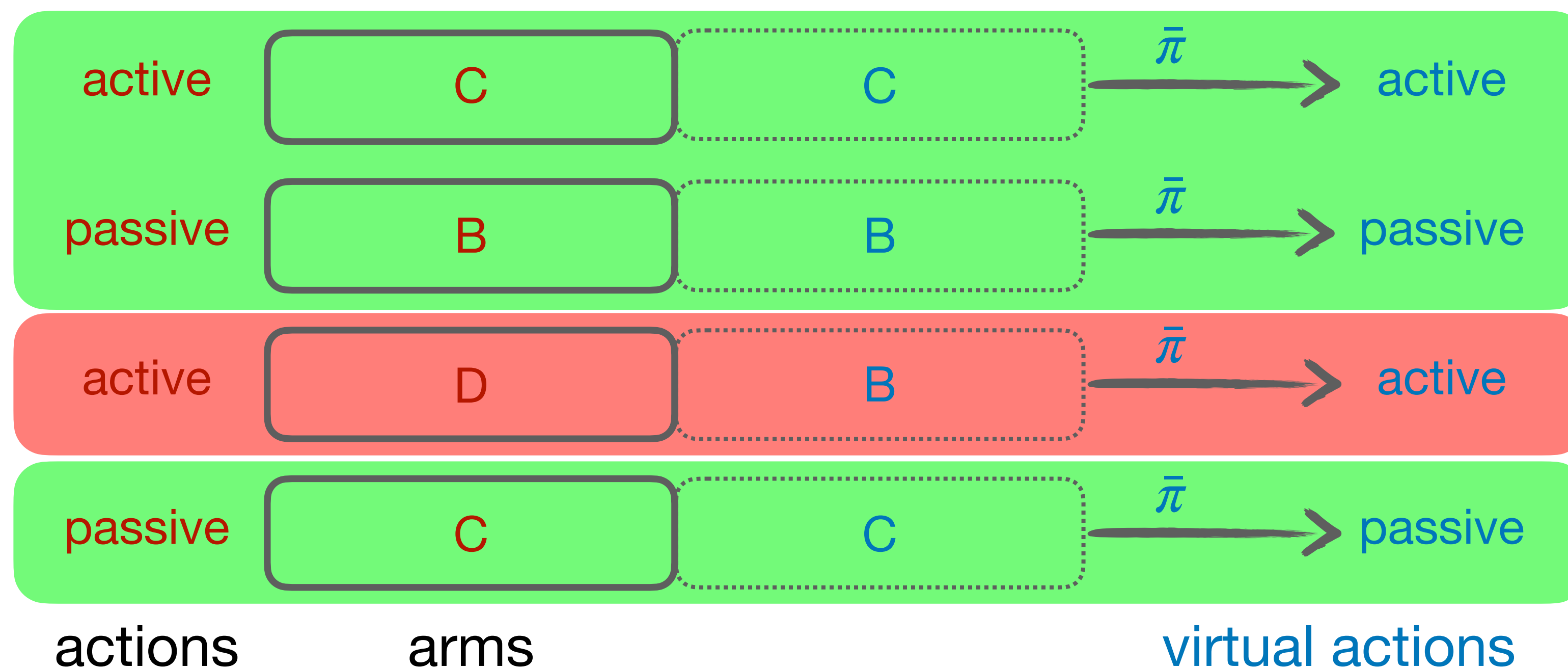
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

- Input: a single-armed policy $\bar{\pi}$
- Each arm simulates a virtual single-armed system following $\bar{\pi}$
- Take actions following virtual actions as much as possible

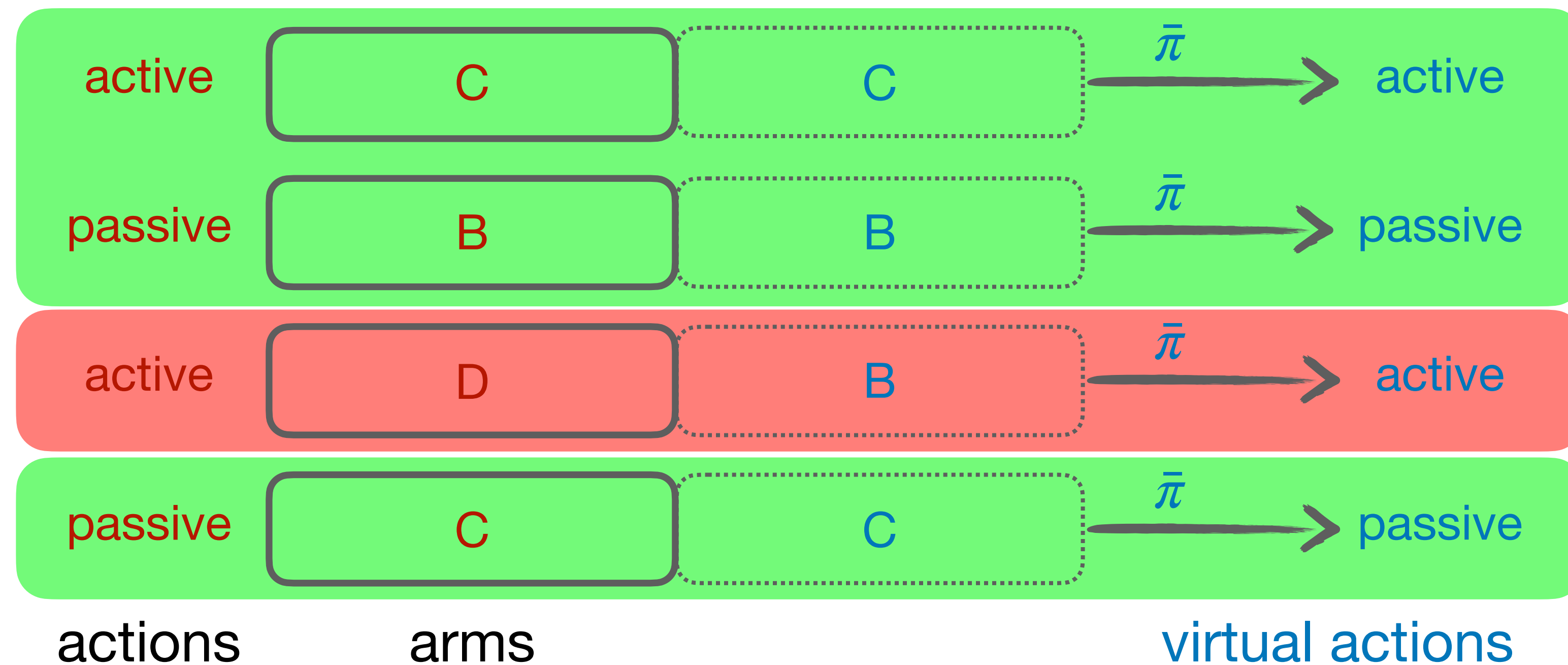
Budget:
 $\alpha = \frac{1}{2}$



Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

Follow the virtual action even when an arm is a bad arm

Budget:
 $\alpha = \frac{1}{2}$

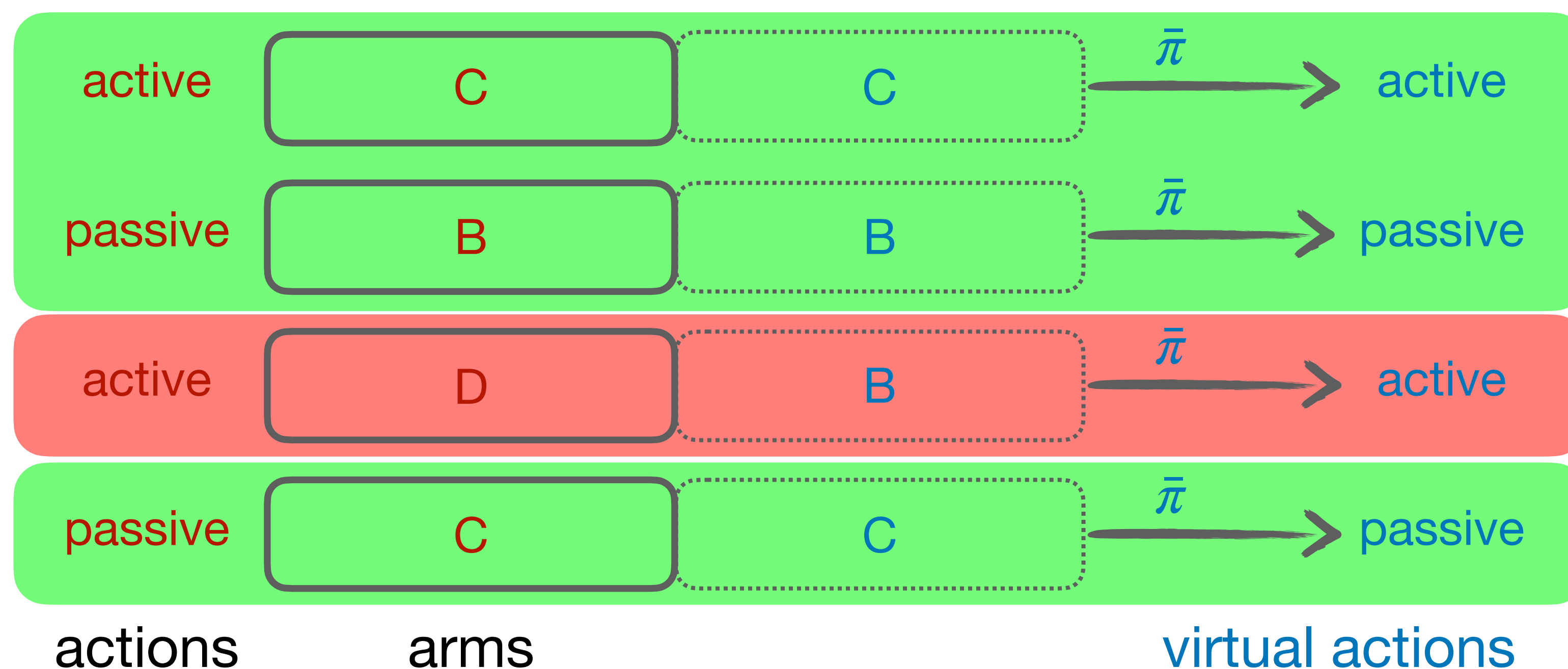


Our policy: FOLLOW-THE-VIRTUAL-ADVICE (FTVA)

The SA assumption assumes that the states will couple again within a finite time

Follow the virtual action even when an arm is a bad arm

Budget:
 $\alpha = \frac{1}{2}$



Proof of optimality gap guarantee

Proof of optimality gap guarantee

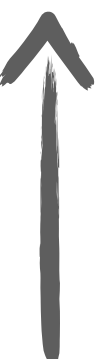
Little's law:

$$\mathbb{E}[\# \text{ bad arms}] = \text{rate of generating bad arms} \times \mathbb{E}[\text{time being a bad arm}]$$

Proof of optimality gap guarantee

Little's law:

$\mathbb{E}[\# \text{ bad arms}] = \text{rate of generating bad arms} \times \mathbb{E}[\text{time being a bad arm}]$

$$\Theta(\sqrt{N})$$


Proof of optimality gap guarantee

Little's law:

$$\mathbb{E}[\# \text{ bad arms}] = \text{rate of generating bad arms} \times \mathbb{E}[\text{time being a bad arm}]$$

$$\Theta(\sqrt{N})$$


constant



Summary

- We considered the restless bandit problem with average reward in the large N regime
- We propose a policy named **Follow-The-Virtual-Advice (FTVA)**, which achieves an $O(1/\sqrt{N})$ optimality gap without UGAP
- Discrete-time setting: our result needs an intuitive synchronization assumption
- Continuous-time setting: our result does not need any assumptions beyond the standard unichain

Restless bandits

