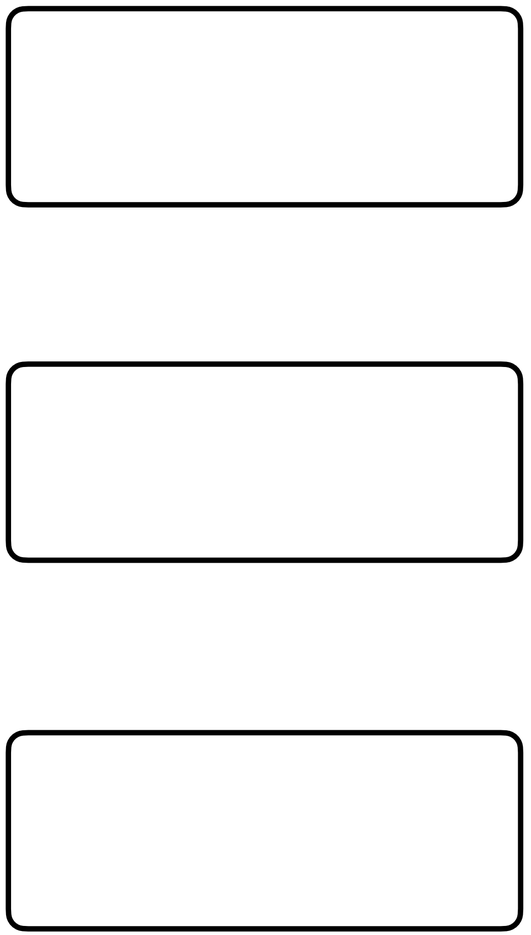
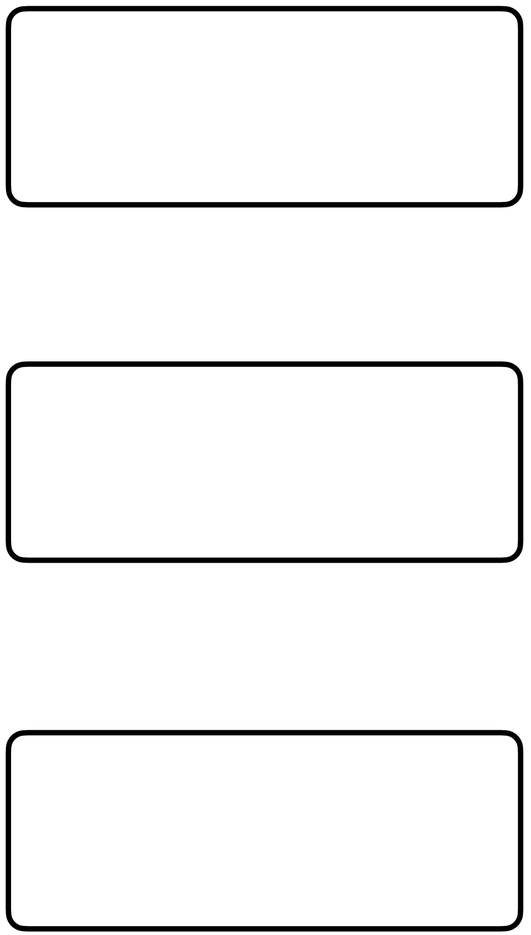
Restless Bandits with Average Reward: Breaking the Uniform Global Attractor Assumption

Presenter: Yige Hong (CMU)
Joint work with Weina Wang (CMU), Qiaomin Xie (UW-Madison), and Yudong Chen (UW-Madison)

Setting: restless bandits

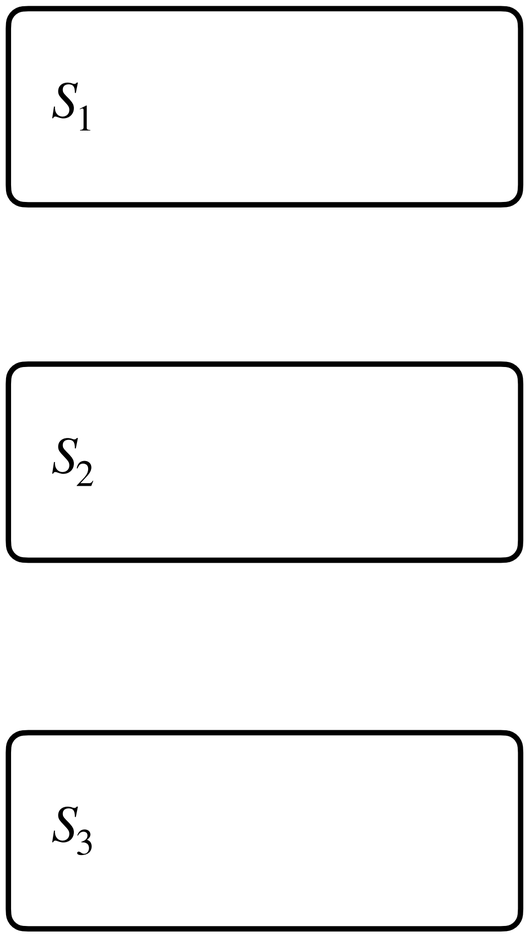


Setting: restless bandits



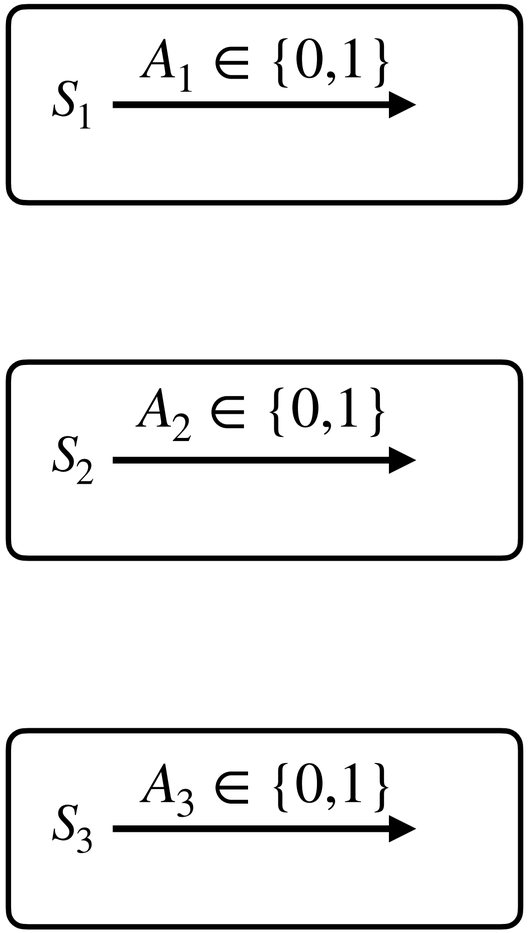
$N = 3$ arms

Setting: restless bandits



$N = 3$ arms

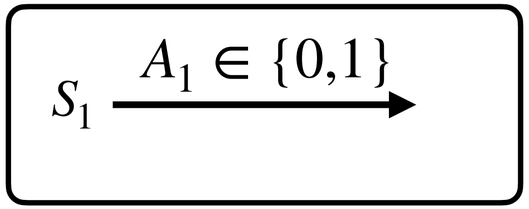
Setting: restless bandits



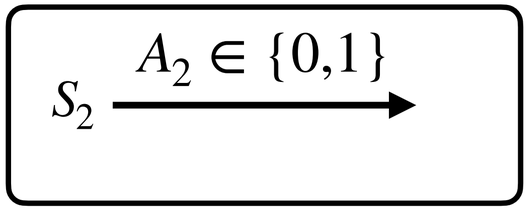
$N = 3$ arms

Setting: restless bandits

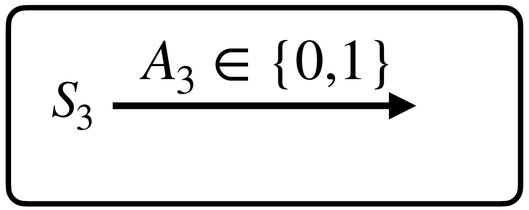
$ = $r(S_1, A_1)$



$ = $r(S_2, A_2)$



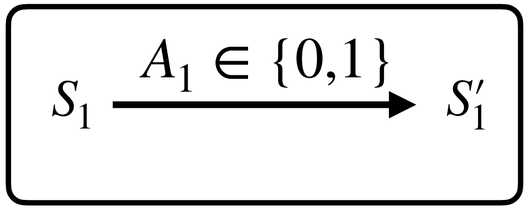
$ = $r(S_3, A_3)$



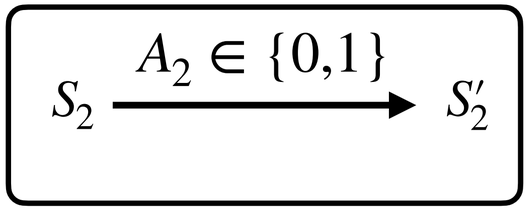
$N = 3$ arms

Setting: restless bandits

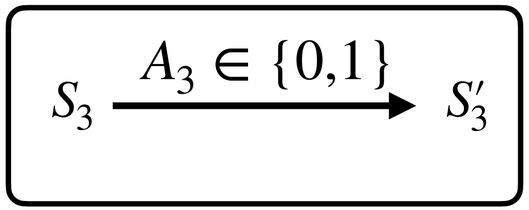
$\$ = r(S_1, A_1)$



$\$ = r(S_2, A_2)$



$\$ = r(S_3, A_3)$



$N = 3$ arms

Setting: restless bandits

$\$ = r(S_1, A_1)$

$$S_1 \xrightarrow[P(\cdot \mid S_1, A_1)]{A_1 \in \{0,1\}} S_1'$$

$\$ = r(S_2, A_2)$

$$S_2 \xrightarrow[P(\cdot \mid S_2, A_2)]{A_2 \in \{0,1\}} S_2'$$

$\$ = r(S_3, A_3)$

$$S_3 \xrightarrow[P(\cdot \mid S_3, A_3)]{A_3 \in \{0,1\}} S_3'$$

$N = 3$ arms

Setting: restless bandits

$\$ = r(S_1, A_1)$

$$S_1 \xrightarrow[P(\cdot \mid S_1, A_1)]{A_1 \in \{0,1\}} S_1'$$

$\$ = r(S_2, A_2)$

$$S_2 \xrightarrow[P(\cdot \mid S_2, A_2)]{A_2 \in \{0,1\}} S_2'$$

$\$ = r(S_3, A_3)$

$$S_3 \xrightarrow[P(\cdot \mid S_3, A_3)]{A_3 \in \{0,1\}} S_3'$$

$N = 3$ arms

$\max_\pi V_N^\pi \triangleq$ long run average reward under policy π

Setting: restless bandits

$\$ = r(S_1, A_1)$

$$S_1 \xrightarrow[P(\cdot \mid S_1, A_1)]{A_1 \in \{0,1\}} S_1'$$

$\$ = r(S_2, A_2)$

$$S_2 \xrightarrow[P(\cdot \mid S_2, A_2)]{A_2 \in \{0,1\}} S_2'$$

$\$ = r(S_3, A_3)$

$$S_3 \xrightarrow[P(\cdot \mid S_3, A_3)]{A_3 \in \{0,1\}} S_3'$$

$N = 3$ arms

$\max_\pi V_N^\pi \triangleq$ long run average reward under policy π

$\pi: (S_1, S_2, \ldots S_N) \mapsto (A_1, A_2, \ldots A_N)$

Setting: restless bandits

$\$ = r(S_1, A_1)$

$$S_1 \xrightarrow[P(\cdot \mid S_1, A_1)]{A_1 \in \{0,1\}} S_1'$$

$\$ = r(S_2, A_2)$

$$S_2 \xrightarrow[P(\cdot \mid S_2, A_2)]{A_2 \in \{0,1\}} S_2'$$

$\$ = r(S_3, A_3)$

$$S_3 \xrightarrow[P(\cdot \mid S_3, A_3)]{A_3 \in \{0,1\}} S_3'$$

$N = 3$ arms

$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

$\pi: (S_1, S_2, \ldots S_N) \mapsto (A_1, A_2, \ldots A_N)$

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

Setting: restless bandits

$\$ = r(S_1, A_1)$

$$S_1 \xrightarrow[P(\cdot \mid S_1, A_1)]{A_1 \in \{0,1\}} S_1'$$

$\$ = r(S_2, A_2)$

$$S_2 \xrightarrow[P(\cdot \mid S_2, A_2)]{A_2 \in \{0,1\}} S_2'$$

$\$ = r(S_3, A_3)$

$$S_3 \xrightarrow[P(\cdot \mid S_3, A_3)]{A_3 \in \{0,1\}} S_3'$$

$N = 3$ arms

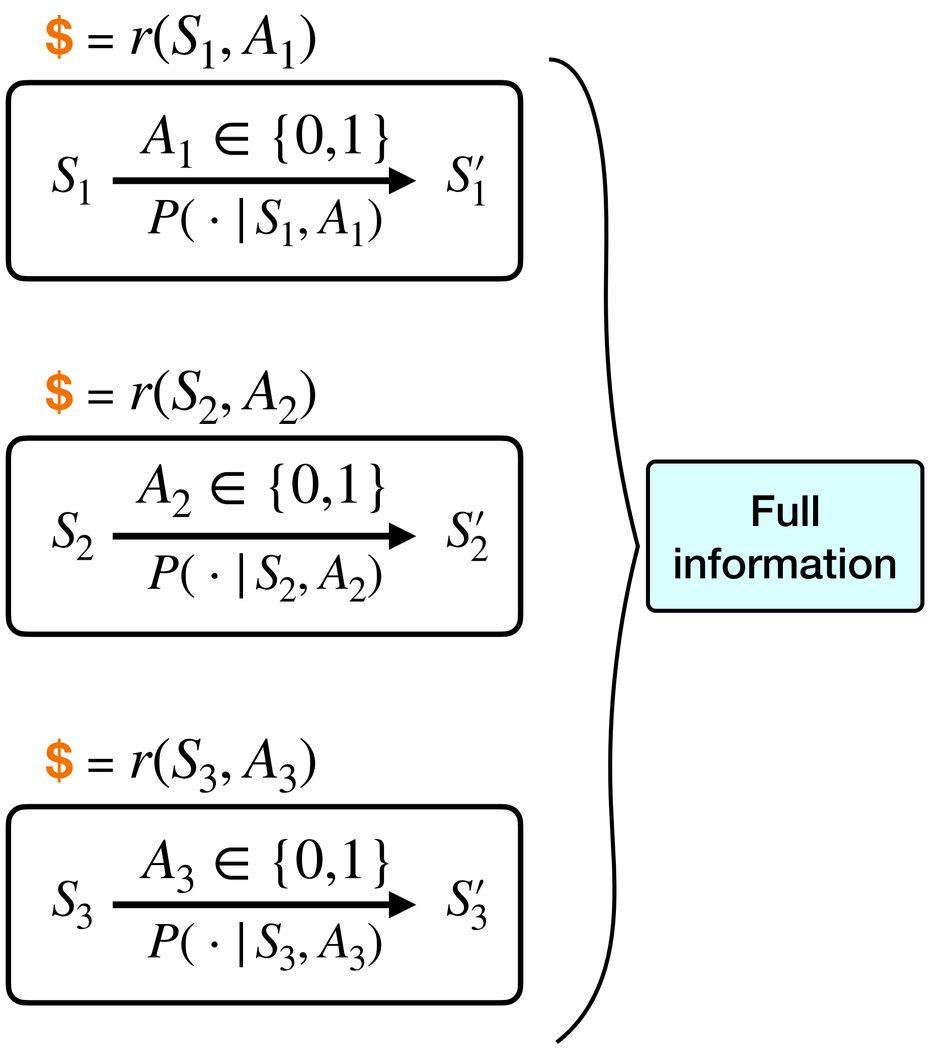
$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

$\pi: (S_1, S_2, \ldots S_N) \mapsto (A_1, A_2, \ldots A_N)$

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

$0 < \alpha < 1$

Setting: restless bandits



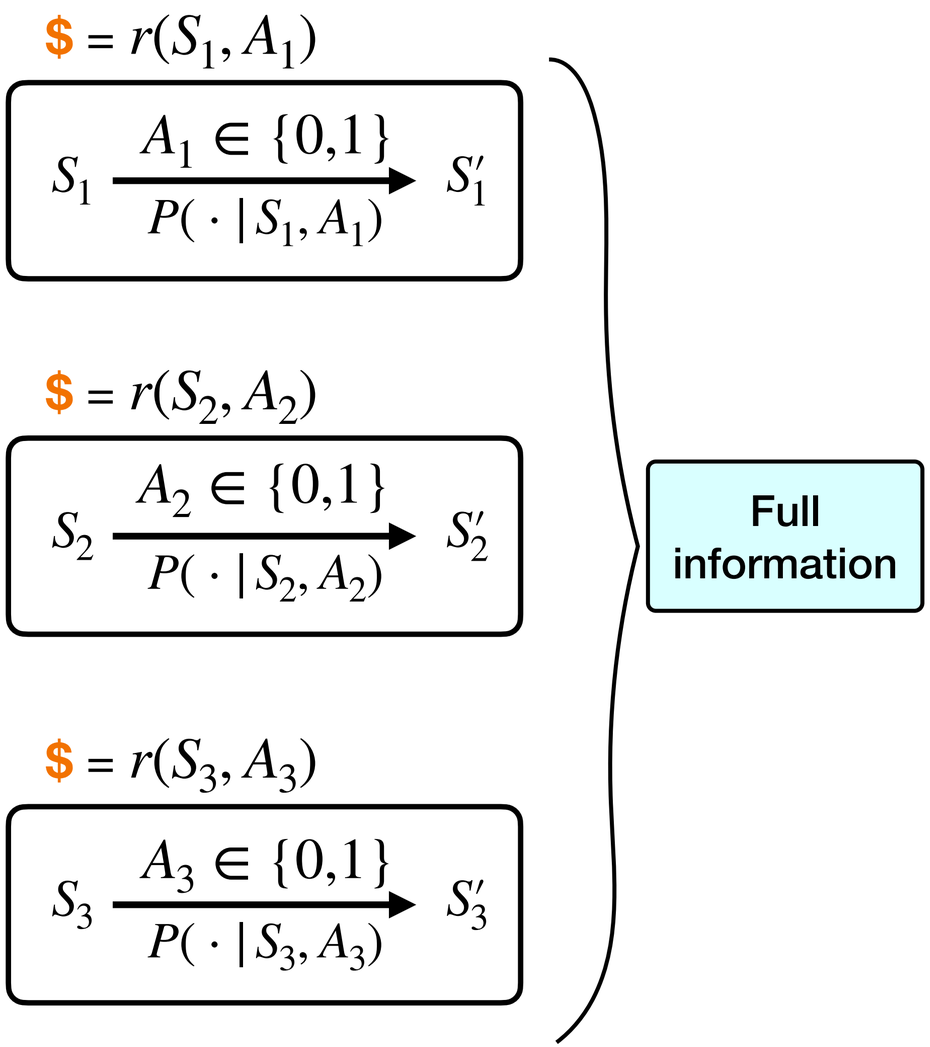
$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

$\pi: (S_1, S_2, \ldots S_N) \mapsto (A_1, A_2, \ldots A_N)$

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

$0 < \alpha < 1$

Setting: restless bandits



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

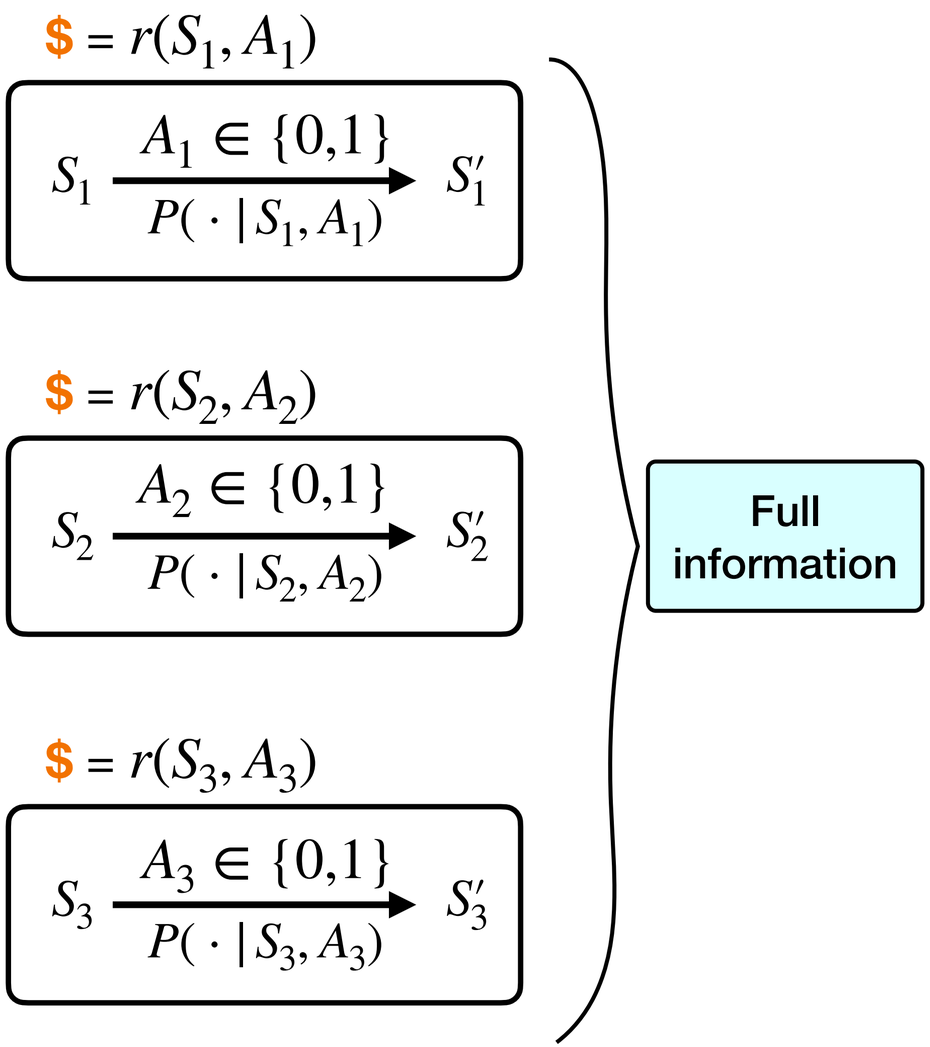
$\pi: (S_1, S_2, \ldots S_N) \mapsto (A_1, A_2, \ldots A_N)$

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

$0 < \alpha < 1$

If N large, high dimensional;
PSPACE hard to solve

Goal: asymptotic optimality for large N



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

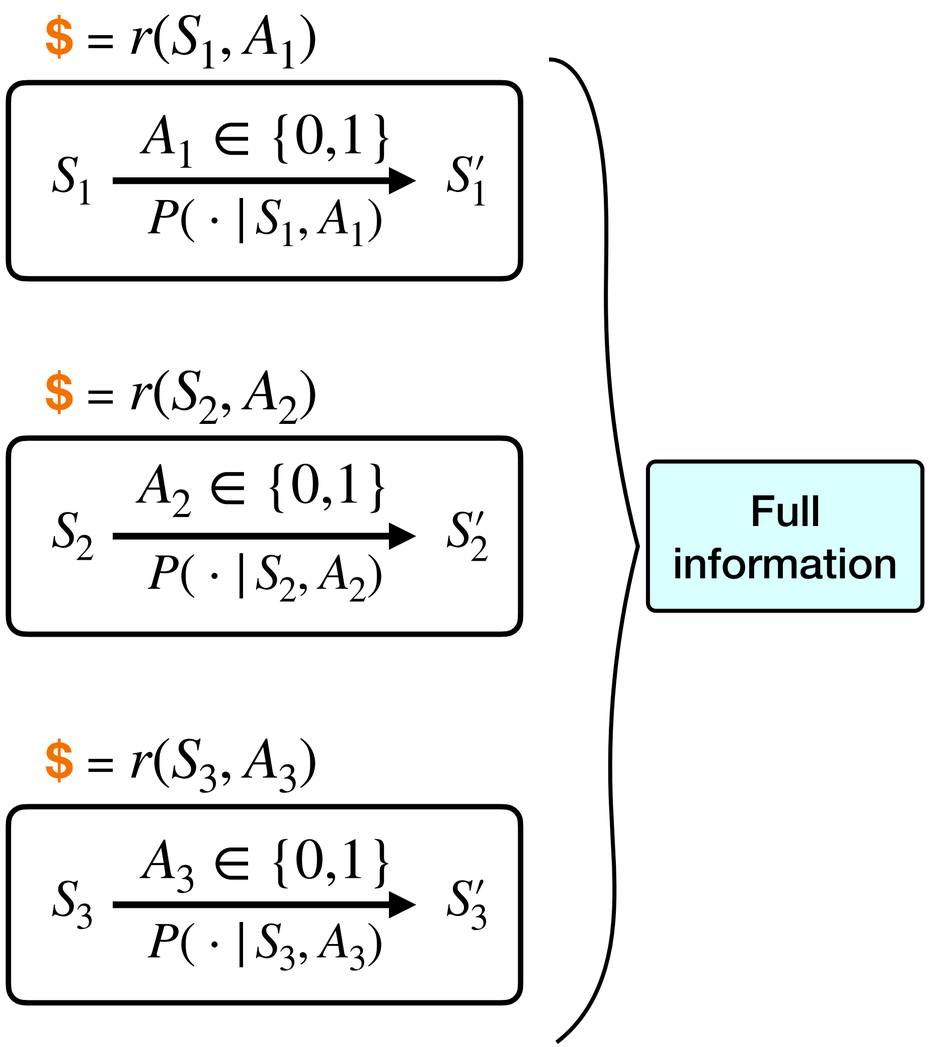
$\pi: (S_1, S_2, \ldots S_N) \mapsto (A_1, A_2, \ldots A_N)$

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

$0 < \alpha < 1$

If N large, high dimensional;
PSPACE hard to solve

Goal: asymptotic optimality for large N



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

$\pi: (S_1, S_2, \ldots S_N) \mapsto (A_1, A_2, \ldots A_N)$

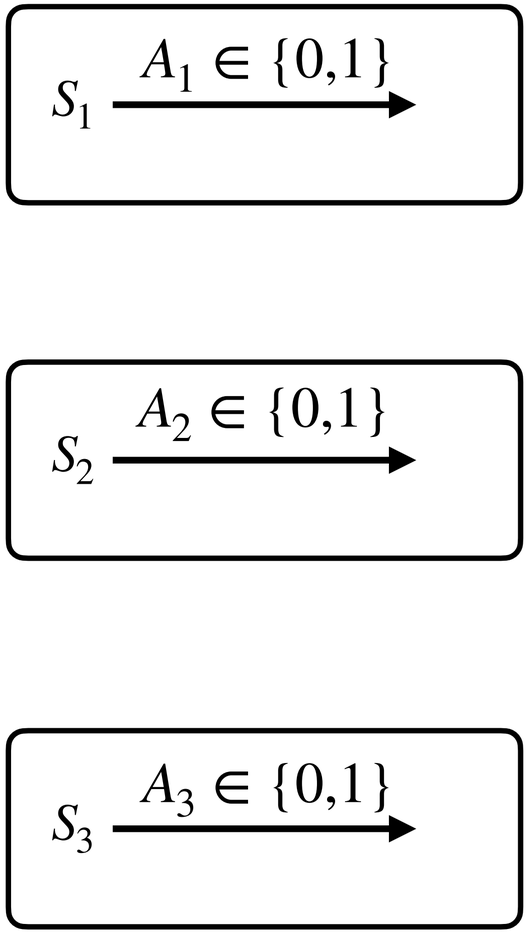
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

$0 < \alpha < 1$

If N large, high dimensional;
PSPACE hard to solve

Goal: computationally efficient algorithm for finding π such that $V_N^* - V_N^{\pi} \to 0$ as $N \to \infty$

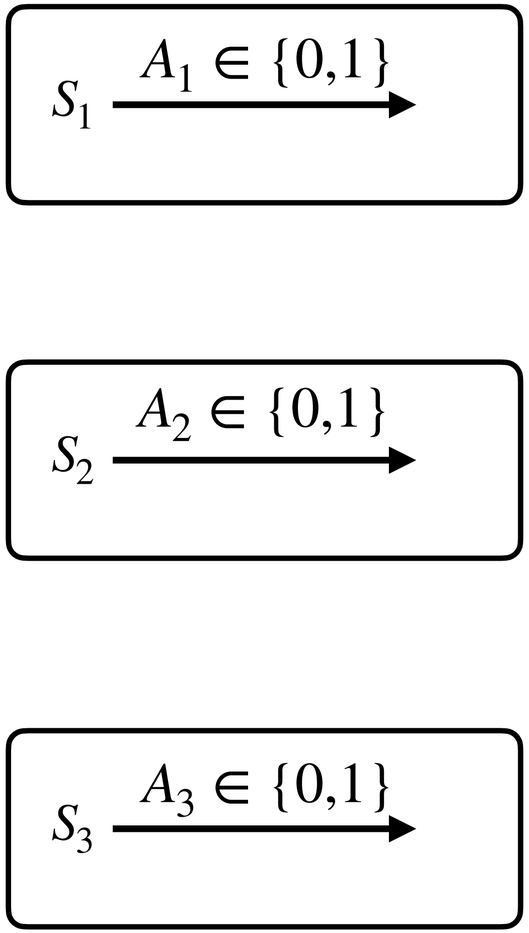
Prior work: priority policies based on relaxation



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

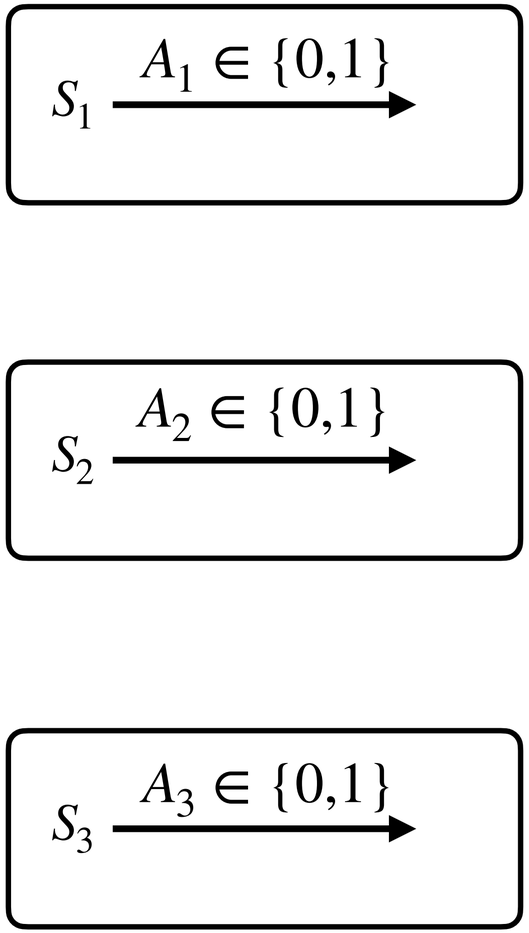
Prior work: priority policies based on relaxation



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

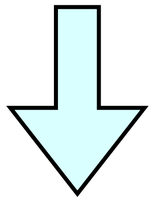
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

Prior work: priority policies based on relaxation



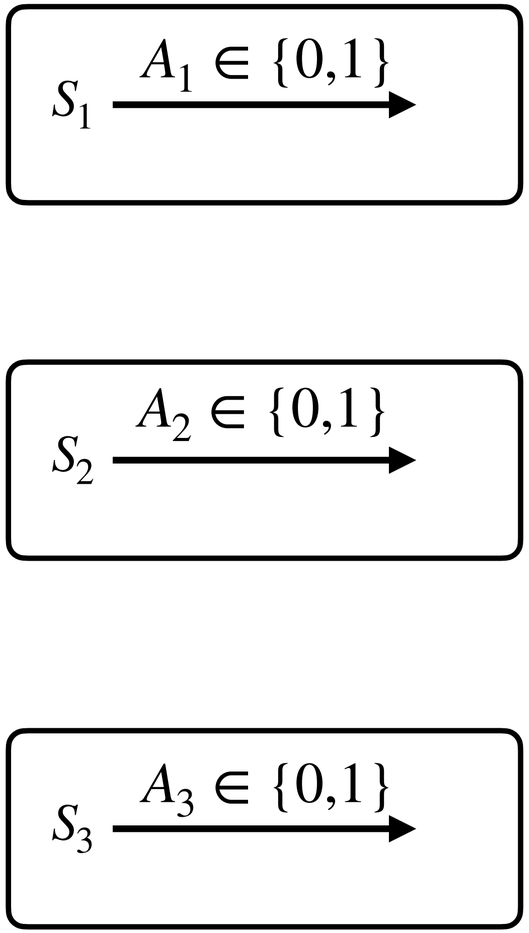
$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot



relax

Prior work: priority policies based on relaxation



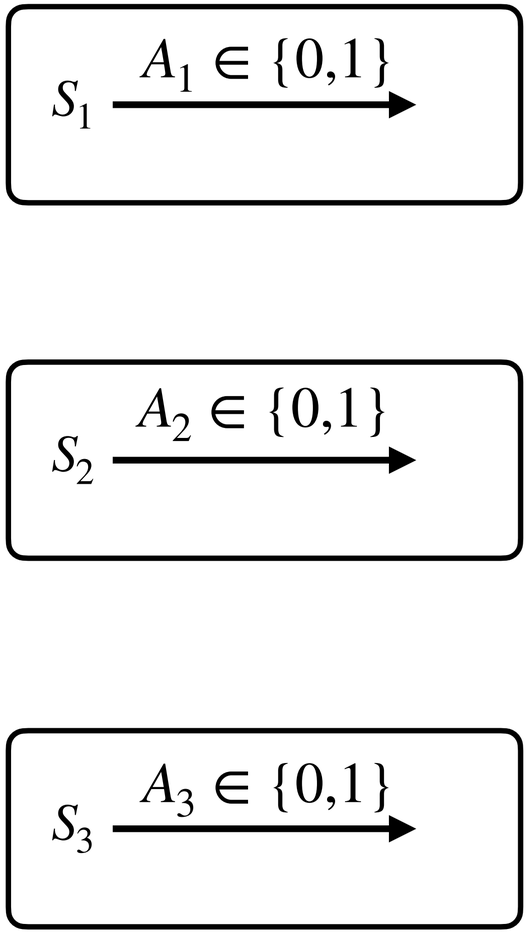
$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

Prior work: priority policies based on relaxation



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

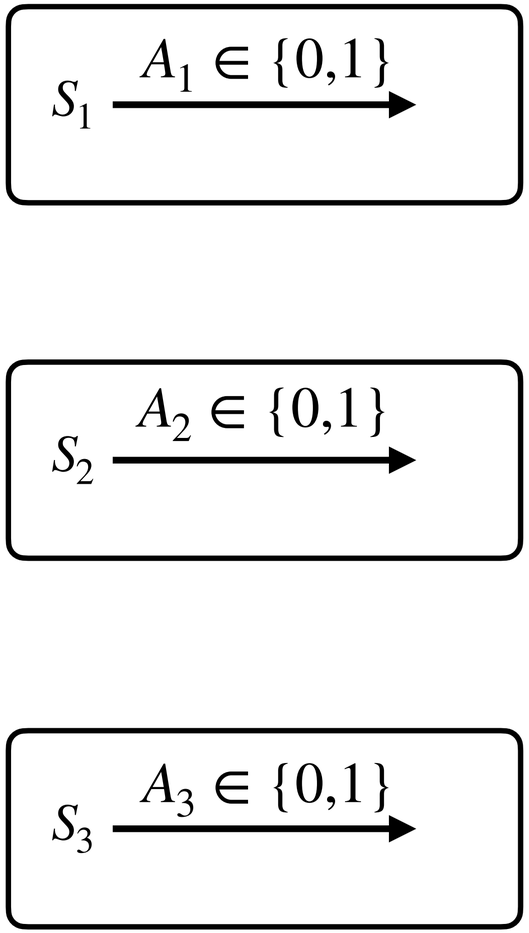
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

independent of N

Prior work: priority policies based on relaxation



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

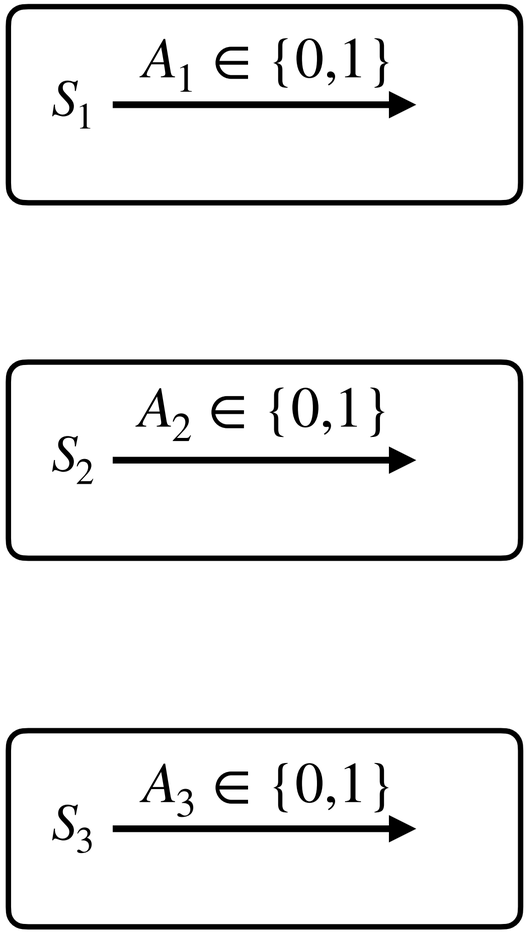
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

independent of N

Prior work: priority policies based on relaxation



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

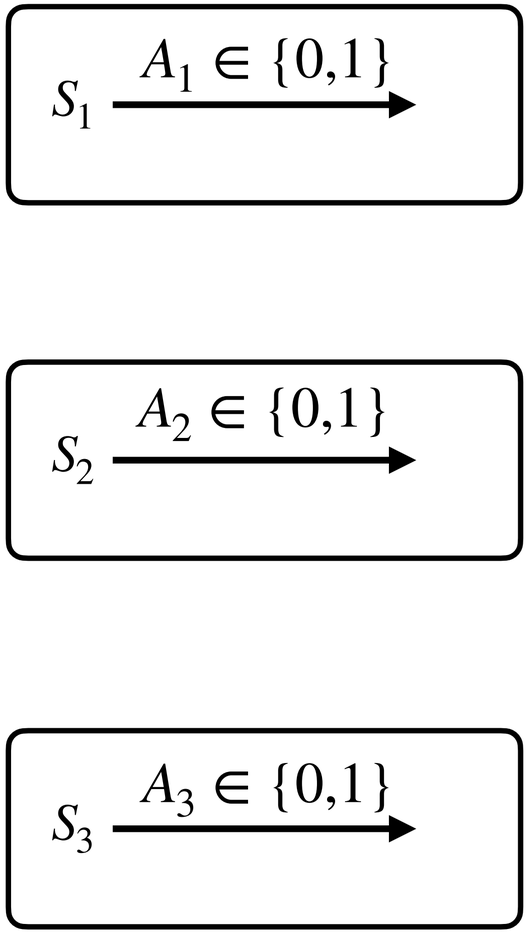
relax

single-armed problem

independent of N

priorities for activating arms

Prior work: priority policies based on relaxation



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

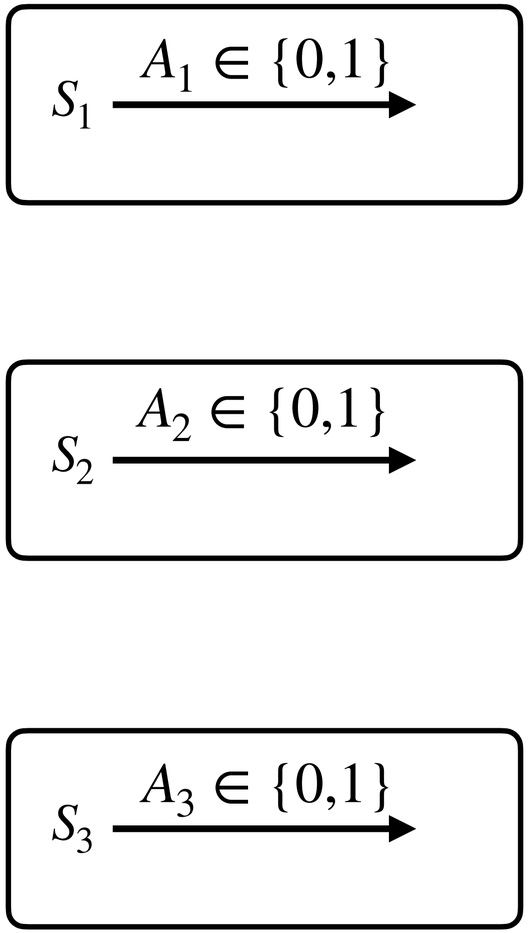
independent of N

priorities for activating arms

Whittle index [Whi88, GGY20]

LP Priority [Ver16, GGY22]

Prior work: priority policies based on relaxation



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

independent of N

priorities for activating arms

Whittle index [Whi88, GGY20]
LP Priority [Ver16, GGY22]

Asymptotically optimal?

Key limitation: require UGAP assumption

Key limitation: require UGAP assumption

All prior policies [WW90][Ver16][GGY20][GGY22]
need *Uniform Global Attractor Property (UGAP)*
to be asymptotic optimal

Key limitation: require UGAP assumption

All prior policies [WW90][Ver16][GGY20][GGY22]
need *Uniform Global Attractor Property (UGAP)*
to be asymptotic optimal

UGAP is tricky and hard to verify

Key limitation: require UGAP assumption

All prior policies [WW90][Ver16][GGY20][GGY22]
need *Uniform Global Attractor Property (UGAP)*
to be asymptotic optimal

UGAP is tricky and hard to verify

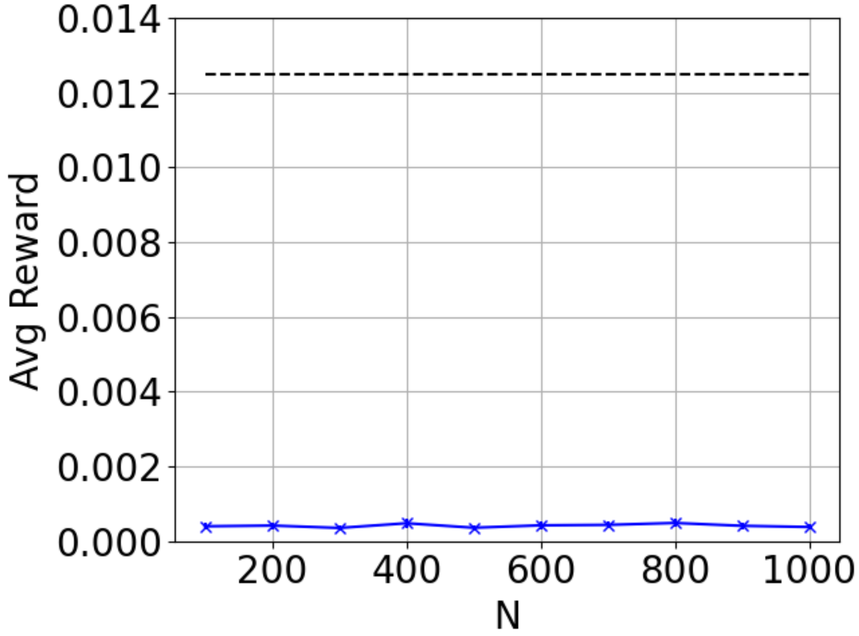
UGAP: no bad “local optimum”

Key limitation: require UGAP assumption

All prior policies [WW90][Ver16][GGY20][GGY22]
need *Uniform Global Attractor Property (UGAP)*
to be asymptotic optimal

UGAP is tricky and hard to verify

UGAP: no bad “local optimum”

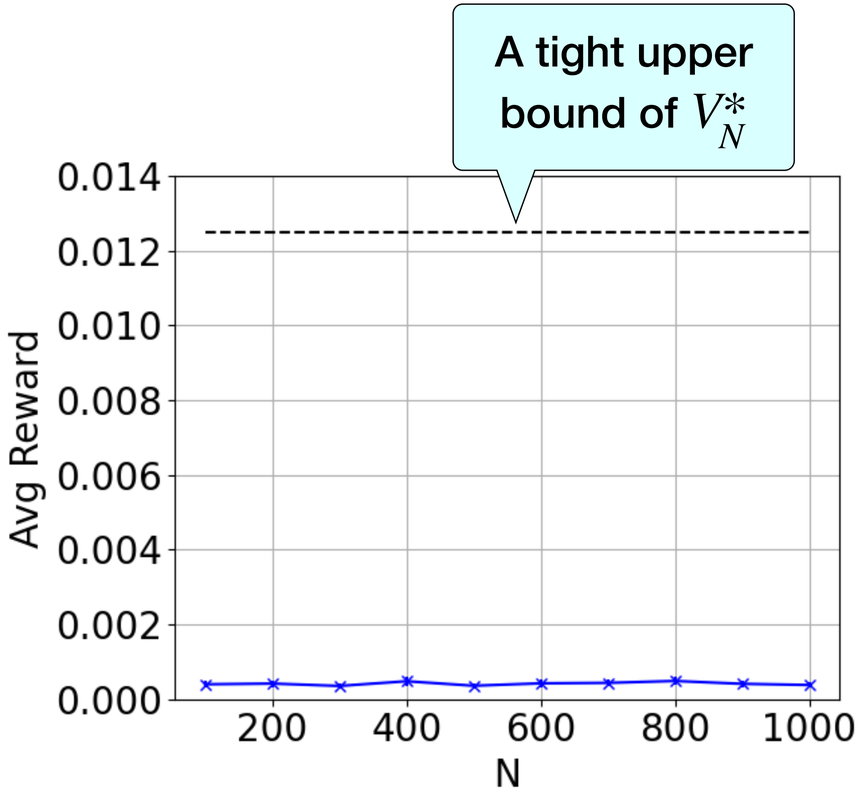


Key limitation: require UGAP assumption

All prior policies [WW90][Ver16][GGY20][GGY22]
need *Uniform Global Attractor Property (UGAP)*
to be asymptotic optimal

UGAP is tricky and hard to verify

UGAP: no bad “local optimum”

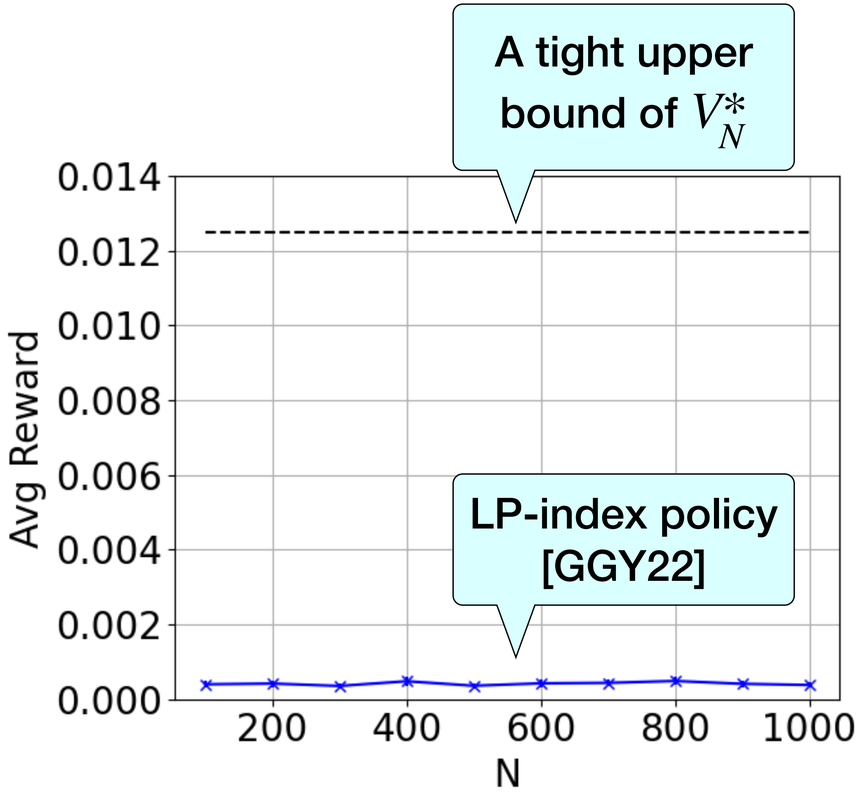


Key limitation: require UGAP assumption

All prior policies [WW90][Ver16][GGY20][GGY22]
need *Uniform Global Attractor Property (UGAP)*
to be asymptotic optimal

UGAP is tricky and hard to verify

UGAP: no bad “local optimum”



Can we remove UGAP?

Can we remove UGAP?

YES!

Can we remove UGAP?

YES!

We propose the first policy that is asymptotically optimal without UGAP.

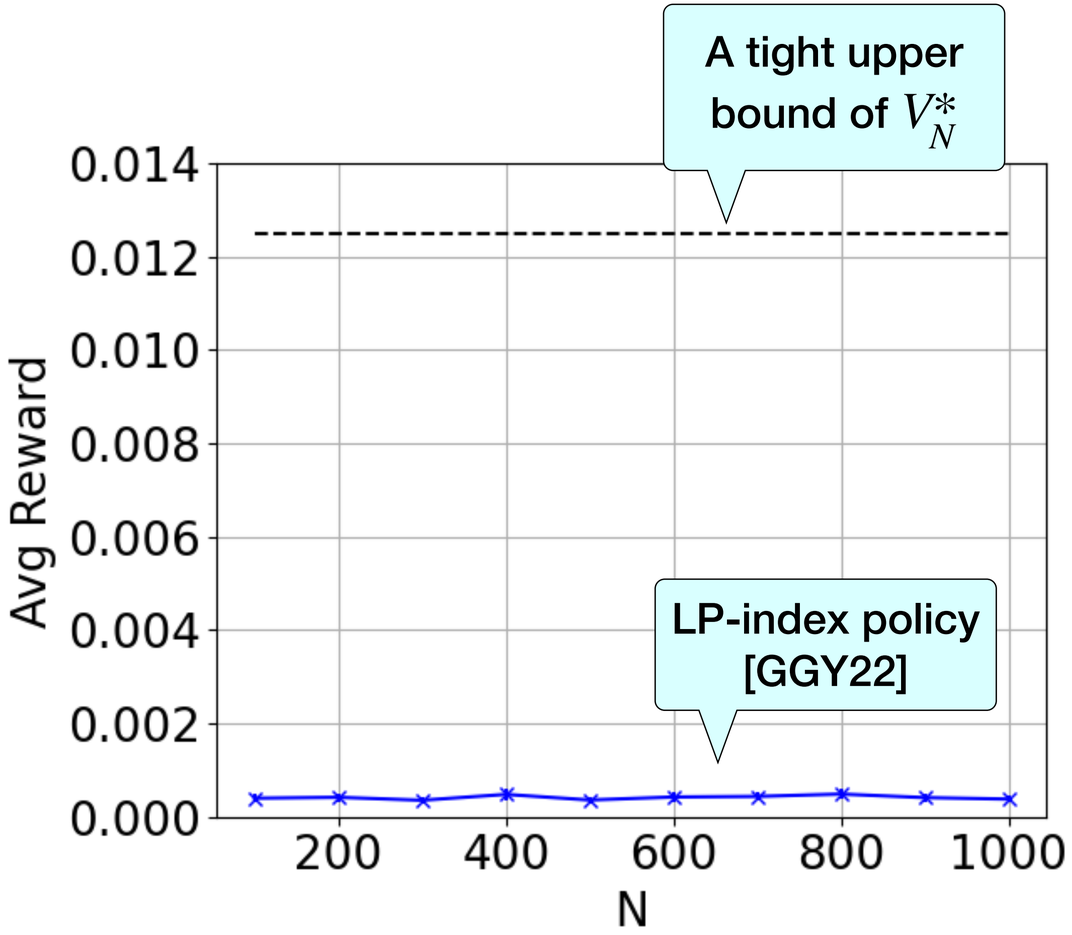
Can we remove UGAP?

YES!

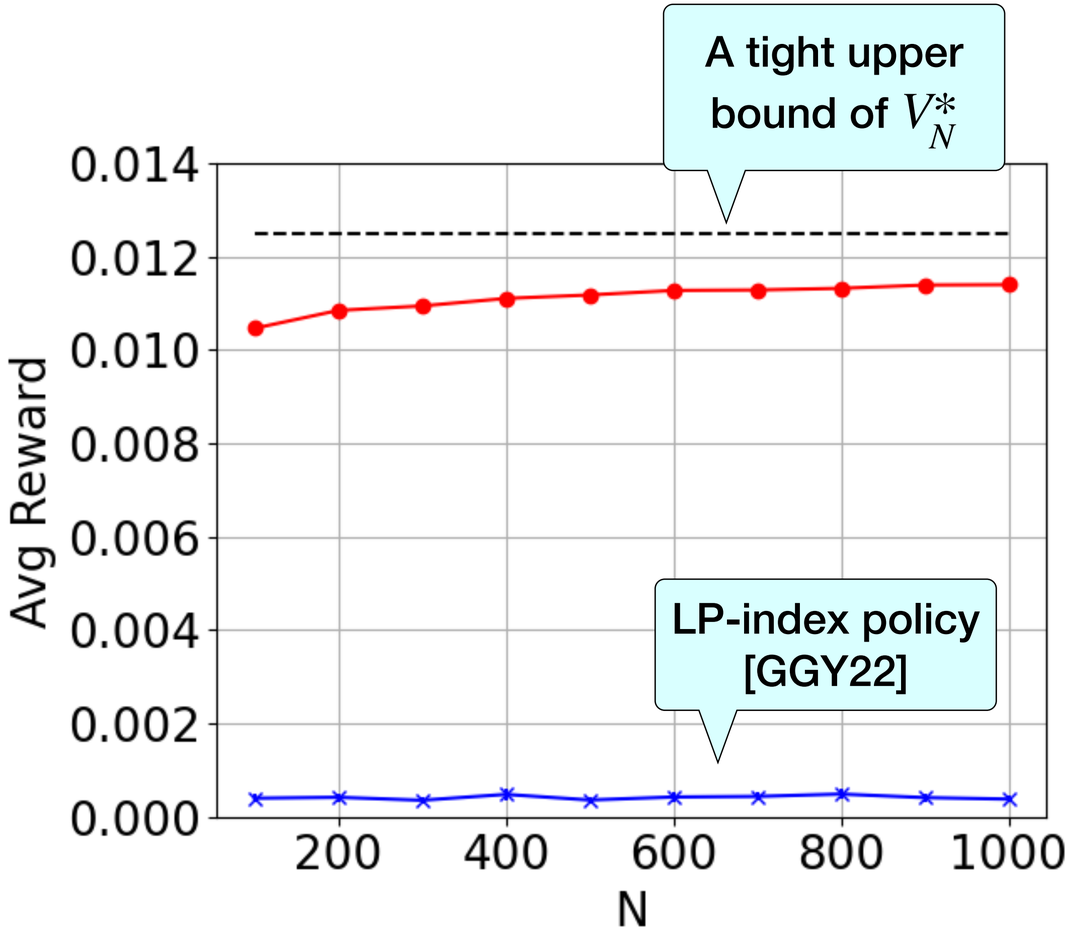
We propose the first policy that is asymptotically optimal without UGAP.

(See our paper for specific conditions.)

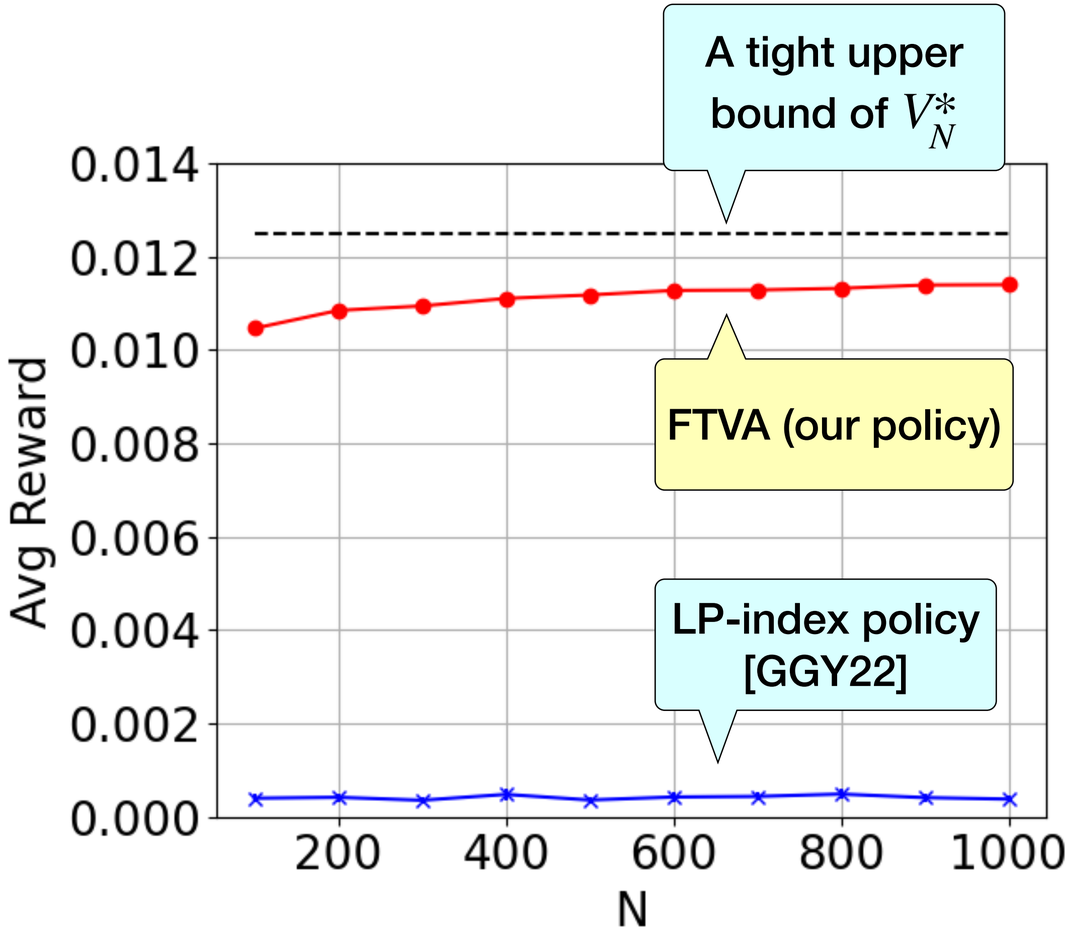
Our policy: Follow-the-Virtual-Advice (FTVA)



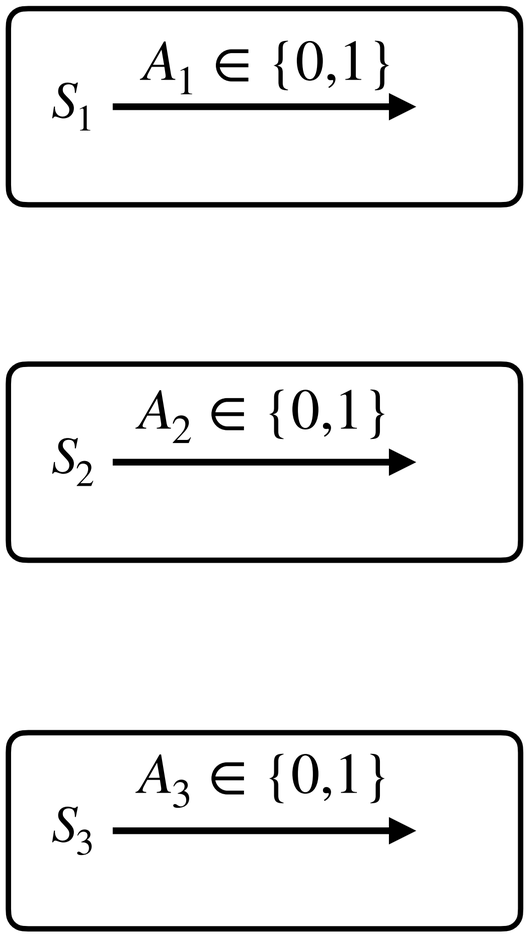
Our policy: Follow-the-Virtual-Advice (FTVA)



Our policy: Follow-the-Virtual-Advice (FTVA)



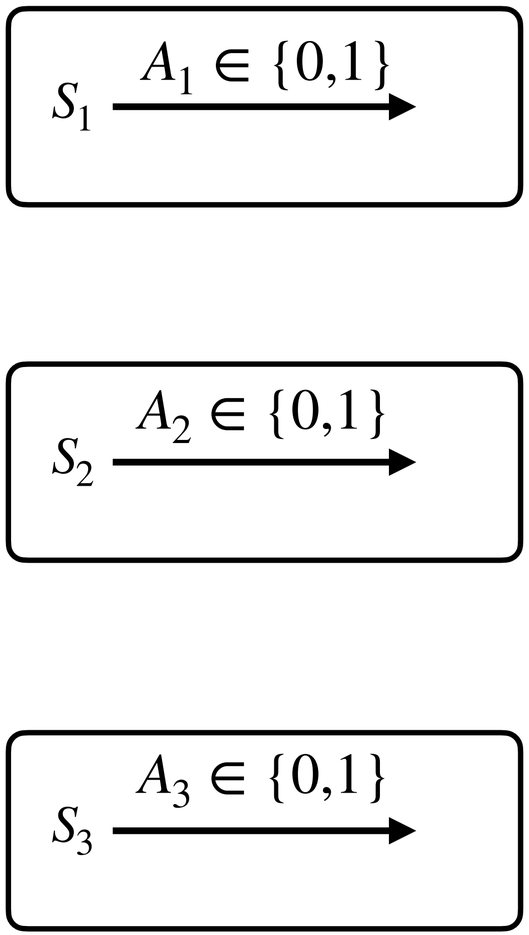
Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

Our policy: Follow-the-Virtual-Advice (FTVA)

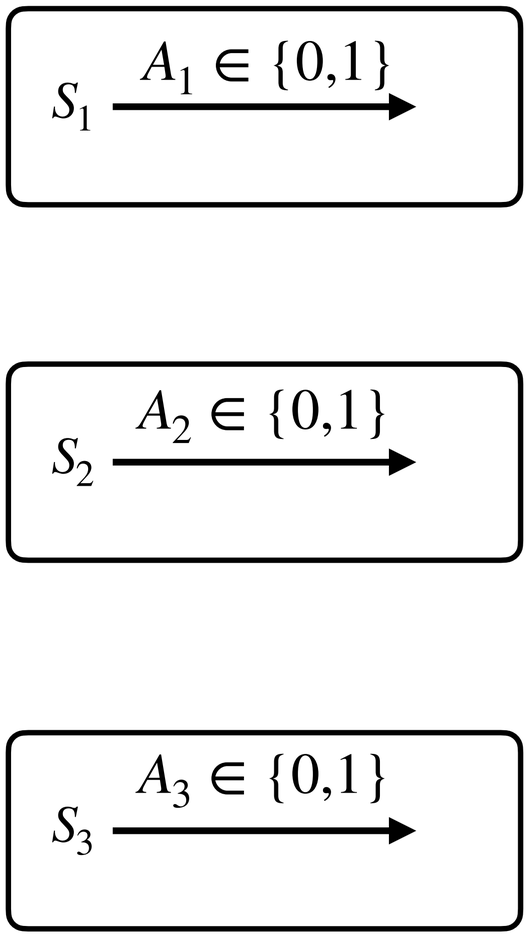


$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

Our policy: Follow-the-Virtual-Advice (FTVA)



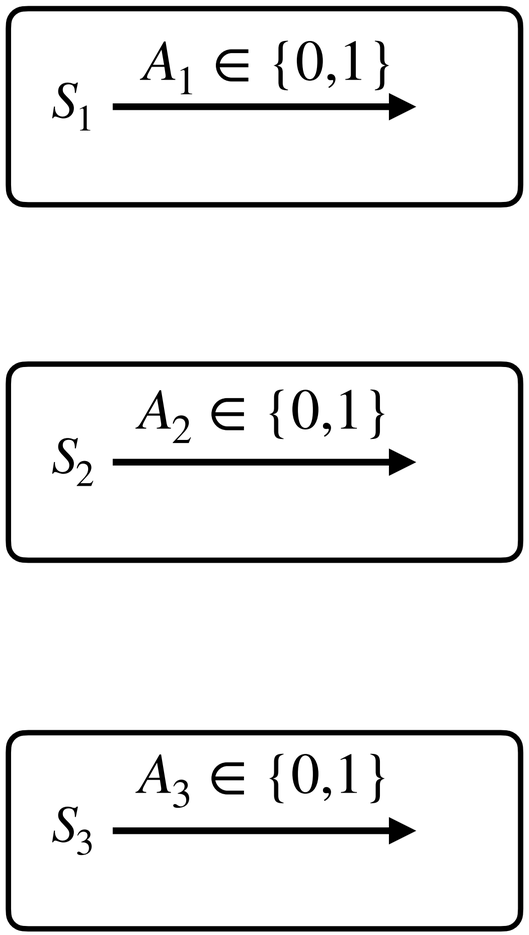
$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

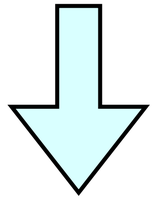
single-armed problem

Our policy: Follow-the-Virtual-Advice (FTVA)



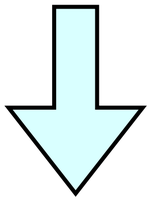
$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

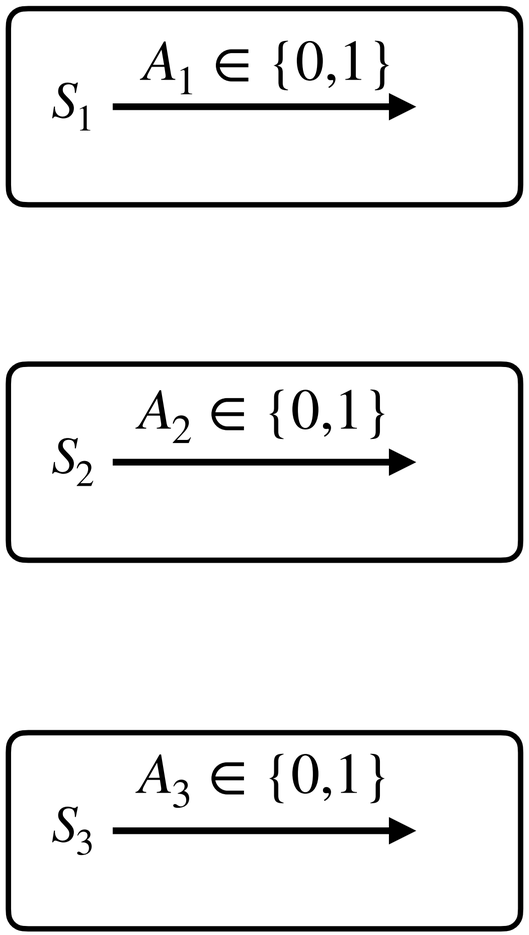


relax

single-armed problem



Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

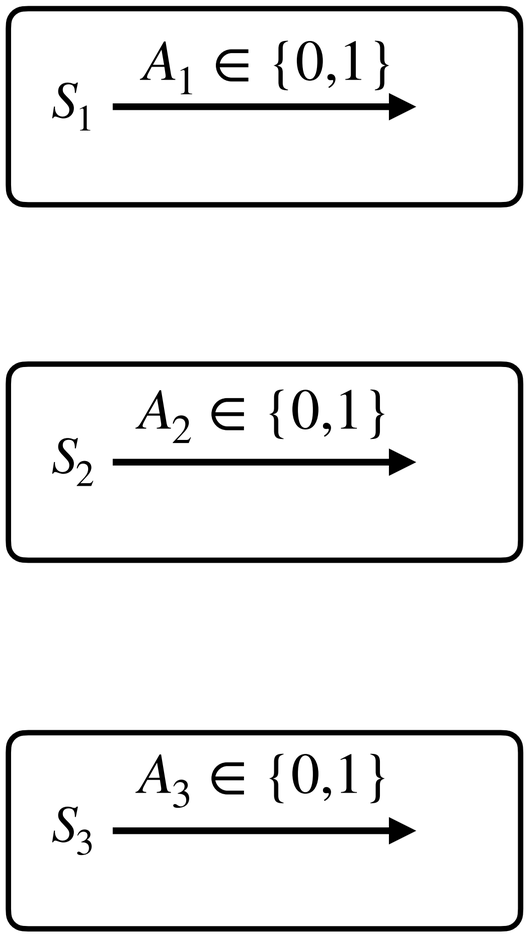
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\overline{\pi}(a \mid s)$

Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

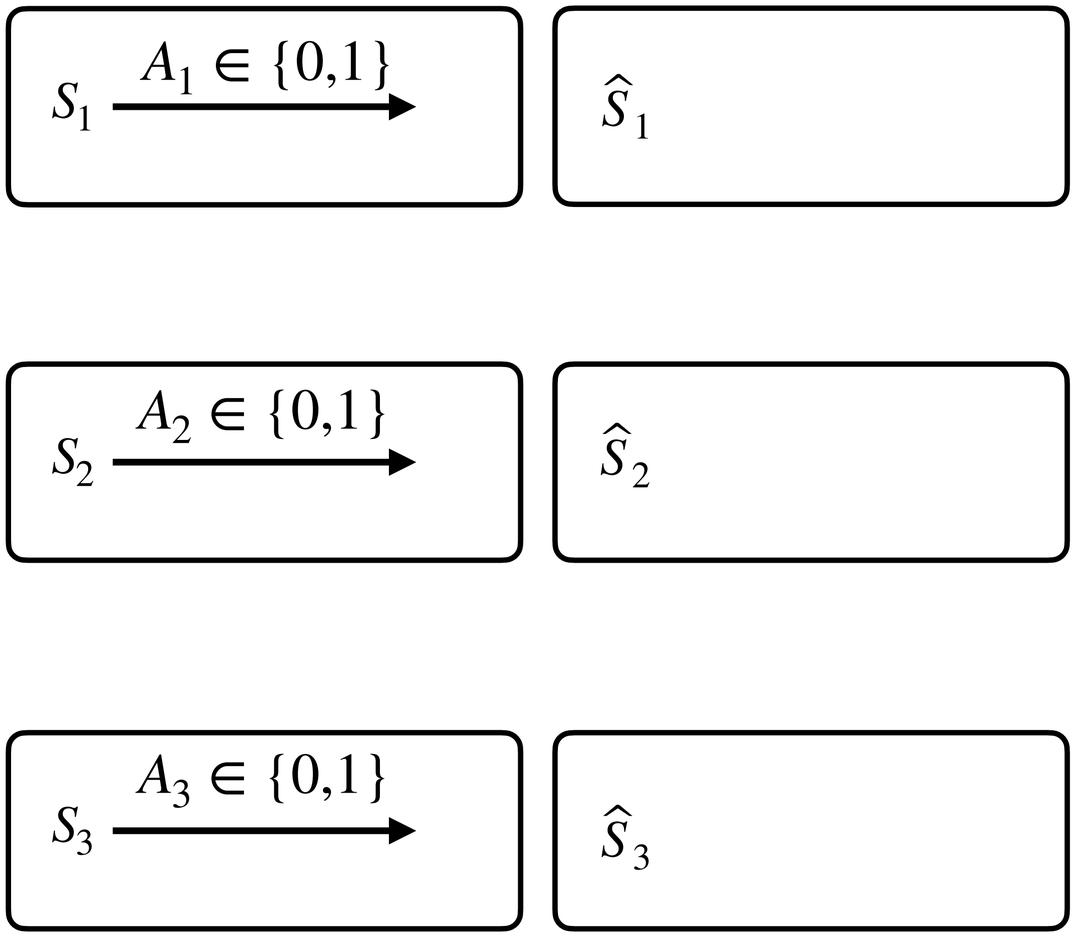
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\overline{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

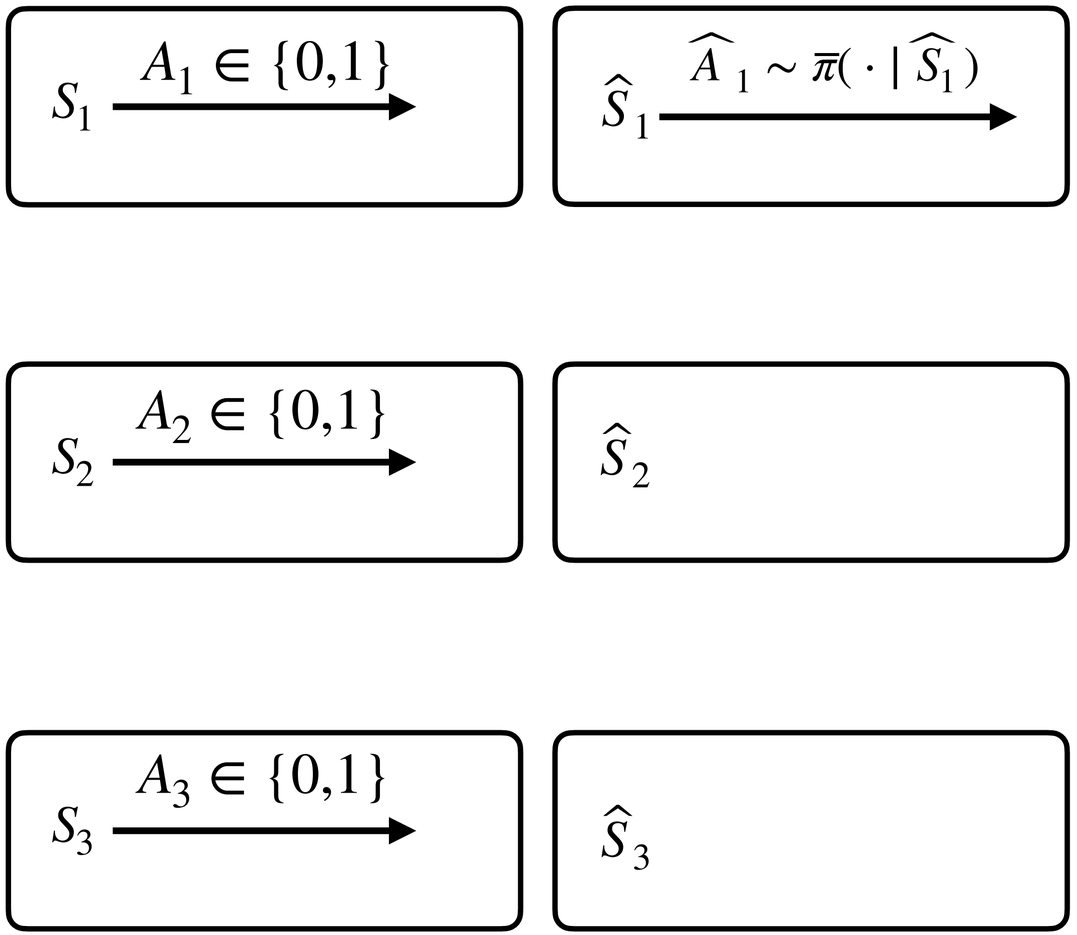
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\overline{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

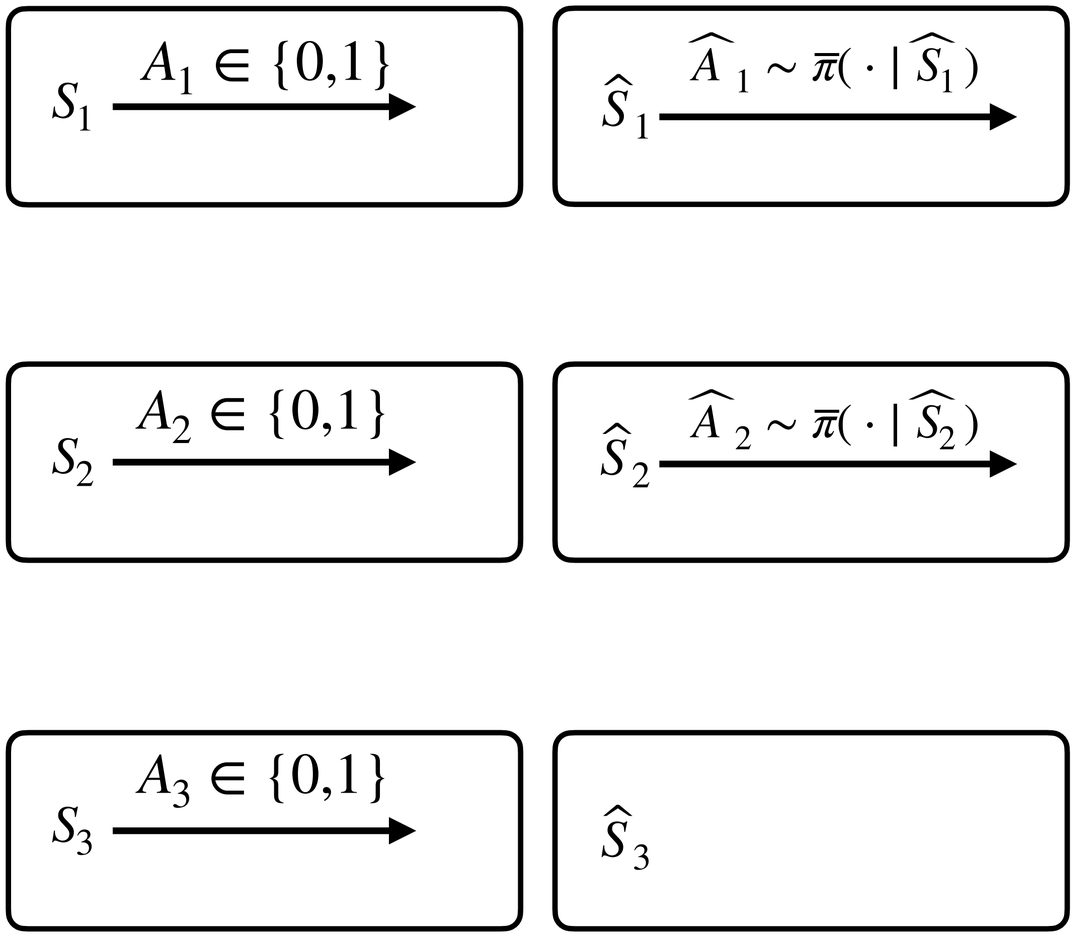
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\bar{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

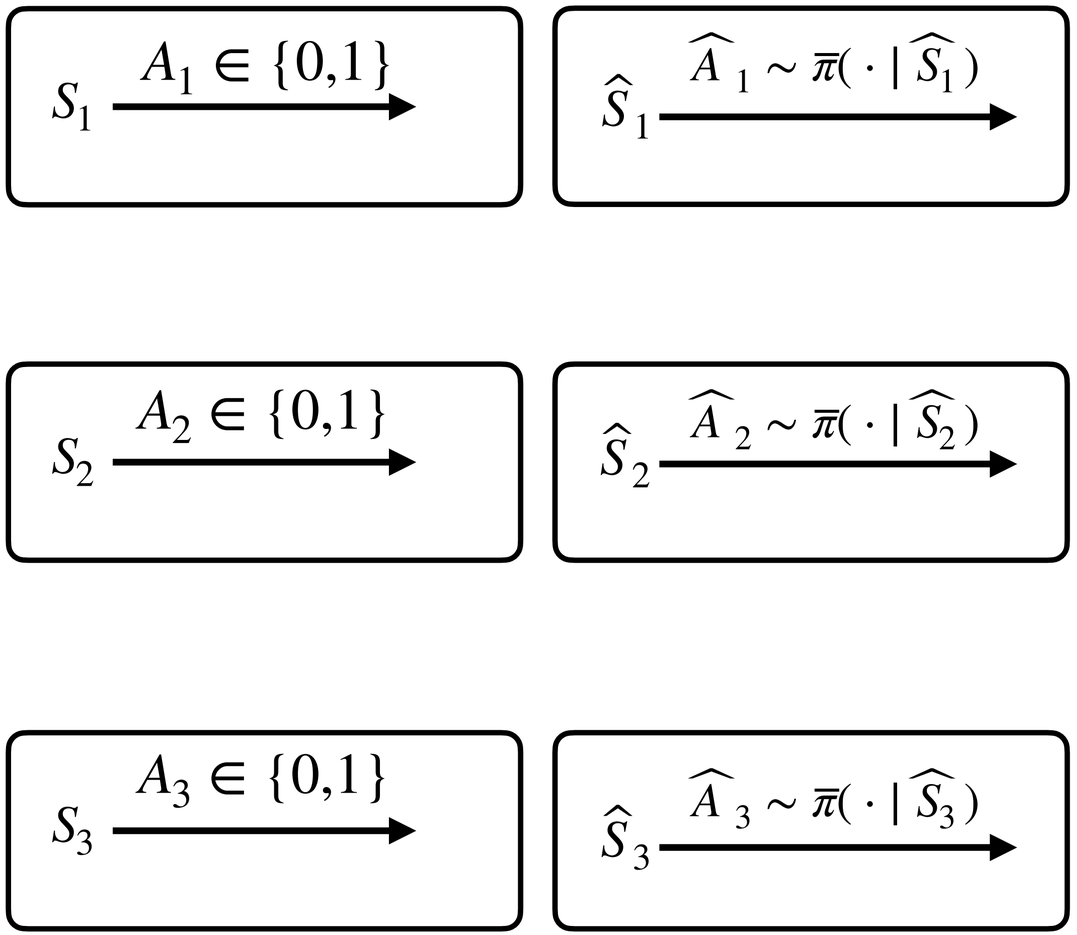
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\overline{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

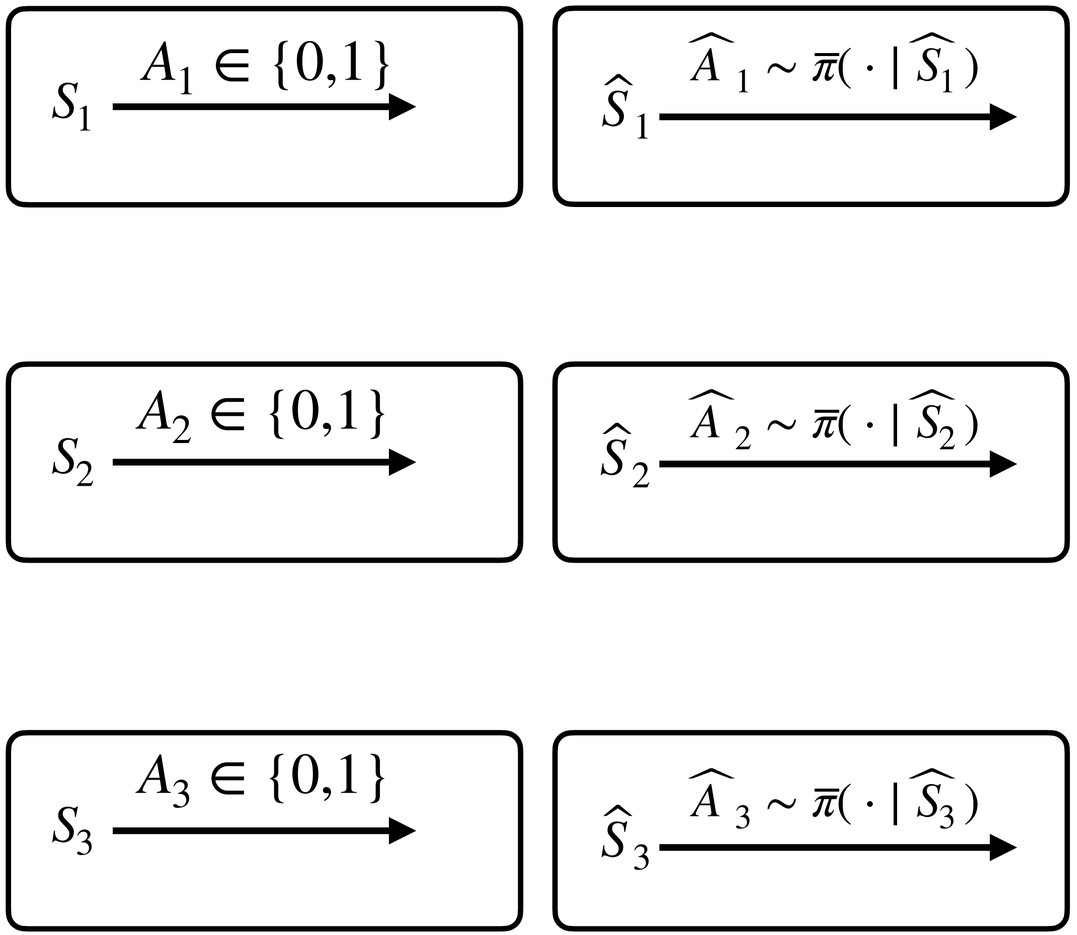
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\overline{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



no constraint

$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

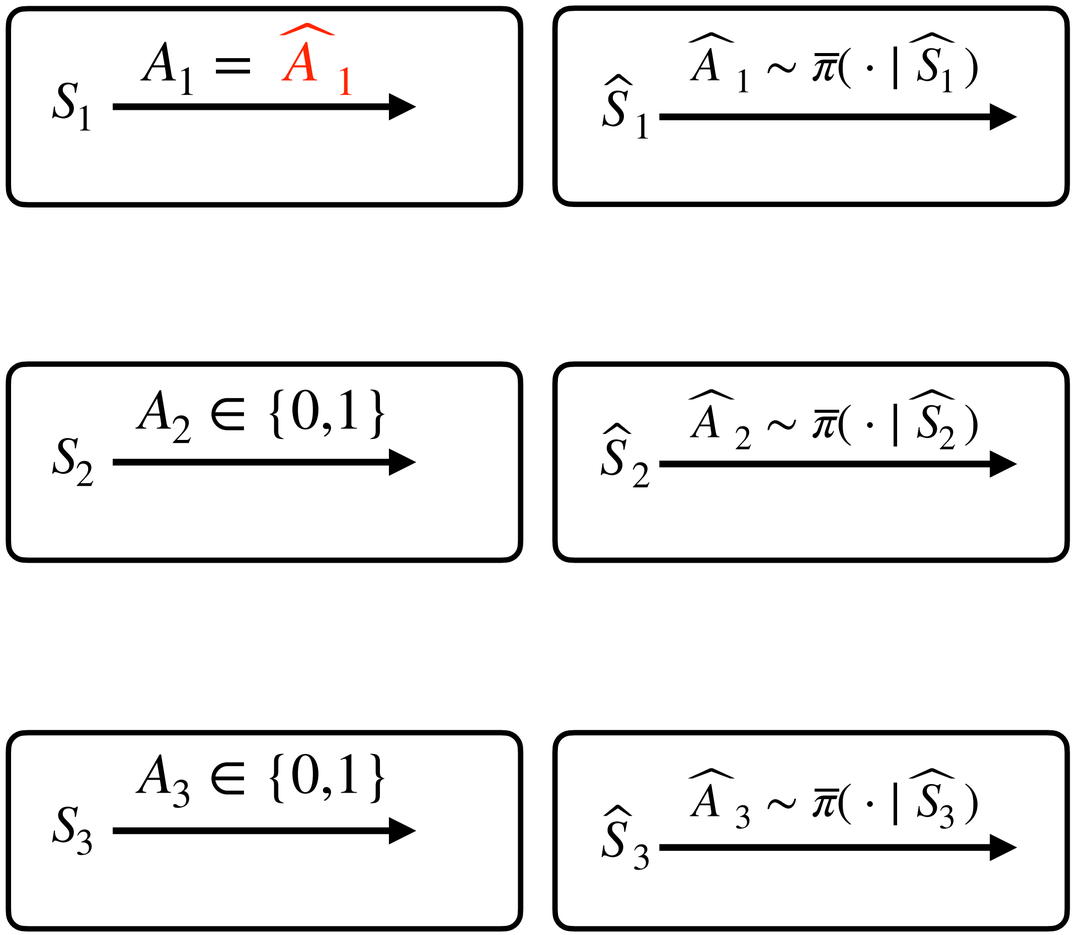
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\overline{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

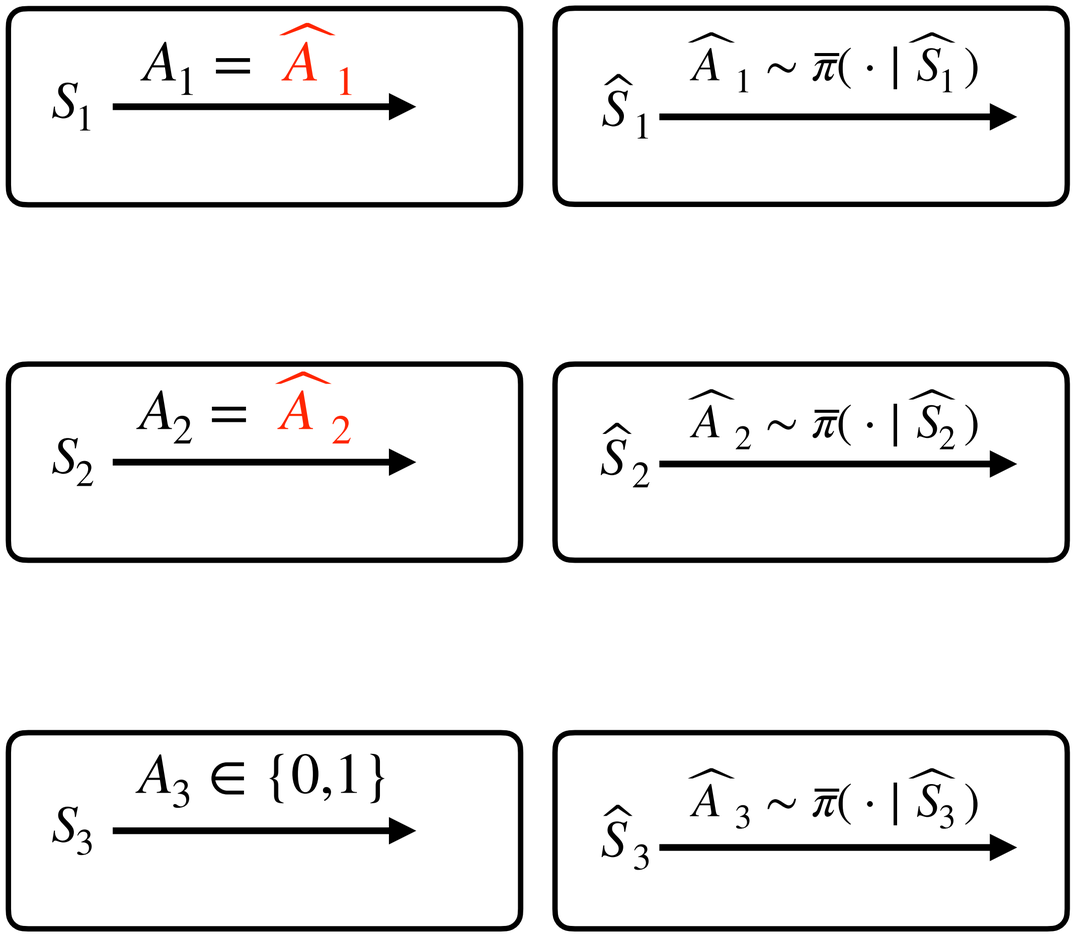
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\bar{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



no constraint

$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

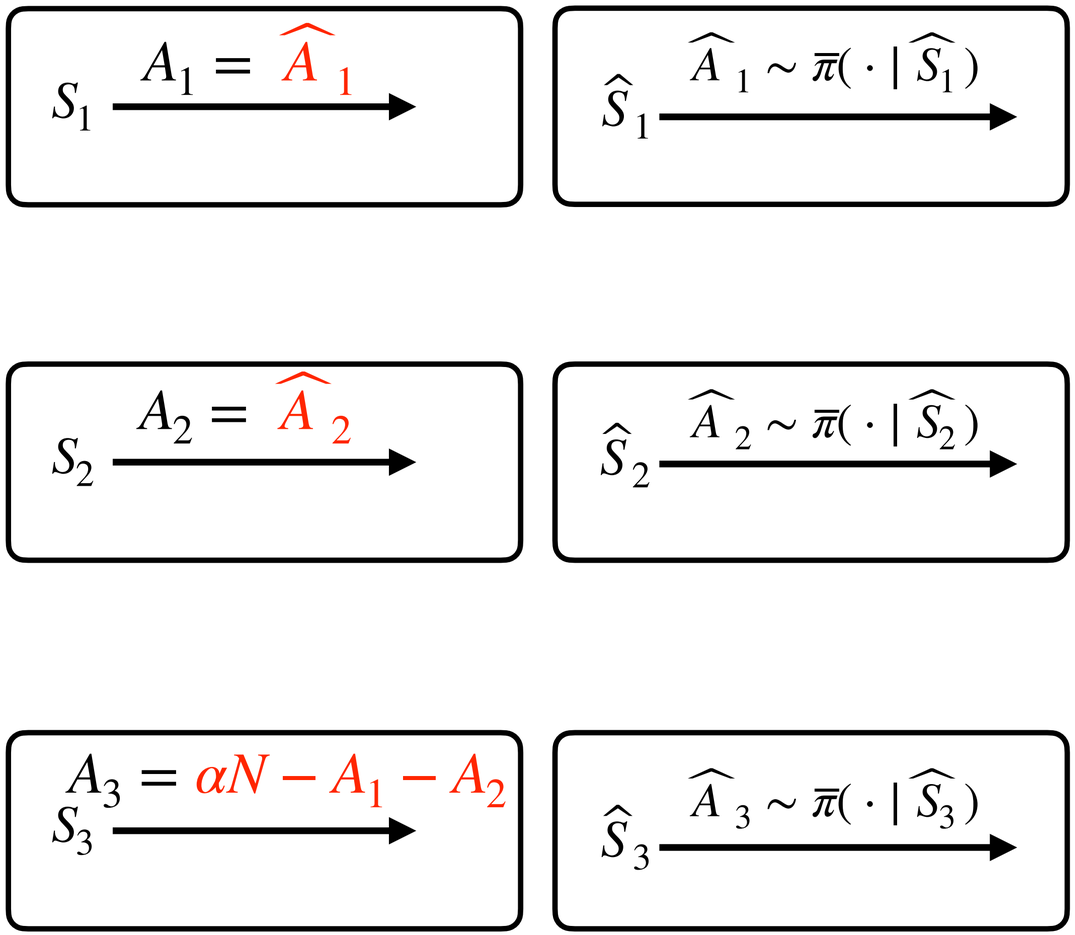
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\bar{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



no constraint

$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

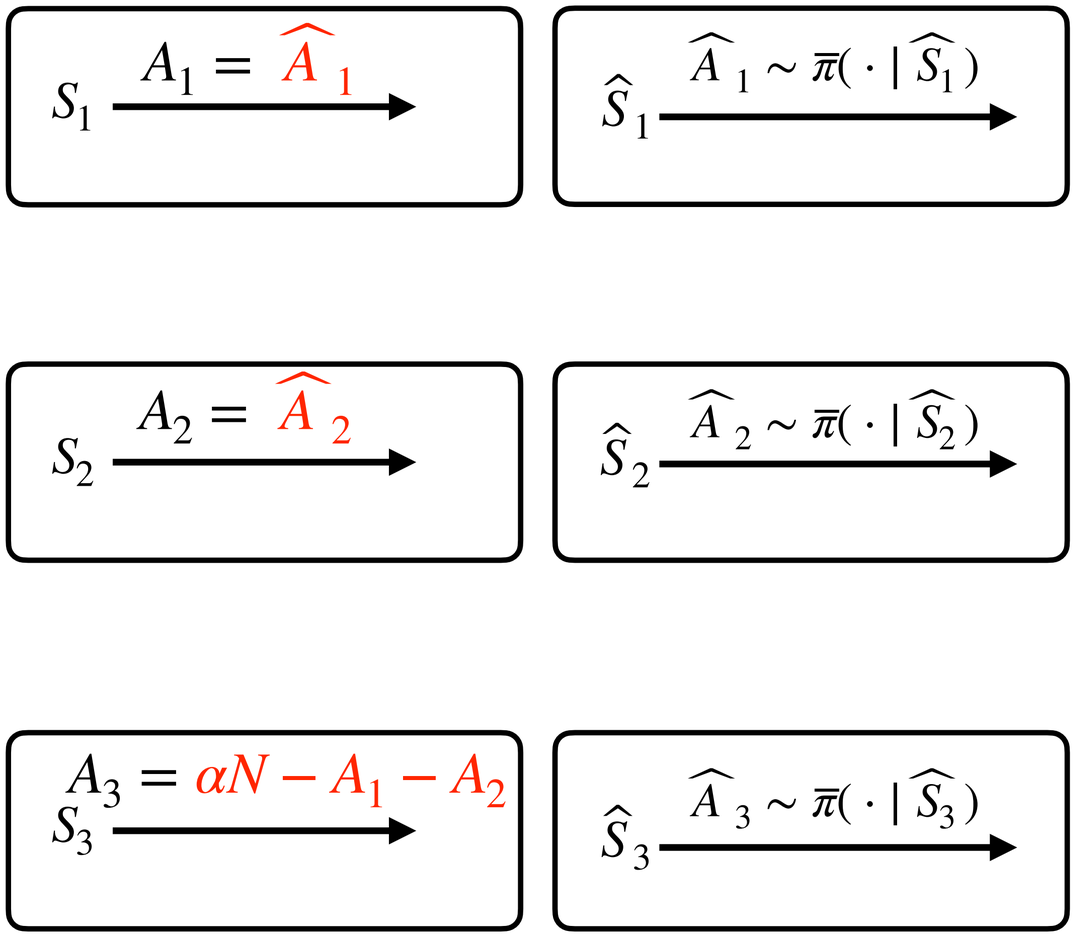
s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\bar{\pi}(a \mid s)$ — generate ideal actions

Our policy: Follow-the-Virtual-Advice (FTVA)



with constraint no constraint

$\max_{\pi} V_N^{\pi} \triangleq$ long run average reward under policy π

s.t. $\sum_{i=1}^{N} A_i = \alpha N$, any time slot

relax

single-armed problem

single-armed policy $\bar{\pi}(a \mid s)$ — generate ideal actions