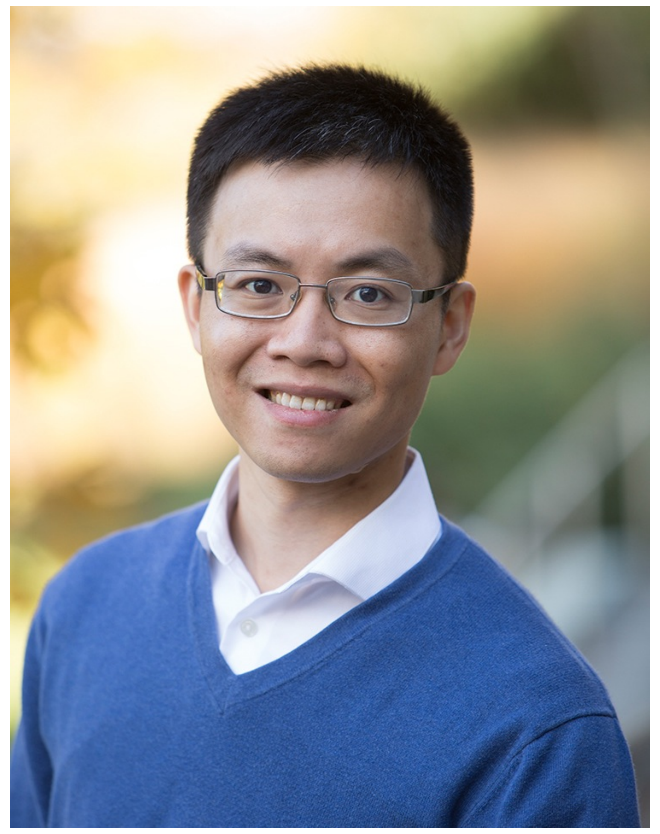
Average-Reward Restless Bandits:
Unichain and Aperiodicity are Sufficient for Asymptotic Optimality

Yige Hong
Carnegie Mellon University

INFORMS 2024

Weina Wang
CMU

Qiaomin Xie
UW—Madison



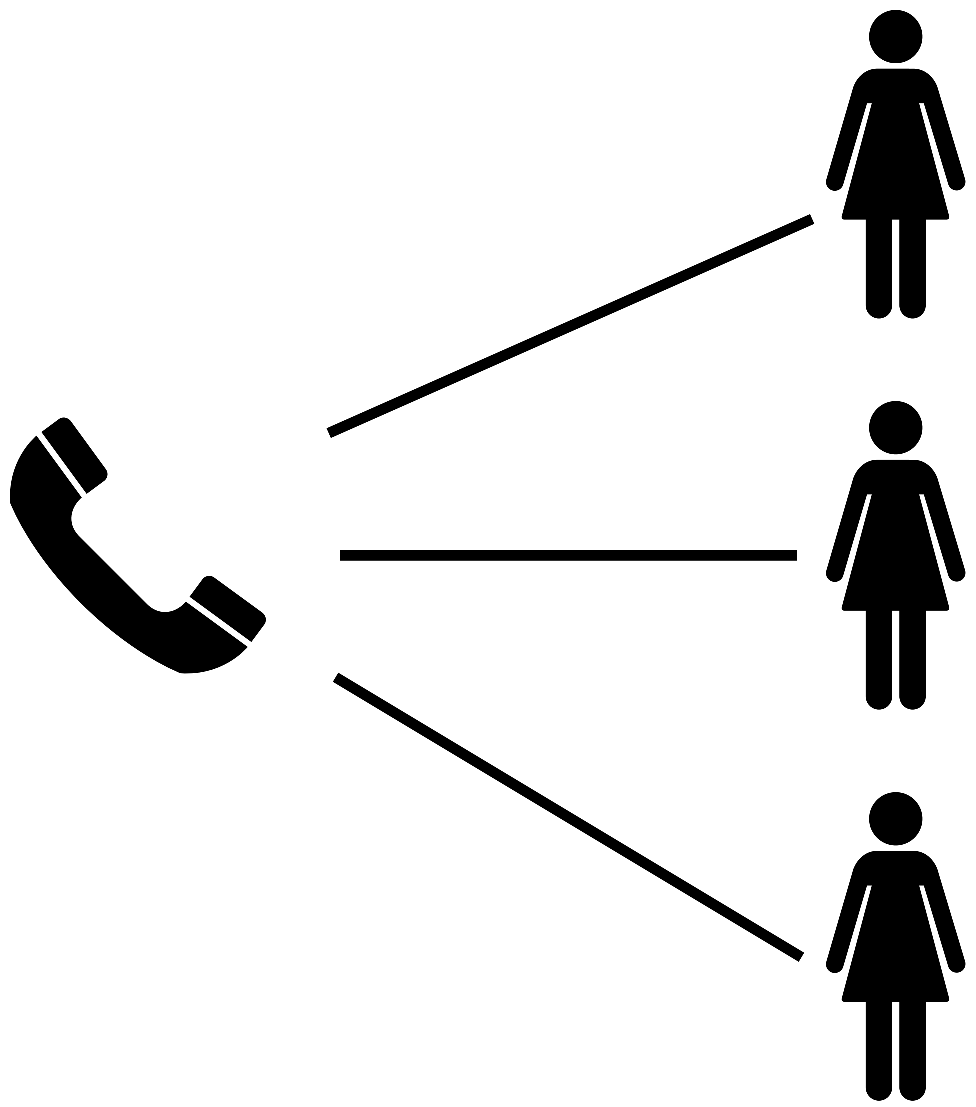
Yudong Chen
UW—Madison

Motivation

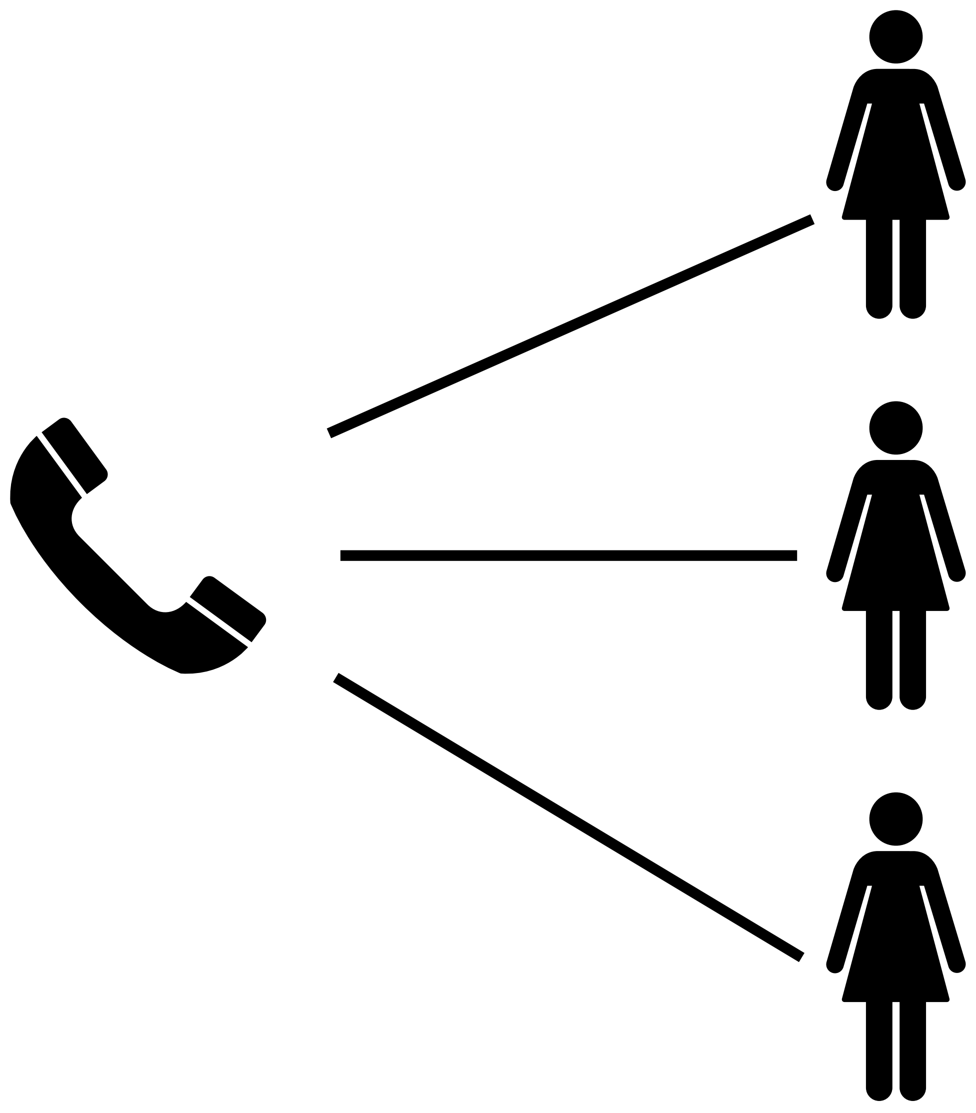
- Restless bandits:

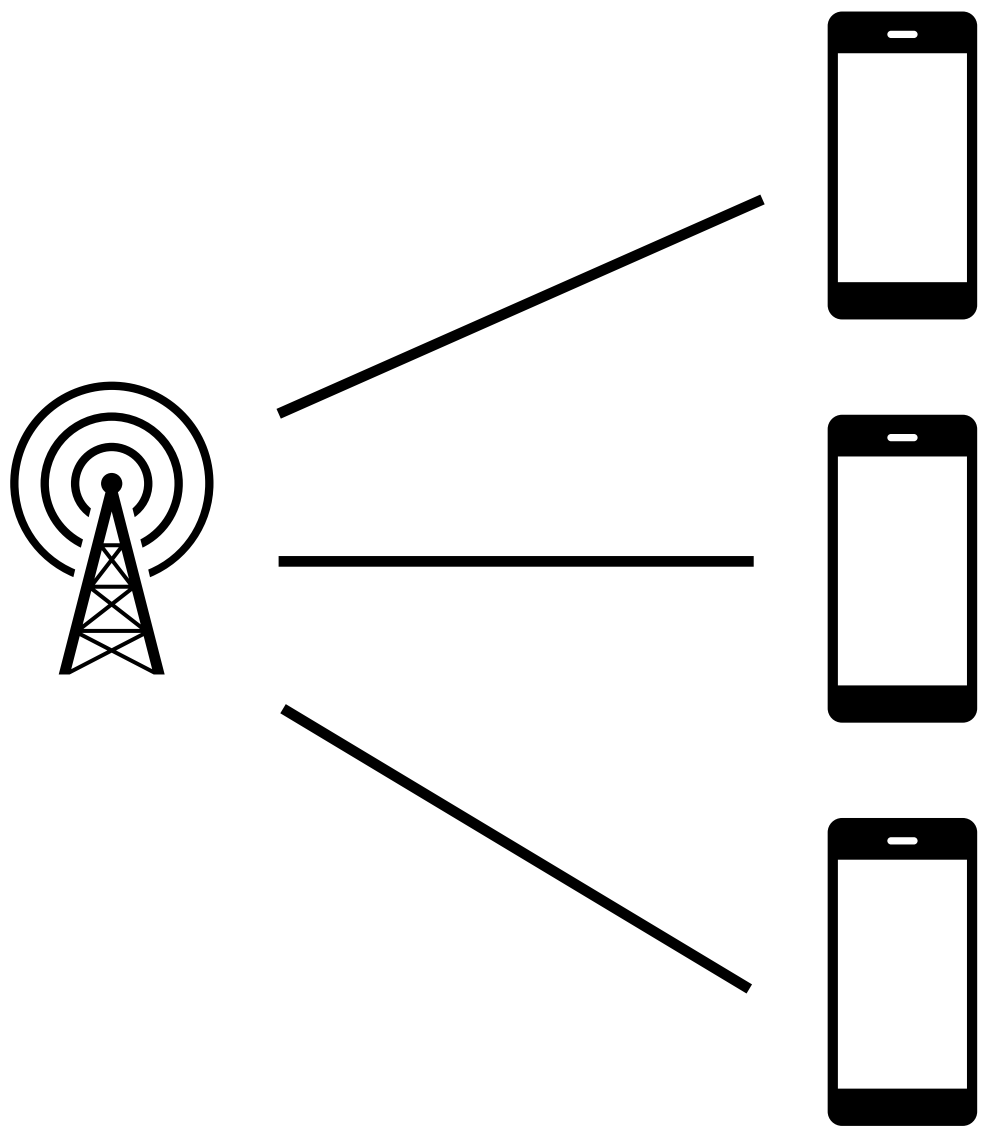
Motivation

- Restless bandits:
  - Public health



Motivation

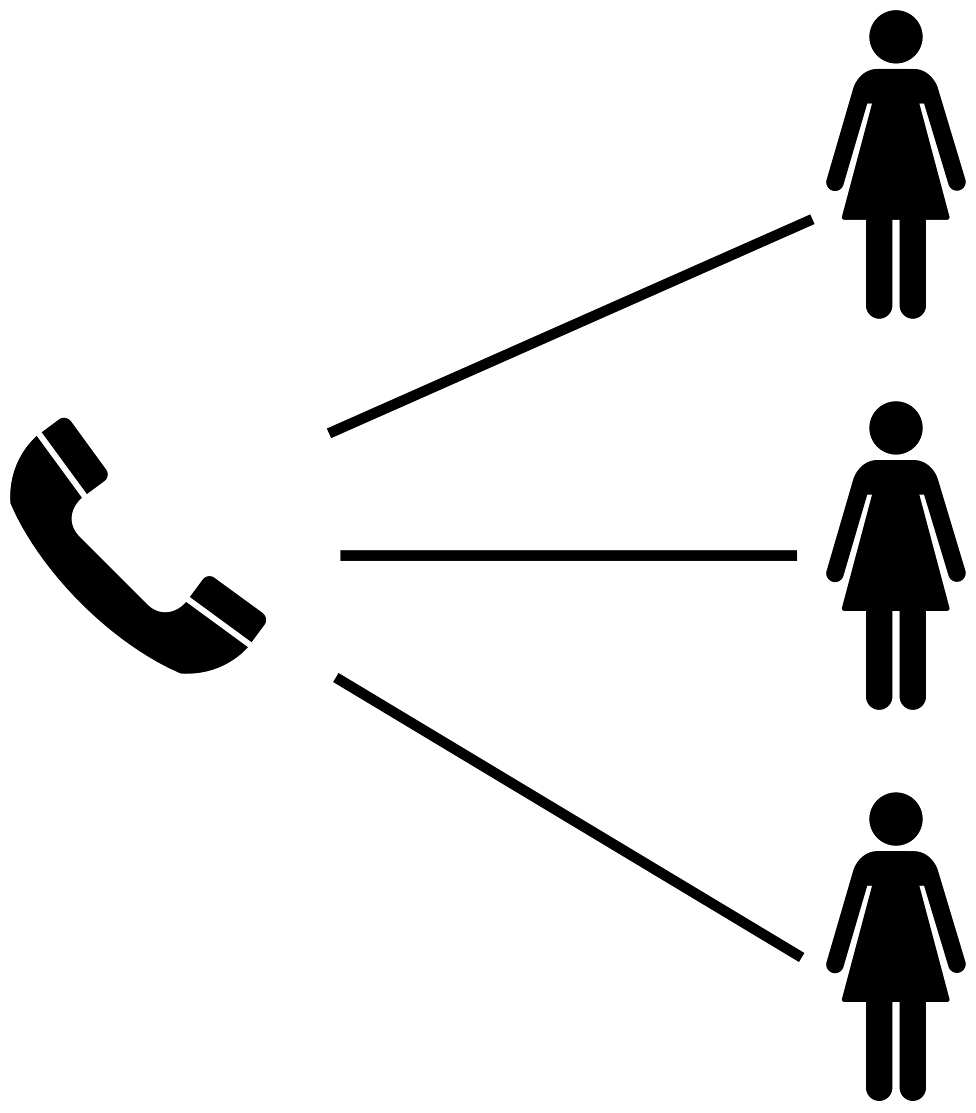


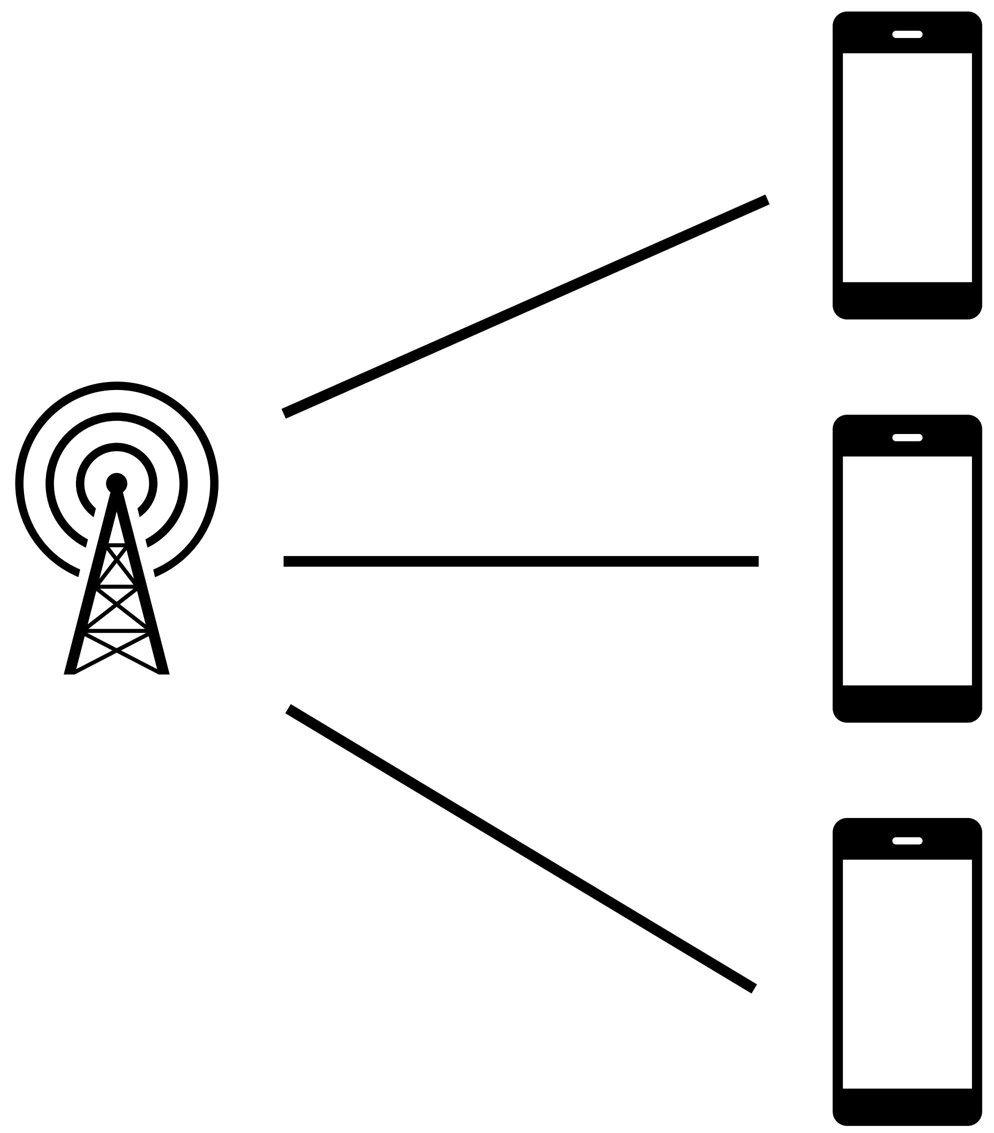


- Restless bandits:
  - Public health
  - Wireless communications

Motivation

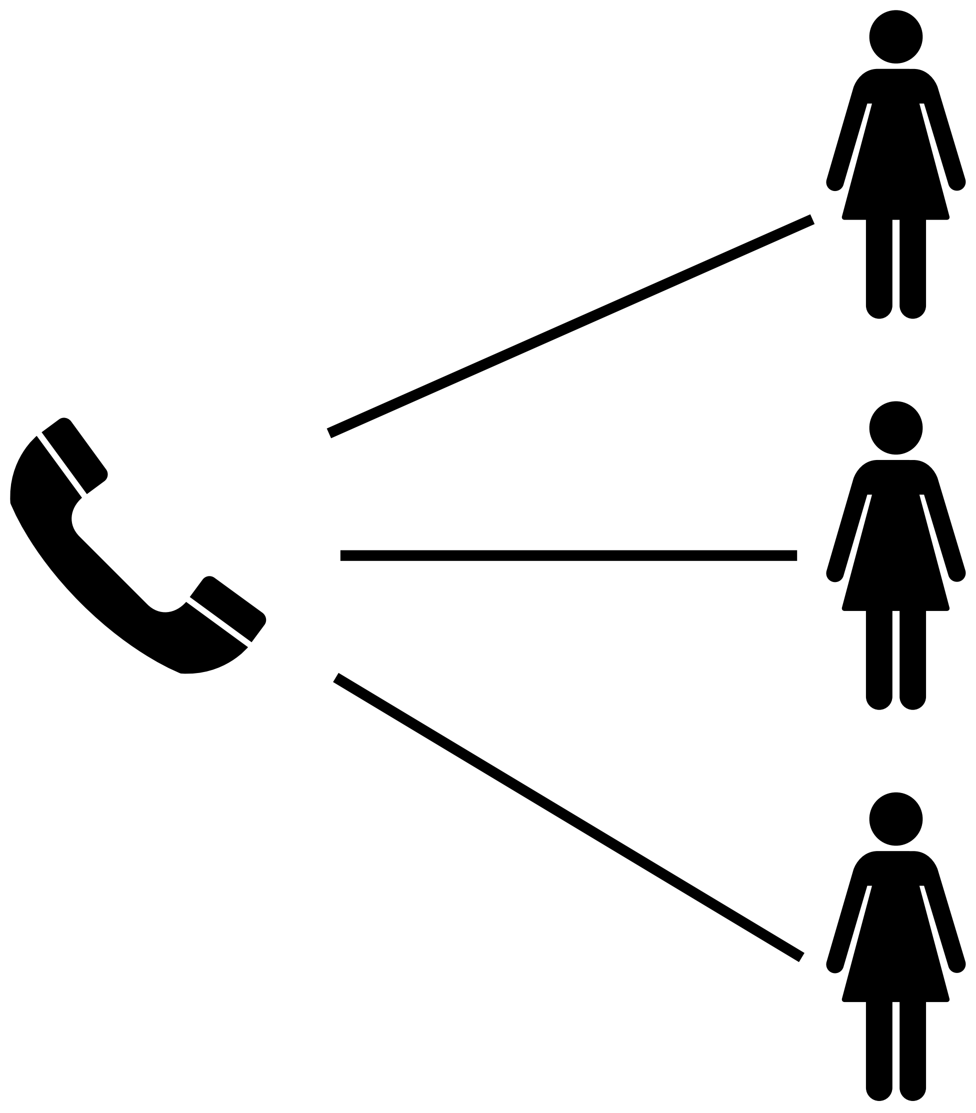
- Restless bandits:
  - Public health
  - Wireless communications
  - Machine maintenance scheduling

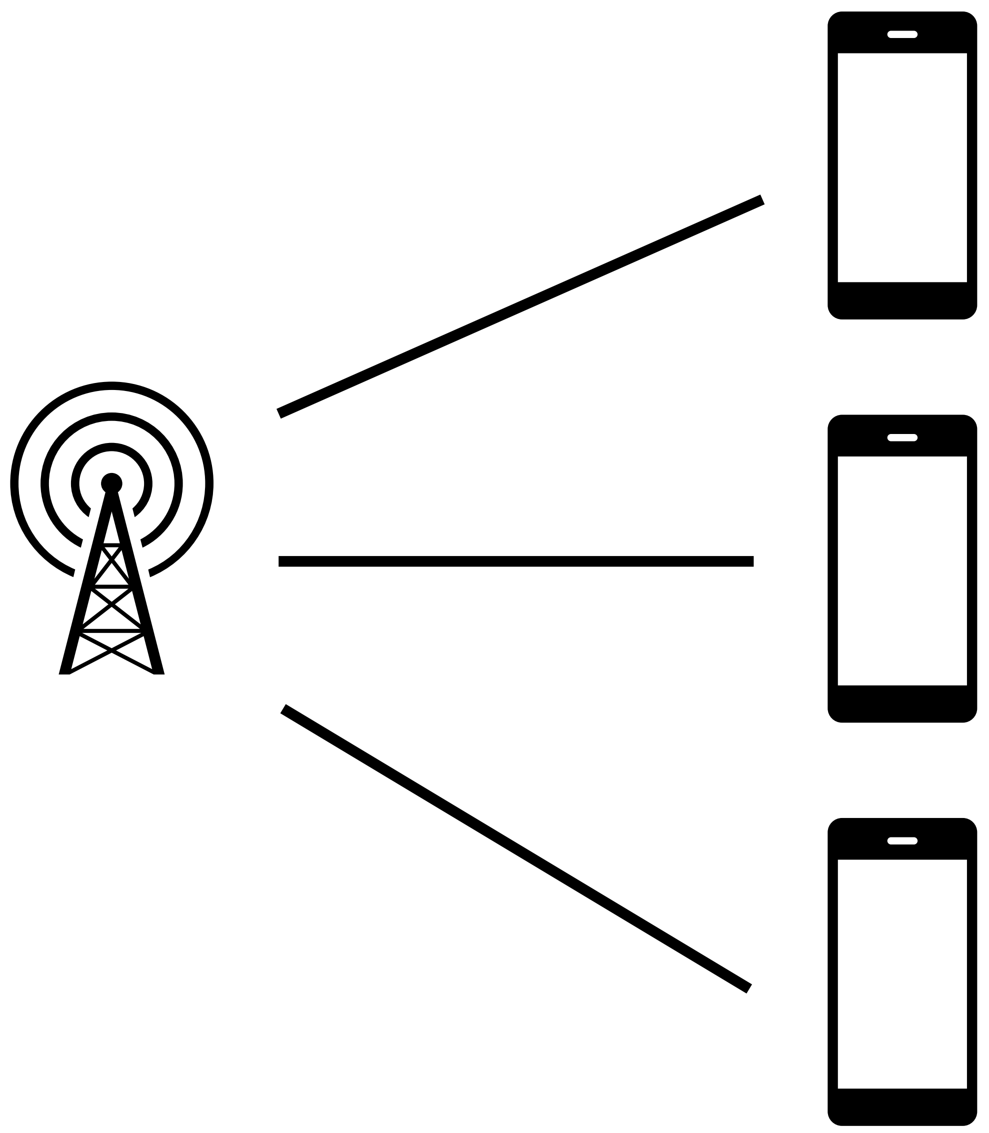




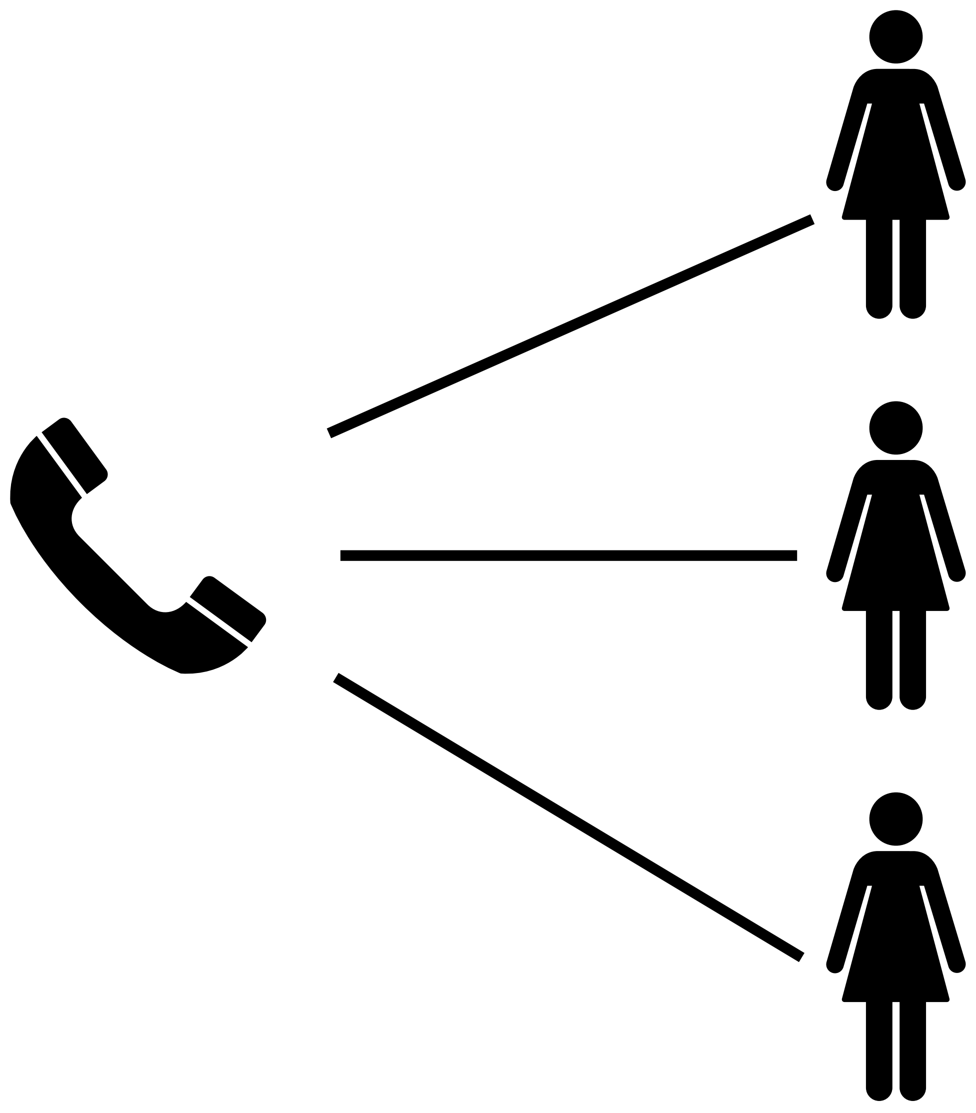
Motivation

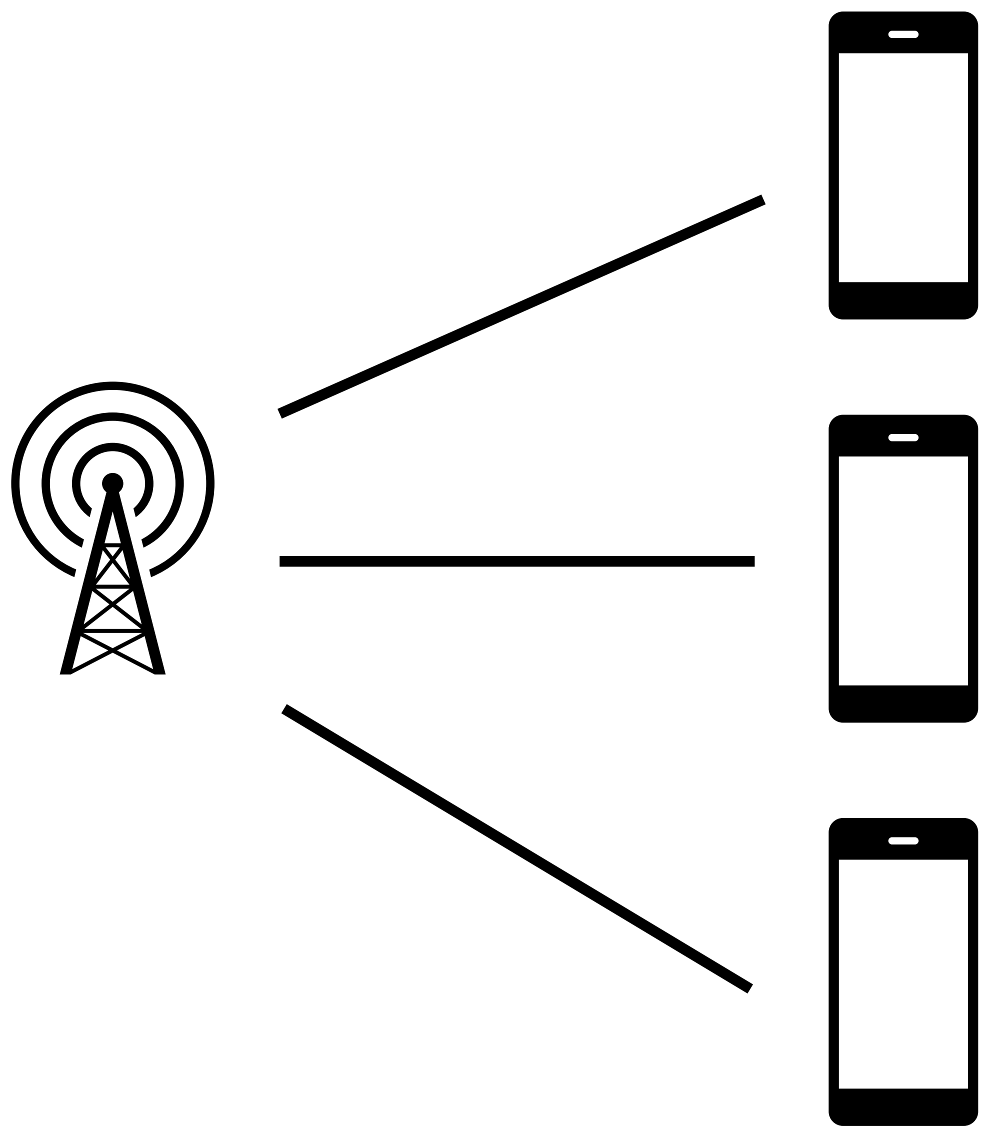
- Restless bandits:
  - Public health
  - Wireless communications
  - Machine maintenance scheduling
  - Machine learning





Motivation

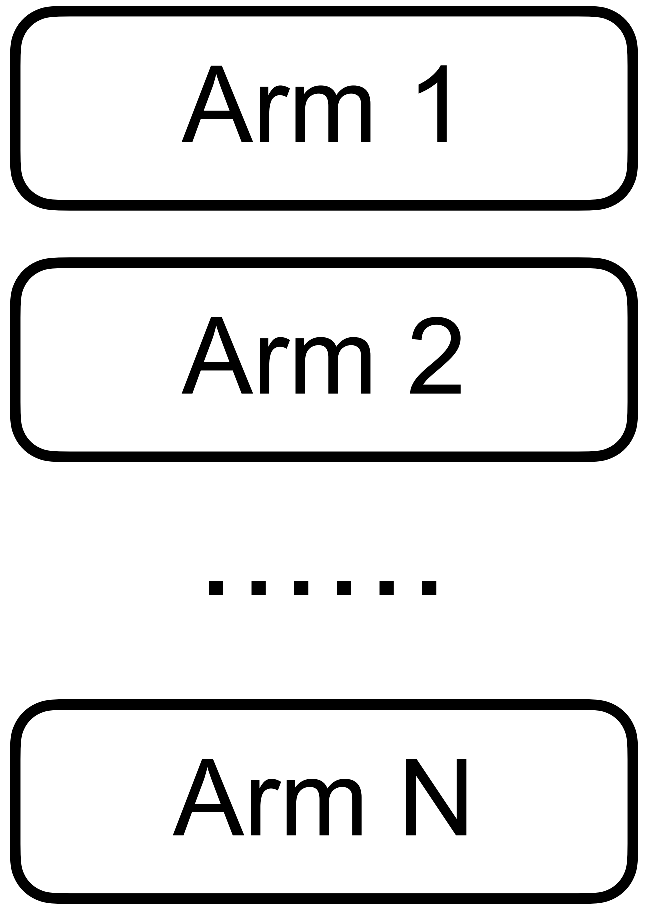




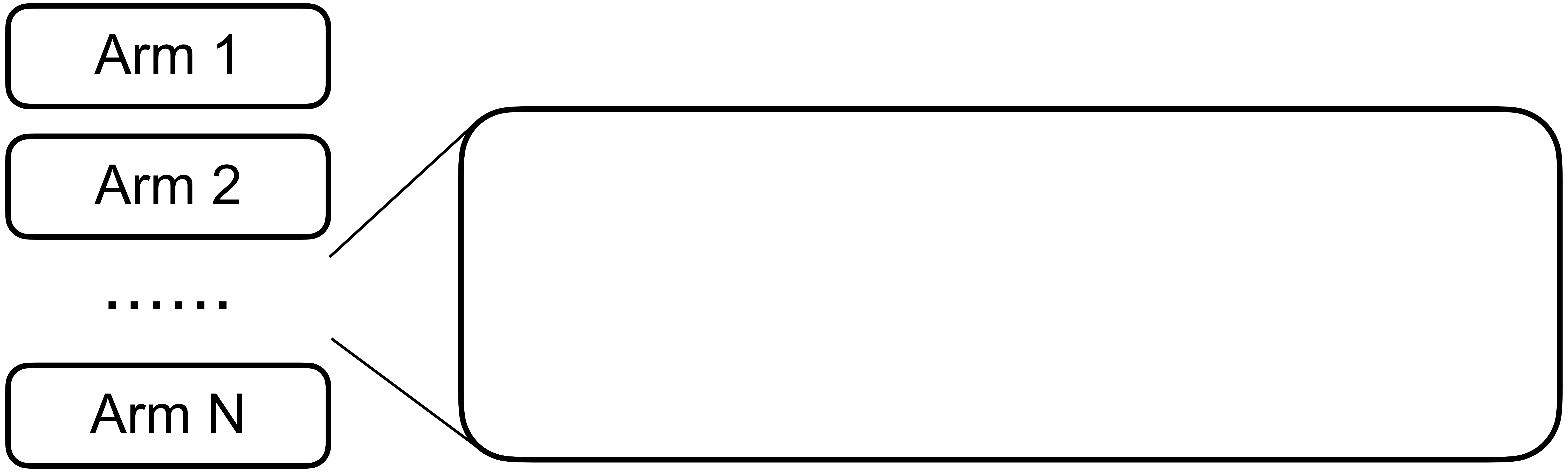
- Restless bandits:
  - Public health
  - Wireless communications
  - Machine maintenance scheduling
  - Machine learning

How to optimally allocate resources in a large system consisting of multiple dynamic components?

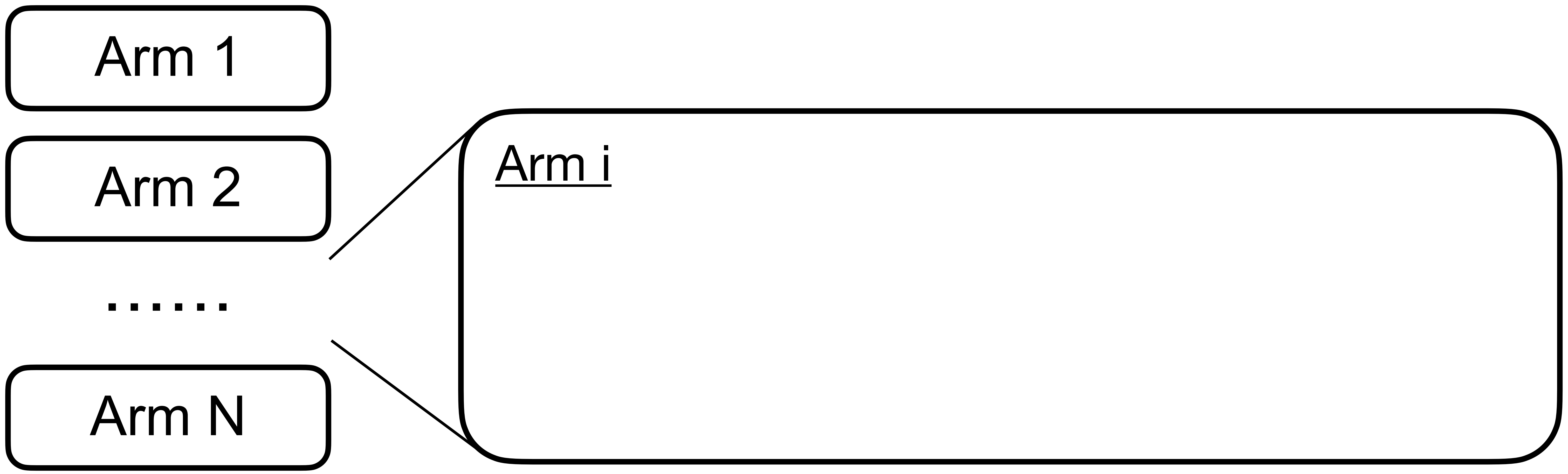
Restless bandits: problem definition



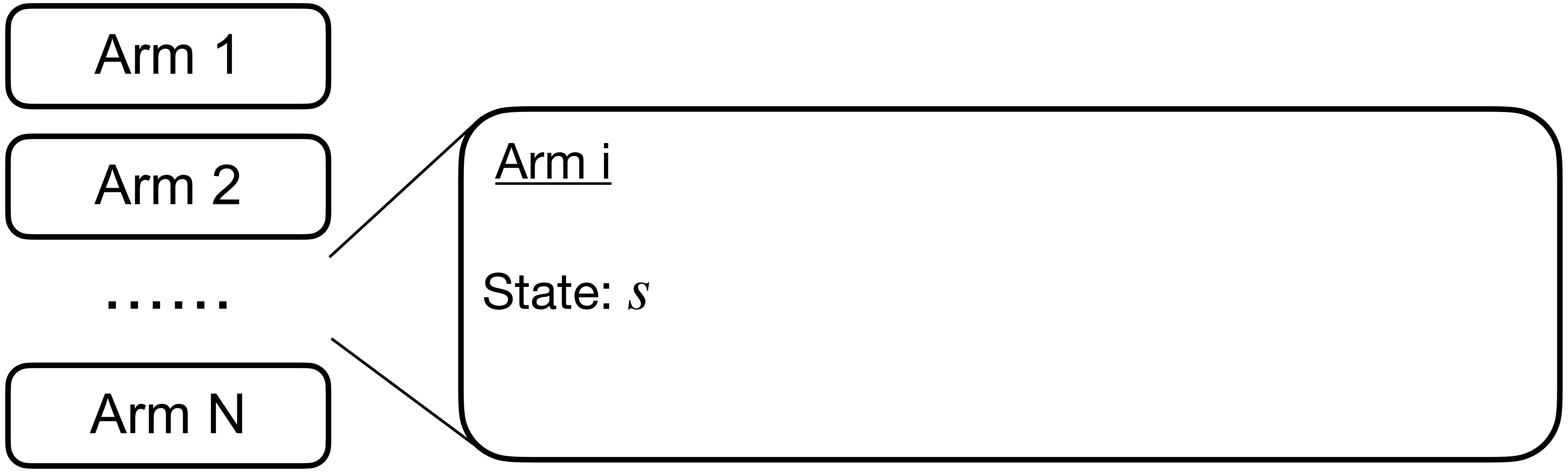
Restless bandits: problem definition



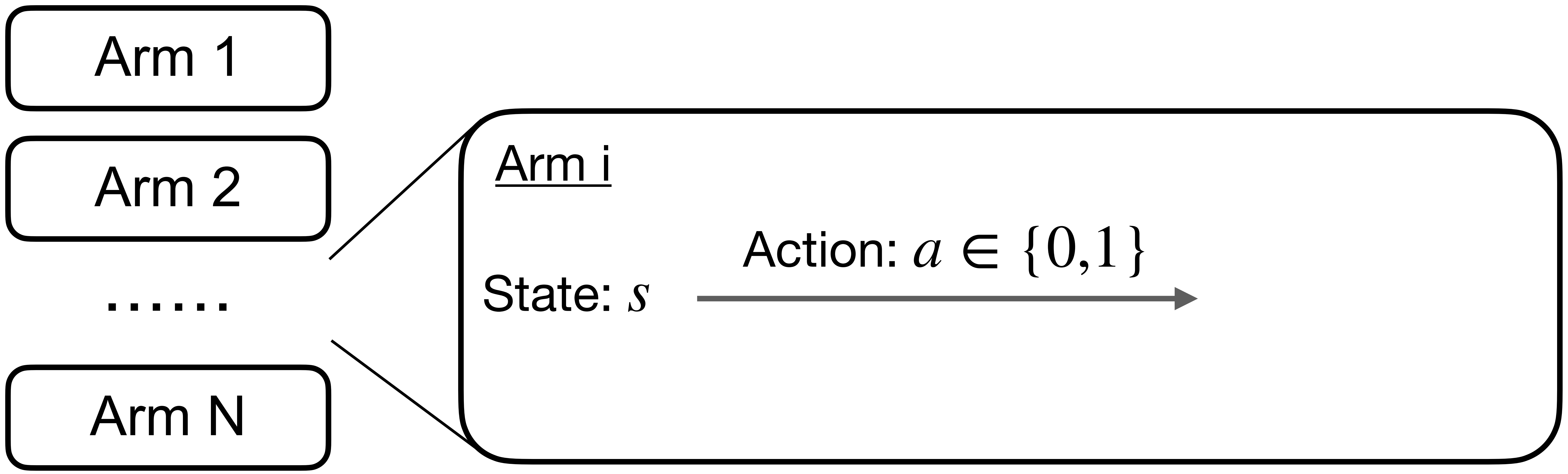
Restless bandits: problem definition



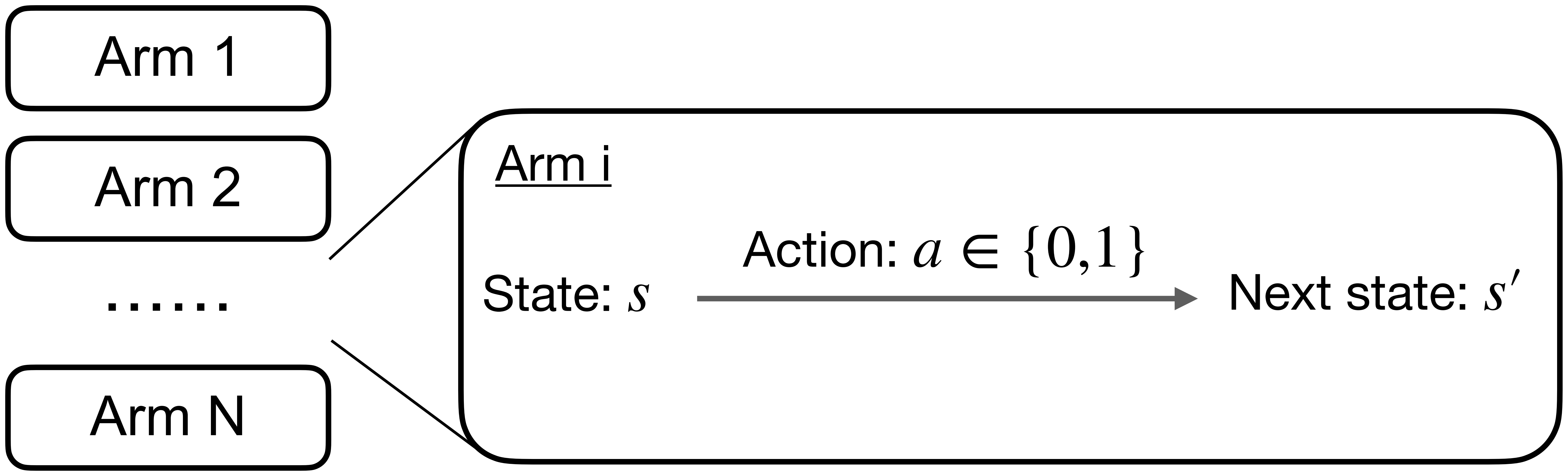
Restless bandits: problem definition



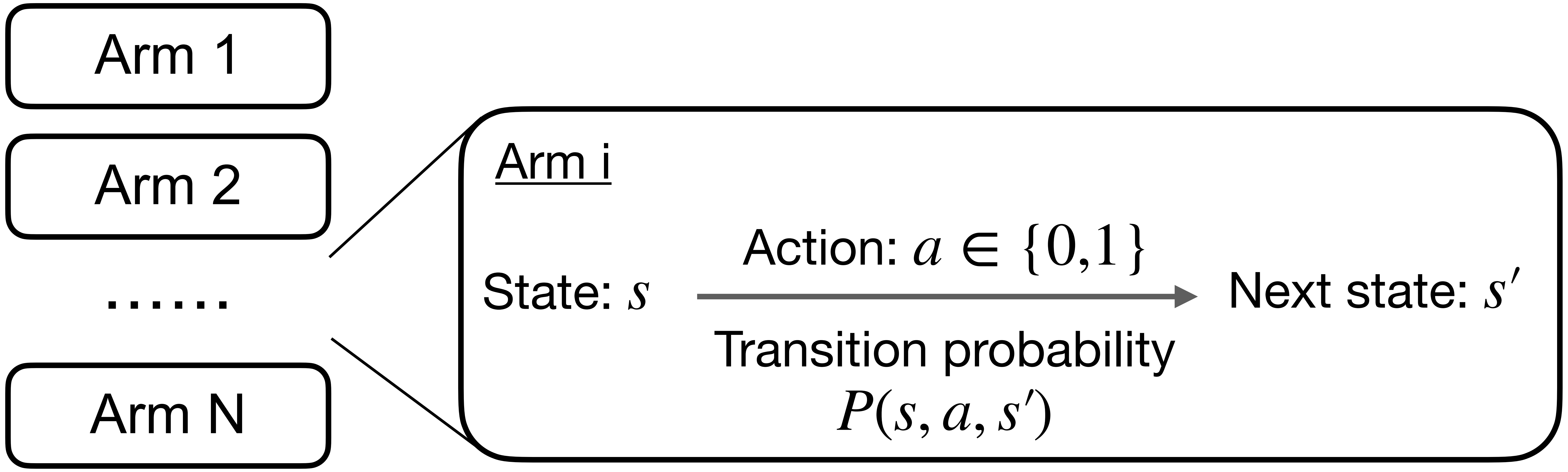
Restless bandits: problem definition



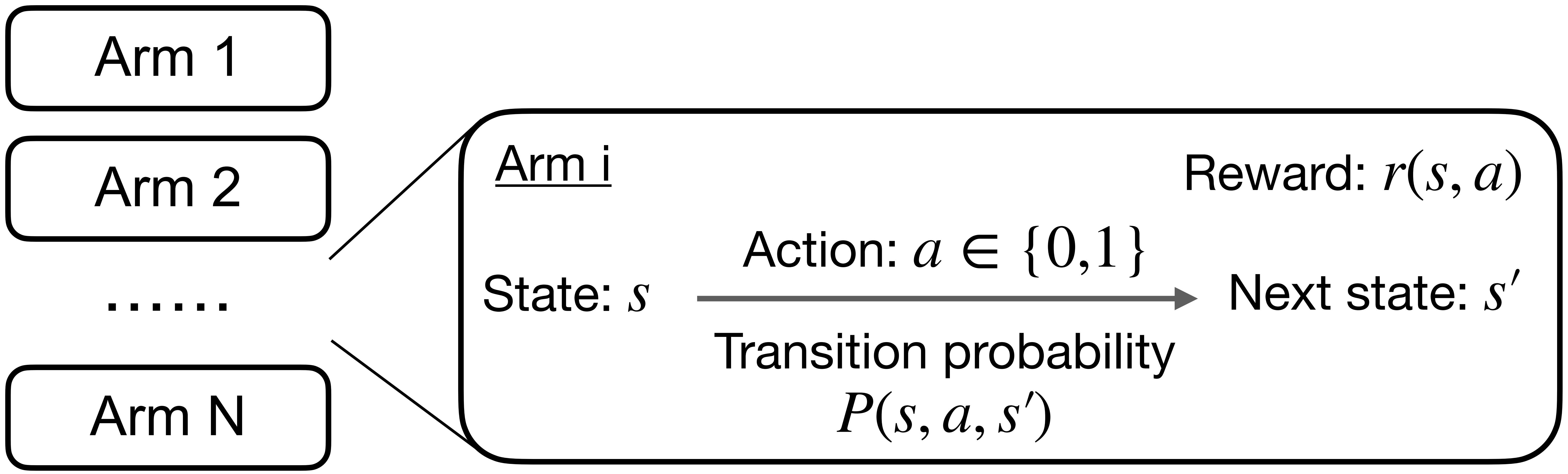
Restless bandits: problem definition



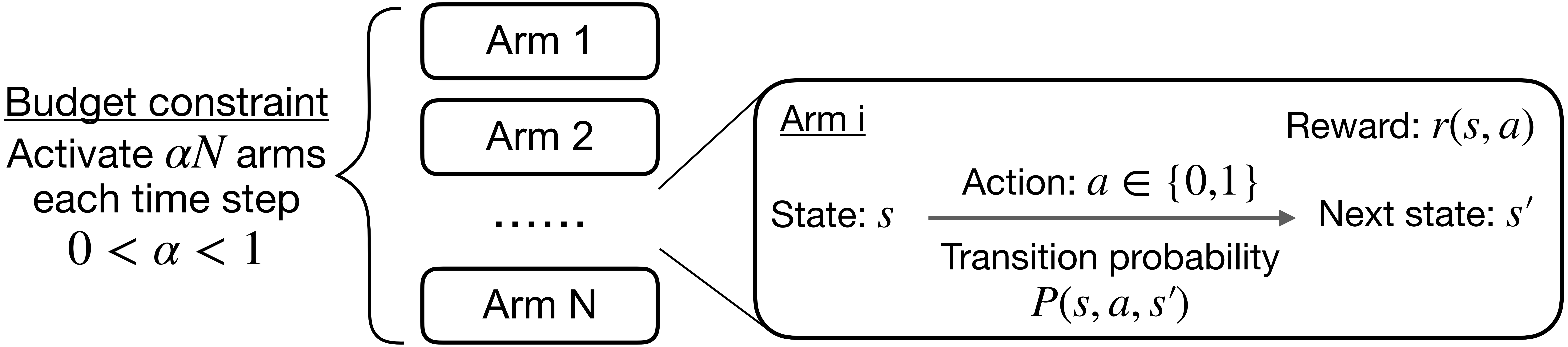
Restless bandits: problem definition



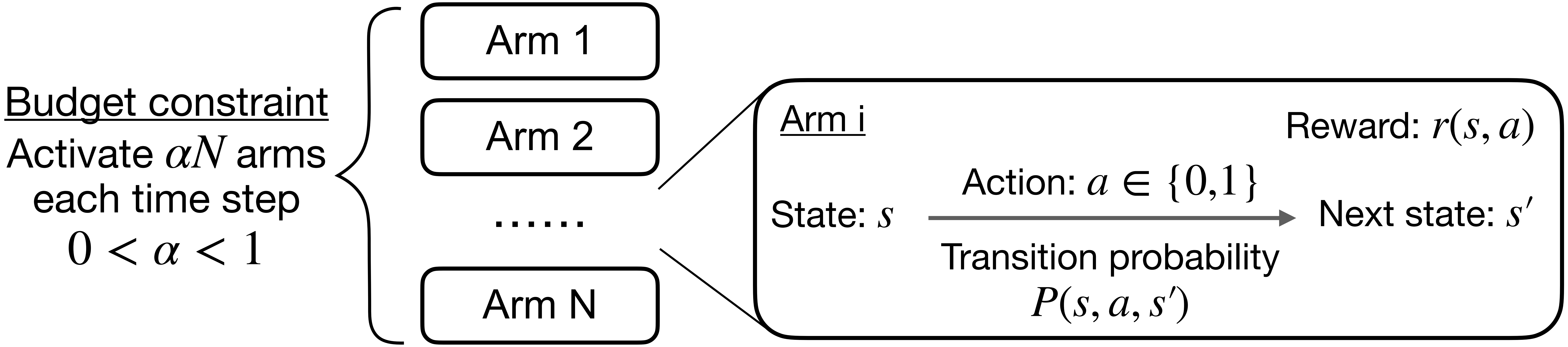
Restless bandits: problem definition



Restless bandits: problem definition

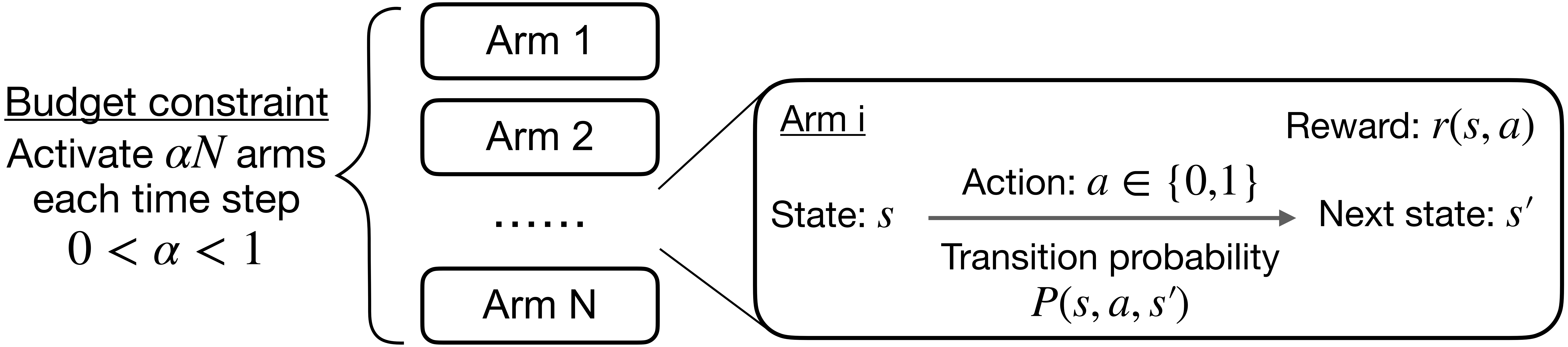


Restless bandits: problem definition



Objective: $\max_{\pi} R_N(\pi) \triangleq$ long-run average reward per time step and per arm

Restless bandits: problem definition



Objective: $\max_{\pi} R_N(\pi) \triangleq$ long-run average reward per time step and per arm

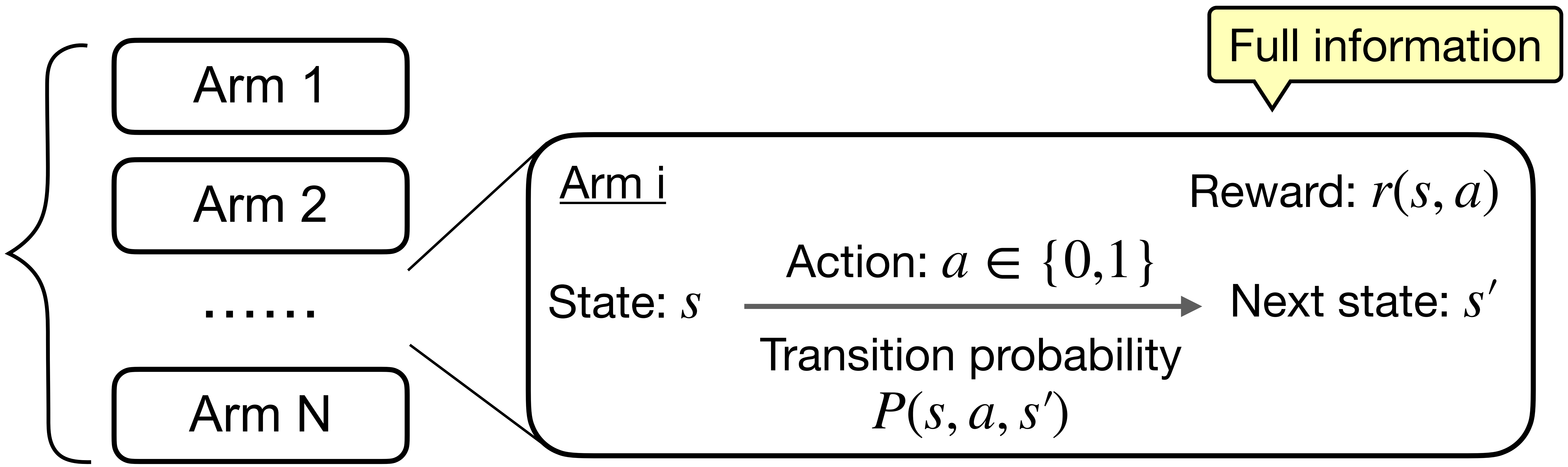
Policy π can see all states

Restless bandits: problem definition

Budget constraint

Activate αN arms each time step

$0 < \alpha < 1$

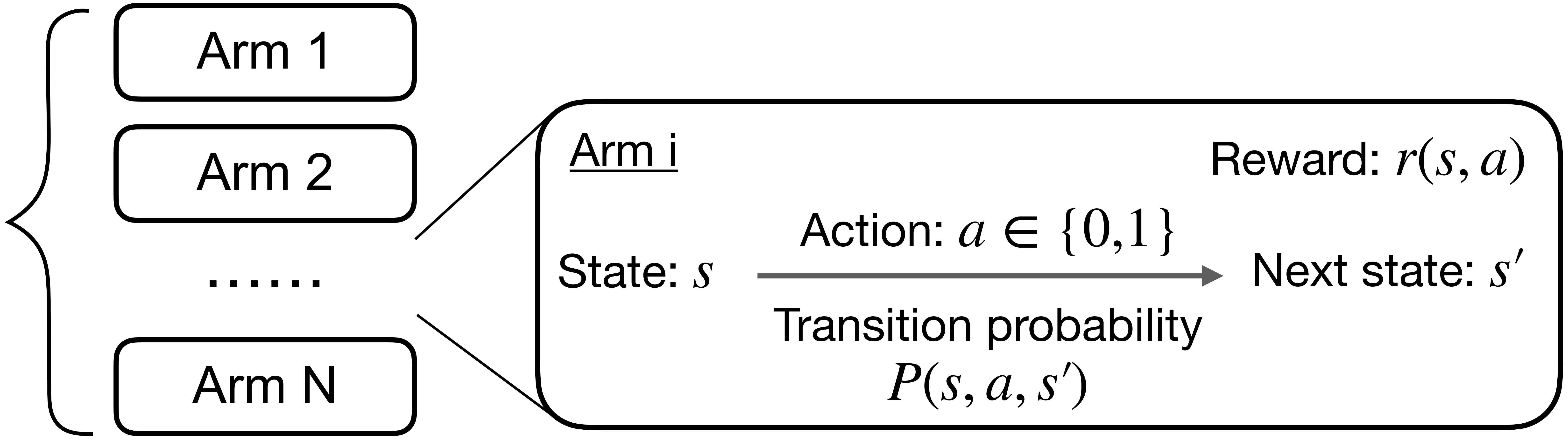


Objective: $\max_{\pi} R_N(\pi) \triangleq$ long-run average reward per time step and per arm

Policy π can see all states

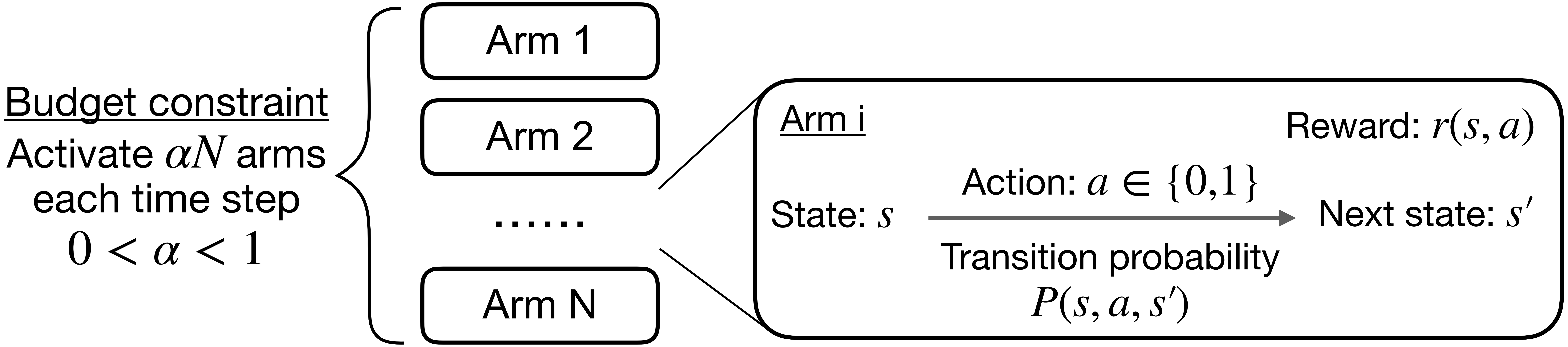
Restless bandits: asymptotic optimality

Budget constraint
Activate αN arms each time step
$0 < \alpha < 1$



Objective: $\max_{\pi} R_N(\pi) \triangleq$ long-run average reward per time step and per arm

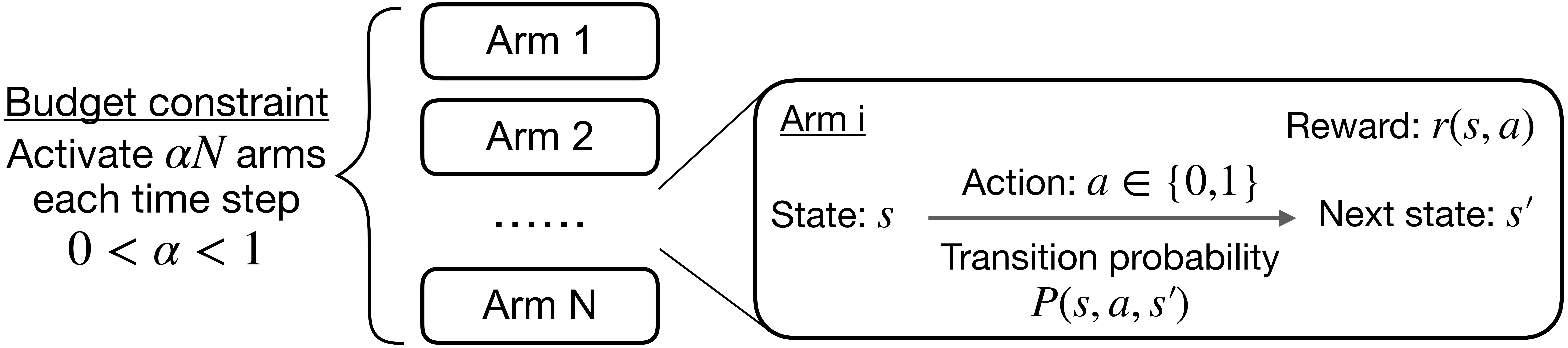
Restless bandits: asymptotic optimality



Objective: $\max_{\pi} R_N(\pi) \triangleq$ long-run average reward per time step and per arm

- Huge joint state space, when N is large; finding exact optimal policy is in general intractable

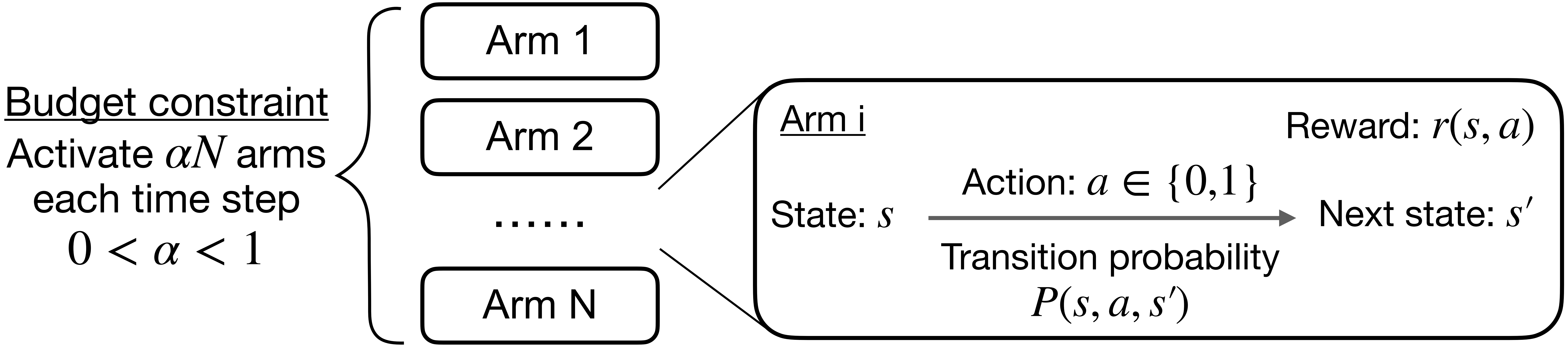
Restless bandits: asymptotic optimality



Objective: $\max_{\pi} R_N(\pi) \triangleq$ long-run average reward per time step and per arm

- Huge joint state space, when N is large; finding exact optimal policy is in general intractable
- Goal: find π s.t. $R_N^* - R_N(\pi) \to 0$ as $N \to \infty$

Restless bandits: asymptotic optimality



Objective: $\max_{\pi} R_N(\pi) \triangleq$ long-run average reward per time step and per arm

- Huge joint state space, when N is large; finding exact optimal policy is in general intractable

- Goal: find π s.t. $\underbrace{R_N^* - R_N(\pi)}_{\text{Optimality gap}} \to 0$ as $N \to \infty$

Prior work

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate

* aperiodic, unichain conditions

Prior work

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate

All require GAP or UGAP to be asymptotically optimal

* aperiodic, unichain conditions

Mystery about global behavior of restless bandits

Mystery about global behavior of restless bandits

GAP: empirical state distribution $\approx \mu^*$ in steady state

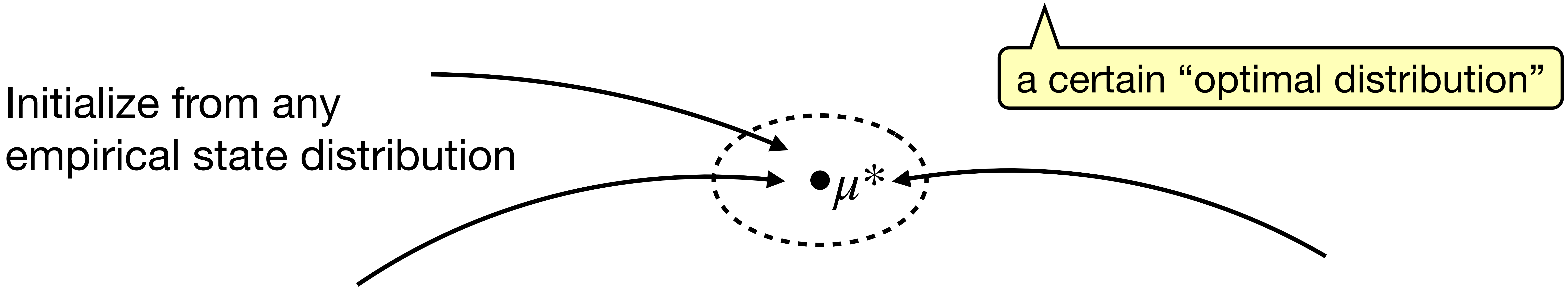
Mystery about global behavior of restless bandits

GAP: empirical state distribution $\approx \mu^*$ in steady state

a certain "optimal distribution"

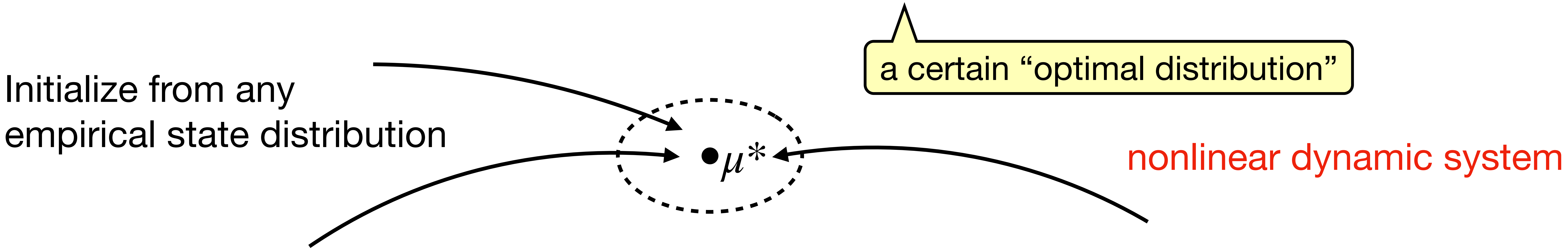
Mystery about global behavior of restless bandits

GAP: empirical state distribution $\approx \mu^*$ in steady state



Mystery about global behavior of restless bandits

GAP: empirical state distribution $\approx \mu^*$ in steady state

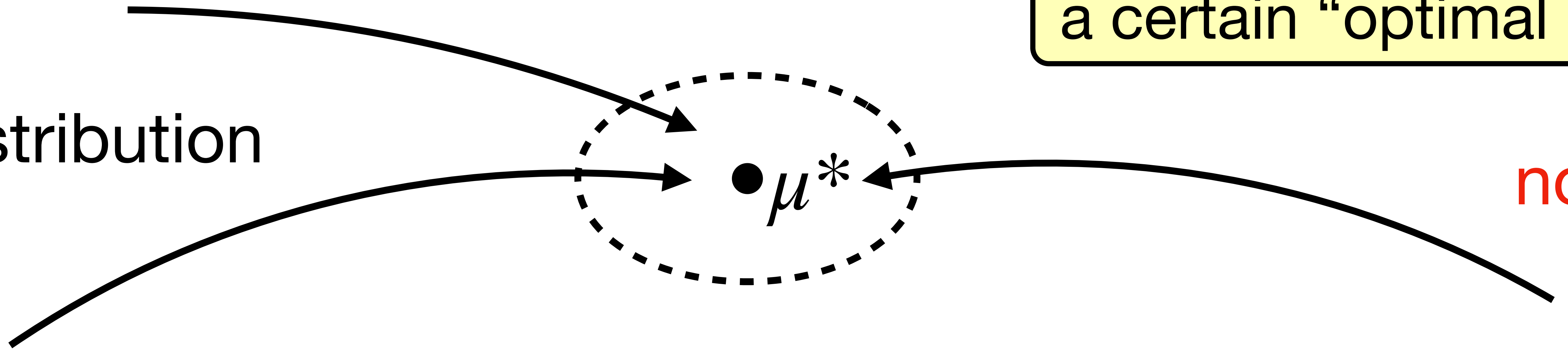


Mystery about global behavior of restless bandits

GAP: empirical state distribution $\approx \mu^*$ in steady state

a certain "optimal distribution"

Initialize from any empirical state distribution



nonlinear dynamic system

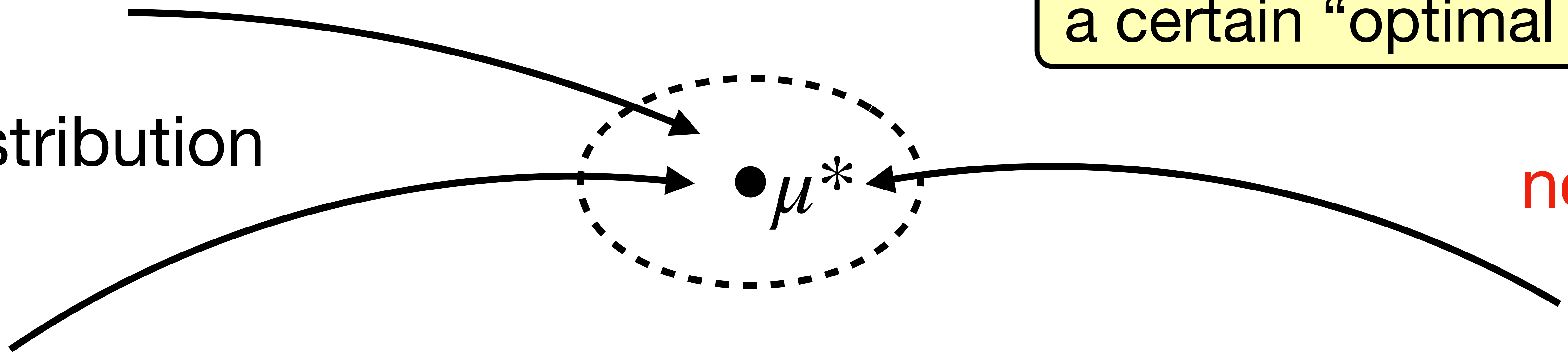
- Previous policies assuming GAP do not inherently guarantee the behavior outside the neighborhood of μ^*; in particular, global convergence may not hold

Mystery about global behavior of restless bandits

GAP: empirical state distribution $\approx \mu^*$ in steady state

a certain "optimal distribution"

Initialize from any empirical state distribution



nonlinear dynamic system

- Previous policies assuming GAP do not inherently guarantee the behavior outside the neighborhood of μ^*; in particular, global convergence may not hold

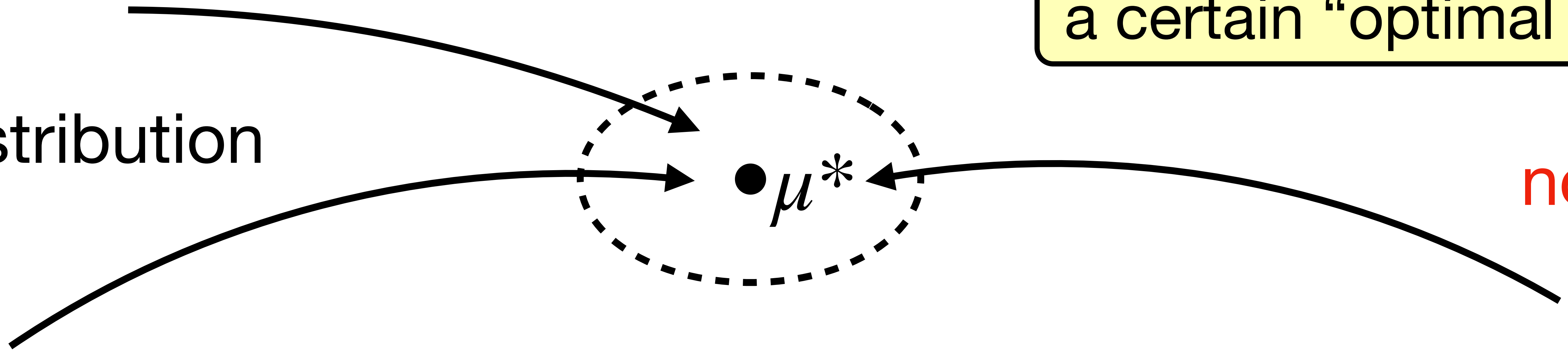
- How should one control the empirical state distribution when far away from μ^*?

Mystery about global behavior of restless bandits

GAP: empirical state distribution $\approx \mu^*$ in steady state

a certain "optimal distribution"

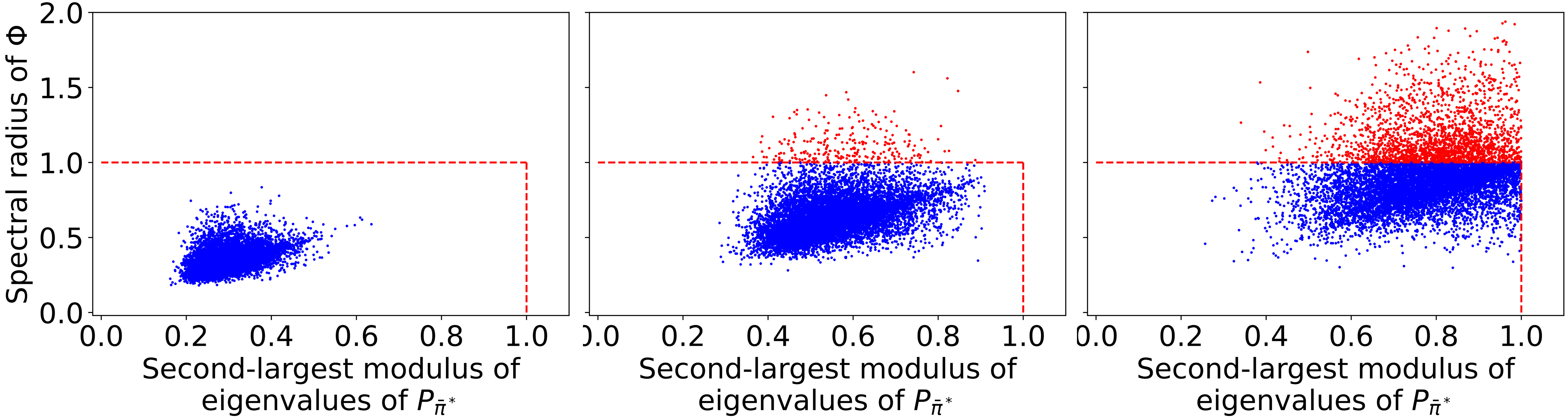
Initialize from any
empirical state distribution



nonlinear dynamic system

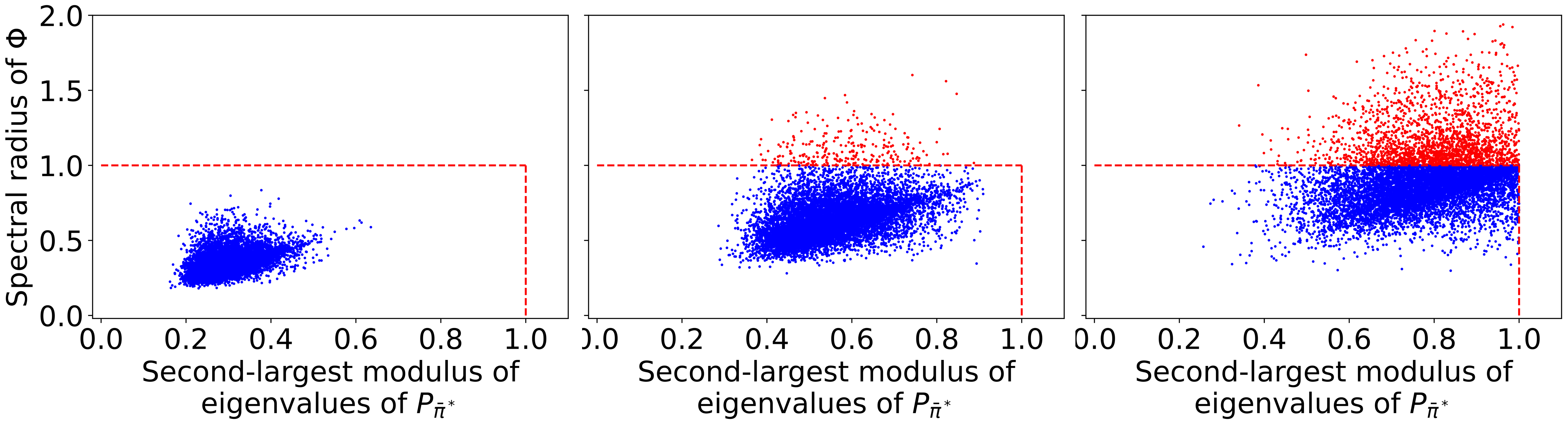
- Previous policies assuming GAP do not inherently guarantee the behavior outside the neighborhood of μ^*; in particular, global convergence may not hold
- How should one control the empirical state distribution when far away from μ^*?
- How complicated does a "globally convergent" policy need to be in the system with lots of "weakly-coupled" components?

How frequently does GAP fail?



Uniform distribution (=Dirichlet(1)) | Dirichlet(0.2) | Dirichlet(0.05)

How frequently does GAP fail?



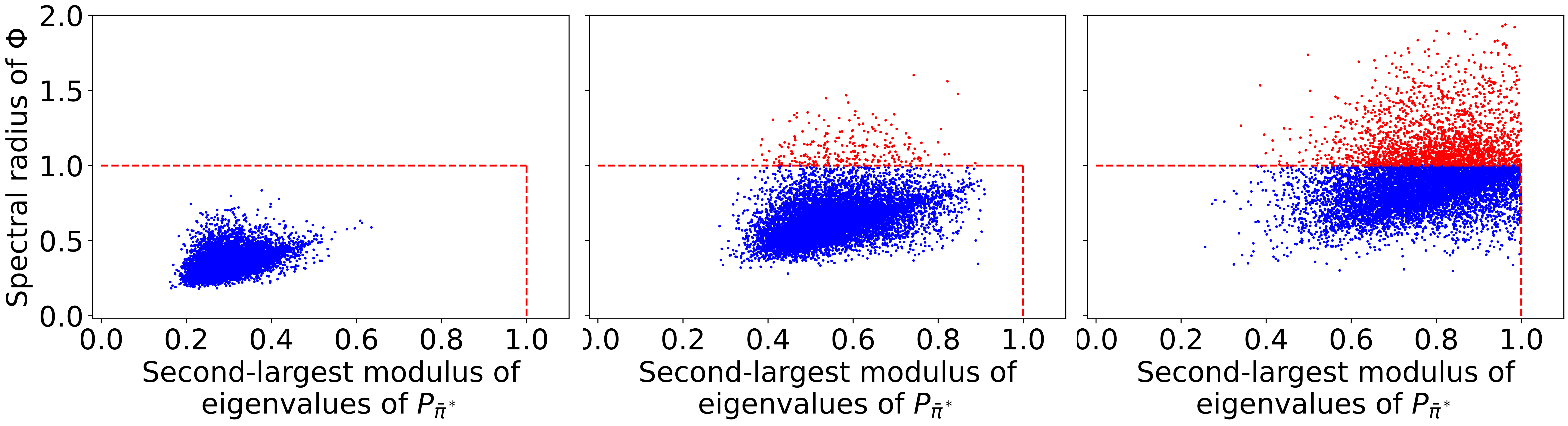
Uniform distribution (=Dirichlet(1))

Dirichlet(0.2)

Dirichlet(0.05)

- Each dot represents a random instance; red dots are non-GAP

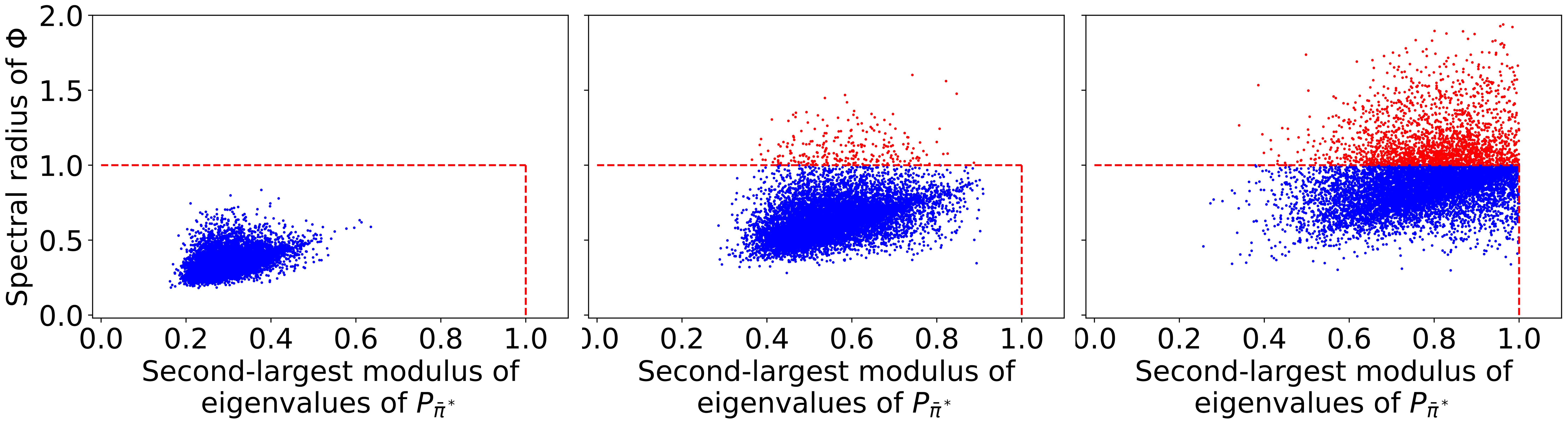
How frequently does GAP fail?



Uniform distribution (=Dirichlet(1)) Dirichlet(0.2) Dirichlet(0.05)

- Each dot represents a random instance; red dots are non-GAP
- More non-UGAP instances when transition matrix is sparse

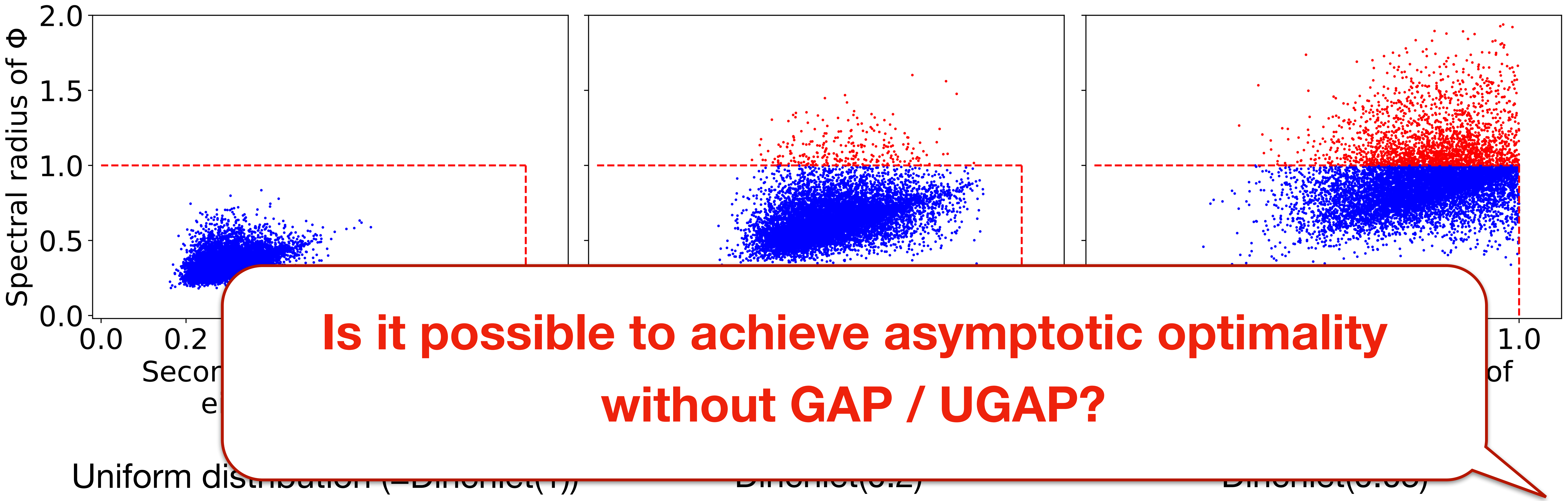
How frequently does GAP fail?



Uniform distribution (=Dirichlet(1)) Dirichlet(0.2) Dirichlet(0.05)

- Each dot represents a random instance; red dots are non-GAP
- More non-UGAP instances when transition matrix is sparse
- At least 20.2% instances violate GAP in Dirichlet(0.05), under any index (priority) policy.

How frequently does GAP fail?



Is it possible to achieve asymptotic optimality without GAP / UGAP?

- Each dot represents a random instance; red dots are non-GAP
- More non-UGAP instances when transition matrix is sparse
- At least 20.2% instances violate GAP in Dirichlet(0.05), under any index (priority) policy.

Our results

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate

* aperiodic, unichain conditions

Our results

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
Hong et al. 23	**FTVA**	$O(1/\sqrt{N})$	**Synchronization Assumption**

* aperiodic, unichain conditions

Our results

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
Hong et al. 23	FTVA	$O(1/\sqrt{N})$	Synchronization Assumption
Hong et al. 24a	Focus-set	$O(1/\sqrt{N})$	—

* aperiodic, unichain conditions

Our results

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
Hong et al. 23	**FTVA**	$O(1/\sqrt{N})$	**Synchronization Assumption**
Hong et al. 24a	**Focus-set**	$O(1/\sqrt{N})$	—
Hong et al. 24b	**Two-Set**	$O(e^{-cN})$	**Local stability** & Non-degenerate

* aperiodic, unichain conditions

Our results

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
Hong et al. 23	**FTVA**	$O(1/\sqrt{N})$	**Synchronization Assumption**
Hong et al. 24a	**Focus-set**	$O(1/\sqrt{N})$	**—**
Hong et al. 24b	**Two-Set**	$O(e^{-cN})$	**Local stability** & Non-degenerate

[Yan 24] [Goldsztajn and Avrachenkov 24]: further relaxes unichain

* aperiodic, unichain conditions

Our results

GAP = Global Attractor Property
UGAP = Uniform Global Attractor Property

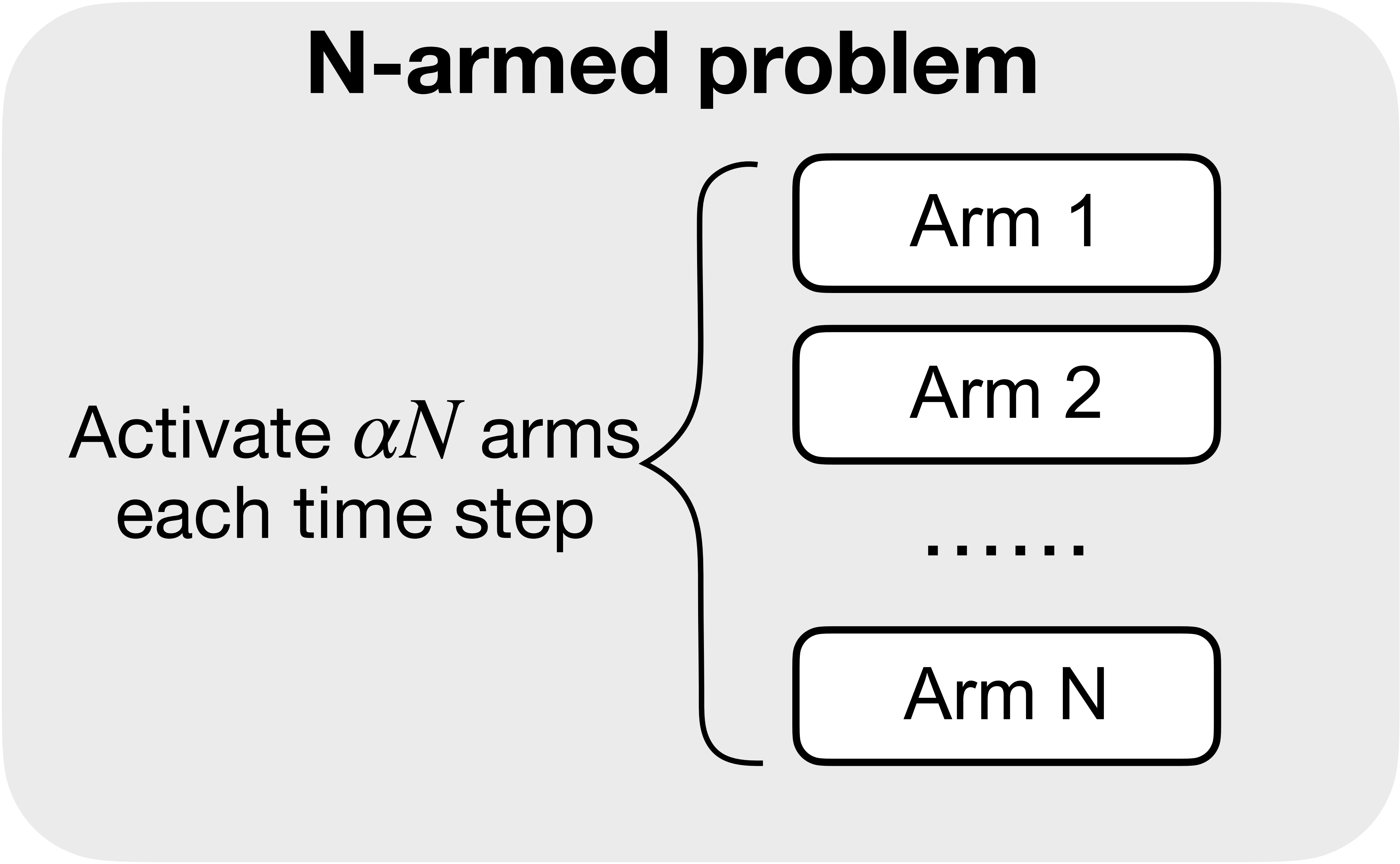
Paper	Policy	Optimality Gap	Conditions*
Weber and Weiss 90	Whittle Index	$o(1)$	Indexable & GAP
Verloop 16	LP-Priority	$o(1)$	GAP
Gast, Gaujal, and Yan 23a	Whittle Index	$O\left(e^{-cN}\right)$	Indexable, UGAP, Non-singular
Gast, Gaujal, and Yan 23b	LP-Priority	$O\left(e^{-cN}\right)$	UGAP & Non-degenerate
Hong et al. 23	FTVA	$O(1/\sqrt{N})$	Synchronization Assumption
Hong et al. 24a	Focus-set	$O(1/\sqrt{N})$	—
Hong et al. 24b	Two-Set	$O(e^{-cN})$	Local stability & Non-degenerate

rest of the talk (→ Hong et al. 24a)

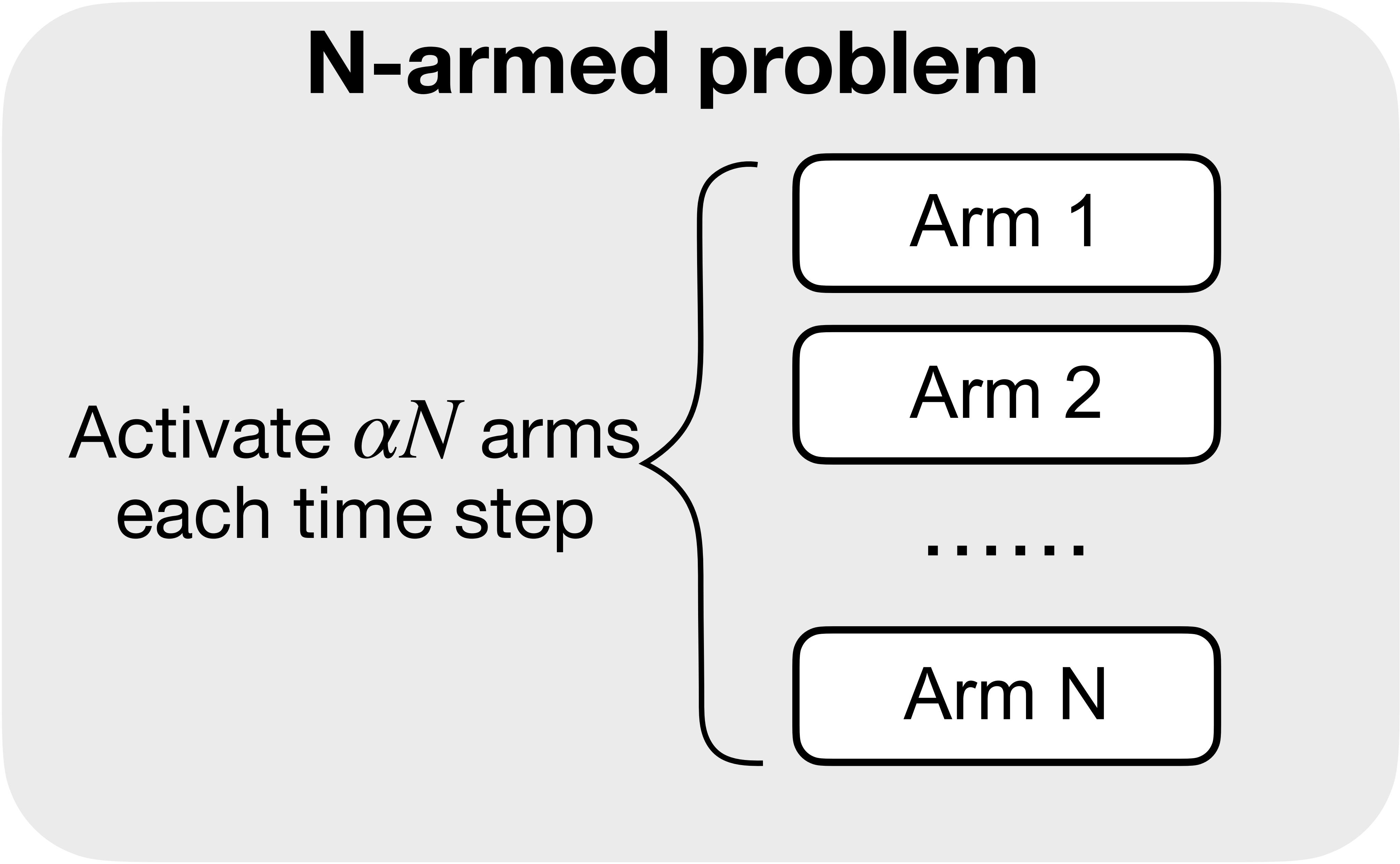
[Yan 24] [Goldsztajn and Avrachenkov 24]: further relaxes unichain

* aperiodic, unichain conditions

Preliminary: single-armed relaxation

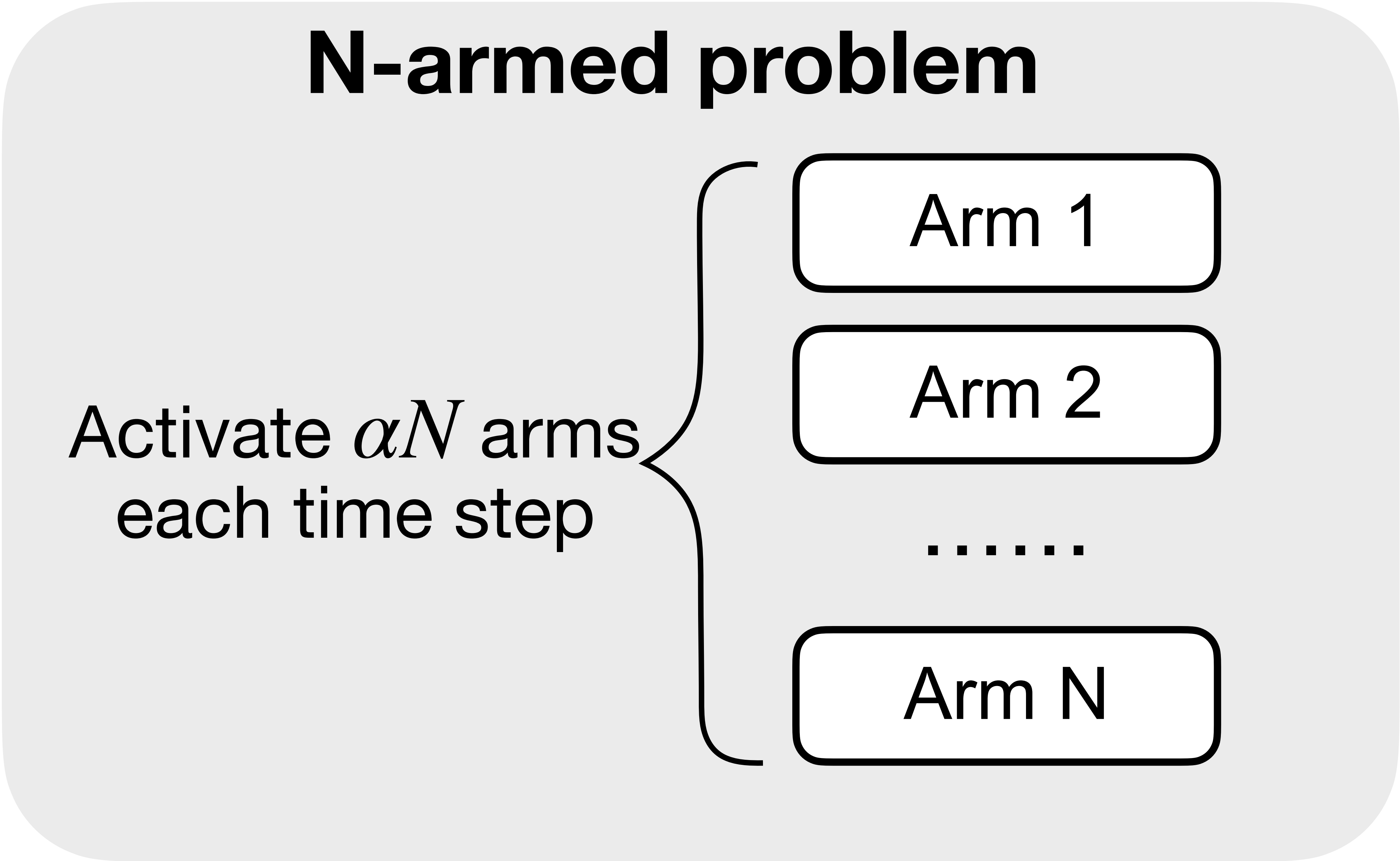


Preliminary: single-armed relaxation



$$R_N(\pi) \leq R_N^*$$

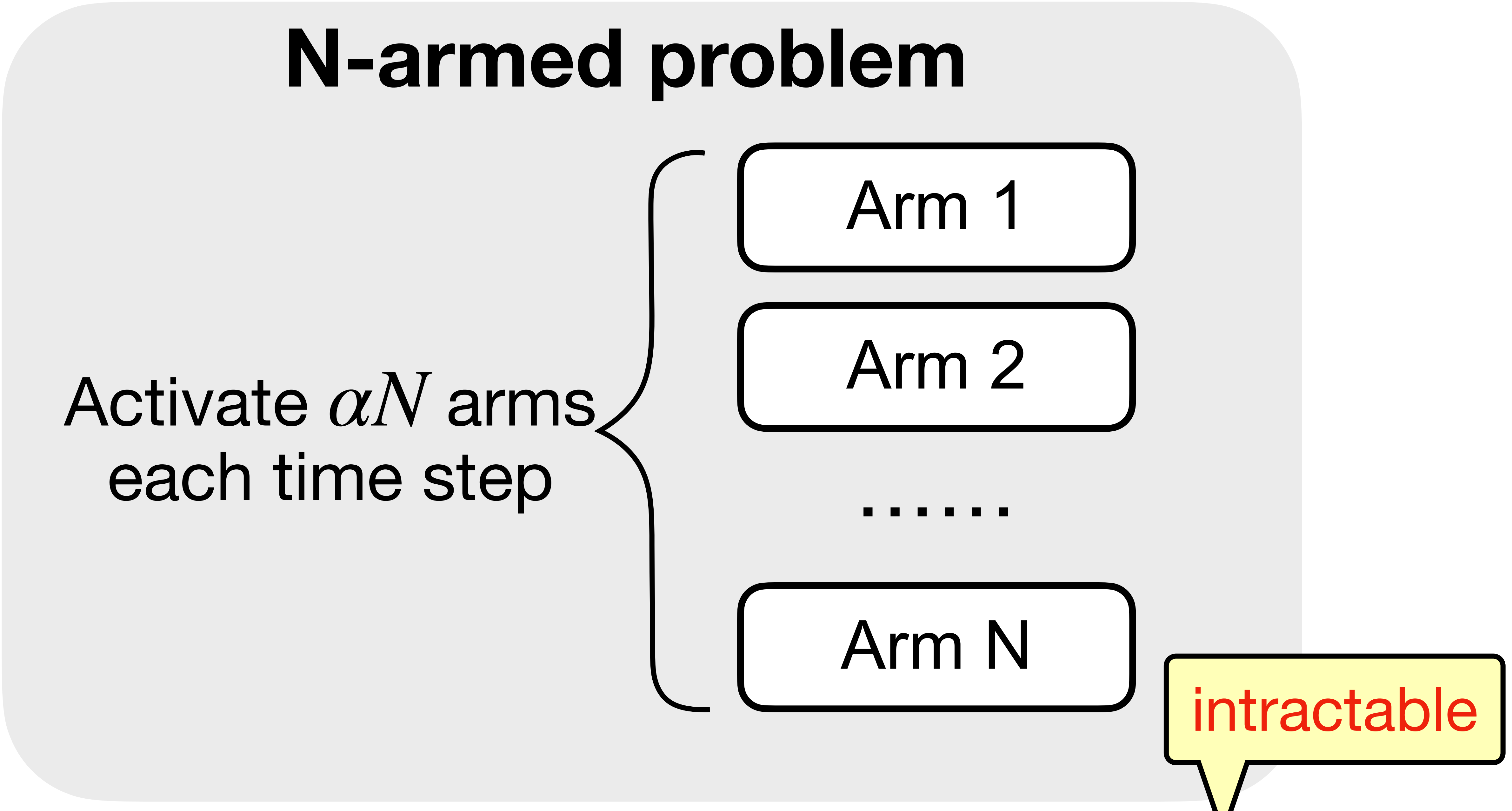
Preliminary: single-armed relaxation



$$R_N(\pi) \quad \leq \quad R_N^*$$

Optimality gap

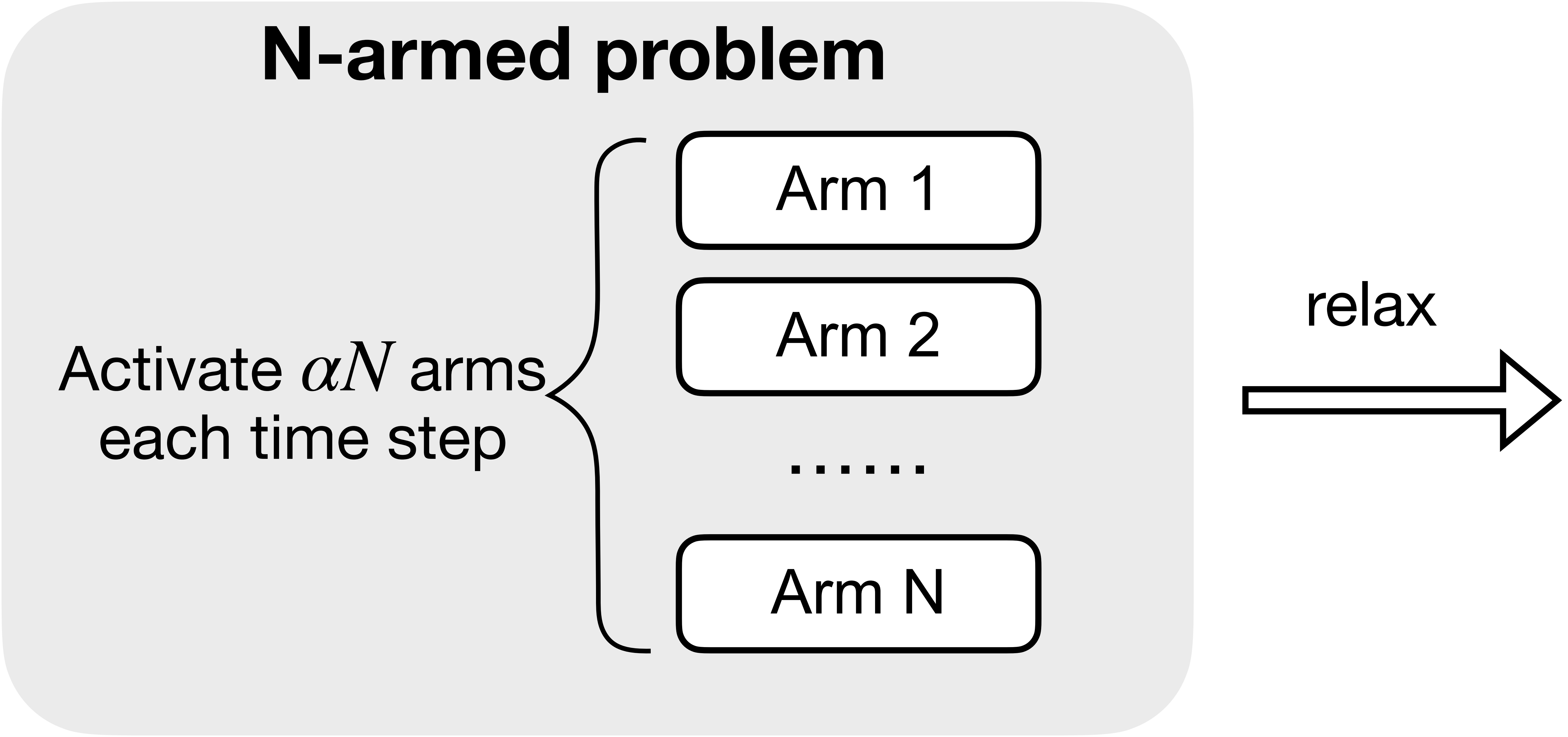
Preliminary: single-armed relaxation



$$R_N(\pi) \leq R_N^*$$

Optimality gap

Preliminary: single-armed relaxation



$R_N(\pi) \quad \leq \quad R_N^*$

Optimality gap

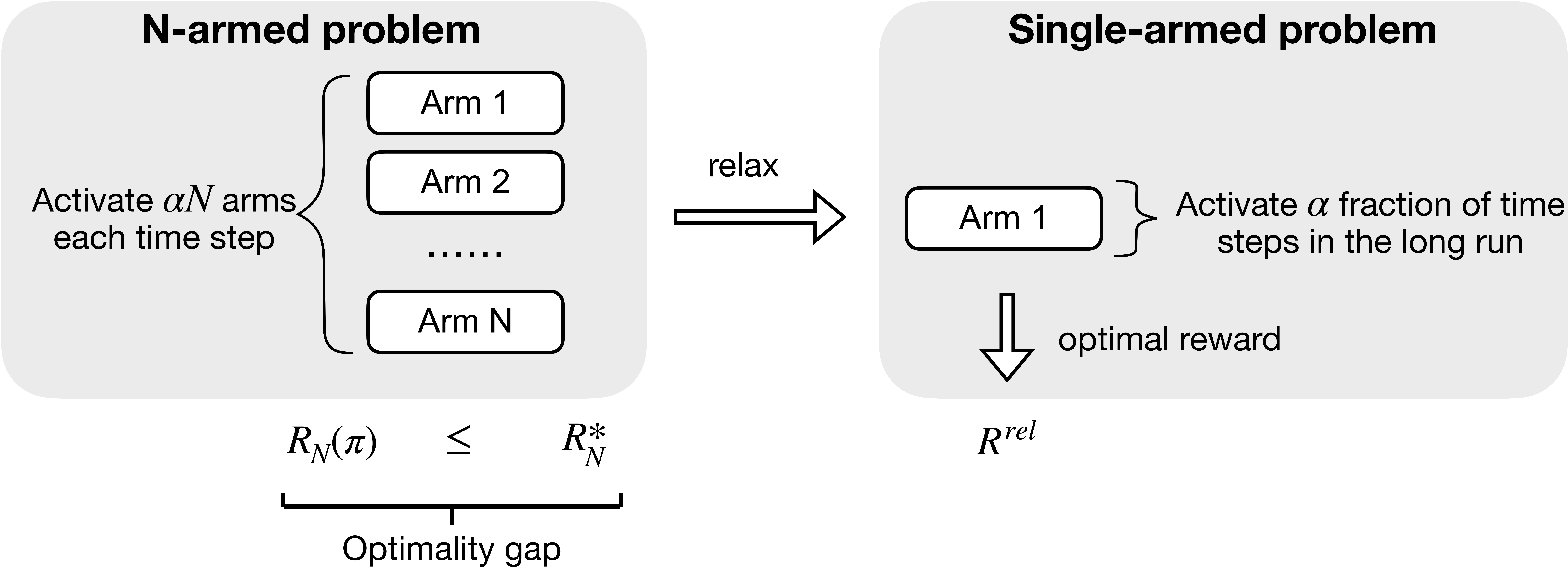
Preliminary: single-armed relaxation

$$R_N(\pi) \quad \leq \quad R_N^*$$

Optimality gap

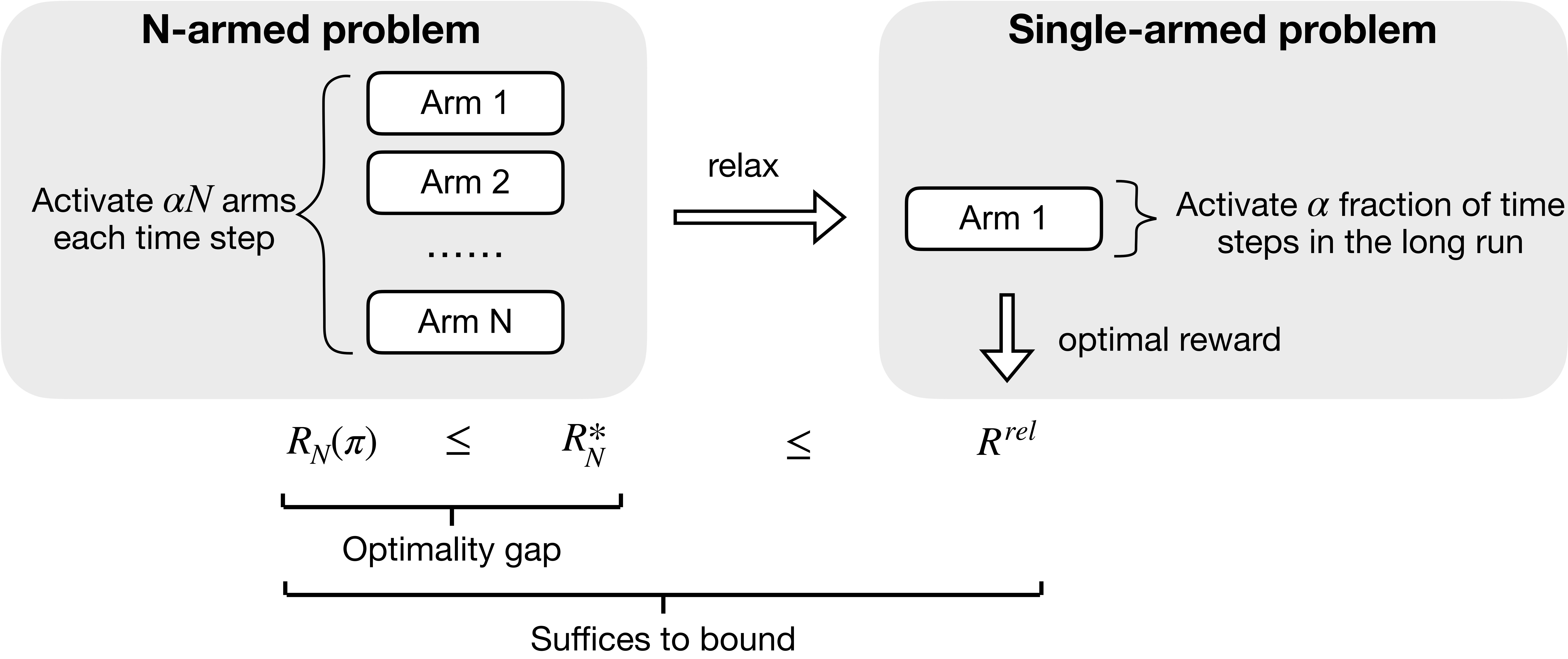
Preliminary: single-armed relaxation

Preliminary: single-armed relaxation



Preliminary: single-armed relaxation

Preliminary: single-armed relaxation



When is $R_N(\pi)$ close to R^{rel}?

When is $R_N(\pi)$ close to R^{rel}?

$Y^{\pi}(s, a)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π $\quad Y^{\pi}(s, a)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$Y^\pi(s, a)$

$y^*(s, a)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$Y^\pi(s, a)$

$y^*(s, a)$

Optimal state-action frequency

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$Y^{\pi}(s, a)$

$y^*(s, a)$

Optimal state-action frequency

$R_N(\pi) = \sum_{s,a} r(s, a) Y^{\pi}(s, a)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π — $Y^\pi(s,a)$

$y^*(s,a)$ — Optimal state-action frequency

$$R_N(\pi) = \sum_{s,a} r(s,a) Y^\pi(s,a)$$

$$R^{rel} = \sum_{s,a} r(s,a) y^*(s,a)$$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π — $Y^\pi(s, a)$ $\approx$ $y^*(s, a)$ — Optimal state-action frequency

$R_N(\pi) = \sum_{s,a} r(s, a) Y^\pi(s, a)$

$R^{rel} = \sum_{s,a} r(s, a) y^*(s, a)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π $\quad Y^{\pi}(s,a) \quad \approx \quad y^*(s,a) \quad$ Optimal state-action frequency

$$\Downarrow$$

$$R_N(\pi) = \sum_{s,a} r(s,a) Y^{\pi}(s,a) \quad \approx \quad R^{rel} = \sum_{s,a} r(s,a) y^*(s,a)$$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$$Y^\pi(s,a) \approx y^*(s,a)$$

Optimal state-action frequency

$$\Downarrow$$

$$R_N(\pi) = \sum_{s,a} r(s,a)Y^\pi(s,a) \approx R^{rel} = \sum_{s,a} r(s,a)y^*(s,a)$$

What does this requirement mean for designing a policy π?

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$$Y^\pi(s,a) \approx y^*(s,a)$$

Optimal state-action frequency

$$\Downarrow$$

$$R_N(\pi) = \sum_{s,a} r(s,a) Y^\pi(s,a) \approx R^{rel} = \sum_{s,a} r(s,a) y^*(s,a)$$

What does this requirement mean for designing a policy π?

In the steady state, under π, there should be:

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$Y^\pi(s,a) \approx y^*(s,a)$

Optimal state-action frequency

⇓

$$R_N(\pi) = \sum_{s,a} r(s,a) Y^\pi(s,a) \approx R^{rel} = \sum_{s,a} r(s,a) y^*(s,a)$$

What does this requirement mean for designing a policy π?

In the steady state, under π, there should be:

- **Empirical state distribution** $X^\pi(s) \triangleq \sum_a Y^\pi(s,a) \approx \mu^*(s) \triangleq \sum_a y^*(s,a)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$$Y^\pi(s,a) \approx y^*(s,a)$$

Optimal state-action frequency

$$\Downarrow$$

$$R_N(\pi) = \sum_{s,a} r(s,a) Y^\pi(s,a) \quad \approx \quad R^{rel} = \sum_{s,a} r(s,a) y^*(s,a)$$

What does this requirement mean for designing a policy π?

In the steady state, under π, there should be:

"Optimal stationary distribution"

- **Empirical state distribution** $X^\pi(s) \triangleq \sum_a Y^\pi(s,a) \approx \mu^*(s) \triangleq \sum_a y^*(s,a)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$$Y^\pi(s,a) \approx y^*(s,a)$$

Optimal state-action frequency

$$\Downarrow$$

$$R_N(\pi) = \sum_{s,a} r(s,a) Y^\pi(s,a) \approx R^{rel} = \sum_{s,a} r(s,a) y^*(s,a)$$

What does this requirement mean for designing a policy π?

In the steady state, under π, there should be:

"Optimal stationary distribution"

- **Empirical state distribution** $X^\pi(s) \triangleq \sum_a Y^\pi(s,a) \approx \mu^*(s) \triangleq \sum_a y^*(s,a)$
- **Given an arm in state s,** prob. of action a approximates $\bar{\pi}^*(a \mid s) \triangleq y^*(s,a)/\mu^*(s)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π | $Y^\pi(s,a) \approx y^*(s,a)$ | Optimal state-action frequency

$\Downarrow$

$$R_N(\pi) = \sum_{s,a} r(s,a)Y^\pi(s,a) \approx R^{rel} = \sum_{s,a} r(s,a)y^*(s,a)$$

What does this requirement mean for designing a policy π?

In the steady state, under π, there should be:

- **Empirical state distribution** $X^\pi(s) \triangleq \sum_a Y^\pi(s,a) \approx \mu^*(s) \triangleq \sum_a y^*(s,a)$
- **Given an arm in state** s**,** prob. of action a approximates $\bar{\pi}^*(a \mid s) \triangleq y^*(s,a)/\mu^*(s)$

"Optimal stationary distribution" ($\mu^*(s)$)

"Optimal single-armed policy" ($\bar{\pi}^*(a \mid s)$)

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$$Y^\pi(s,a) \approx y^*(s,a)$$

Optimal state-action frequency

$$R_N(\pi) = \sum_{s,a} r(s,a)Y^\pi(s,a) \approx R^{rel} = \sum_{s,a} r(s,a)y^*(s,a)$$

What does this requirement mean for designing a policy π?

In the steady state, under π, there should be:

- **Empirical state distribution** $X^\pi(s) \triangleq \sum_a Y^\pi(s,a) \approx \mu^*(s) \triangleq \sum_a y^*(s,a)$
- **Given an arm in state s,** prob. of action a approximates $\bar{\pi}^*(a \mid s) \triangleq y^*(s,a)/\mu^*(s)$

When is $R_N(\pi)$ close to R^{rel}?

State-action frequency under policy π

$$Y^\pi(s,a) \quad \approx \quad y^*(s,a)$$

Optimal state-action frequency

$$\Downarrow$$

$$R_N(\pi) = \sum_{s,a} r(s,a) Y^\pi(s,a) \quad \approx \quad R^{rel} = \sum_{s,a} r(s,a) y^*(s,a)$$

What does this requirement mean for designing a policy π?

In the steady state, under π, there should be:

Global convergence

- **Empirical state distribution** $X^\pi(s) \triangleq \sum_a Y^\pi(s,a) \approx \mu^*(s) \triangleq \sum_a y^*(s,a)$
- **Given an arm in state s,** prob. of action a approximates $\bar{\pi}^*(a \mid s) \triangleq y^*(s,a)/\mu^*(s)$

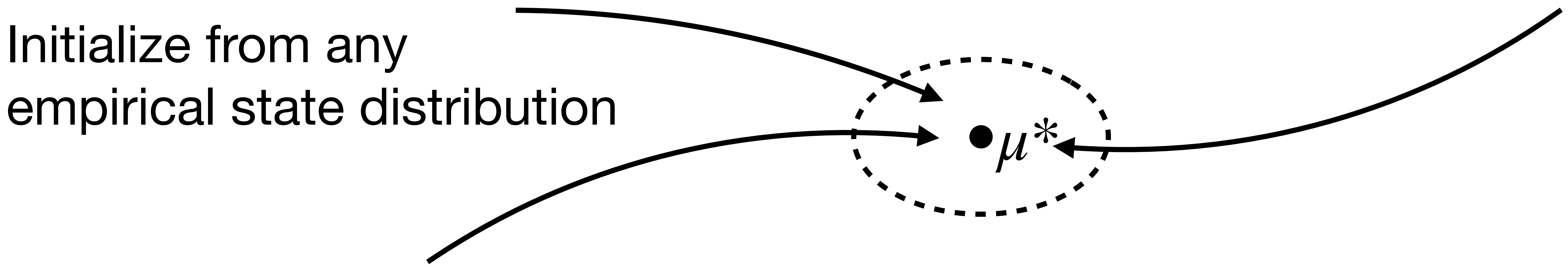
Challenge: global convergence

Challenge: global convergence

Requirement: empirical state distribution $X^\pi \approx \mu^*$ in steady state

Challenge: global convergence

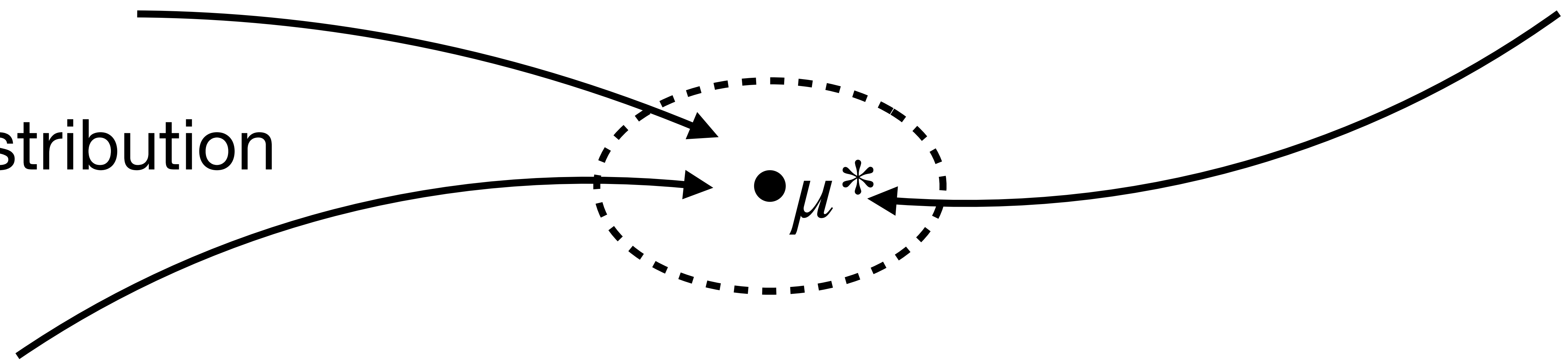
Requirement: empirical state distribution $X^\pi \approx \mu^*$ in steady state



Challenge: global convergence

Requirement: empirical state distribution $X^\pi \approx \mu^*$ in steady state

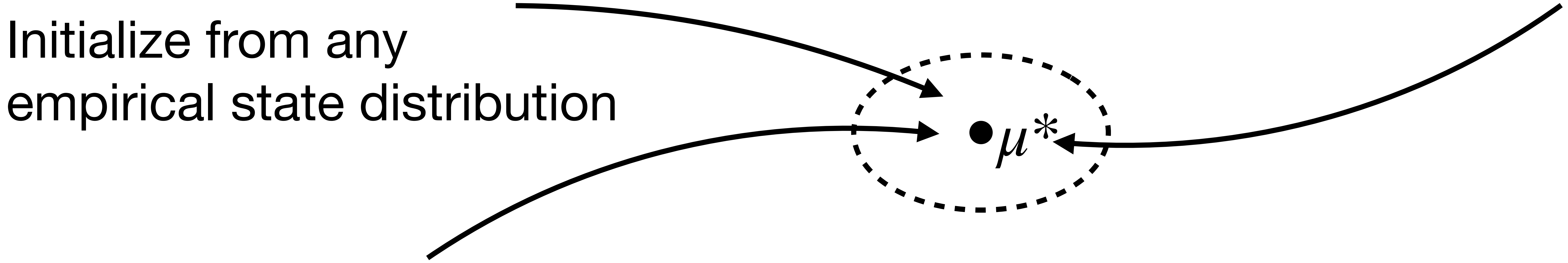
Initialize from any
empirical state distribution



Prior work fixes a policy, and has to *assume* global convergence

Challenge: global convergence

Requirement: empirical state distribution $X^\pi \approx \mu^*$ in steady state



Prior work fixes a policy, and has to *assume* global convergence

Does there exists a policy that achieves global convergence *on its own*?

Clue: convergence of single arm distribution

Clue: convergence of single arm distribution

A single arm under policy $\bar{\pi}^*$ is a *Markov chain* with stationary distribution μ^*

Clue: convergence of single arm distribution

A single arm under policy $\bar{\pi}^*$ is a *Markov chain* with stationary distribution μ^*

assume aperiodic and irreducibility

Clue: convergence of single arm distribution

A single arm under policy $\bar{\pi}^*$ is a *Markov chain* with stationary distribution μ^*

assume aperiodic and irreducibility

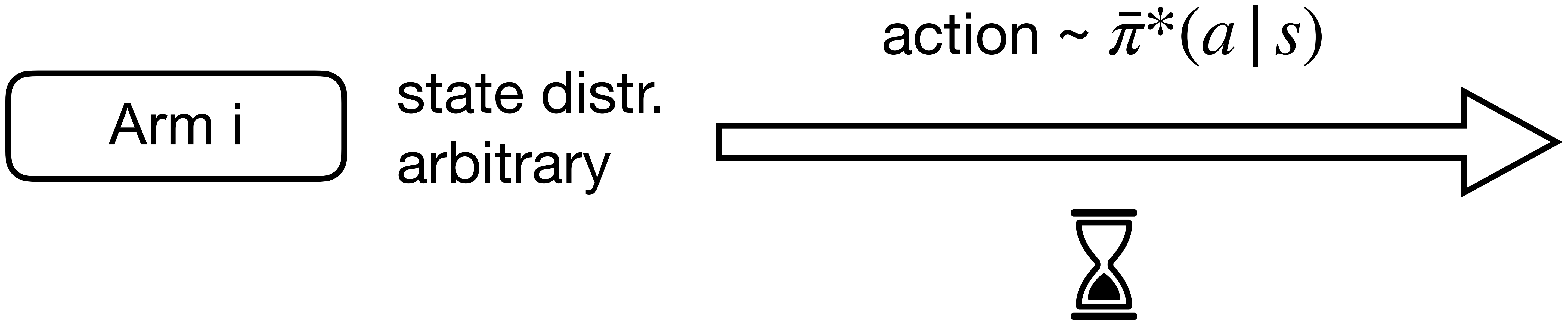
Arm i

state distr.
arbitrary

Clue: convergence of single arm distribution

A single arm under policy $\bar{\pi}^*$ is a *Markov chain* with stationary distribution μ^*

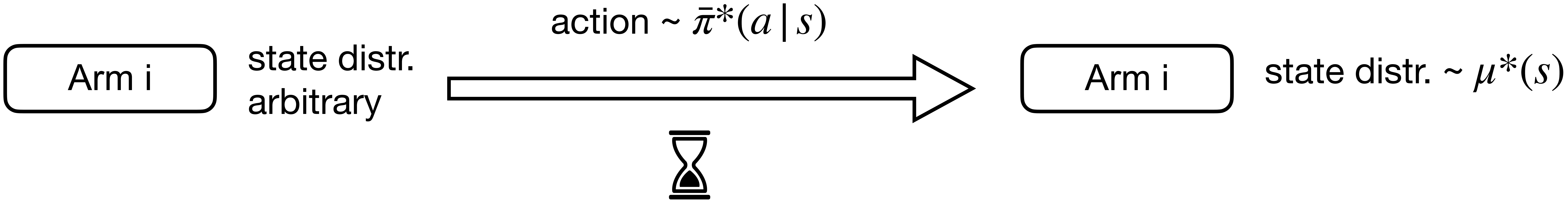
assume aperiodic and irreducibility



Clue: convergence of single arm distribution

A single arm under policy $\bar{\pi}^*$ is a *Markov chain* with stationary distribution μ^*

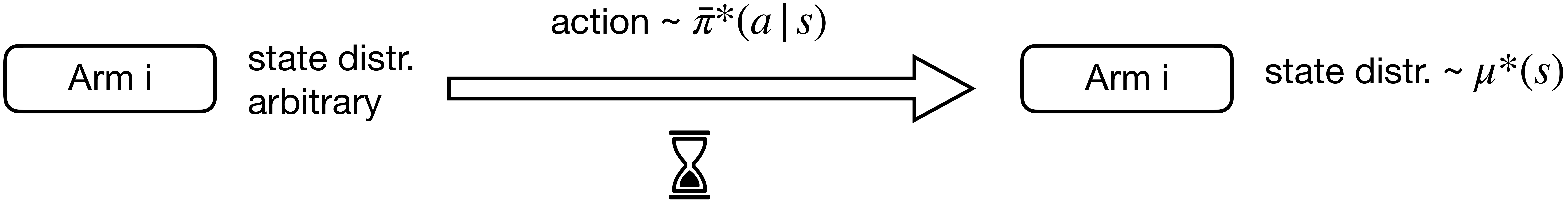
assume aperiodic and irreducibility



Clue: convergence of single arm distribution

A single arm under policy $\bar{\pi}^*$ is a *Markov chain* with stationary distribution μ^*

assume aperiodic and irreducibility

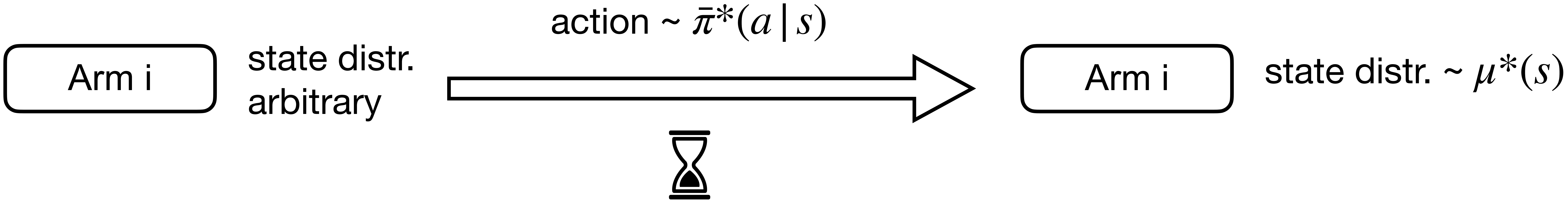


How much is controlling N “weakly coupled” arms harder than controlling one arm?

Clue: convergence of single arm distribution

A single arm under policy $\bar{\pi}^*$ is a *Markov chain* with stationary distribution μ^*

assume aperiodic and irreducibility

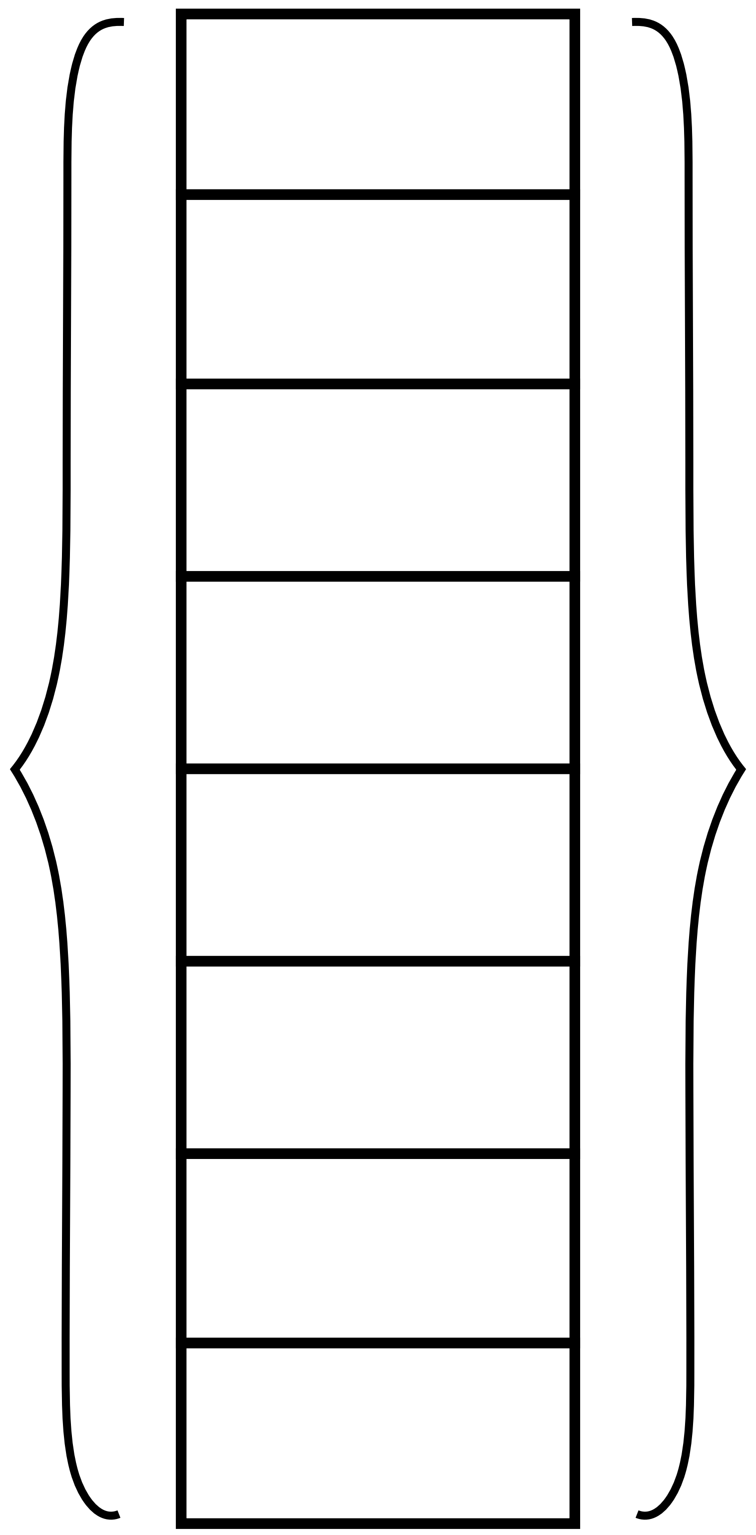


How much is controlling N “weakly coupled” arms harder than controlling one arm?

Can we utilize $\bar{\pi}^*$ to drive the state distr. of each arm to μ^*?

First attempt

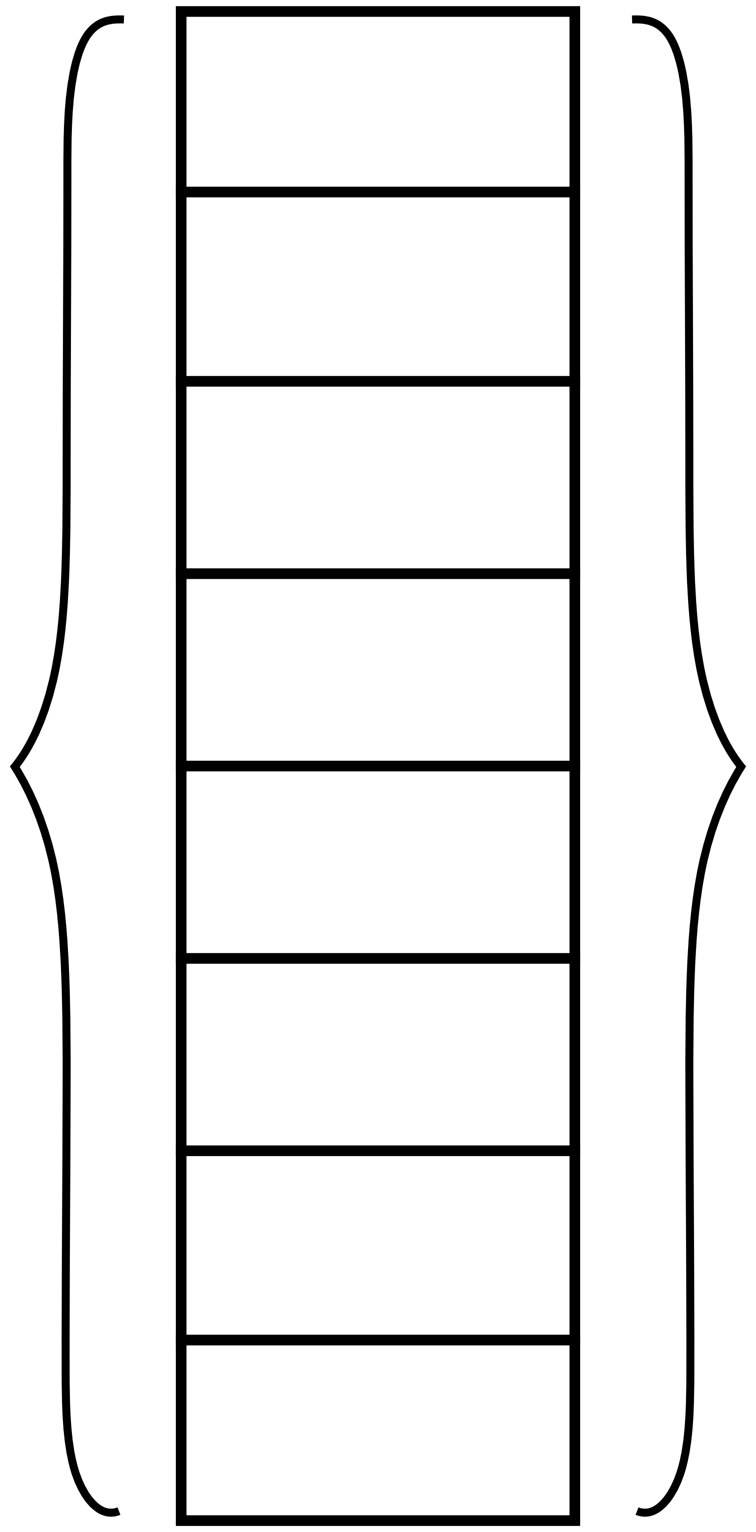
$$\alpha = \frac{1}{2}$$



First attempt

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

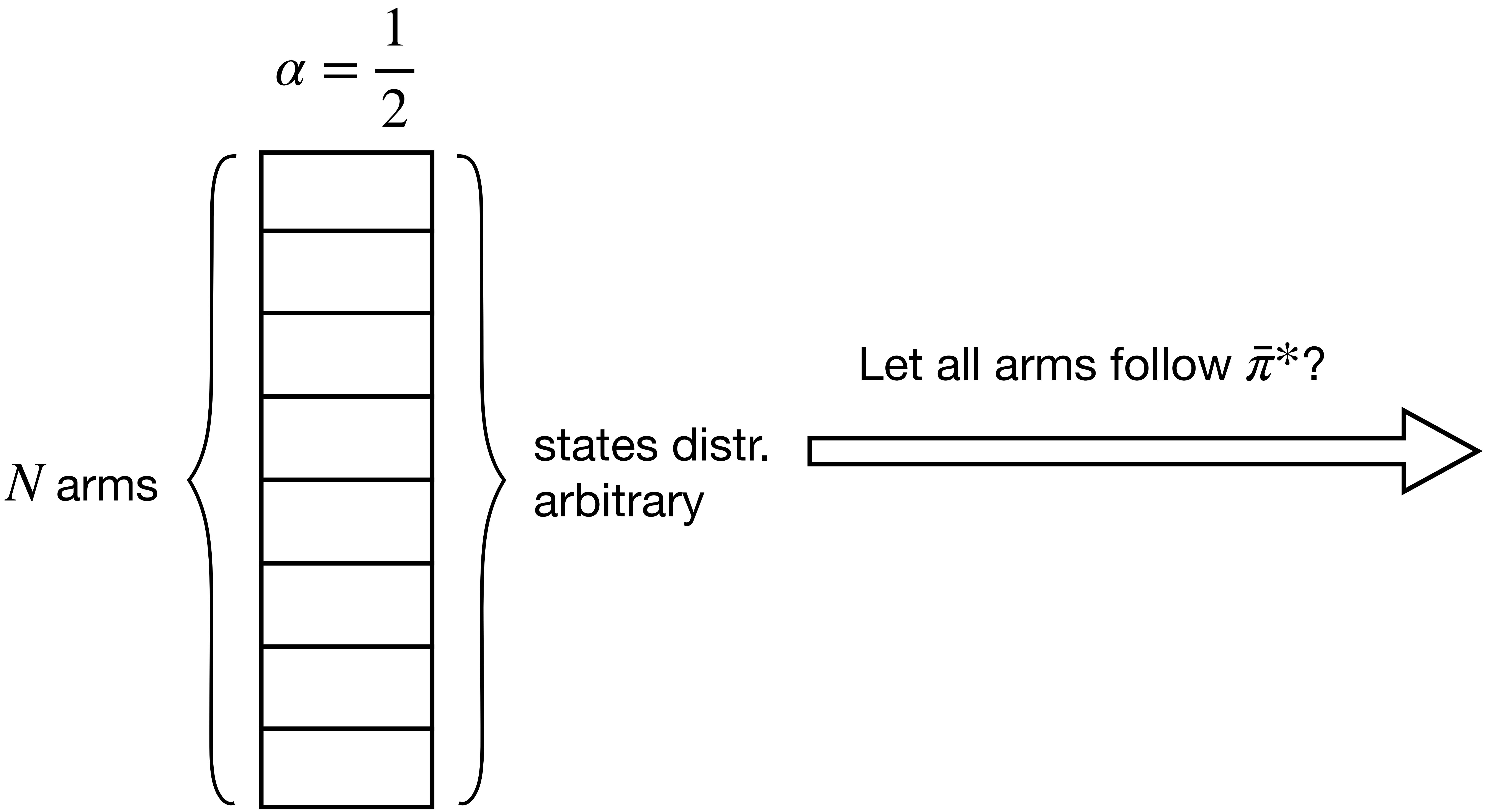
N arms



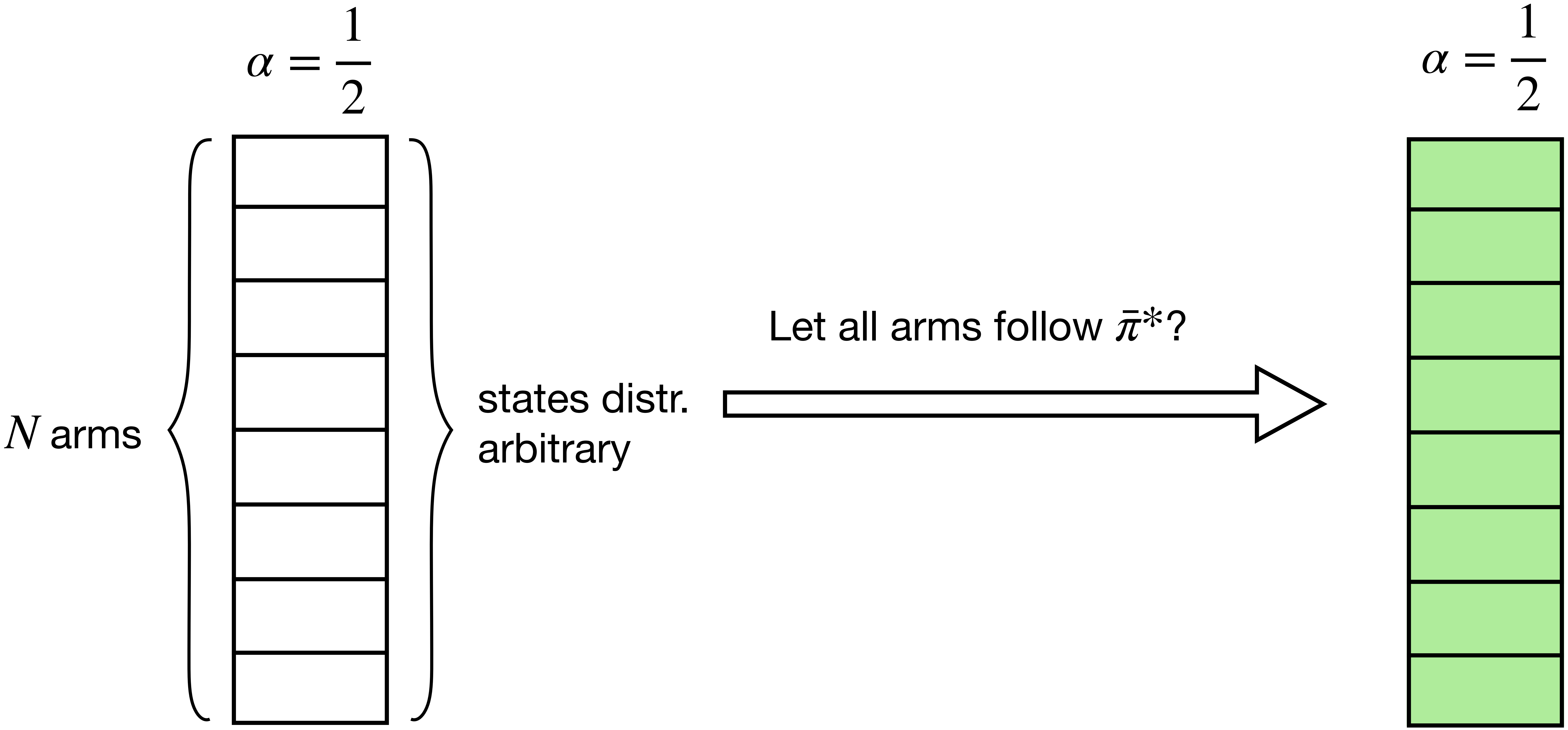
states distr. arbitrary

Let all arms follow $\bar{\pi}^*$?

First attempt



First attempt

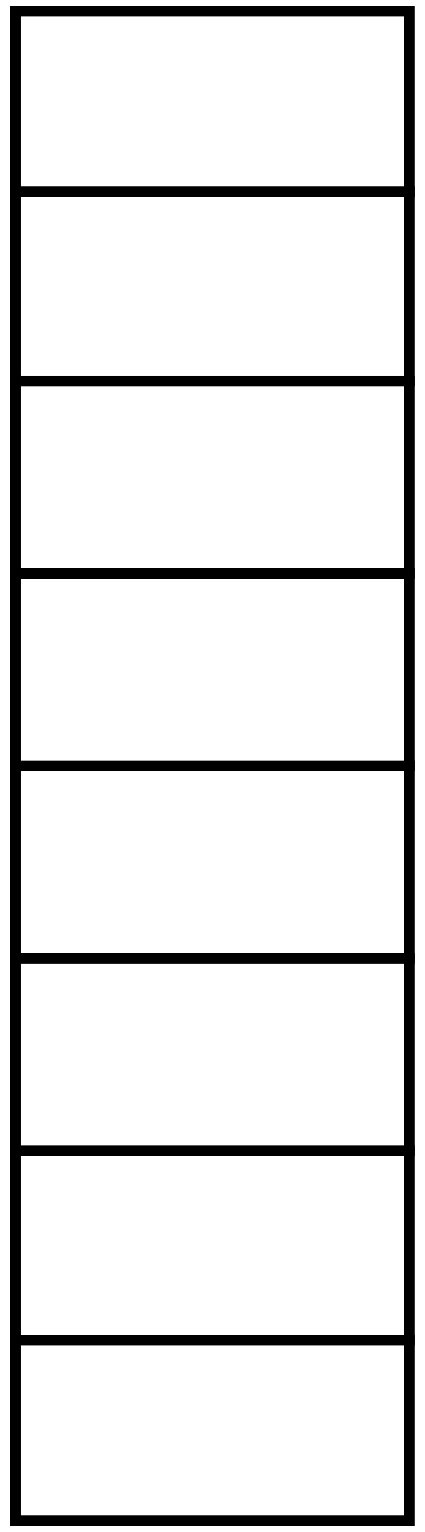


First attempt

First attempt

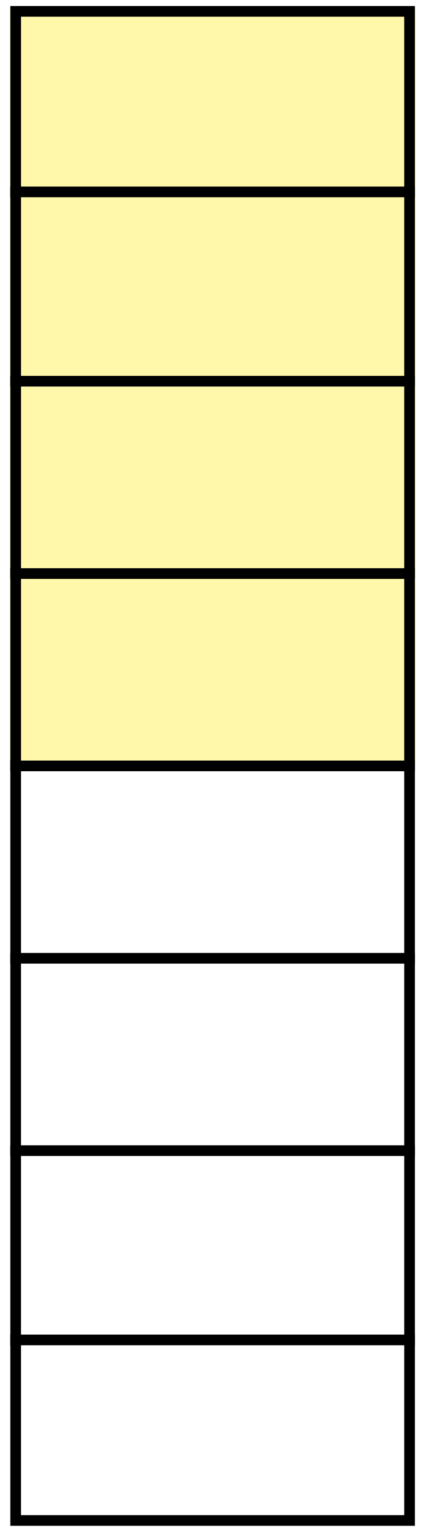
Intuition: start from a subset

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



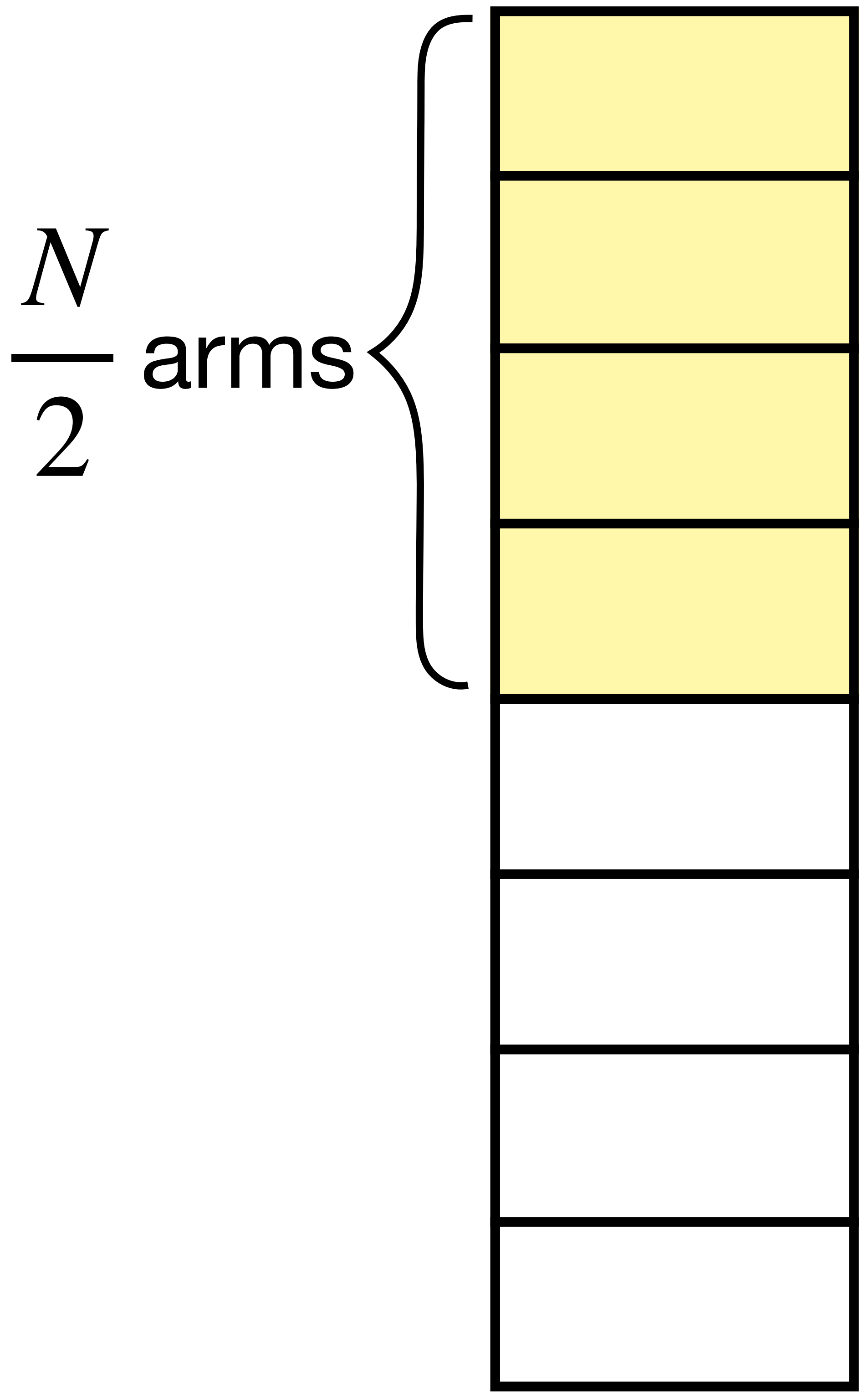
Intuition: start from a subset

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



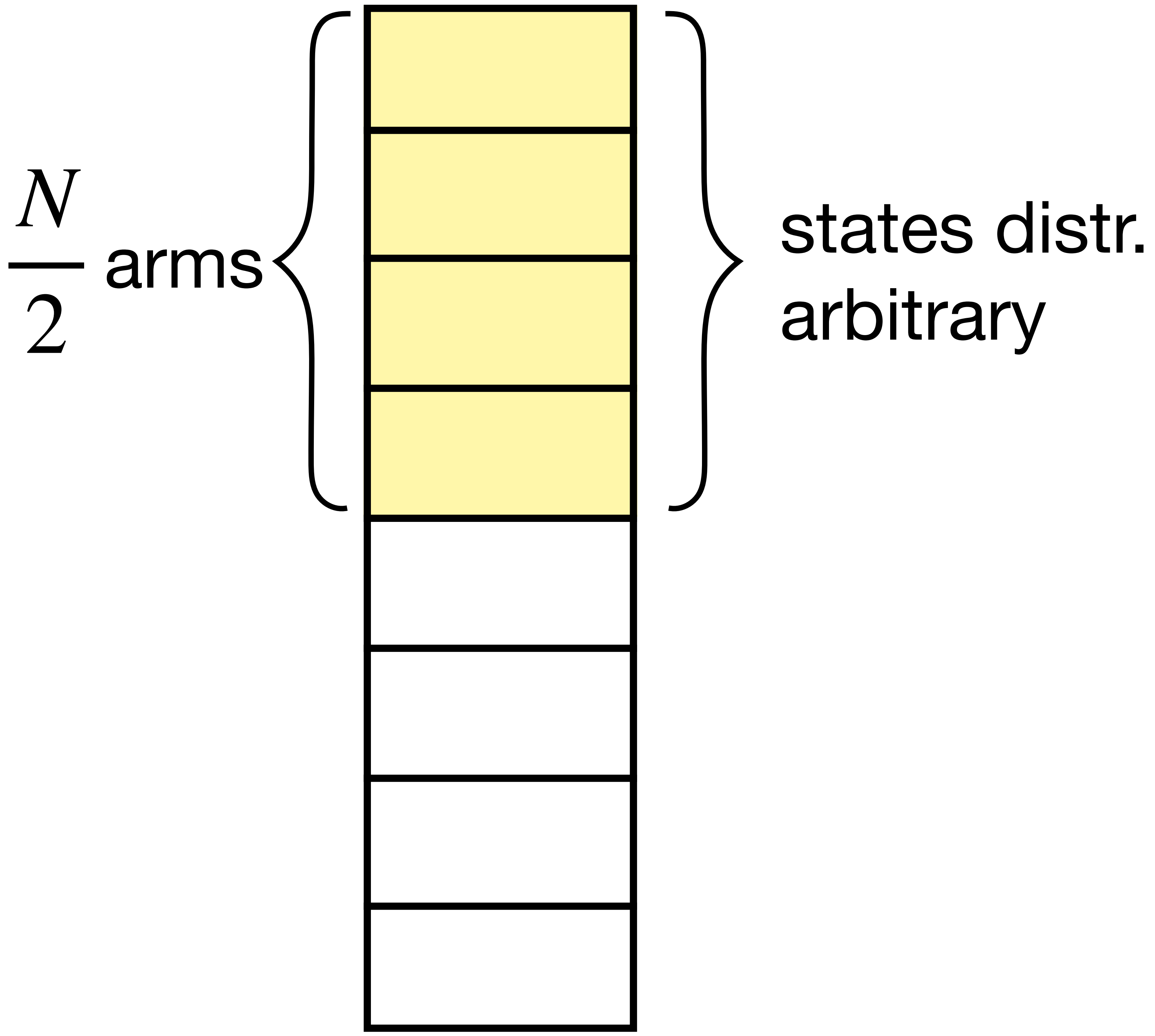
Intuition: start from a subset

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

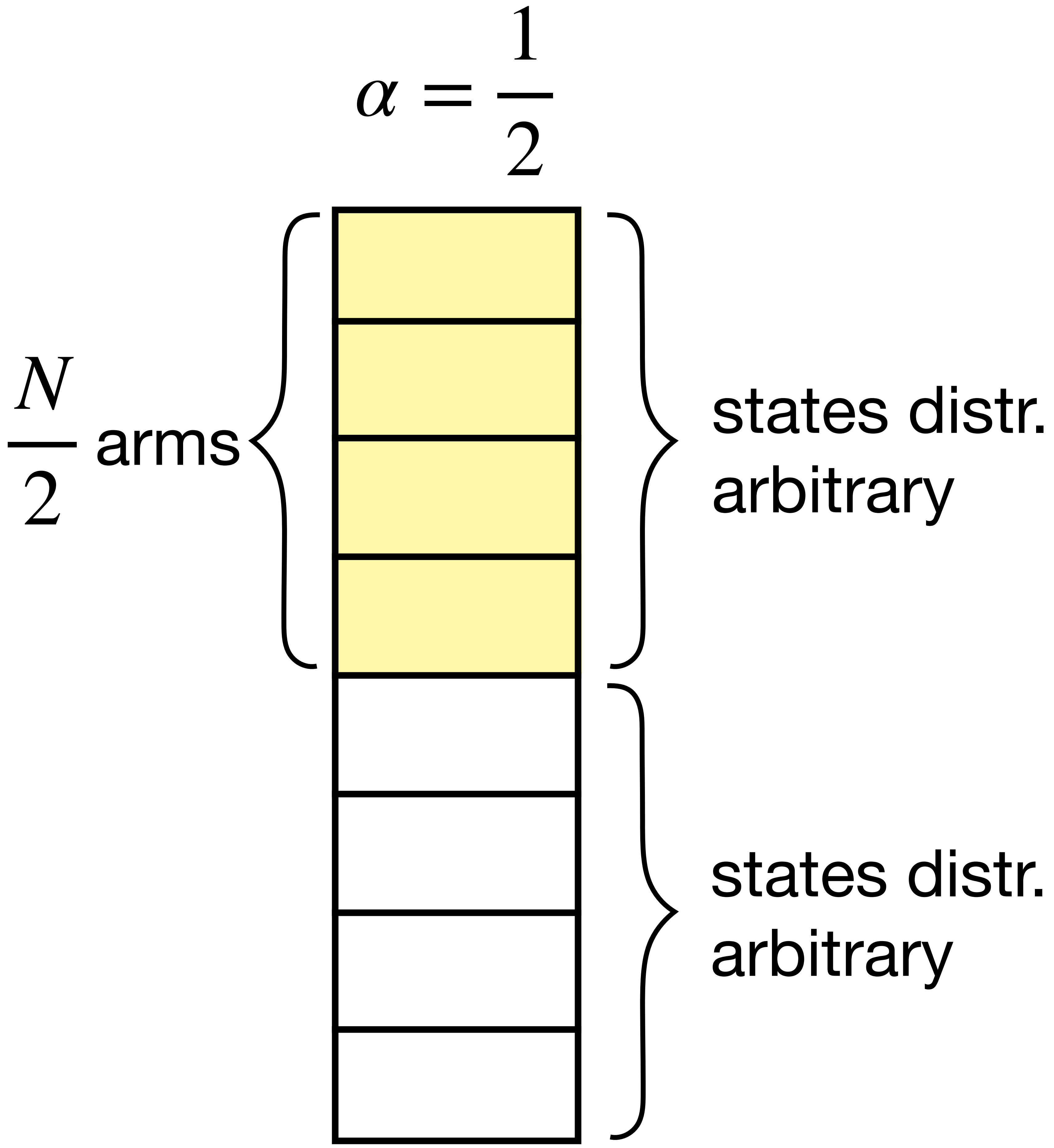


Intuition: start from a subset

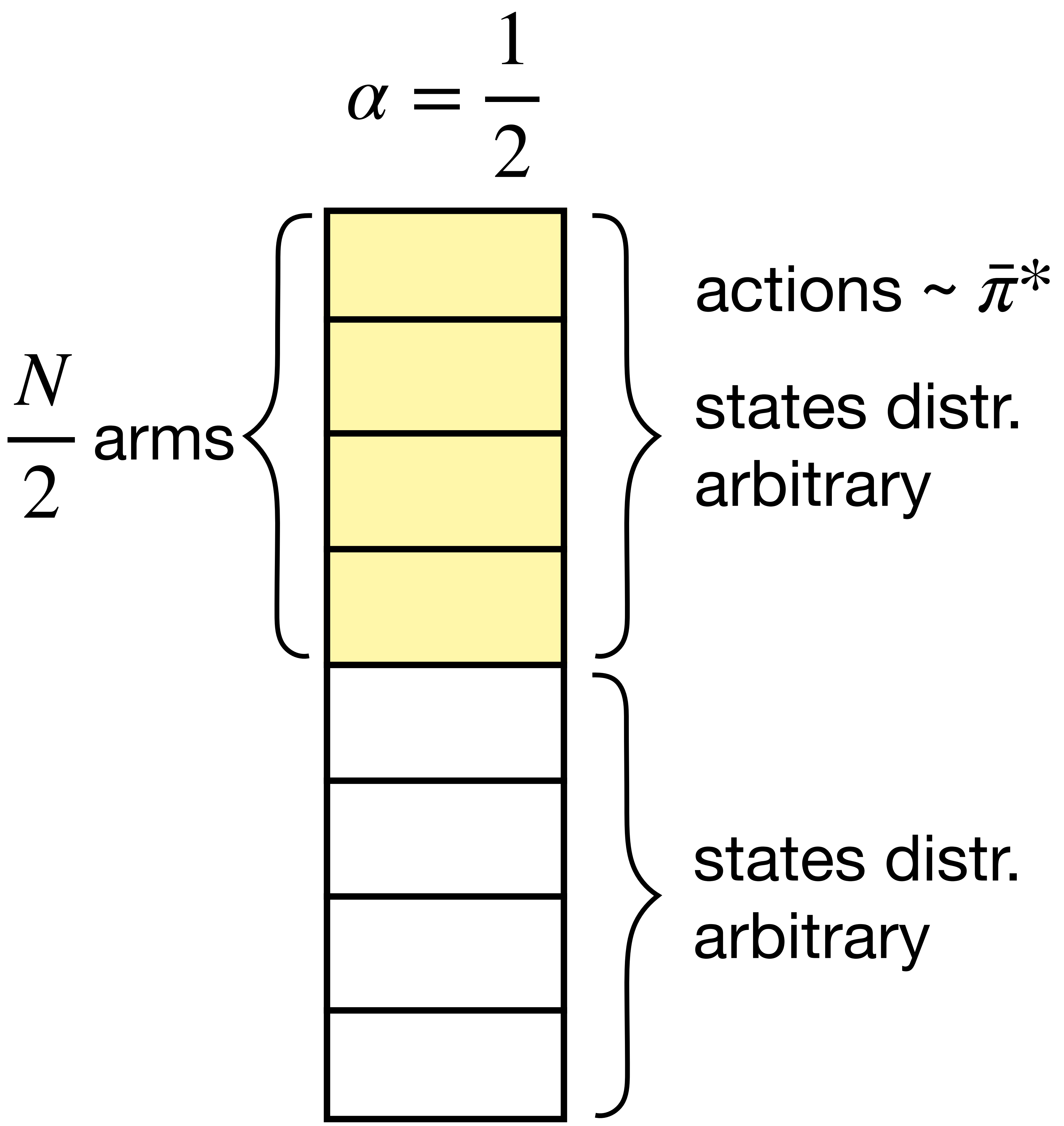
$\alpha = \frac{1}{2}$



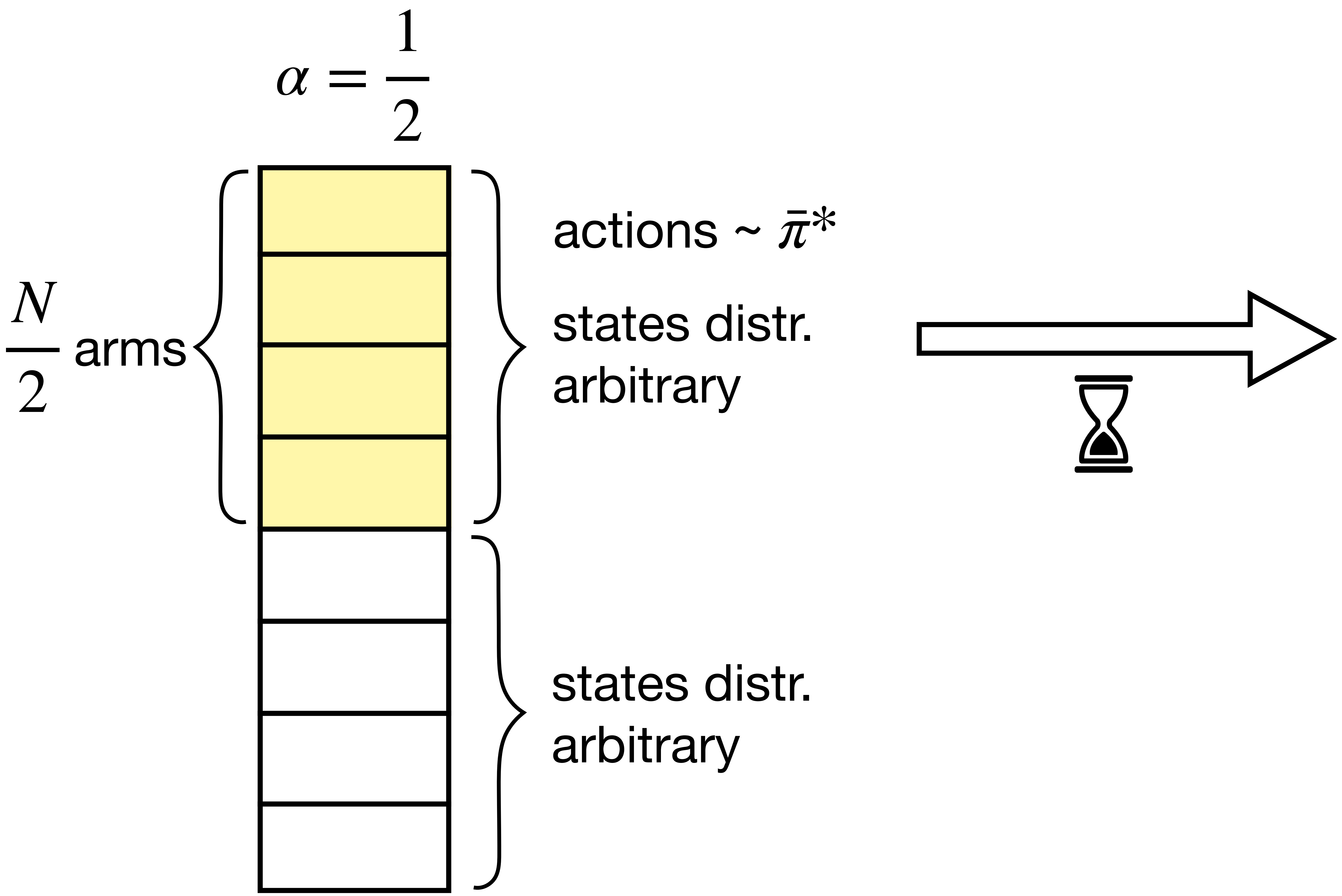
Intuition: start from a subset



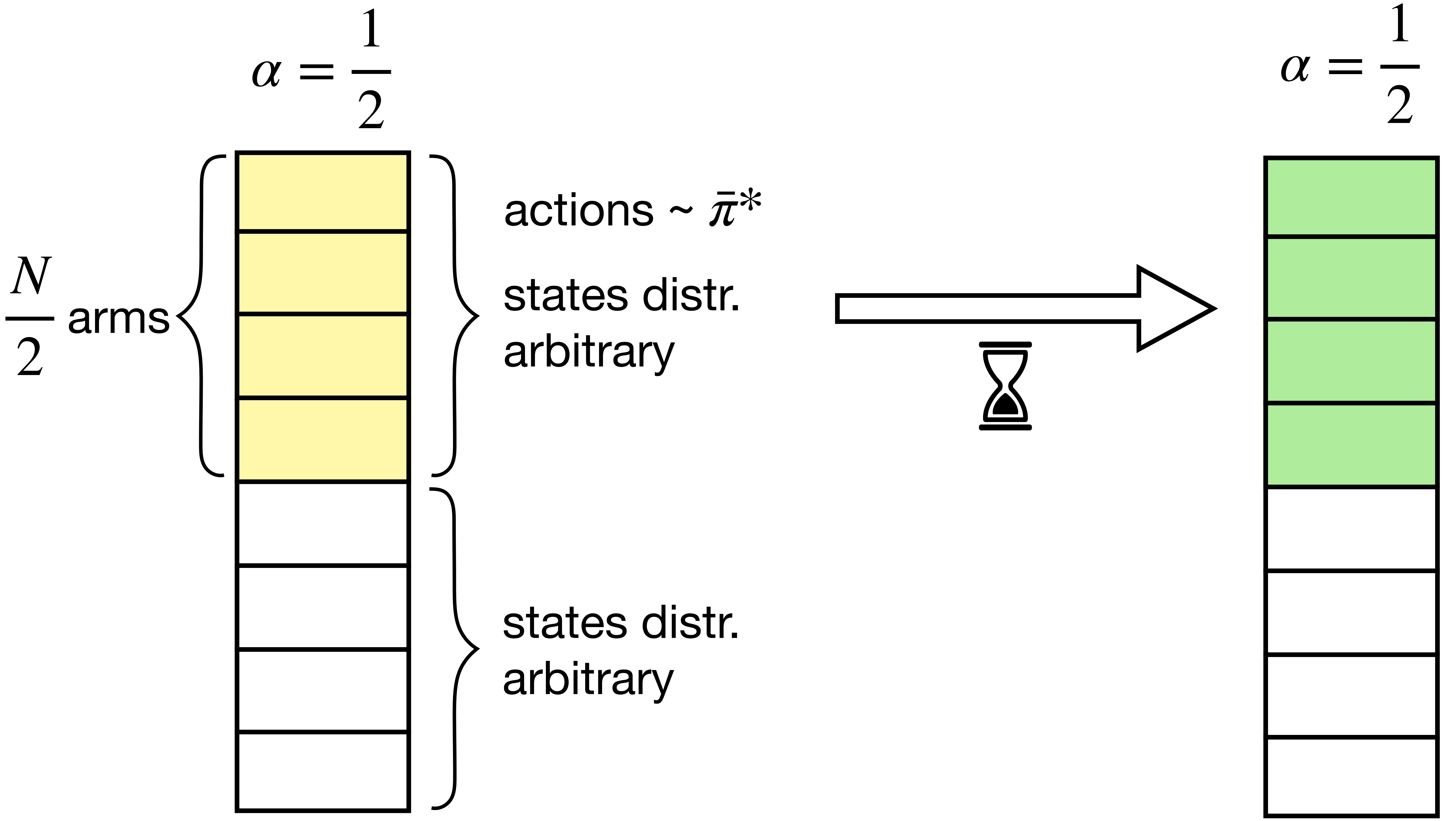
Intuition: start from a subset



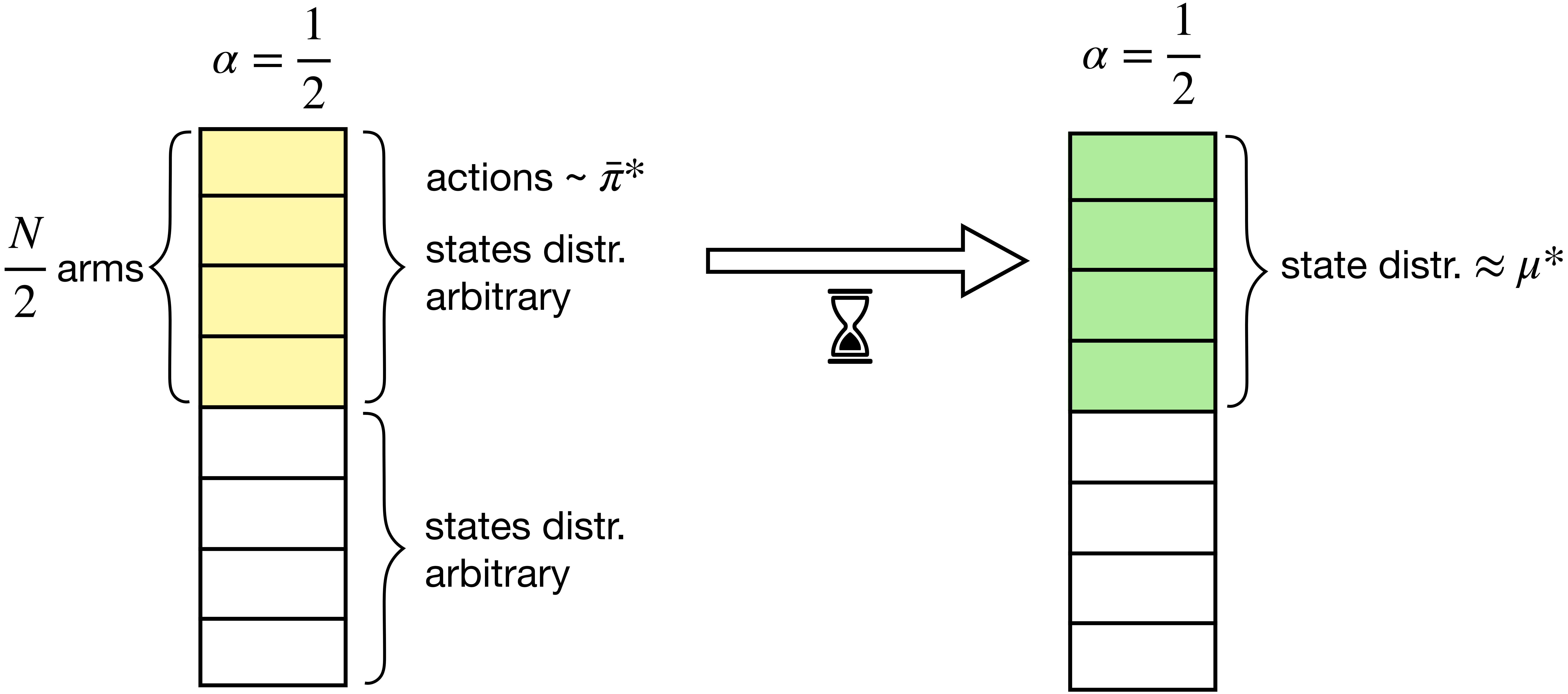
Intuition: start from a subset



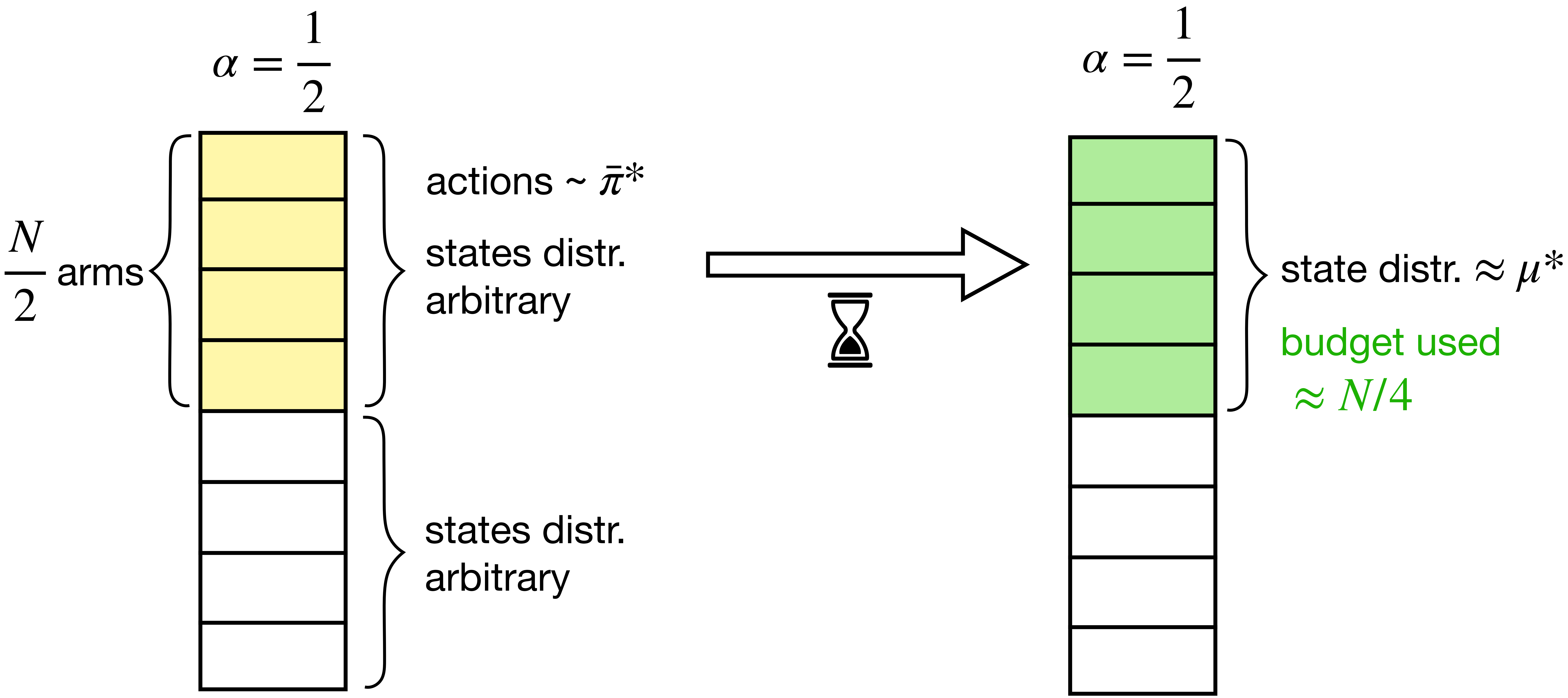
Intuition: start from a subset



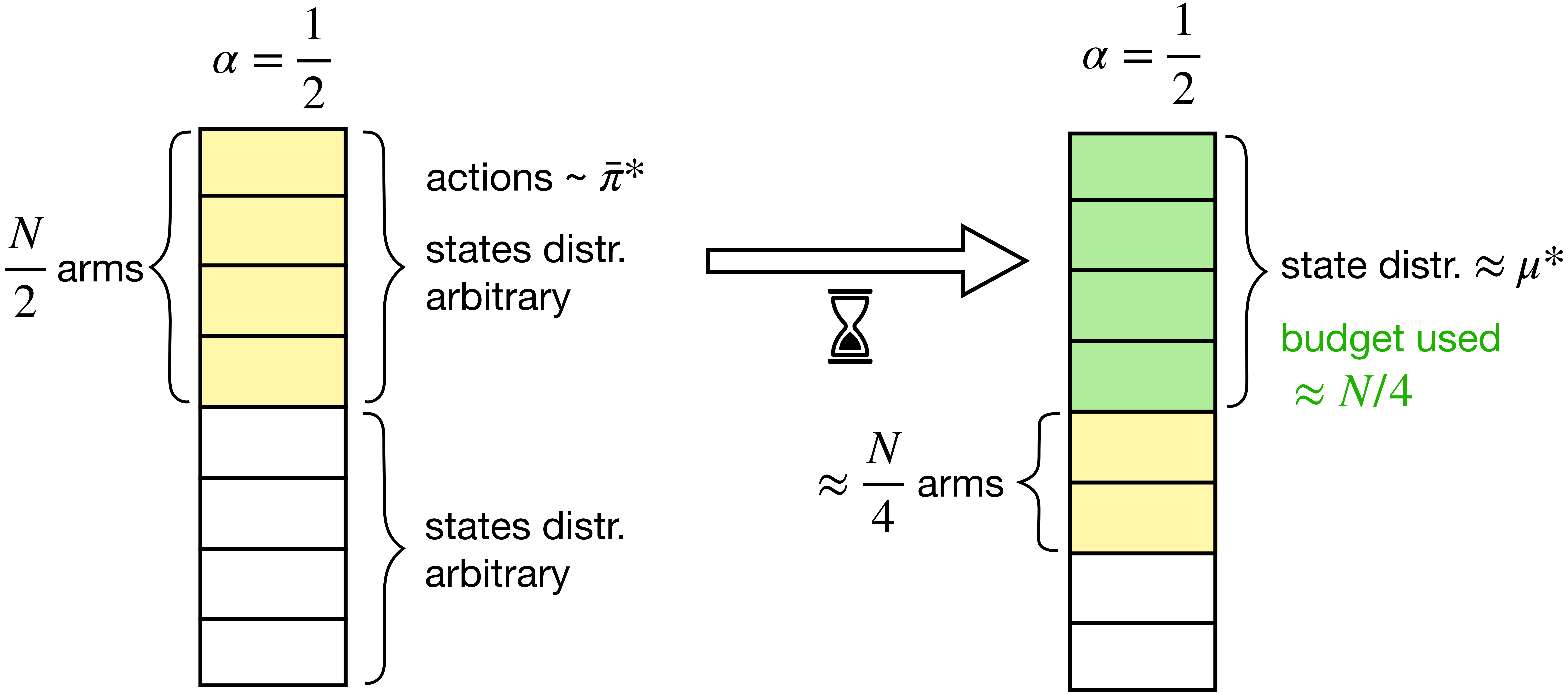
Intuition: start from a subset



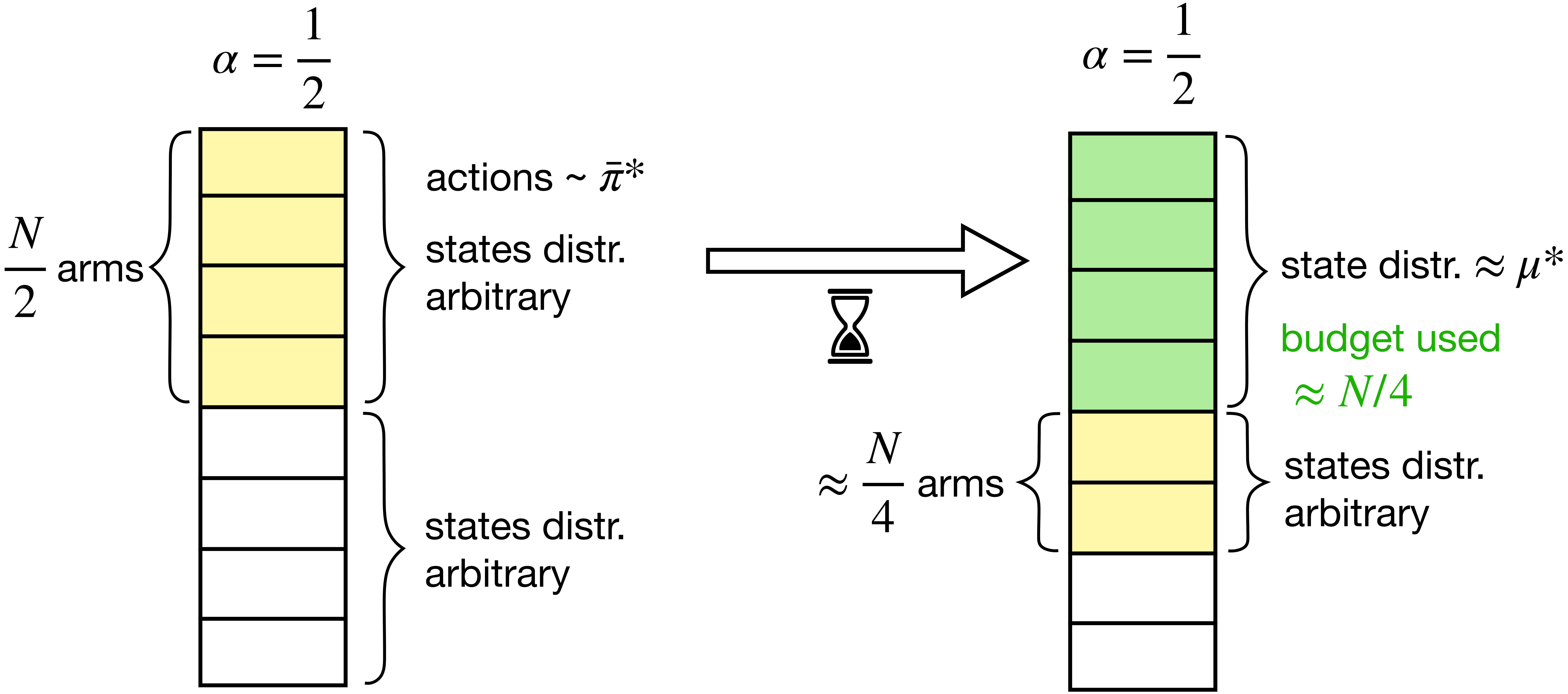
Intuition: start from a subset



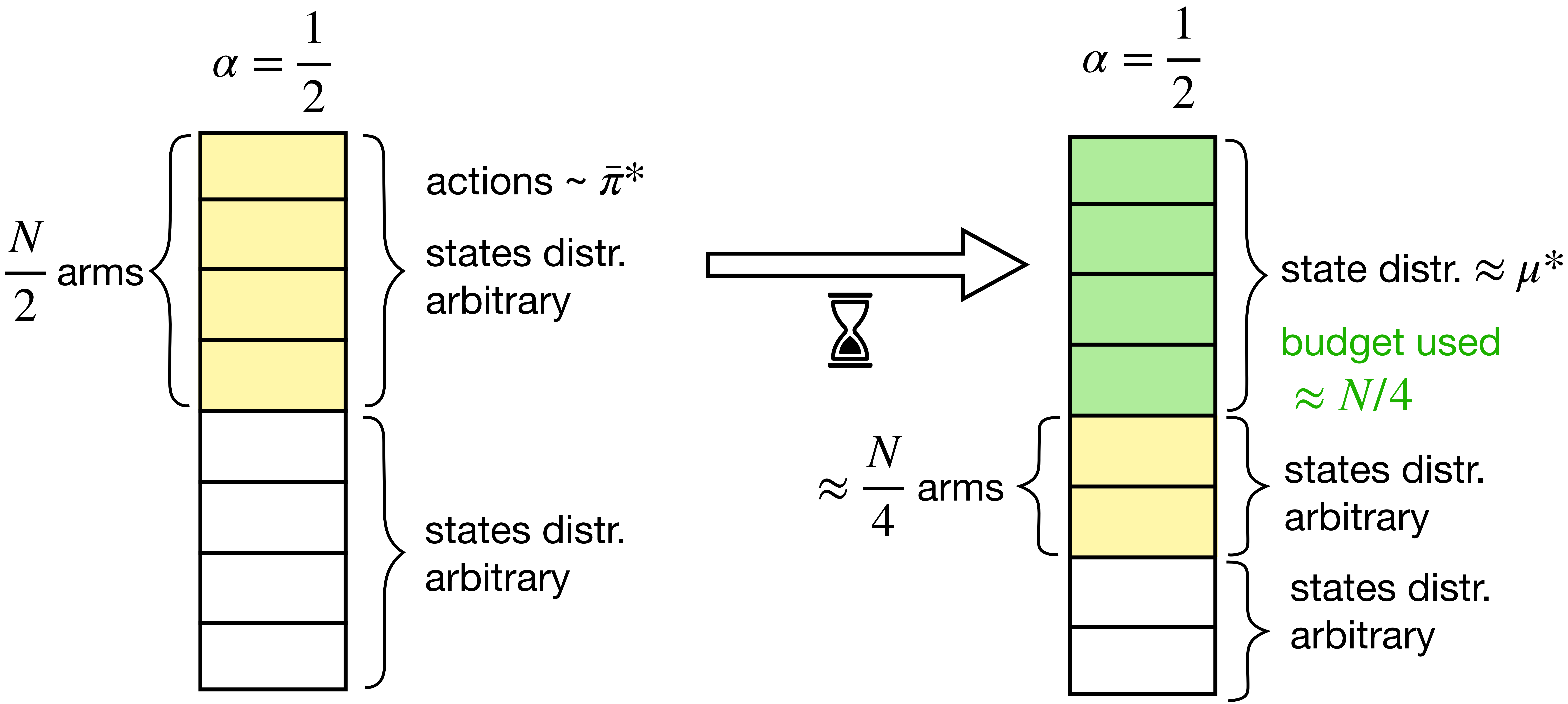
Intuition: start from a subset



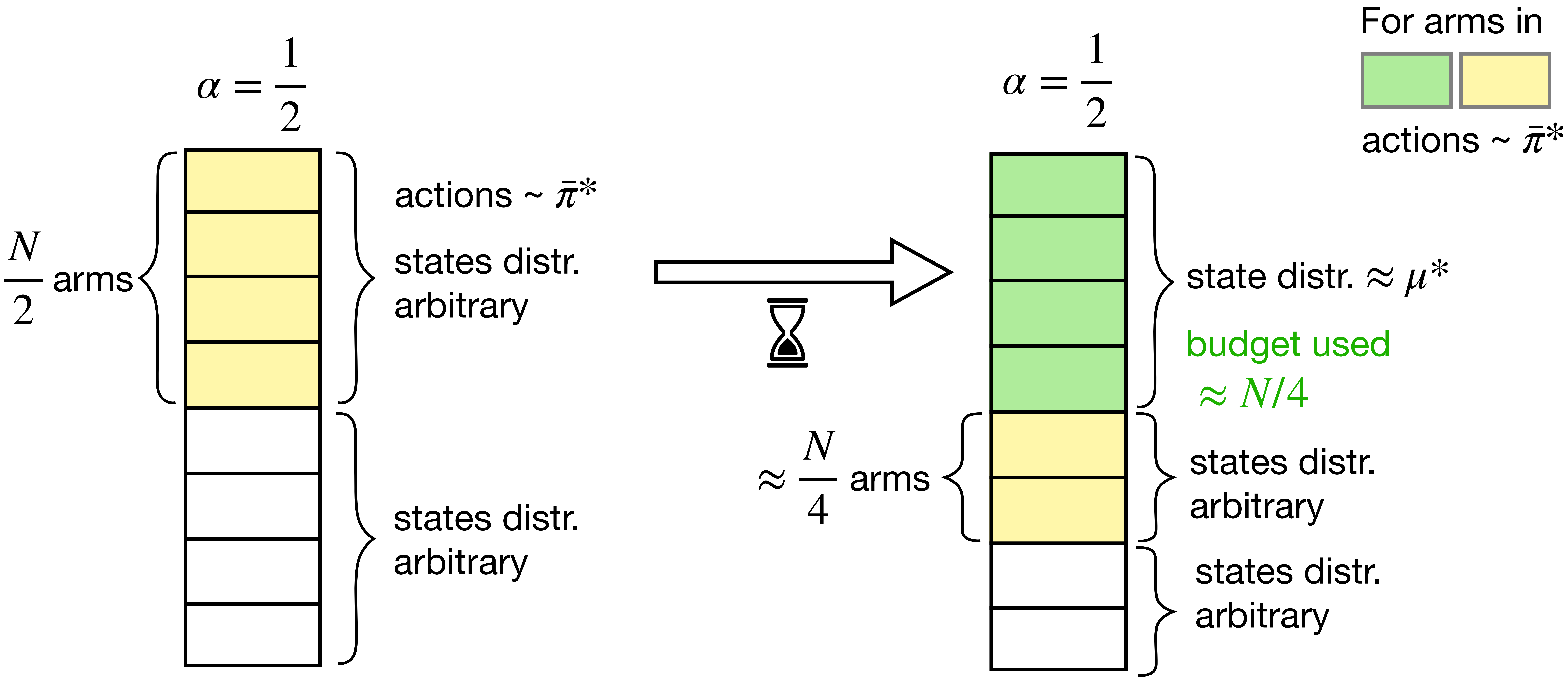
Intuition: start from a subset



Intuition: start from a subset

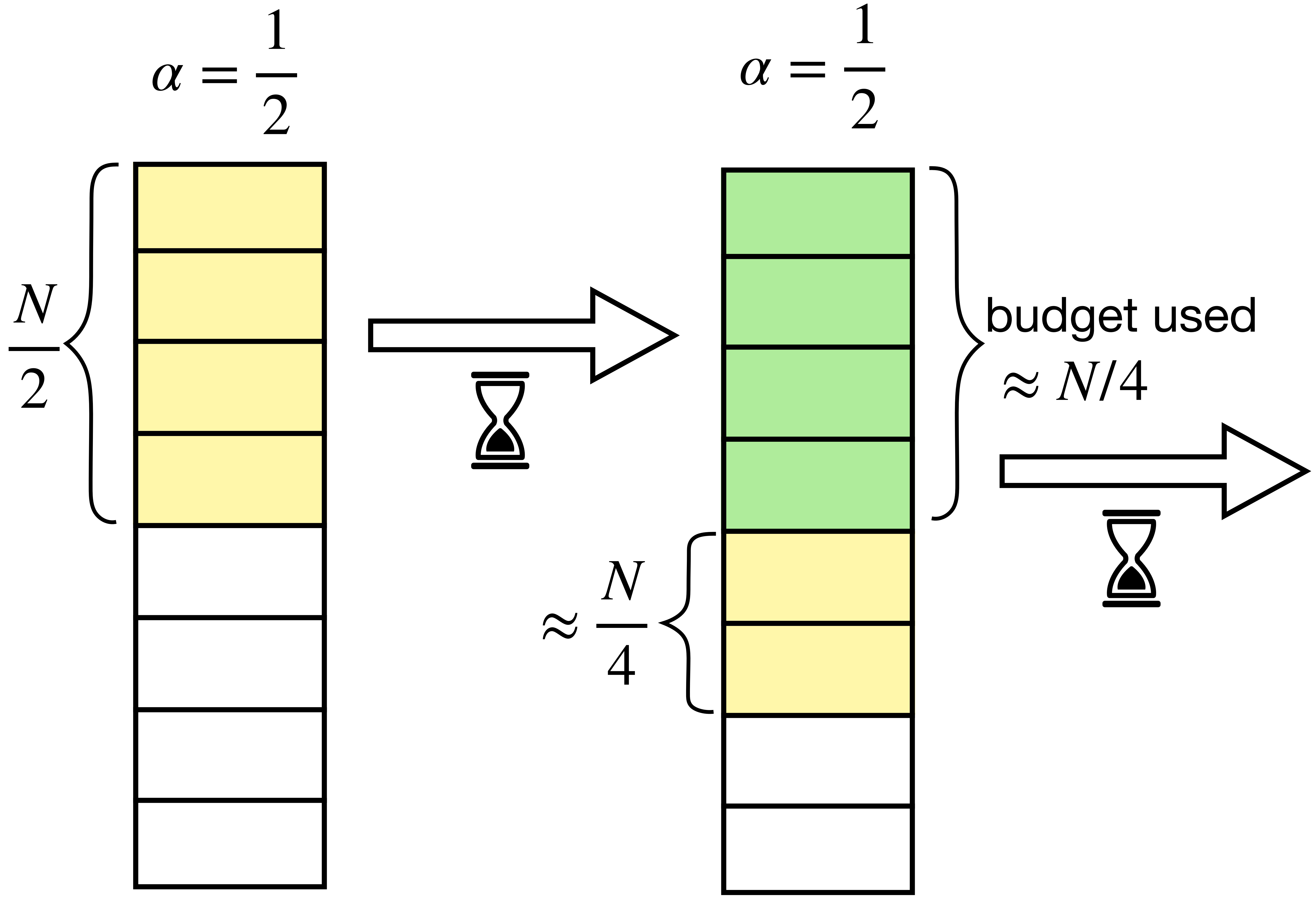


Intuition: start from a subset

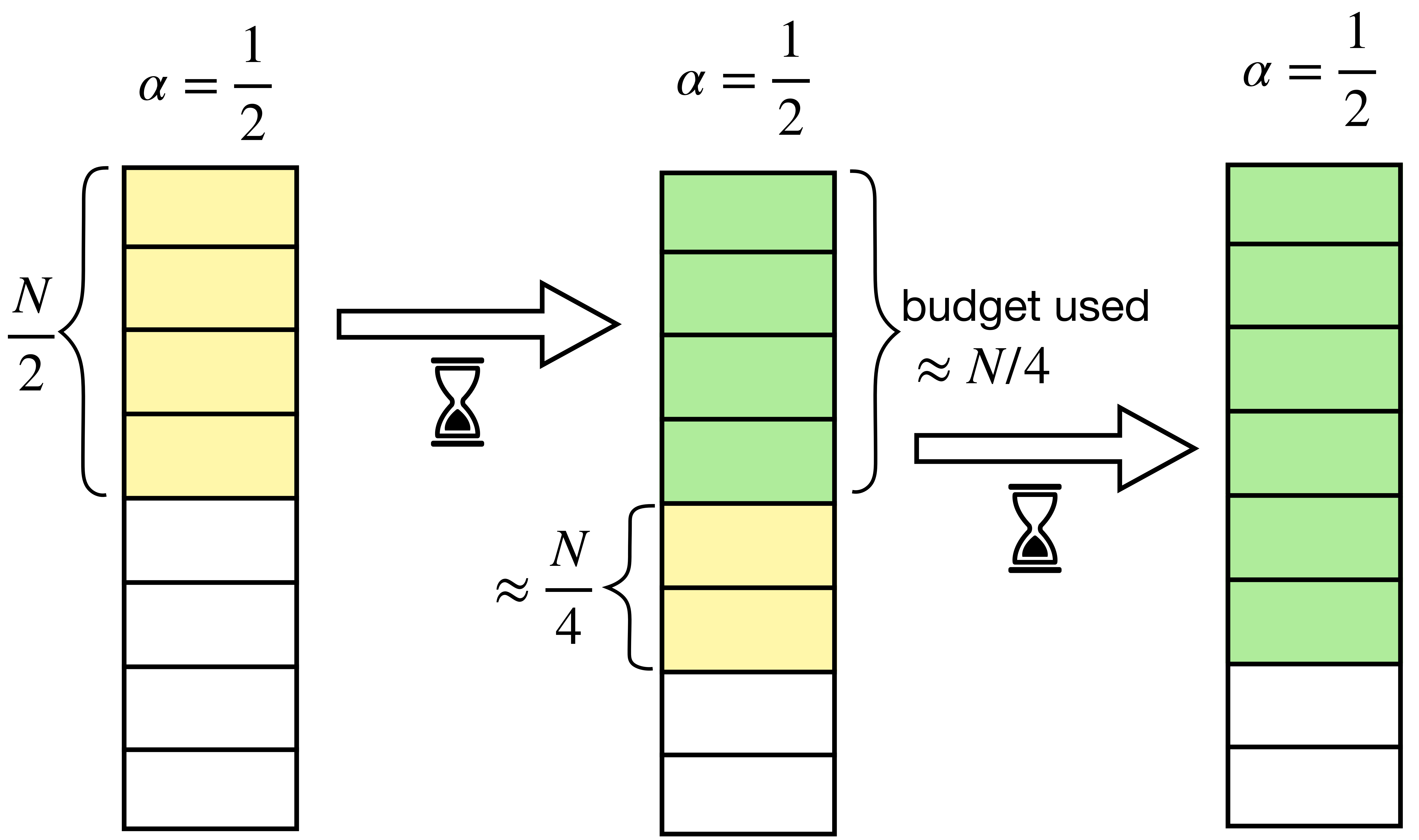


Intuition: start from a subset

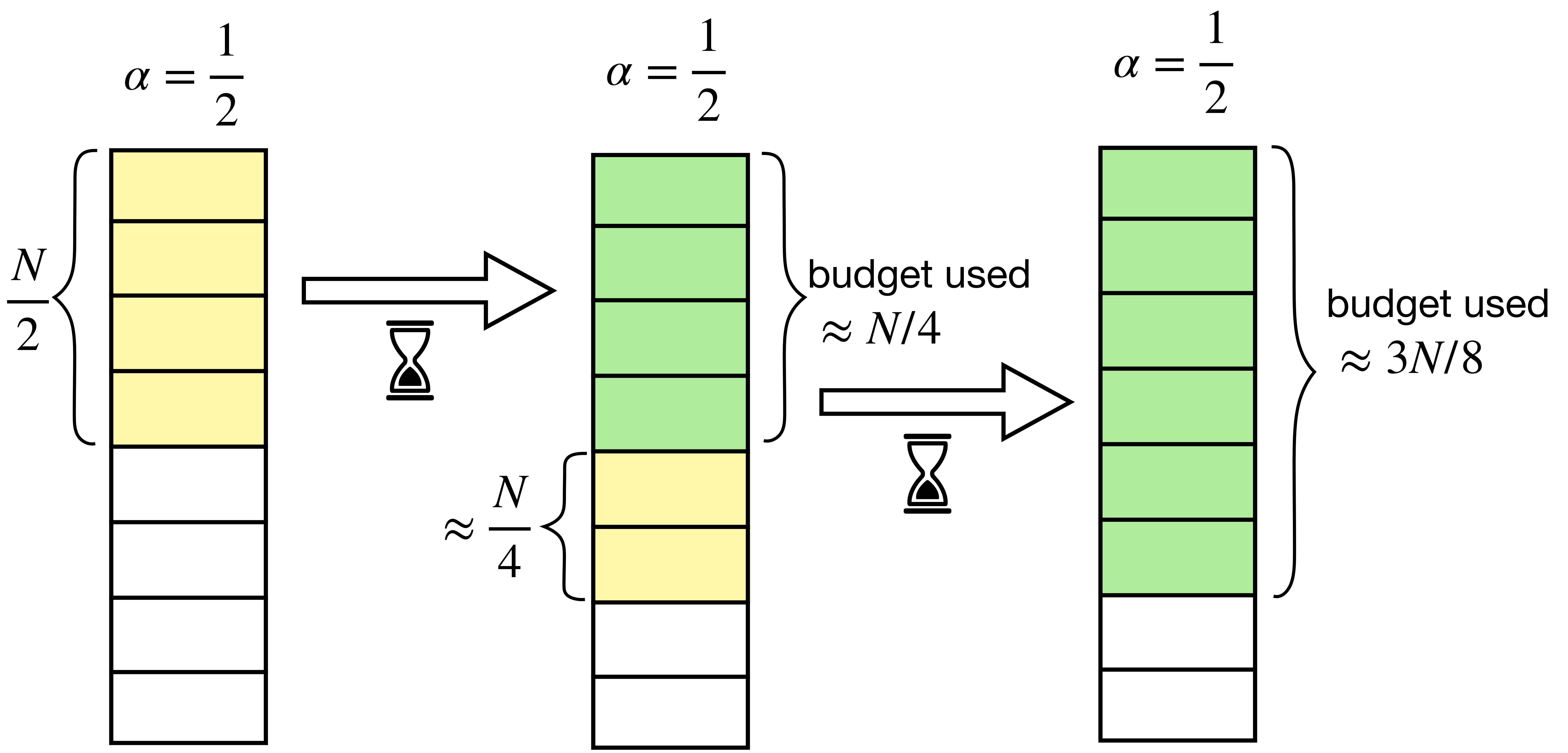
Expand the subset



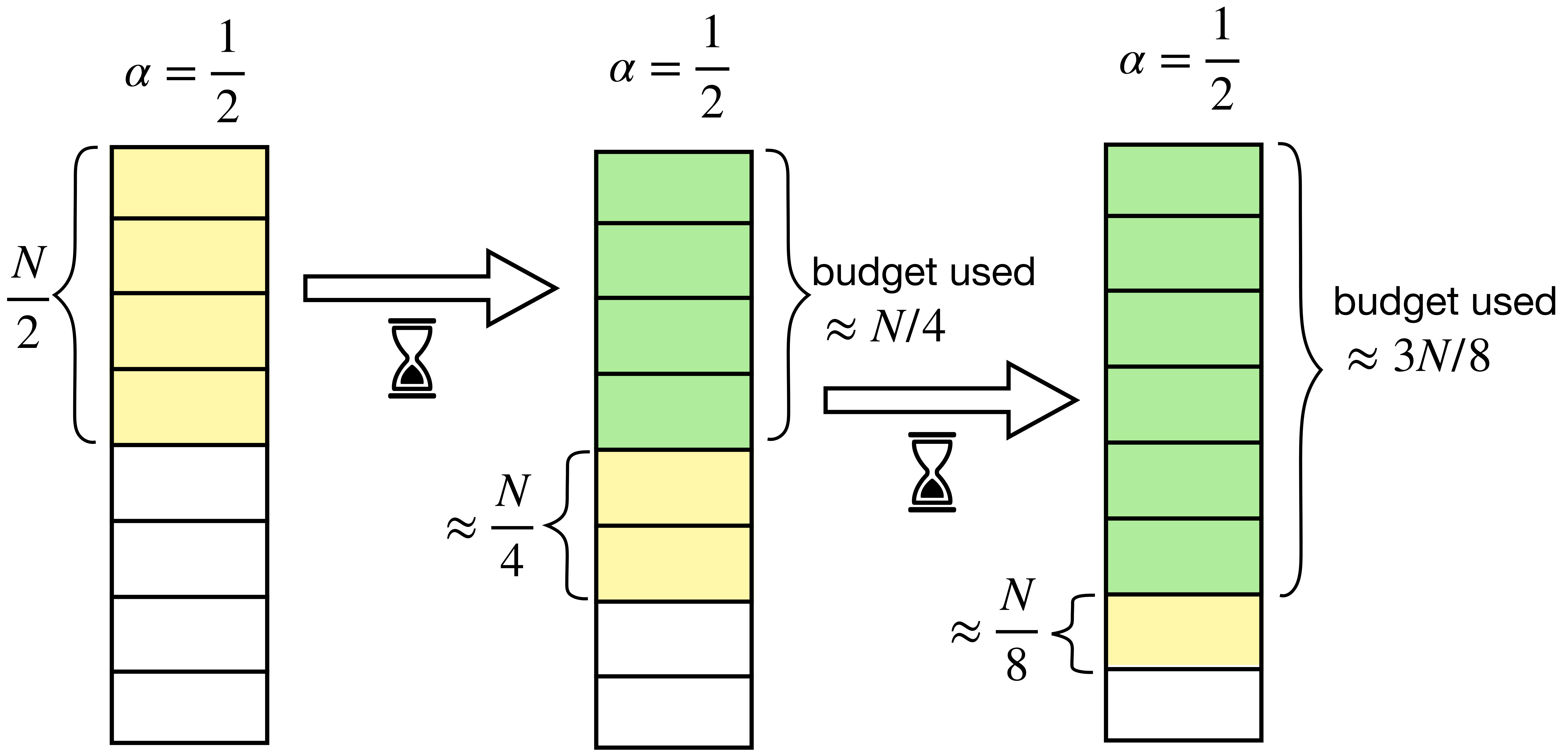
Expand the subset



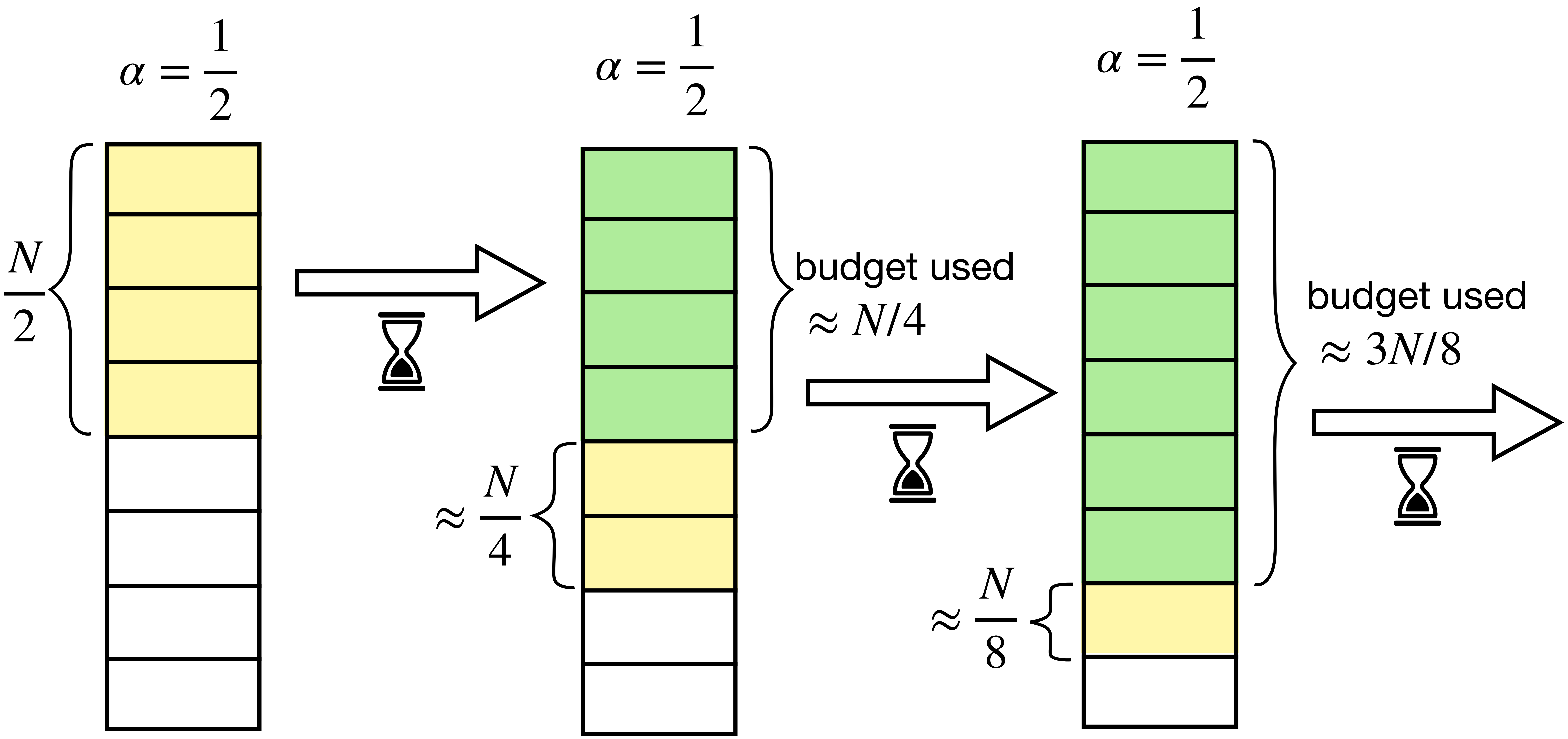
Expand the subset



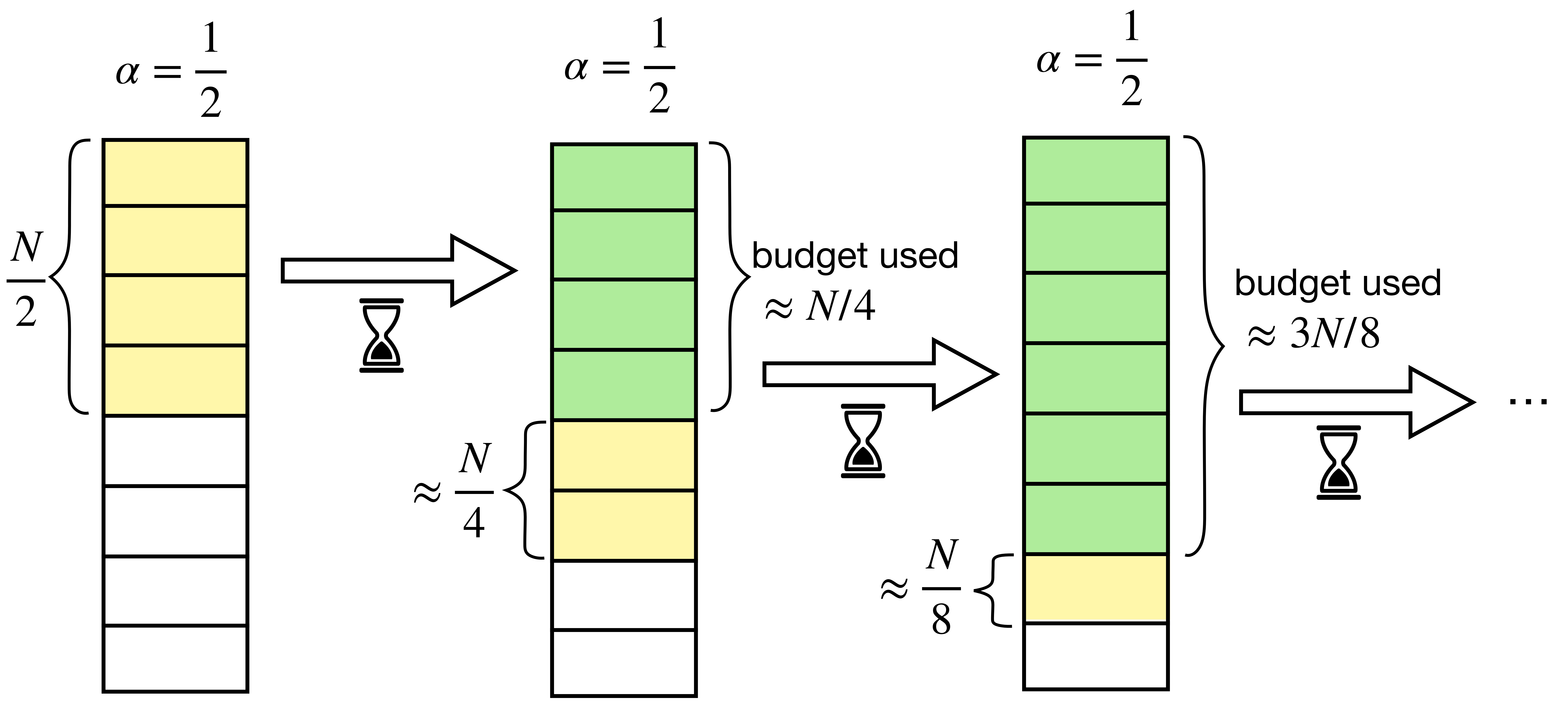
Expand the subset



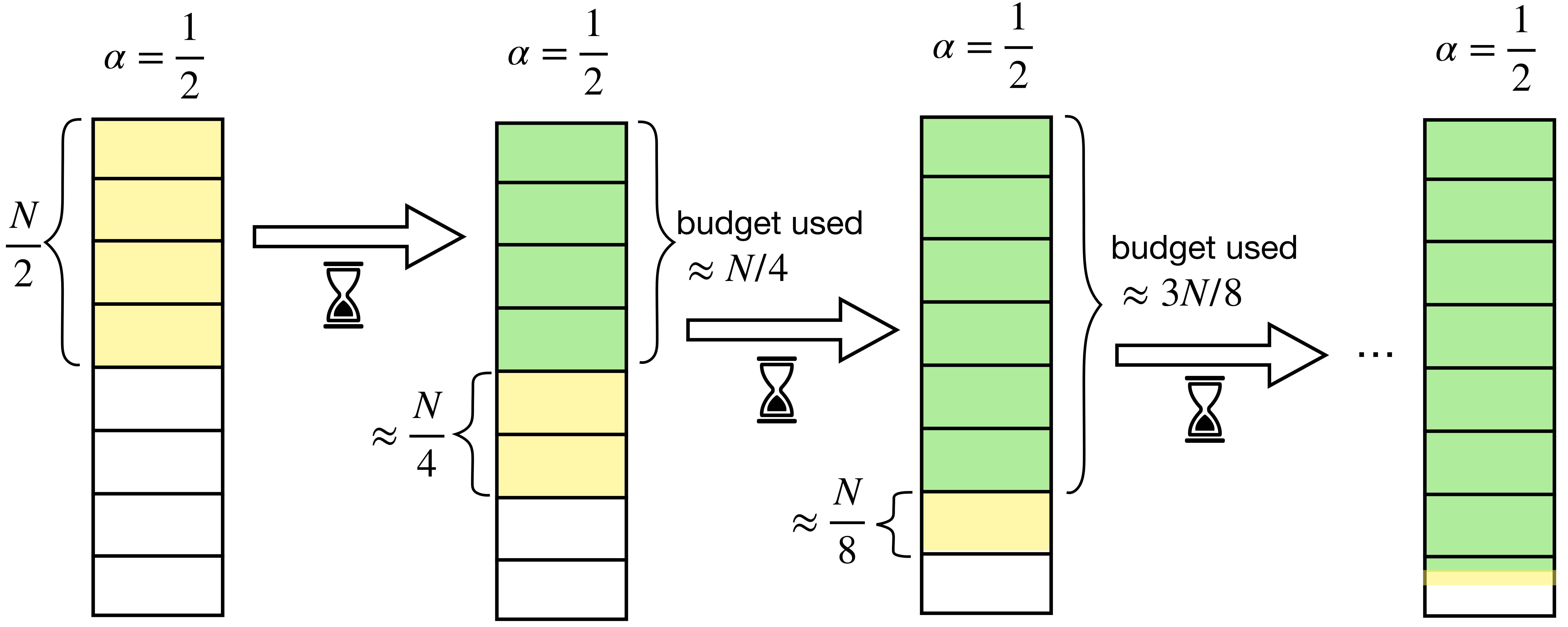
Expand the subset



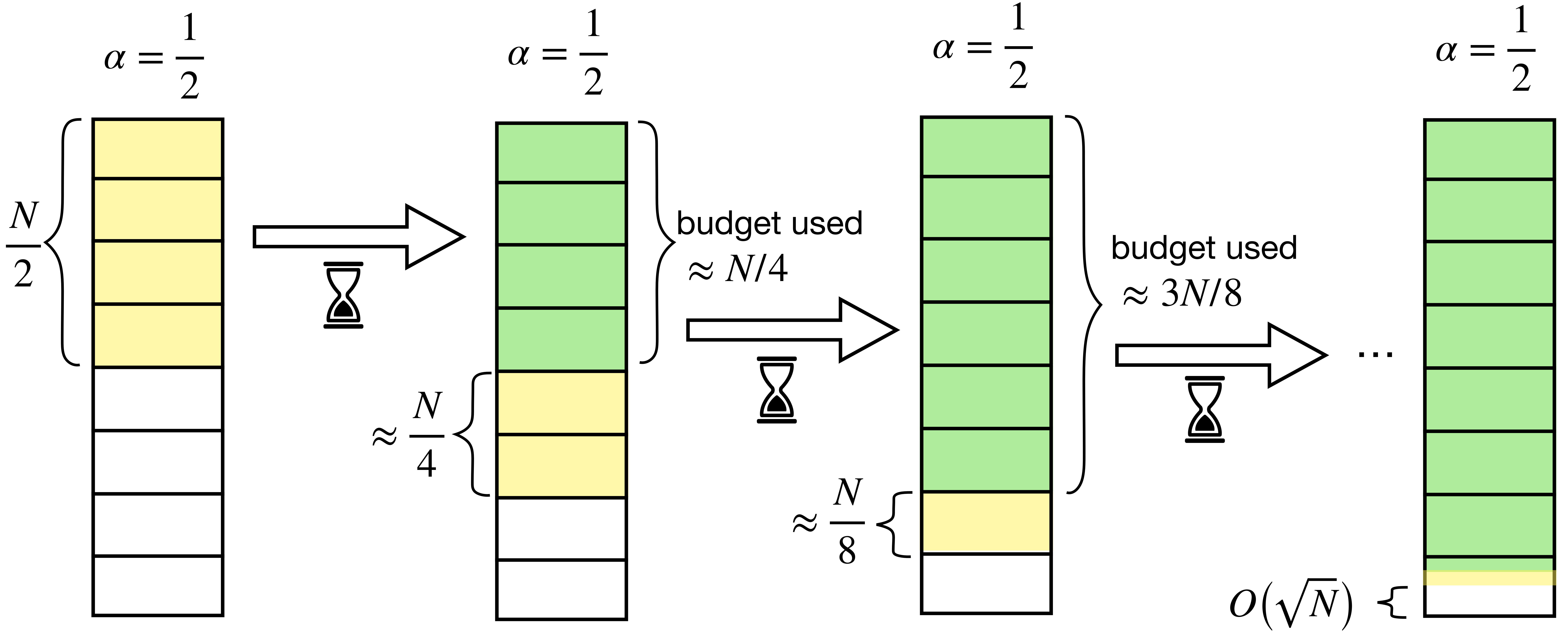
Expand the subset



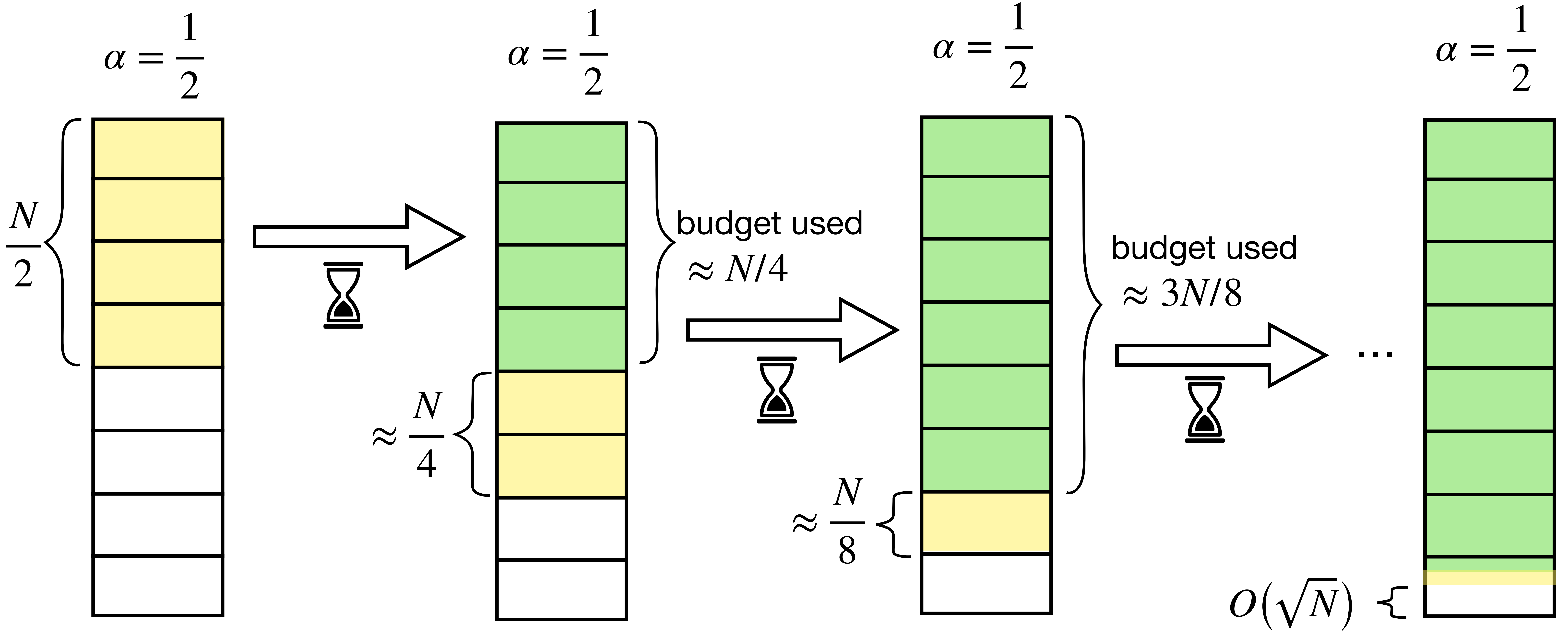
Expand the subset



Expand the subset

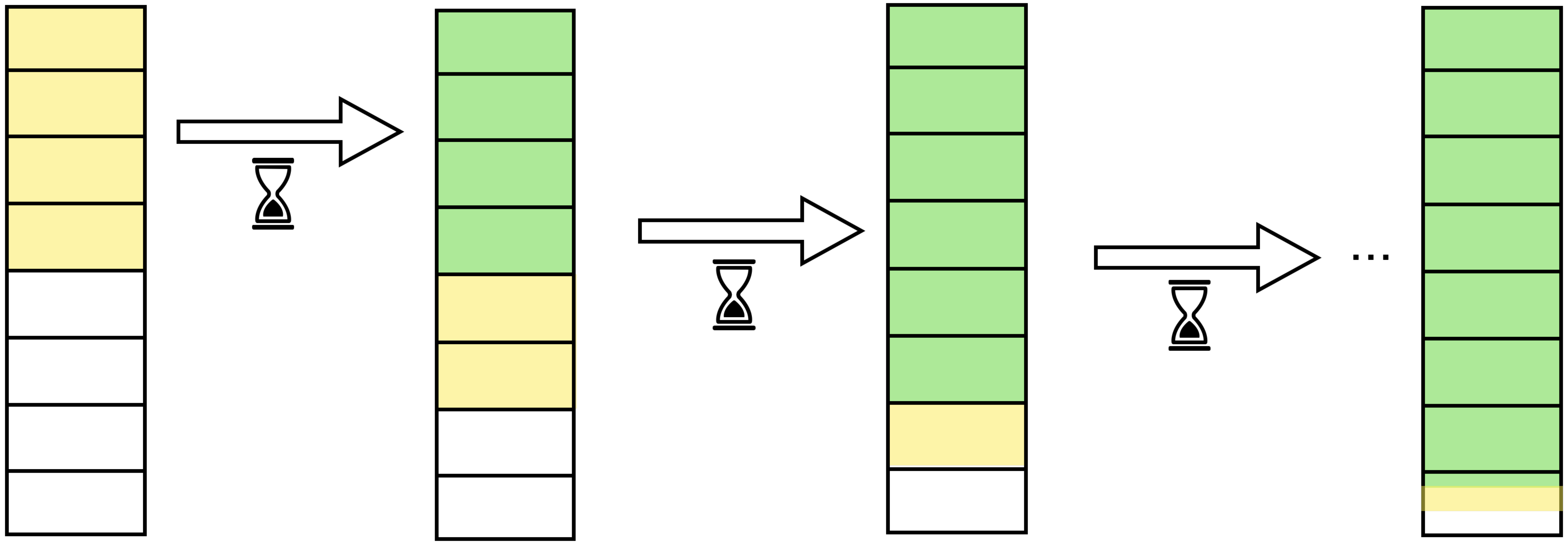


Expand the subset

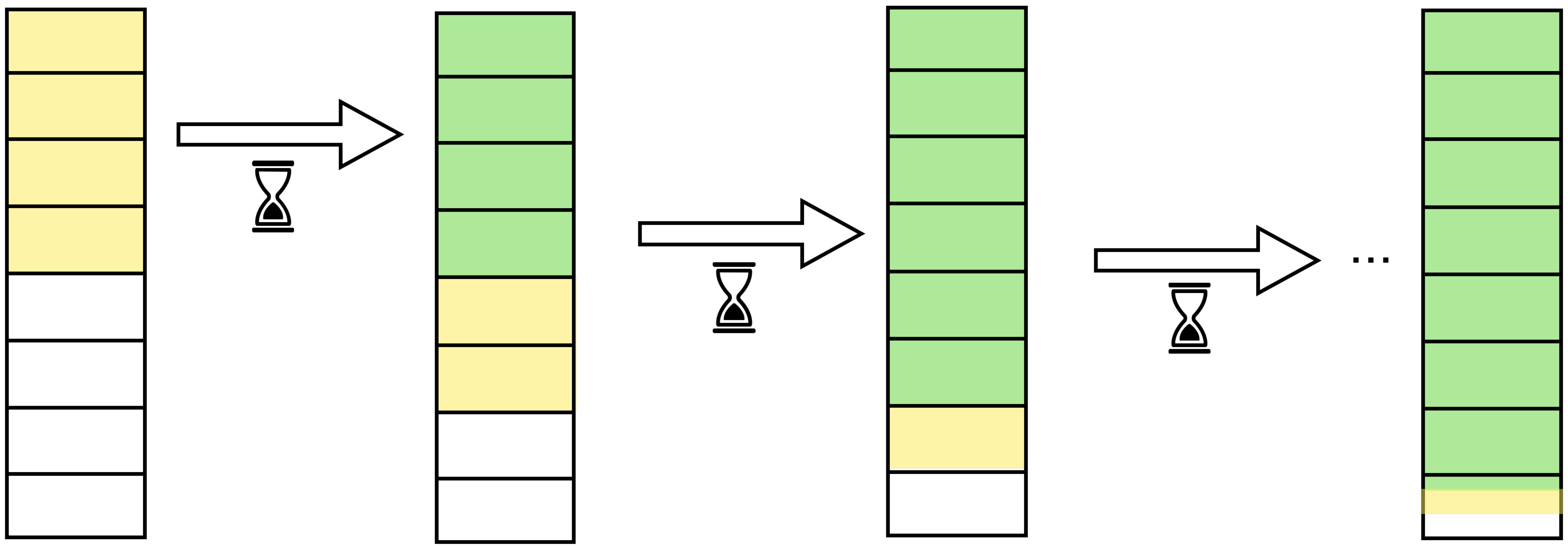


How to design a policy to implement this intuition?

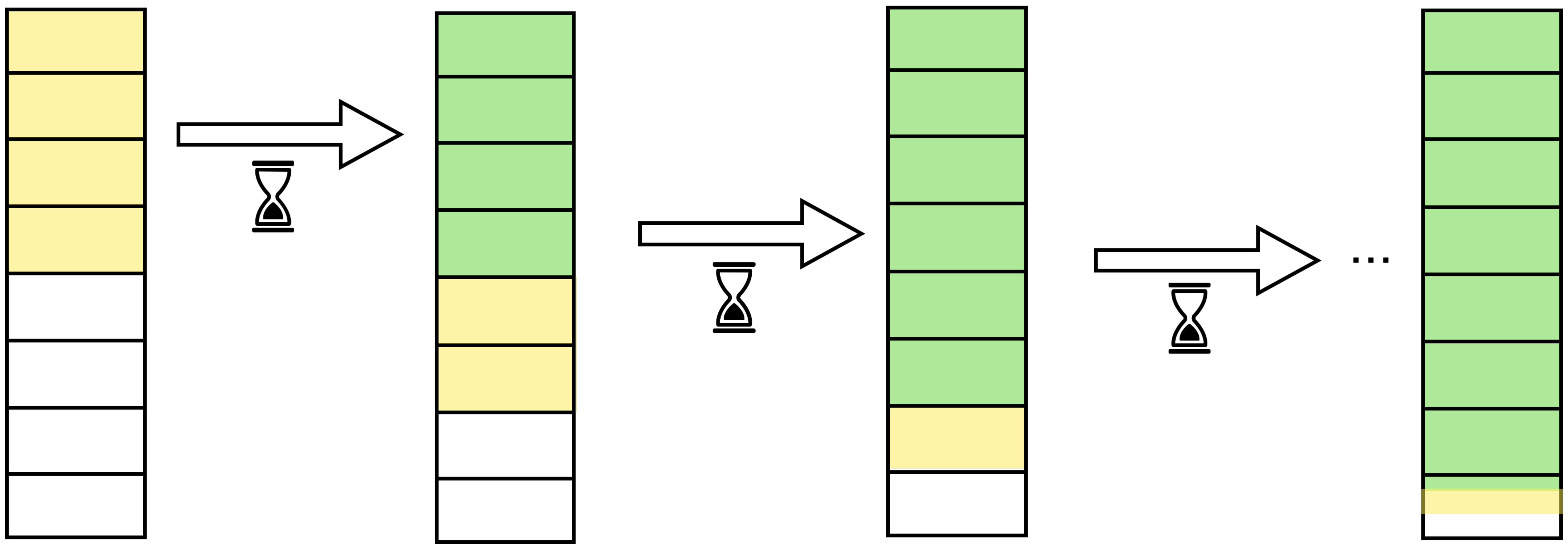
Implement the intuition



Implement the intuition

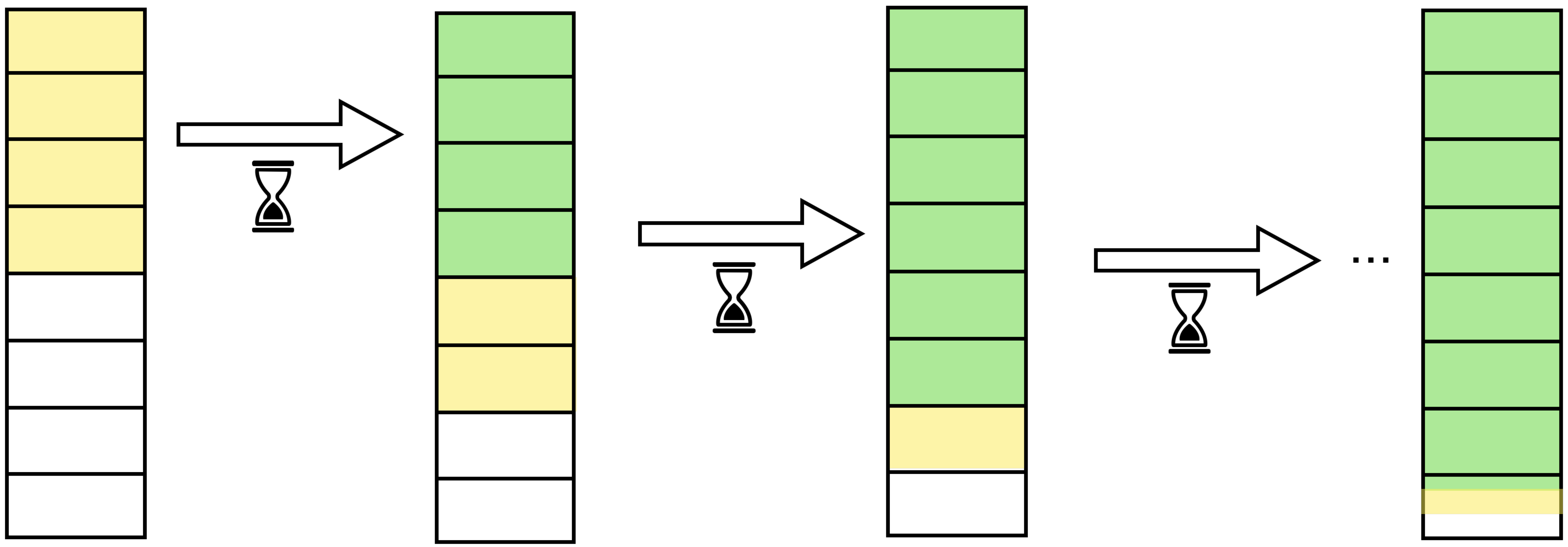


Implement the intuition



Let as many arms follow $\bar{\pi}^*$ as possible, in a fixed order.

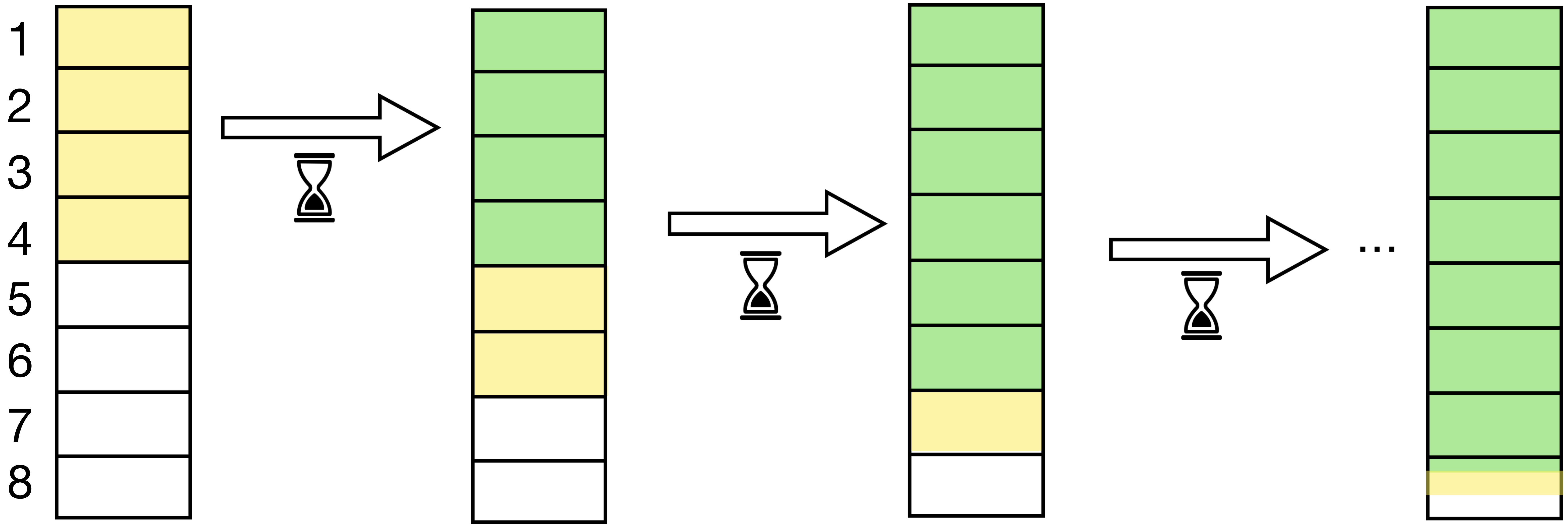
Implement the intuition



Let as many arms follow $\bar{\pi}^*$ as possible, in a fixed order.

ID Policy:

Implement the intuition

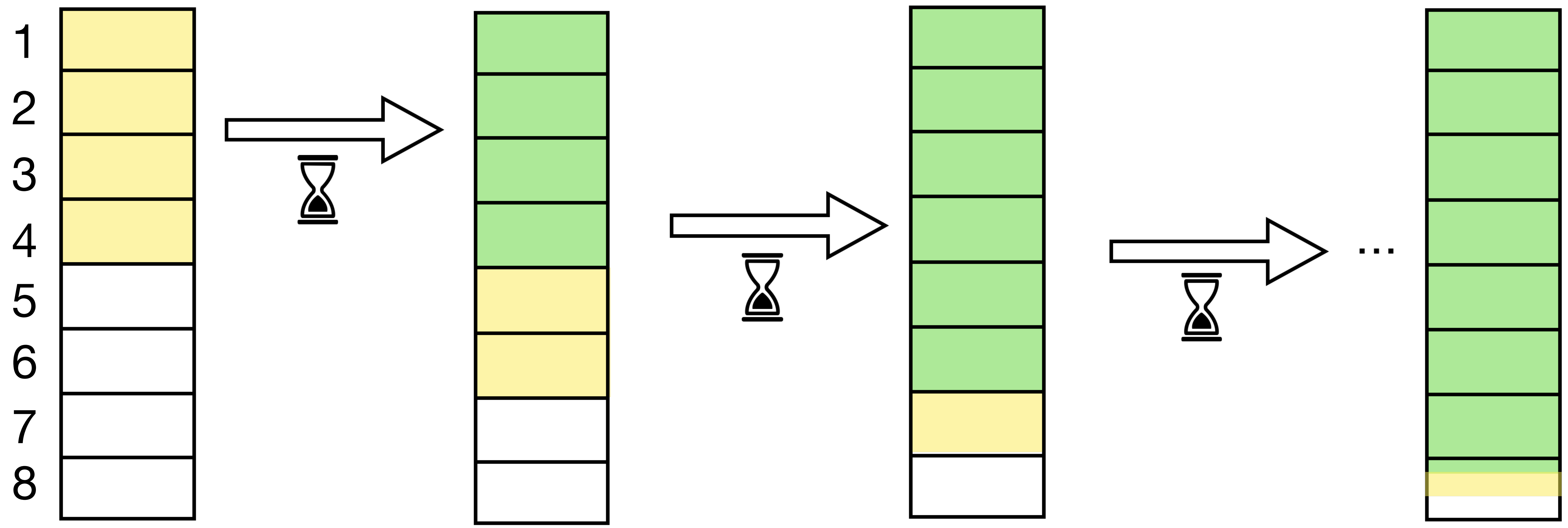


Let as many arms follow $\bar{\pi}^*$ as possible, in a fixed order.

ID Policy:

- Fix an arbitrary IDs for the arms;

Implement the intuition

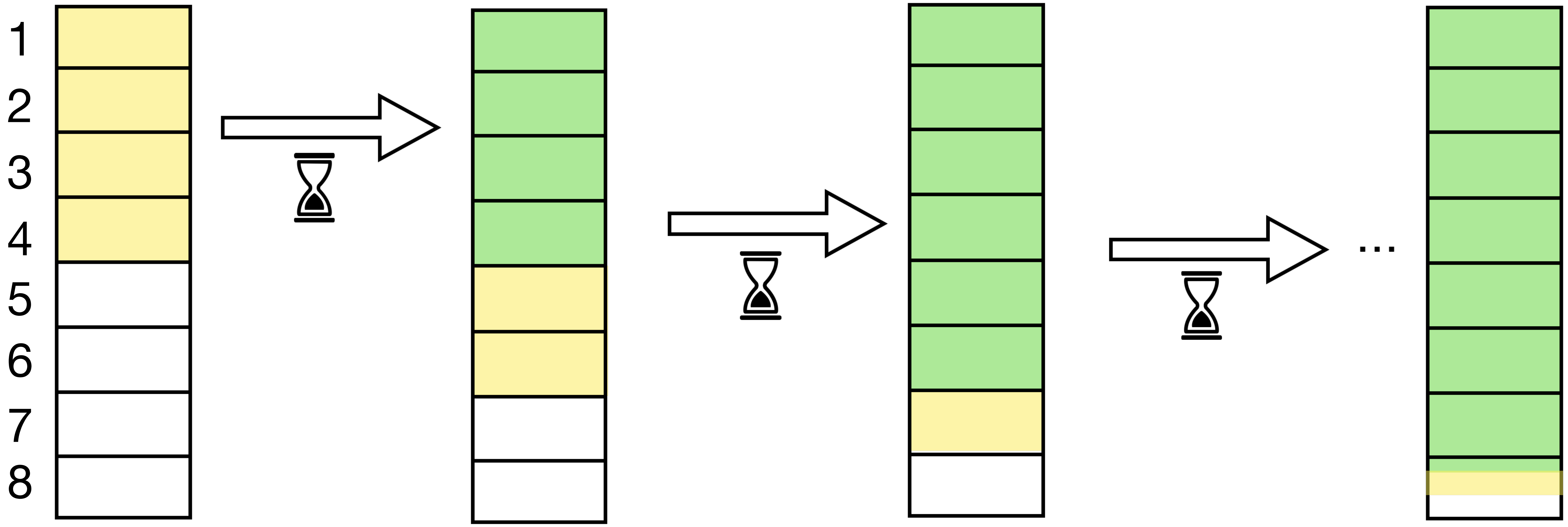


Let as many arms follow $\bar{\pi}^*$ as possible, in a fixed order.

ID Policy:

- Fix an arbitrary IDs for the arms;
- Prioritize arms with smaller IDs to follow $\bar{\pi}^*$

Implement the intuition



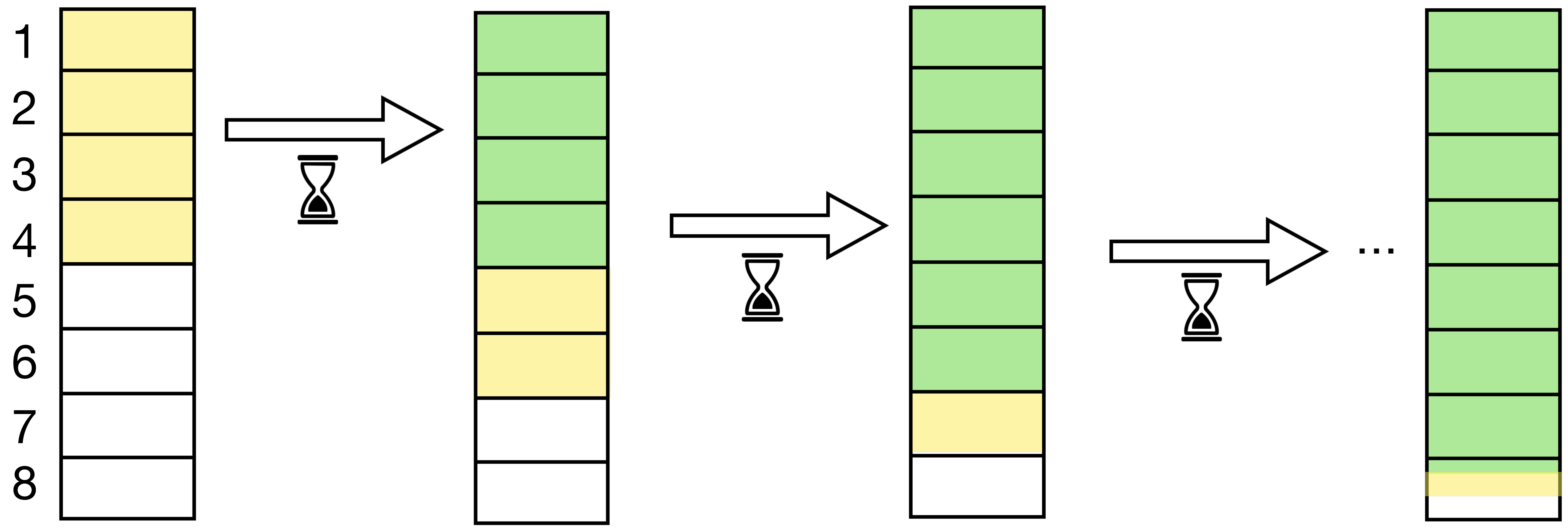
Let as many arms follow $\bar{\pi}^*$ as possible, in a fixed order.

ID Policy:

We can also avoid using IDs;

- Fix an arbitrary IDs for the arms;
- Prioritize arms with smaller IDs to follow $\bar{\pi}^*$

Implement the intuition



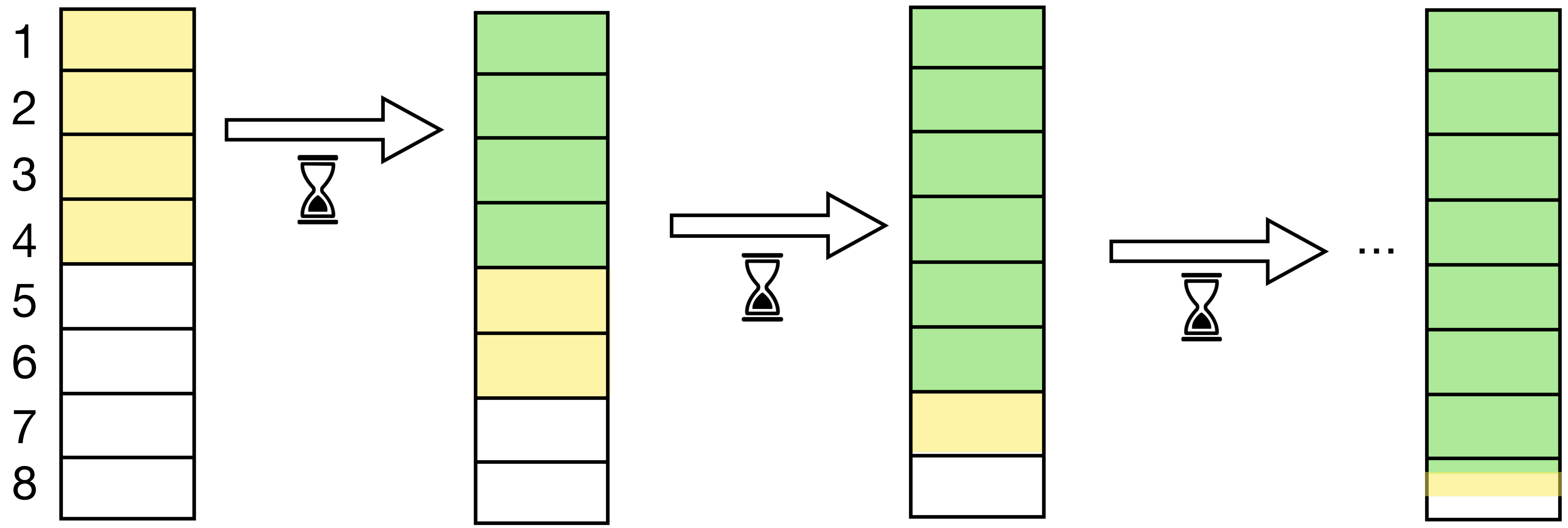
Let as many arms follow $\bar{\pi}^*$ as possible, in a fixed order.

ID Policy:

- Fix an arbitrary IDs for the arms;
- Prioritize arms with smaller IDs to follow $\bar{\pi}^*$

We can also avoid using IDs;
Essentially need persistency;

Implement the intuition



Let as many arms follow $\bar{\pi}^*$ as possible, in a fixed order.

ID Policy:

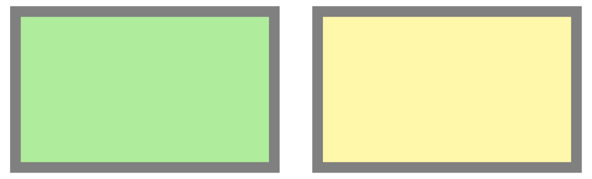
- Fix an arbitrary IDs for the arms;
- Prioritize arms with smaller IDs to follow $\bar{\pi}^*$

We can also avoid using IDs;
Essentially need persistency;
"Focus set policy"

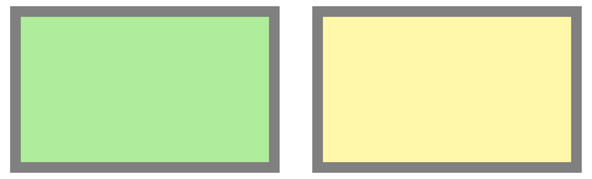
Lyapunov analysis

Lyapunov analysis

- Step 1: Formalize *focus set*: set of arms that will follow $\bar{\pi}^*$ in near future



Lyapunov analysis

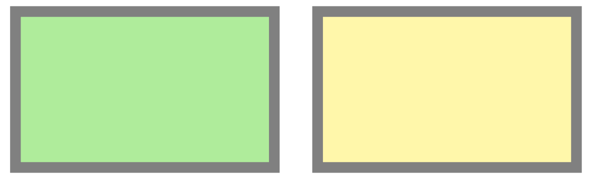
- Step 1: Formalize *focus set*: set of arms that will follow $\bar{\pi}^*$ in near future 
- Step 2: Define Lyapunov function with inputs: $(\text{states in the focus set}, \text{size of the focus set})$

Lyapunov analysis

- Step 1: Formalize *focus set*: set of arms that will follow $\bar{\pi}^*$ in near future
- Step 2: Define Lyapunov function with inputs: $($states in the focus set, size of the focus set$)$
  - Dynamically “focus on” a subsystem with good behaviors, and gradually expand it

Lyapunov analysis

- Step 1: Formalize *focus set*: set of arms that will follow $\bar{\pi}^*$ in near future



- Step 2: Define Lyapunov function with inputs: $($states in the focus set, size of the focus set$)$
  - Dynamically “focus on” a subsystem with good behaviors, and gradually expand it

- For details, see Professor Weina Wang’s talk in the session *Drift Methods for Stochastic Systems* this Tuesday 4 pm. (TE43, Summit 435)

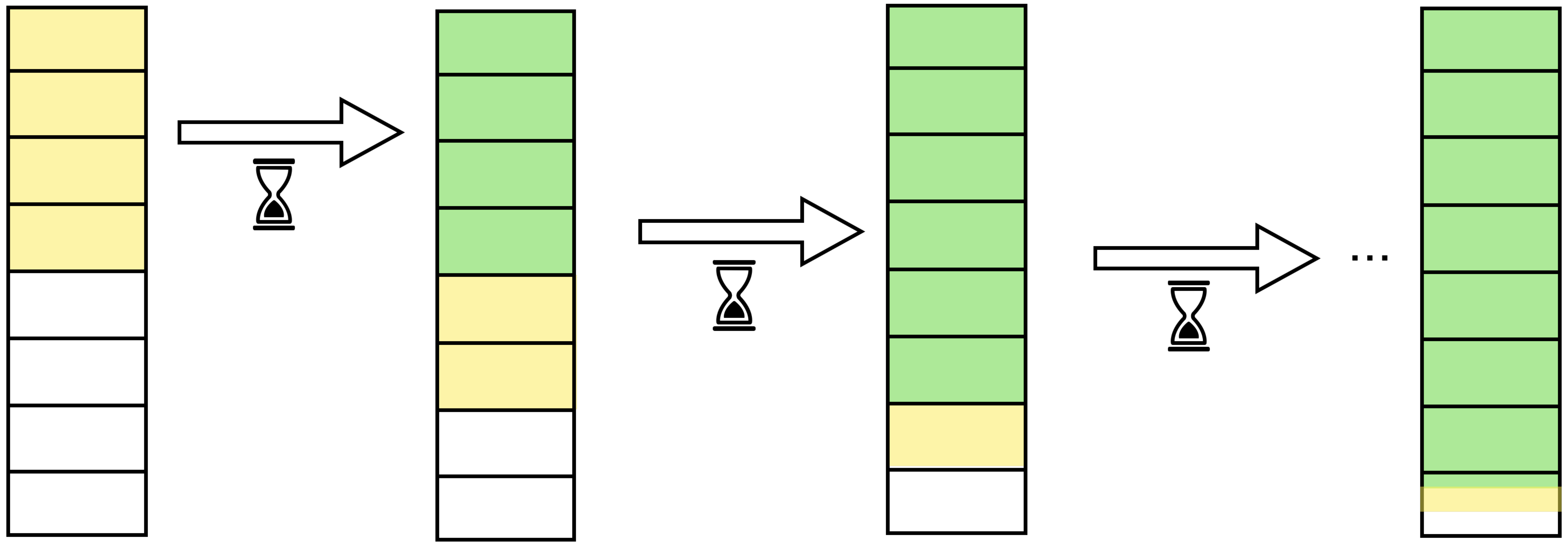
Summary

Summary

- We consider average-reward restless bandits.

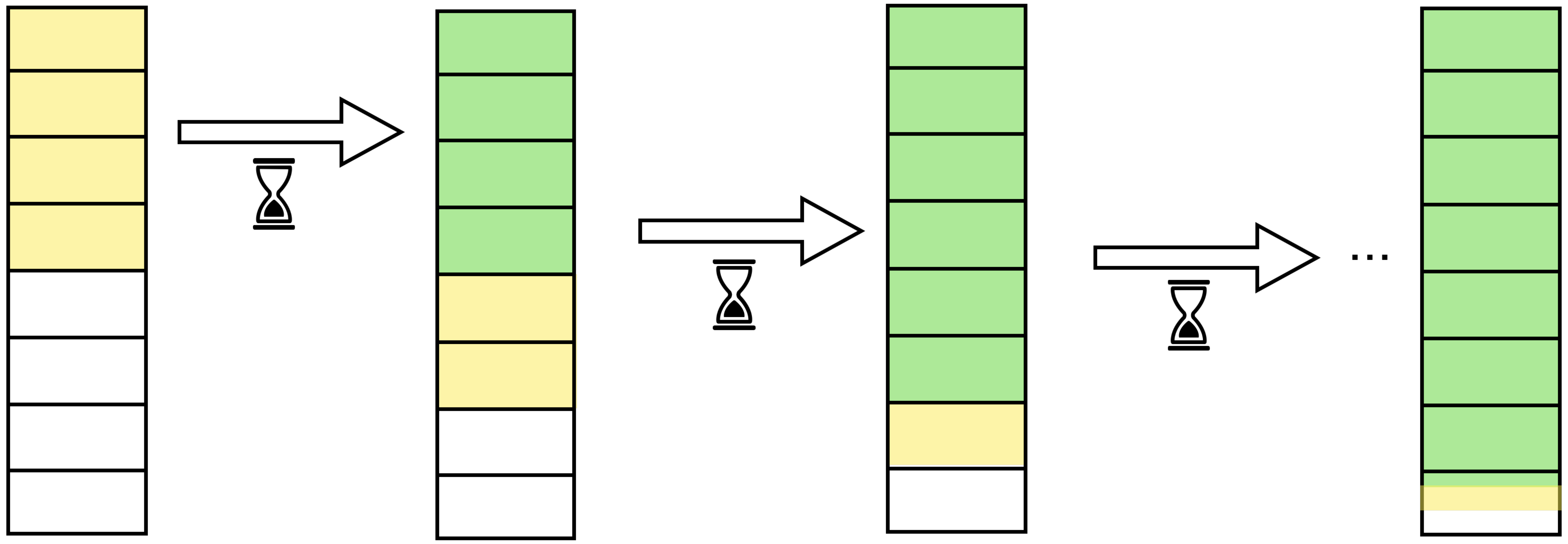
Summary

- We consider average-reward restless bandits.
- We propose asymptotically optimal policy without assuming global attractor. The policy



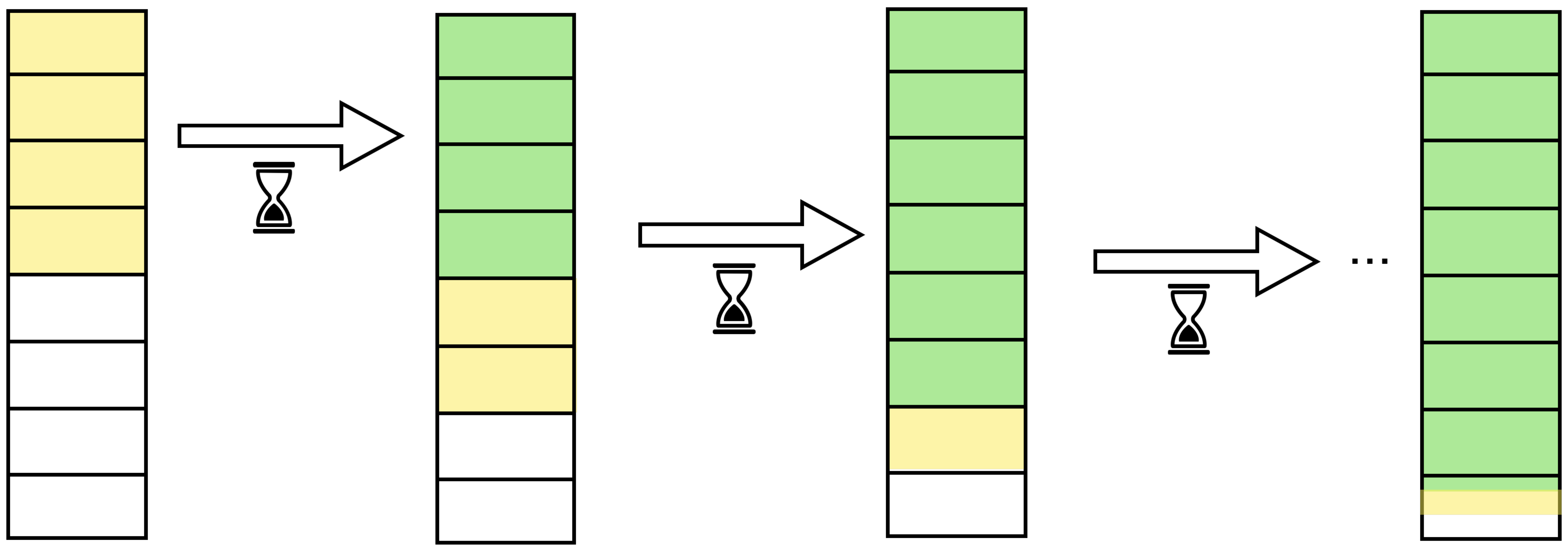
Summary

- We consider average-reward restless bandits.
- We propose asymptotically optimal policy without assuming global attractor. The policy
  - globally drives the distribution to μ^* on its own,



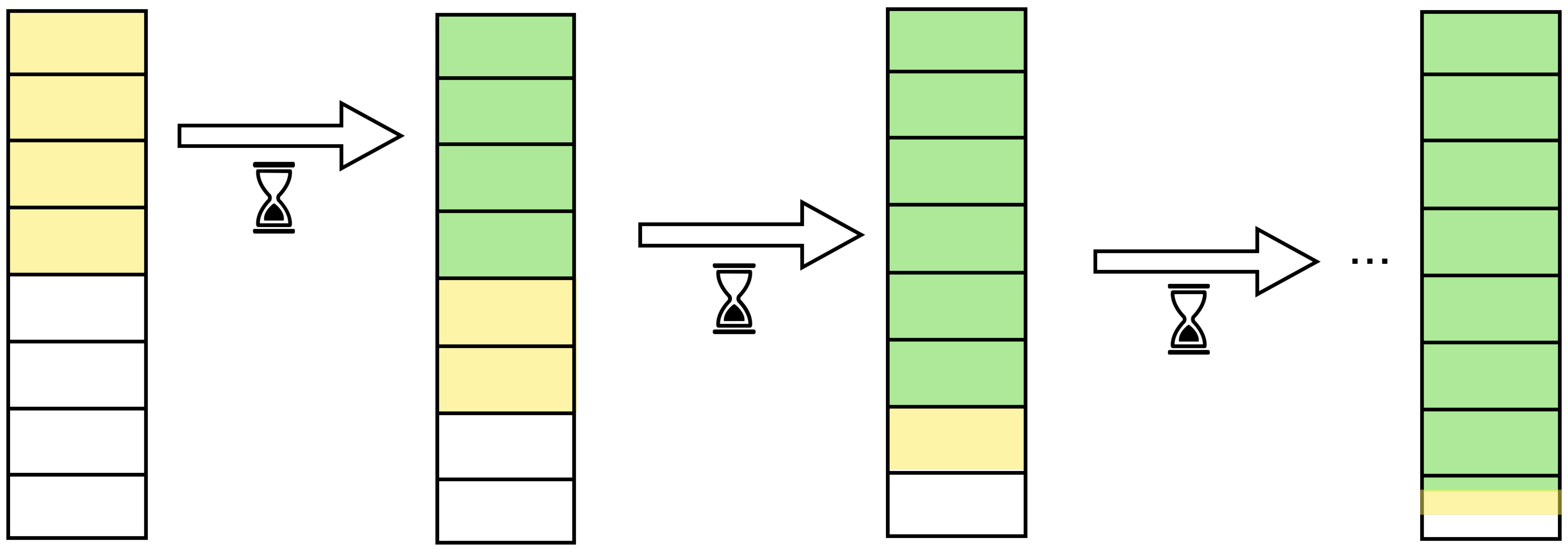
Summary

- We consider average-reward restless bandits.
- We propose asymptotically optimal policy without assuming global attractor. The policy
  - globally drives the distribution to μ^* on its own,
  - utilizing the optimal single-armed policy $\bar{\pi}^*$ to drive each arm to μ^*, following a simple schedule.



Summary

- We consider average-reward restless bandits.
- We propose asymptotically optimal policy without assuming global attractor. The policy
    - globally drives the distribution to μ^* on its own,
    - utilizing the optimal single-armed policy $\bar{\pi}^*$ to drive each arm to μ^*, following a simple schedule.
- We have a novel Lyapunov analysis of the policy using "focus sets" and bivariate Lyapunov functions.



Summary

Thank you!

- We consider average-reward restless bandits.
- We propose asymptotically optimal policy without assuming global attractor. The policy
  - globally drives the distribution to μ^* on its own,
  - utilizing the optimal single-armed policy $\bar{\pi}^*$ to drive each arm to μ^*, following a simple schedule.
- We have a novel Lyapunov analysis of the policy using "focus sets" and bivariate Lyapunov functions.