# The Value Problem for Multiple-Environment MDPs with Parity Objectives

## IFIP Meeting - Aachen - September 2025

originally presented at ICALP 2025, Aarhus, Denmark

Krishnendu Chatterjee, Laurent Doyen, **Jean-François Raskin** and Ocan Sankur

# Markov Decision Processes

## The classical model

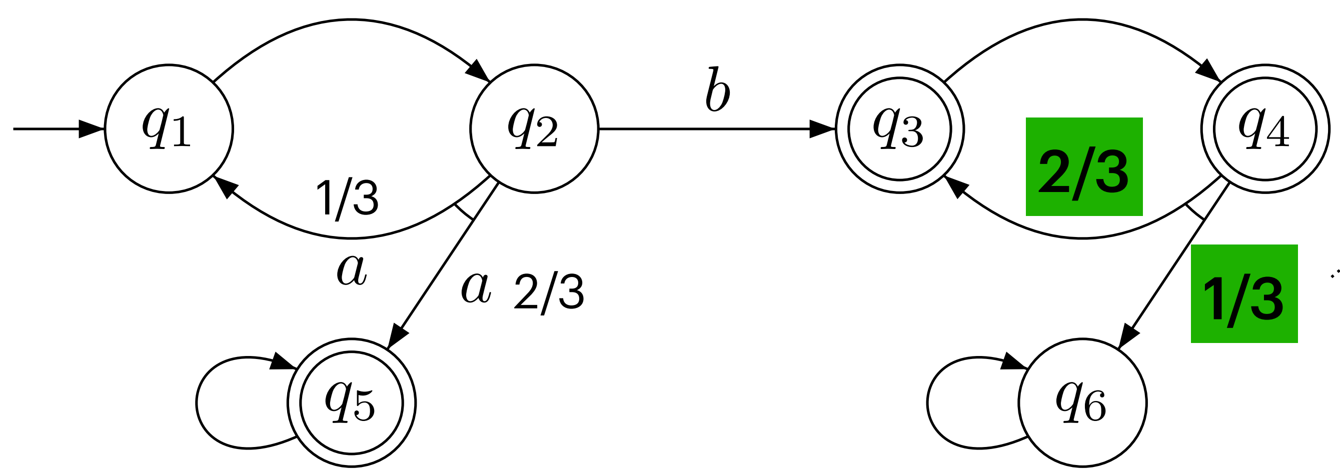
$$M = (Q, (A_q)_{q \in Q}, \delta)$$

$$\delta : Q \times A \rightharpoonup \text{Dist}(Q)$$

States

Available actions

Prob. transition function




**Q** : $\exists$ **strategy** $\sigma : Q^* \to A$ to visit Büchi states (parity condition) $\infty$-often with prob. at least $\alpha$ from $q$ ?

Let $\sigma$ be s.t. $\sigma(h \cdot q_2) = a$ then $M_q \times \sigma \vDash \mathbb{P}\left(\Box \Diamond \{q_3, q_4, q_5\} = 1\right)$

Markov chain

# Markov Decision Processes

## The classical model

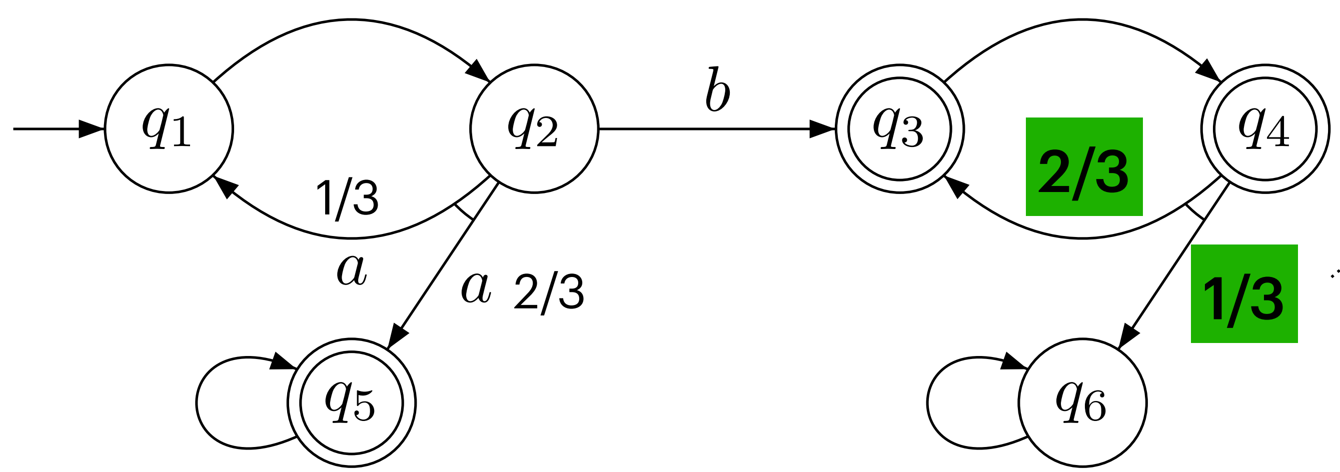
$$M = (Q, (A_q)_{q \in Q}, \delta)$$

$$\delta : Q \times A \rightharpoonup \mathrm{Dist}(Q)$$

States

Available actions

Prob. transition function

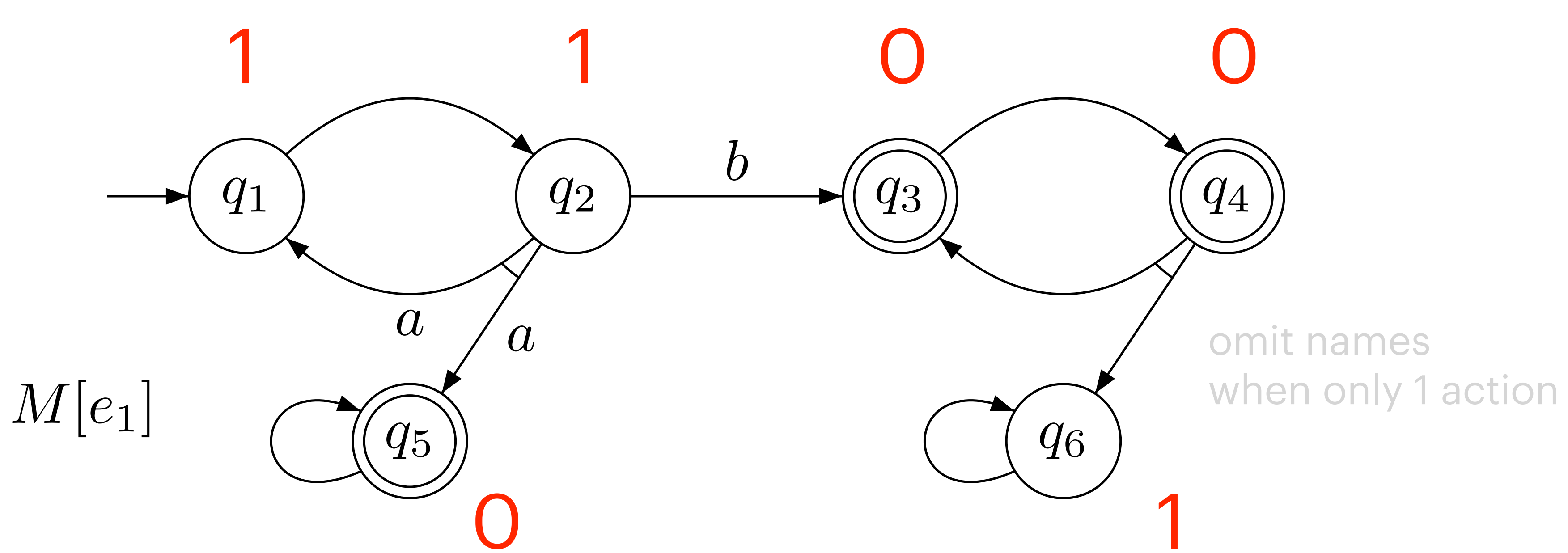



**Q** : $\exists$ **strategy** $\sigma : Q^* \to A$ to visit Büchi states (parity condition) ∞-often with prob. at least $\alpha$ from $q$ ?

Let $\sigma$ be s.t. $\sigma(h \cdot q_2) = a$ then $\mathbb{P}_q^{\sigma}(M, \Box \Diamond \{q_3, q_4, q_5\}) = 1$

# Markov Decision Processes
## Parity objectives

- A parity objective $\varphi$ is defined by a coloring function $p : Q \to \mathbb{N}$
- The color of a path $\rho = q_1 q_2 \dots q_n \dots$, noted $p(\rho)$, is the **minimal** color that appears **infinitely many times** along $\rho$, and $\rho \models_p \varphi$ is winning iff $p(\rho) \in$ Even
- Given a MDP $M$, a state $q$, and a threshold $\alpha$: $\exists^? \sigma \cdot \mathbb{P}_q^\sigma(M, \varphi) \geq \alpha$ can be decided in **PTime**



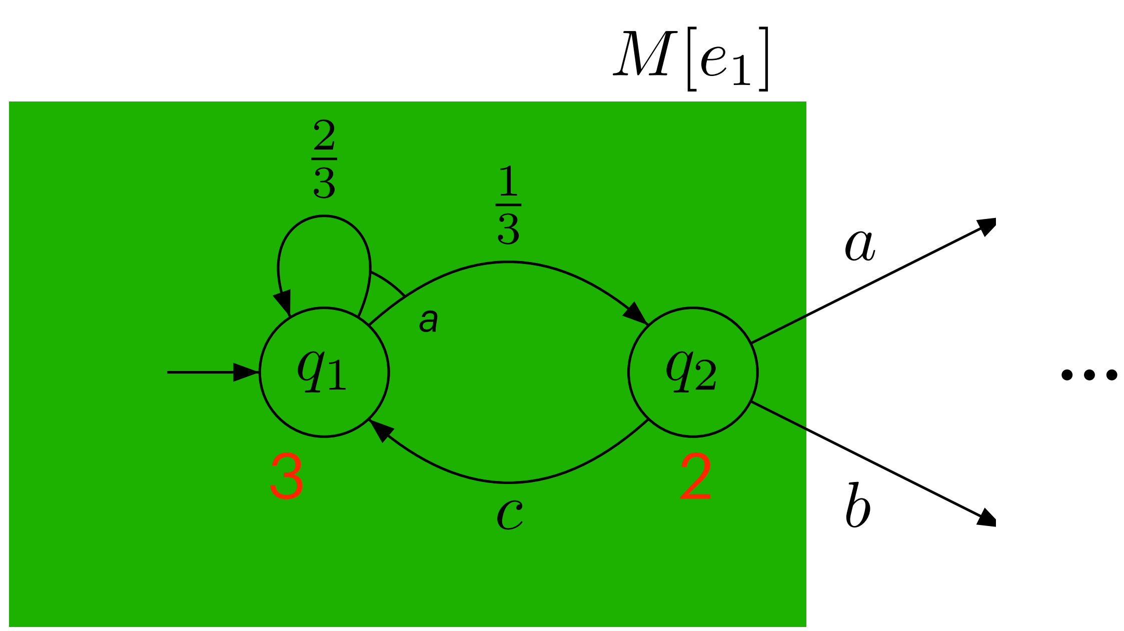
# Markov Decision Processes

## Solving parity objectives - End Components

Let $M = (Q, (A_q)_{q \in Q}, \delta)$, an End-Component (EC) of $M$ is a pair $(S, (B_q)_{q \in S})$ where:

- the graph induced by $S$ and $(B_q)_{q \in S}$ is strongly connected
- $\forall q \in S \cdot \forall a \in B_q$ : Supp$(\delta(s, a)) \subseteq S$

$\left(\{q_1, q_2\}, B_{q_1} = \{a\}, B_{q_2} = \{c\}\right)$ is a EC




Playing all actions in $(B_q)_{q \in S}$ uniformly at random ensures to visit all states of the EC with probability 1.

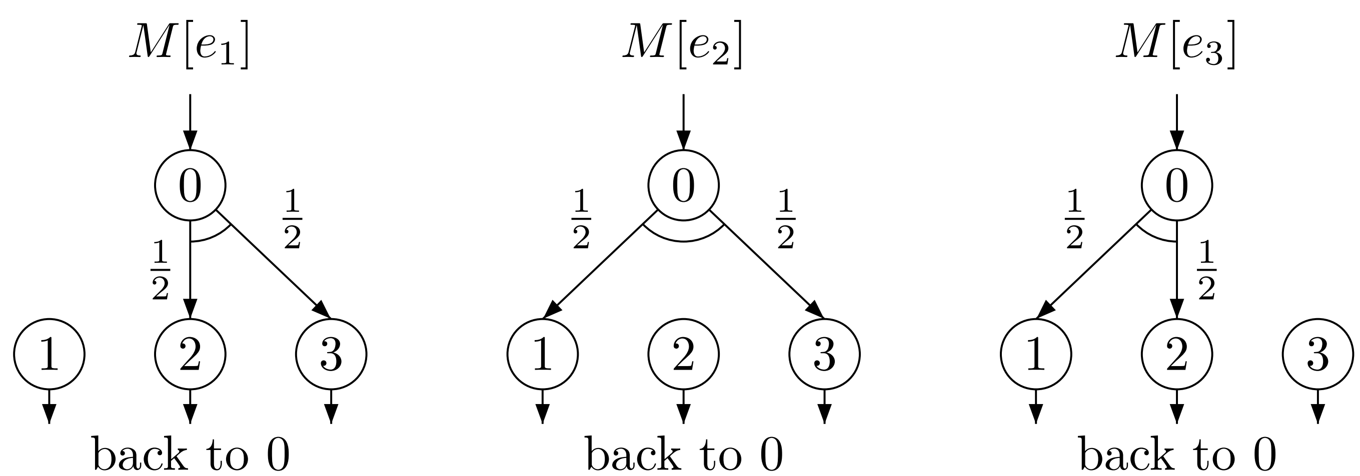
All states in an EC are either winning the parity objective with prob. 1 (if min. color in EC is even) or with prob. 0 (if min. color is odd).

Solving parity objectives in MDP reduces to maximizing the prob. of reaching EC with value 1.

# The Model of Multi-env. MDPs

# Multi-Environment MDPs
## The model

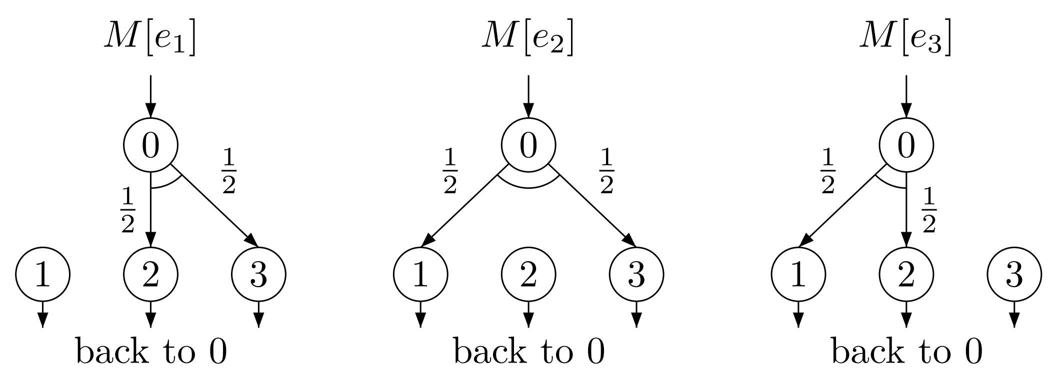
- A multi-env. MDP $M = (Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E})$ = collection of $|E|$ MDPs
- Same state space but **different next state transition functions**
- The controller does **not** observe which $e \in E$ governs the dynamics but **fully observes states**




- **Robust control**: (unique) strategy that enforces a specification (**parity condition**) in **all** environments: $\exists^{?} \sigma \cdot \forall e \in E : \mathbb{P}_q^{\sigma}(M[e], \varphi_p) \geq \alpha$

# Multi-Environment MDPs

## What can be modeled?

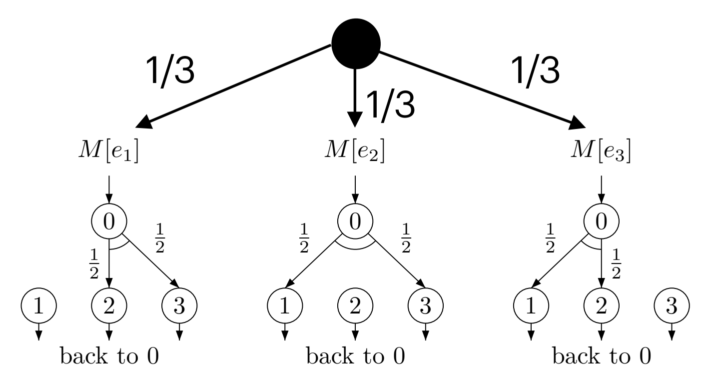



- MEMDPs can be used to model:
  - e.g. several **types** of users, different **types** of patients, etc.
  - finite number of **valuations** for a **parametric** MDP
  - an adversary playing a strategy taken from a **finite set of** finite memory randomized **strategies**

# Multi-Environment MDPs

## vs Partially Observable MDPs

- The states are **fully observable** but environment is **not**
- This is a *variant* of **Partially Observable MDPs (POMPDs)**
  *% env. is chosen adversarially not stochastically*
- Some of the problems that we want to study on MEMDP can be reduced to problems on POMDPs
  *% uniformly choosing the environment*

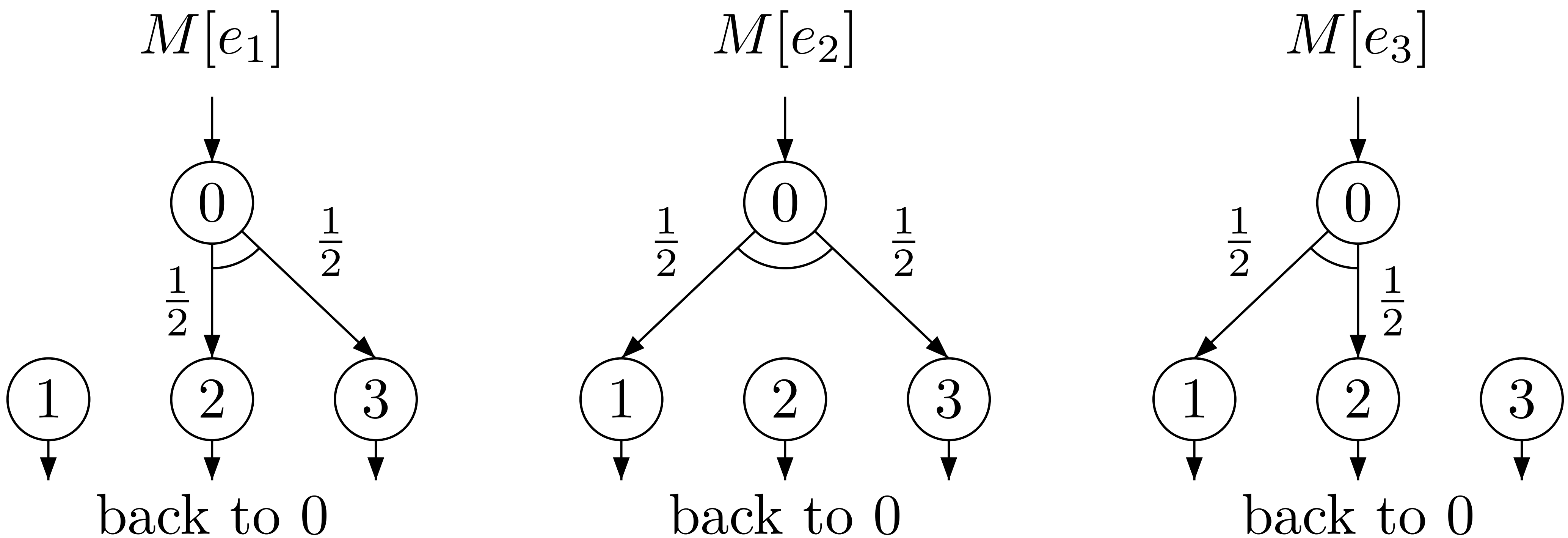



- ... but those problems are **harder** on POMDPs
  *% for instance almost sure reachability is already ExpTime-C and limit-sure reachability is undecidable, almost sure co-Büchi, and so parity, are undecidable*
- So, we study **dedicated algorithms** instead and settle the complexity of the problems

# Modeling examples and decision problems

# Multi-Environment MDPs

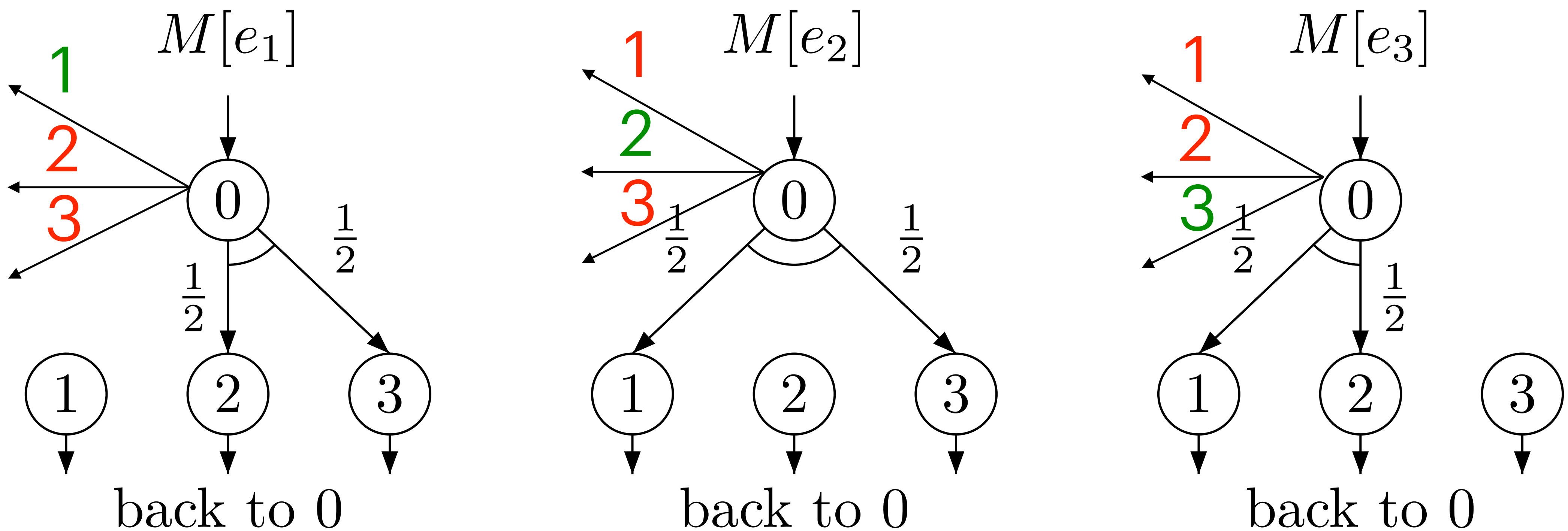
## What can be modeled?




A card deck with one **missing** card

# Multi-Environment MDPs

## What can be modeled?

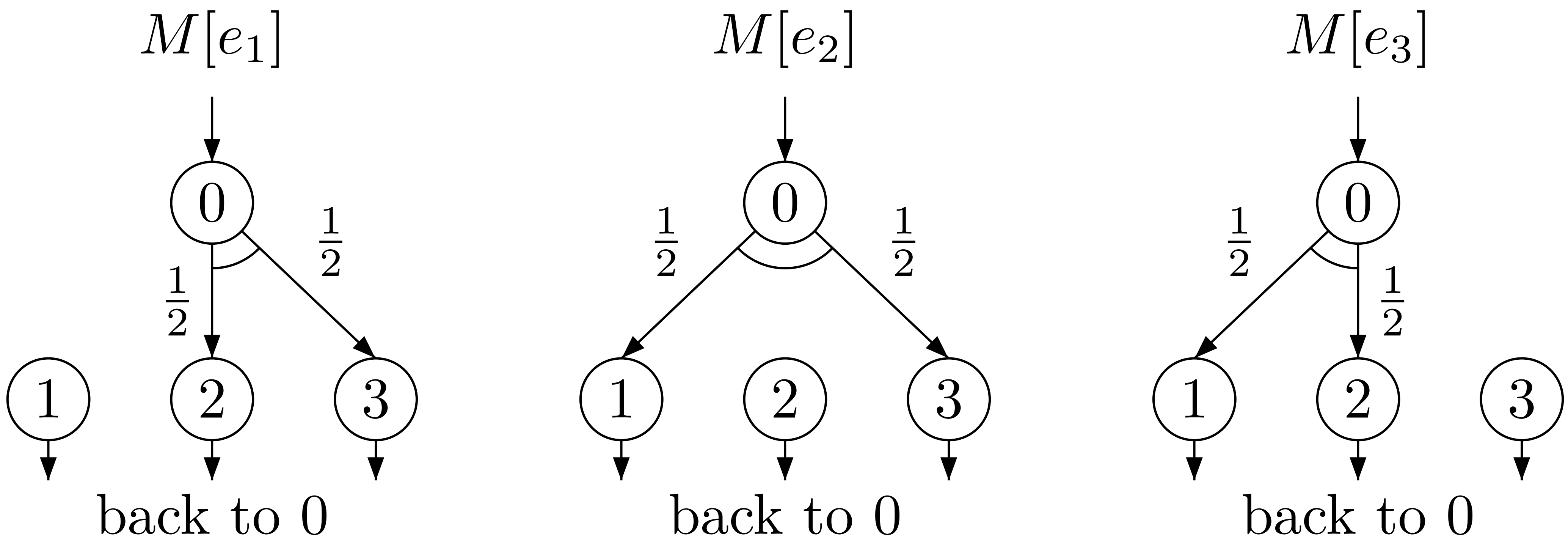



A card deck with one **missing** card

**Q :** Can we discover the missing card? With which probability?

# Multi-Environment MDPs

## What can be modeled?



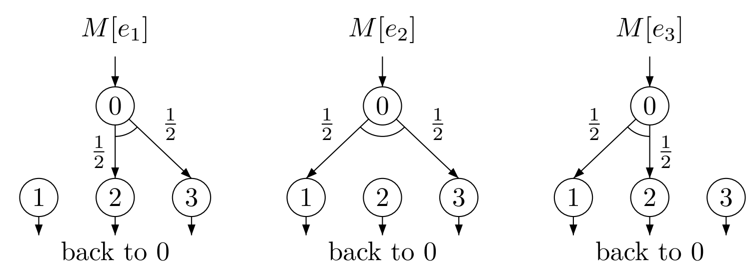

A card deck with one **missing** card

**Q :** Can we discover the missing card? With which probability?

Yes, with probability **one** (**almost surely**) !

# Multi-Environment MDPs

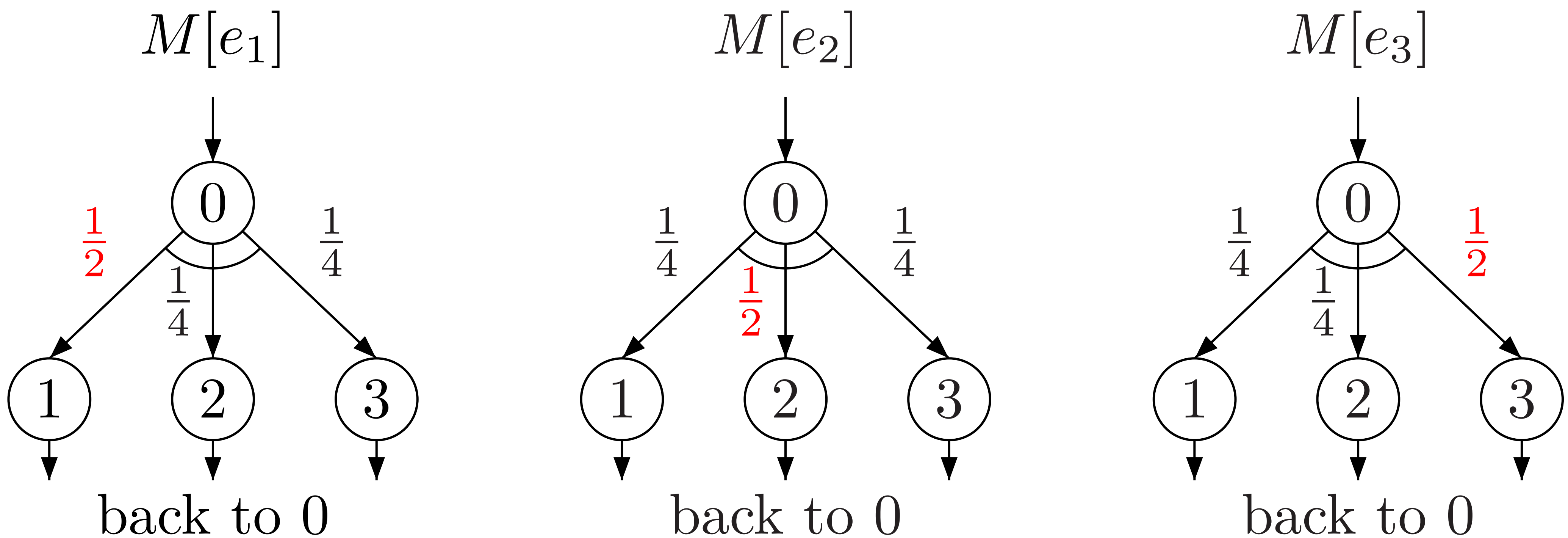
## First example: Modeling a deck of card with one **missing** card




- **Claim**: $\exists\sigma$ to discover missing card with probability 1 (**almost surely**)
  - The MEMDP starts in an **unknown** environment $e \in E$ (chosen adversarially)
  - $\sigma$ draws cards at random and **records** edges seen so far
    Important concept: **revealing edge**
    if $0 \to 1$, then env. in not $e_1$ (knowledge $K \in 2^E \backslash \varnothing$)
  - The two edges appearing in $e \in E$ are eventually revealed, with prob. 1
  - At that time, we know $e \in E$, i.e. what is the missing card !

# Multi-Environment MDPs

## What can be modeled?

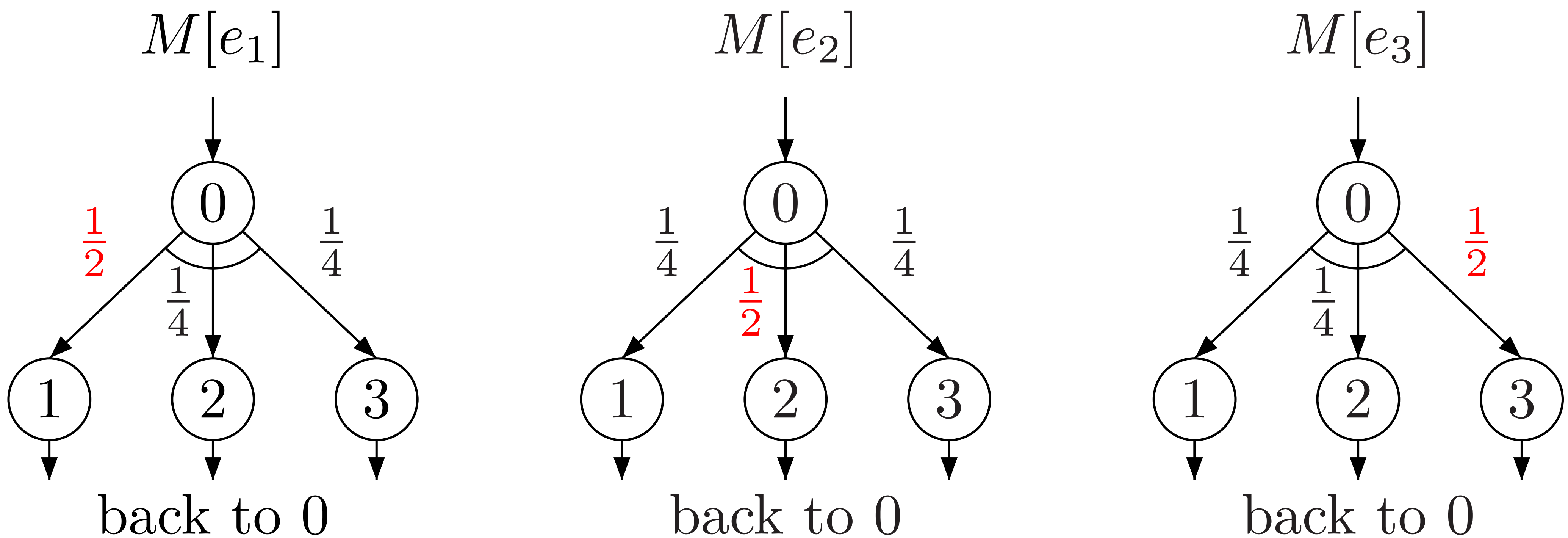



A card deck with **duplicated** card

**Q** : Can we discover the duplicated card? With which probability?

# Multi-Environment MDPs

## What can be modeled?




A card deck with **duplicated** card

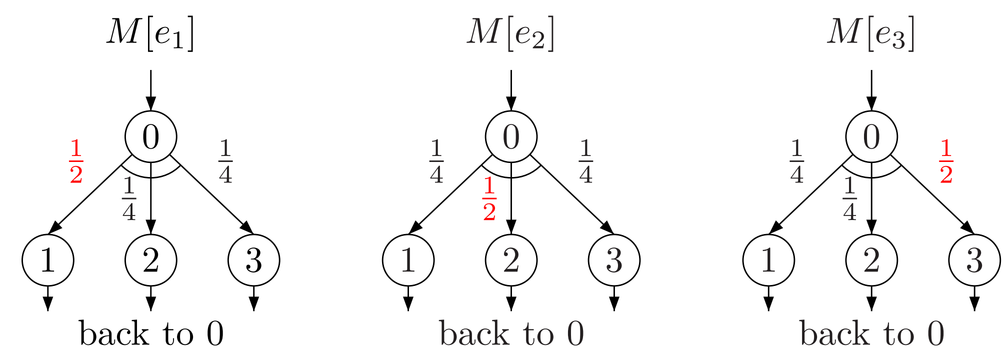
**Q** : Can we discover the duplicated card? With which probability?

Yes, but only with high probability (arbitrarily close to one - **limit surely**)

but **not** with prob. 1

# Multi-Environment MDPs

## Modeling a deck of card with one duplicated card



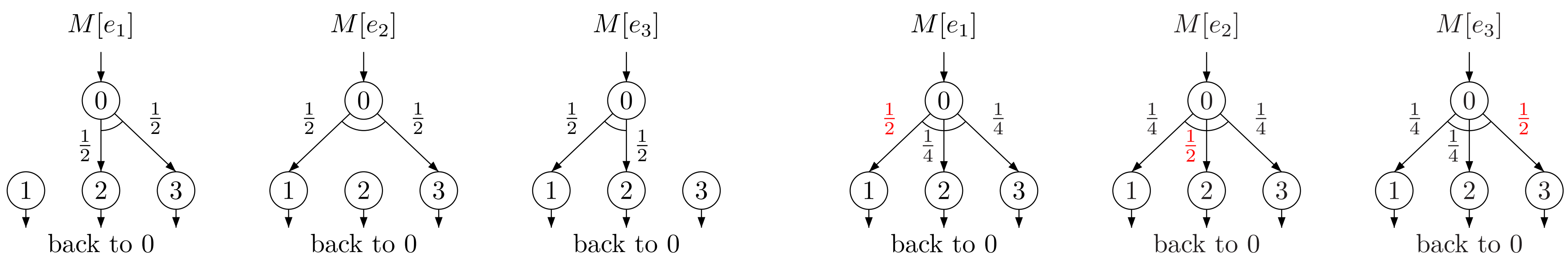

- **Claim**: $\forall \epsilon > 0 \cdot \exists \sigma_\epsilon$ to discover duplicated card with probability $1 - \epsilon$ (**limit surely**)
  - MEMDP starts in an **unknown** environment $e \in E$ (chosen adversarially)
  - $\sigma$ draws cards at random and **records statistics about frequency** of edges seen so far
  - Hoeffding's inequality: $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \delta\right) \leq 2\exp\left(-2n\delta^2\right)$,

    *i.e. Hoeffding's inequality provides a bound on the prob. that the sum of random variables $X_i$ deviates from its expected value*
  - For any given $\epsilon > 0$, we can determine a number of draws $n$ sufficient to determine with probability $p \geq 1 - \epsilon$ the active environment, and so the duplicated card !

    *% similar to PAC learning*

# Multi-Environment MDPs

## The model and decision problems




- Given $M = (Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E})$, and a state $q$, a parity objective $\varphi$ decide:
  - for **almost sure** winning: if there exists a **strategy** $\sigma$ s.t.
    for all $e \in E$, $\mathbb{P}_q^{\sigma}(M[e], \varphi) = 1$
  - for **limit sure** winning: if for all $\epsilon > 0$, there exists a **strategy** $\sigma_\epsilon$ s.t.
    for all $e \in E$, $\mathbb{P}_q^{\sigma_\epsilon}(M[e], \varphi) \geq 1 - \epsilon$
  - for **threshold** $\alpha$: if there exists a **strategy** $\sigma$ s.t.
    for all $e \in E$, $\mathbb{P}_q^{\sigma}(M[e], \varphi) \geq \alpha$

# Main results

# Main results - Almost sure and limit sure

## Complexity

- **Theorem (Almost Sure)** [SVJ24].
  Membership problem is **PSPACE-complete**; solvable in PTime if number of environments is fixed. *Pure exponential-memory strategies suffice*.

- **Theorem (Limit Sure)**.
  Membership problem is **PSPACE-complete**; solvable in PTime if number of environments is fixed. From a LS winning state, for any $\varepsilon > 0$, pure exponential-memory strategies suffice to ensure the objective with probability $\geq 1-\varepsilon$.

# Main results - Threshold problem

## Gap version

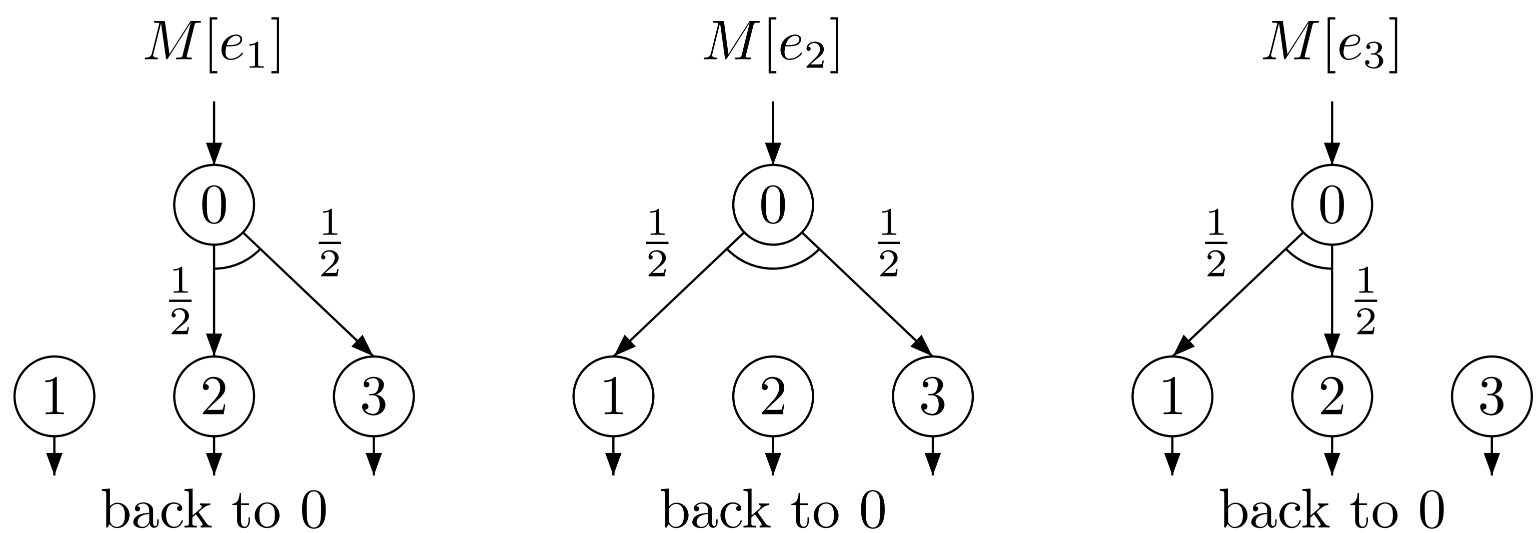
- We leave **open** the decidability of the threshold problem.
- We show how to **solve** a relaxation: the **gap problem** (a.k.a. *promise problem*). Given $0 < \alpha < 1$ and $\epsilon > 0$, a MEMDP $M$, a state $q$, parity objective $\varphi$, answers:
    - **Yes**, if there exists a strategy $\sigma$ such that for all $e \in E$, we have $\mathbb{P}_q^\sigma(M[e], \varphi) \geq \alpha$
    - **No**, if for all strategies $\sigma$, there exists $e \in E$ with $\mathbb{P}_q^\sigma(M[e], \varphi) < \alpha - \epsilon$
    - and **arbitrarily** otherwise
- This can be used to approximate arbitrarily closely the max. realizable threshold

*% see details in the paper*

# Main algorithmic ideas

# Revealing Edges

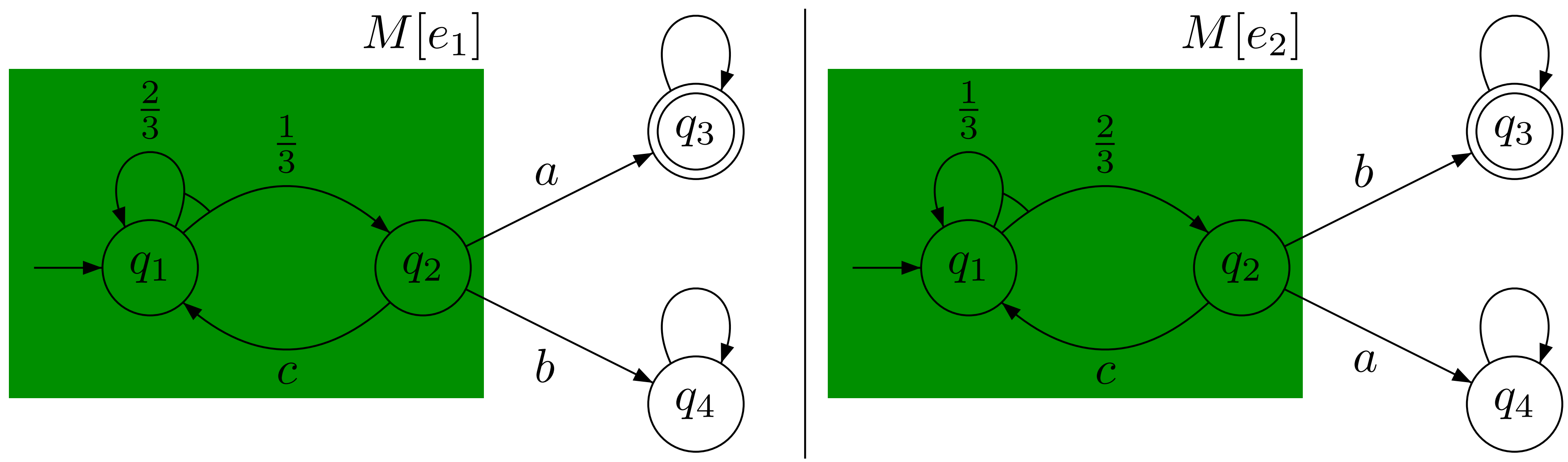
## Improve knowledge about environment **with certainty**




- if $0 \to 1$, then env. in not $e_1$ (knowledge $K \in 2^E \backslash \varnothing$)

# Distinguishing Common End-Component

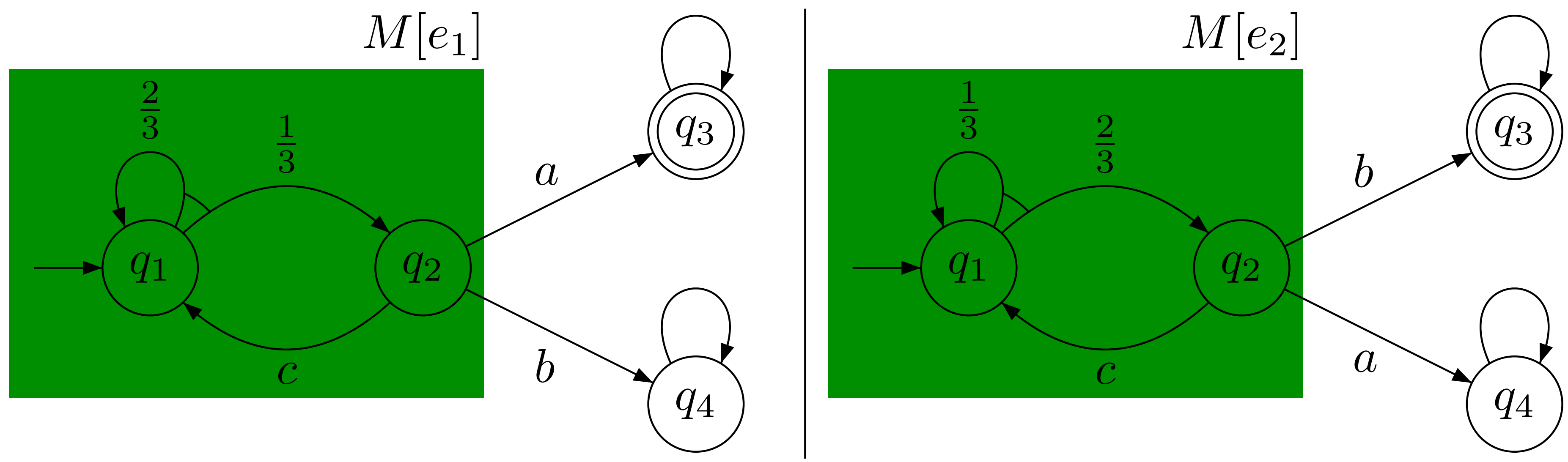
**How to learn about the environment in charge with high probability ?**




Let $M = (Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E})$ be multi-env MDP. A pair $(Q', (A_q)_{q \in Q'})$ is an End-Component (EC) in $e \in E$: that is the subMDP $\left(Q', (A_q)_{q \in Q'}, \delta_e^{(Q', (A_q)_{q \in Q'})}\right)$ is strongly connected.

# Distinguishing Common End-Component

**How to learn about the environment in charge with high probability ?**




A **Common End-Component** (CEC) is a pair $(Q', (A_q)_{q \in Q'})$ that is an EC in all $e \in E$.

It is **distinguishing** if it contains a transition $(q, a, q')$ such that $\delta_e(q,a)(q') \neq \delta_f(q,a)(q')$ for some env. $e, f \in E$.

Distinguishing CEC allows to learn! By observing frequencies of next states, we can guess the correct environment with high probability!

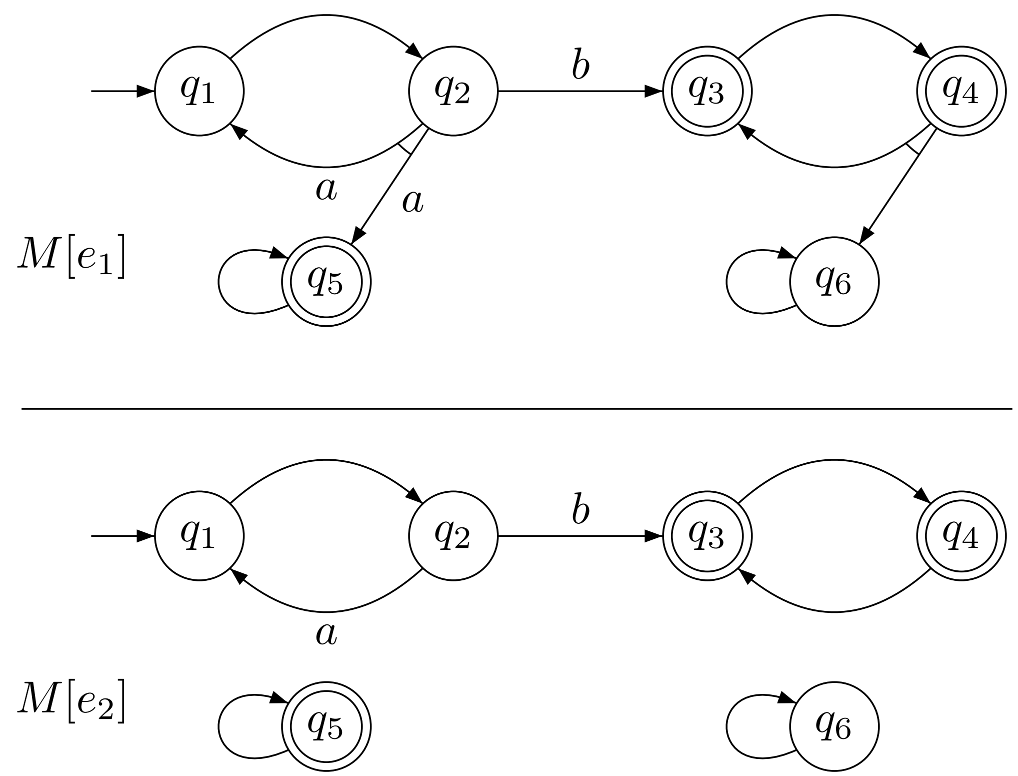
# Algorithmic ideas

## Knowledge and recursion

- The **knowledge** is the set of environment that are potentially active
- Initially: $K_0 = \{e_1, e_2, \ldots, e_n\} = E$ % env. is chosen adversarially
- Two (main) ways to improve the knowledge:
  - when crossing a revealing edge
    % one which is not present in all env. of current knowledge
  - When staying long enough into a distinguishing common-end-component
    % by collecting statistics, we can exclude some environments
- When $K$ is a singleton or all environments share the same future dynamics: we have a plain MDP that we can solve (in PTime) ! (base case)
- **Recurse** when knowledge improves

# Another example

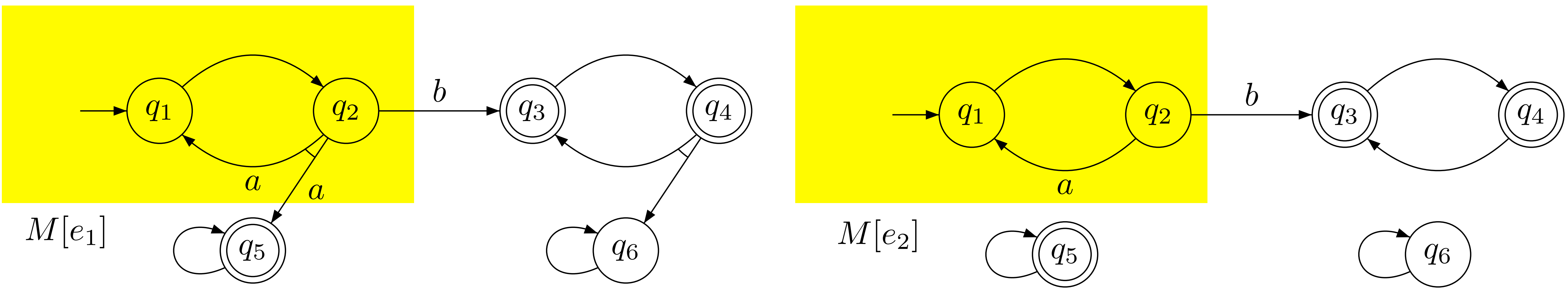
Prob. does not matter




Objective: $\square \Diamond \{q_3, q_4, q_5\}$

# Distinguishing CEC and revealing edge

## … combined …




- If the active env. is $e_2$ then playing action $a$ **never reveals** this
- Still after taking action $a$ many times in $q_2$, we win **or** we can **guess, and being correct with high probability** that active env. is $e_2$
- Playing action $a$ is OK even if $\{q_1, q_2\}$ is not a CEC because action $a$ is « safe » for limit sure winning (every LS state has a « safe » action)
  % see details in the paper

# Multi-Env. MDPs

## Conclusions

- MEMDP=**perfectly observable states** but **unknown**, fixed **environment** dynamics taken in finite set of env., modeling a.o.:
  - several types of users, different types of patients, etc.
  - finite number of valuations for a parametric MDP
  - adversary playing a strategy taken from a finite set of finite memory strategies
- **Variants** (not subclass) of POMDPs with decidable decision problems
- Algorithms for **robust** synthesis across multiple environments
- Almost-sure and limit-sure are solvable in **PSPACE**, and in **PTIME** for **fixed number of environments**. Threshold problem remains **open**, but **gap** problem is **decidable**.