The Importance of Non-Markovianity in Maximum State Entropy Exploration

Mirco Mutti*
Politecnico di Milano
Università di Bologna
@mirco_mutti

Riccardo De Santi*
ETH Zürich
@desariky

Marcello Restelli
Politecnico di Milano

* equal contribution

Motivation

Motivation

Markovian

$$\pi(a|s)$$

Non-Markovian

$$\pi(a|h)$$

$$h = (s_0, a_0, s_1, a_1, \ldots, s)$$

Motivation

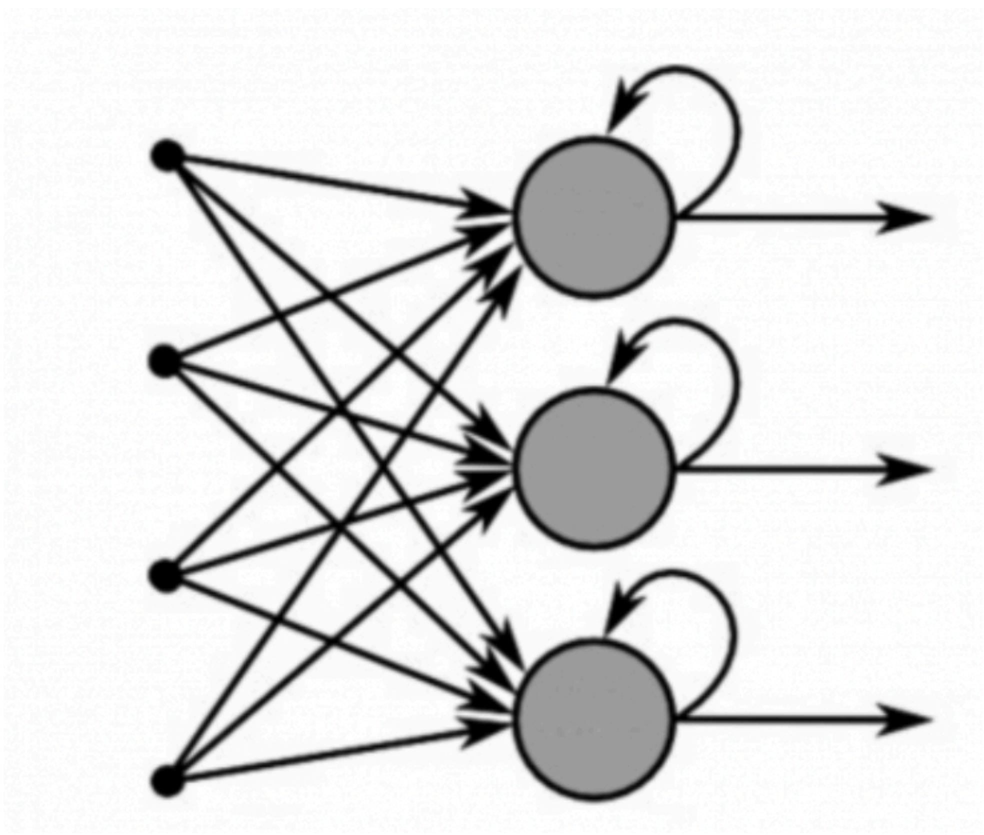
Markovian

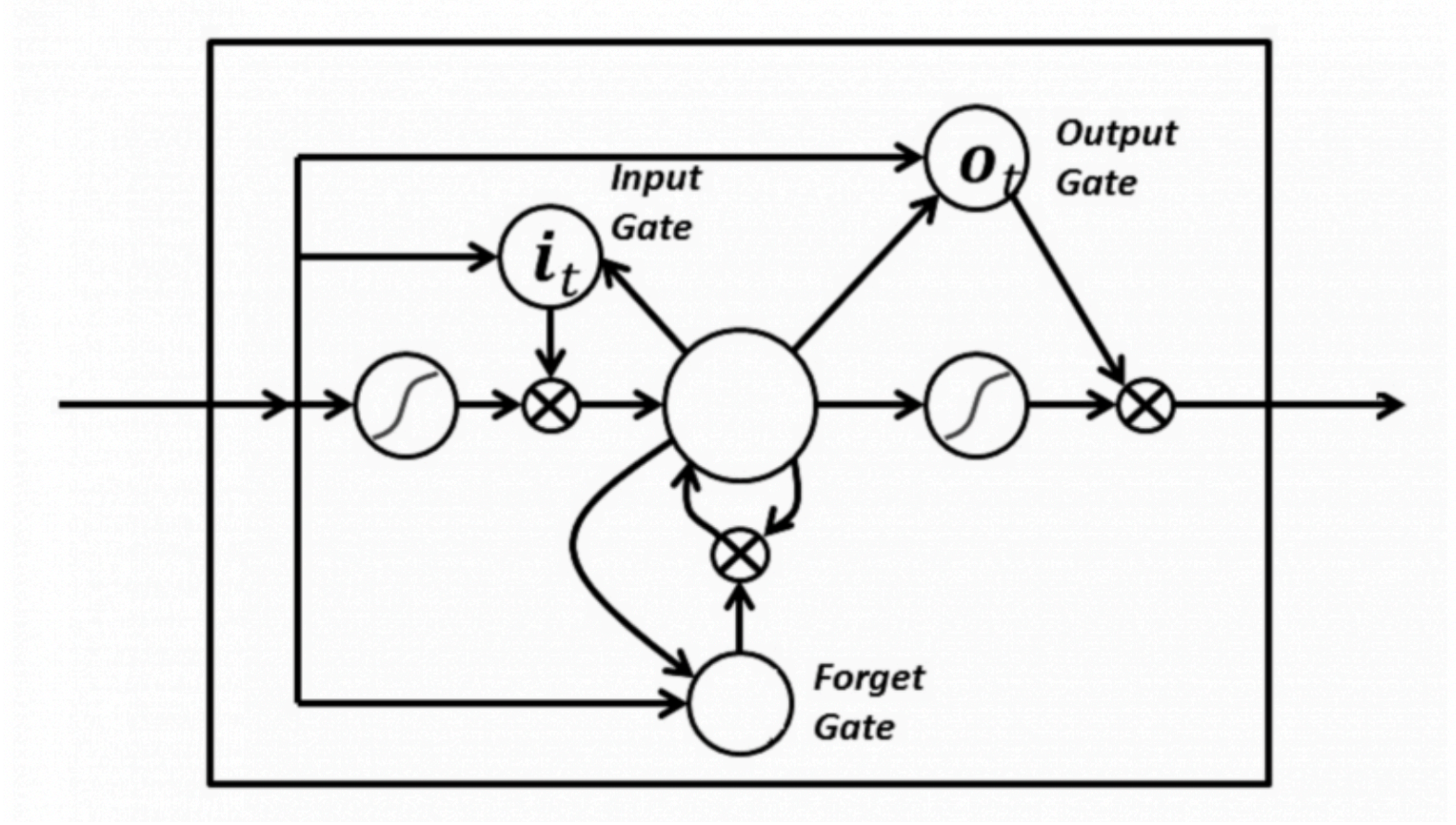
$$\pi(a|s)$$

Non-Markovian

$$\pi(a|h)$$

$$h = (s_0, a_0, s_1, a_1, \dots, s)$$





(Williams & Zipser, 1989), (Hochreiter & Schmidhuber, 1997)

Motivation

Markovian

$\pi(a|s)$

$s_0 \to a_1$

$s_1 \to a_3$

$s_2 \to a_0$

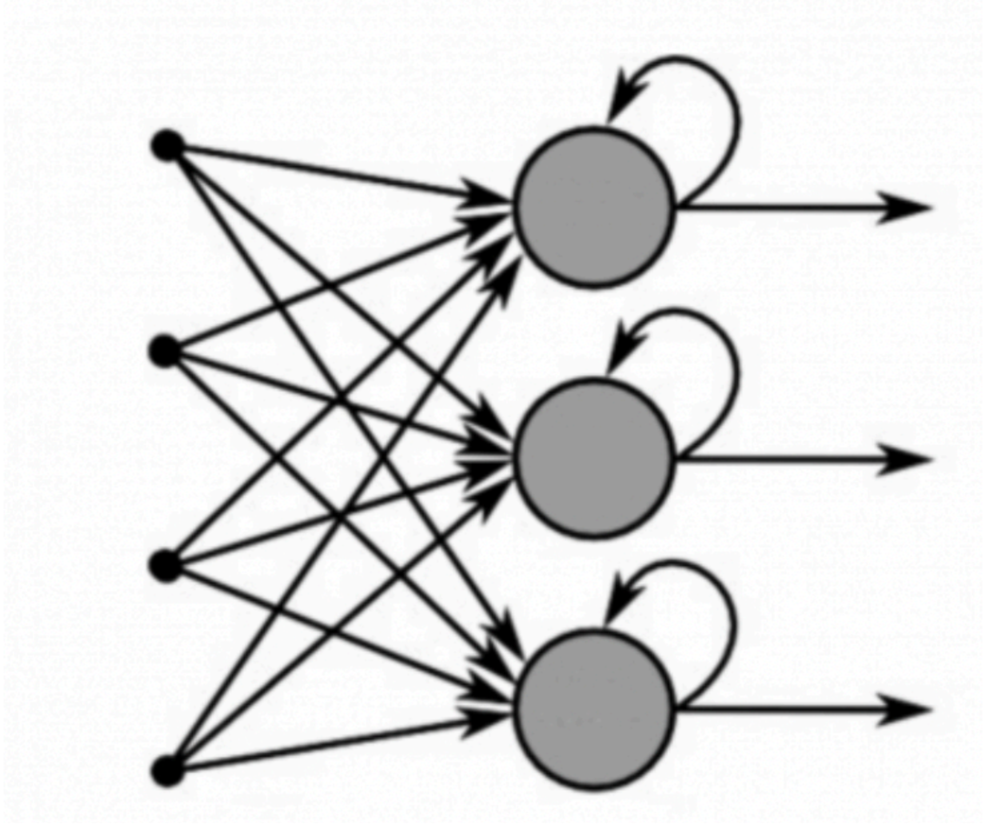
$\dots$

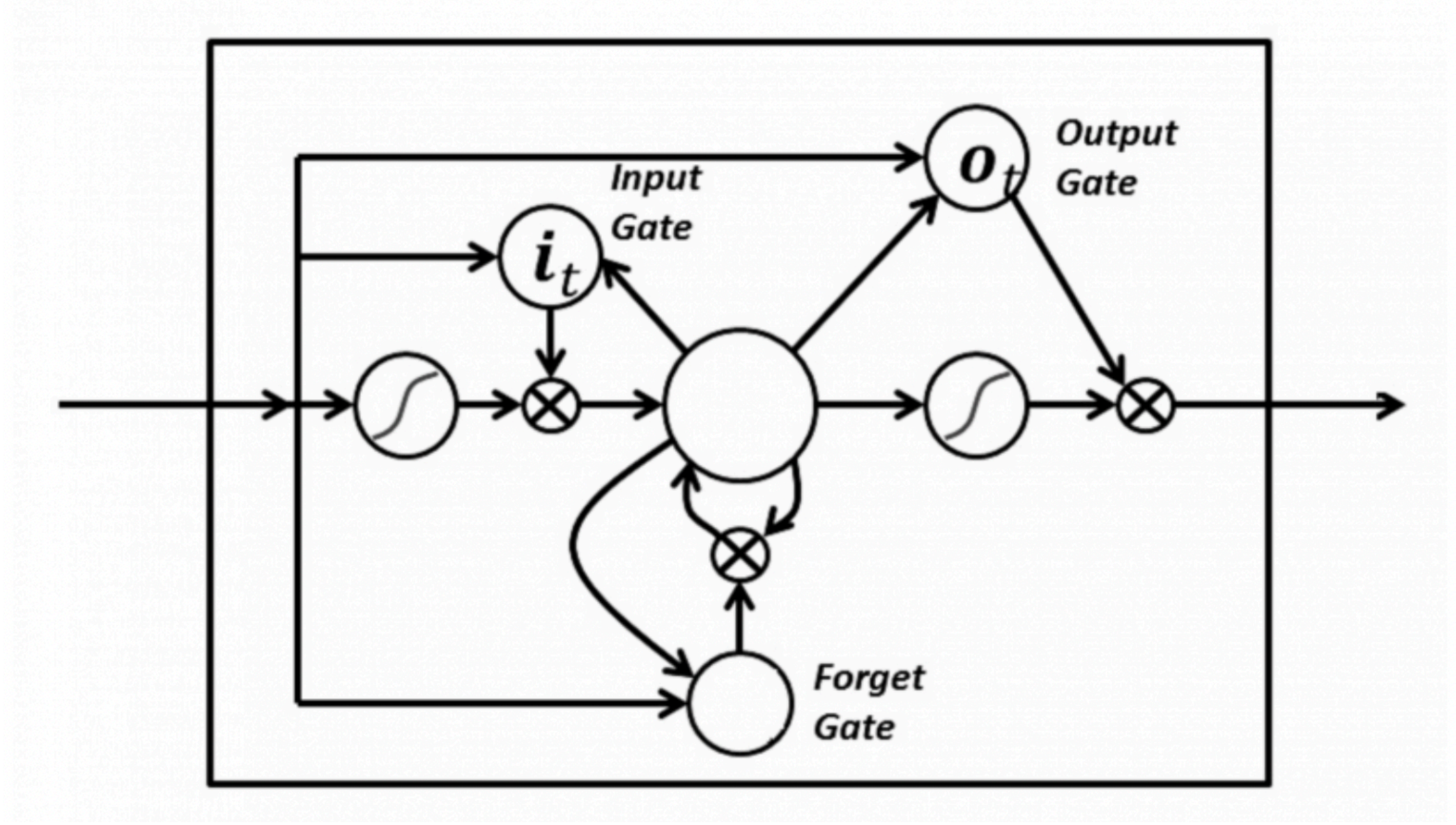
existence of an optimal deterministic Markovian policy[1]

Non-Markovian

$\pi(a|h)$

$h = (s_0, a_0, s_1, a_1, \dots, s)$





(Williams & Zipser, 1989), (Hochreiter & Schmidhuber, 1997), [1](Proposition 4.4.3, Puterman, 2014)

Motivation

Markovian

$\pi(a|s)$

$$s_0 \to a_1$$
$$s_1 \to a_3$$
$$s_2 \to a_0$$
$$\dots$$

existence of an optimal deterministic Markovian policy[1]

Non-Markovian

$\pi(a|h)$

$h = (s_0, a_0, s_1, a_1, \dots, s)$

Who cares about non-Markovian policies?

[1](Proposition 4.4.3, Puterman, 2014)

Motivation

Markovian

$\pi(a|s)$

$$s_0 \rightarrow a_1$$
$$s_1 \rightarrow a_3$$
$$s_2 \rightarrow a_0$$
$$\dots$$

existence of an optimal deterministic Markovian policy[1]

Non-Markovian

$$\pi(a|h)$$

$$h = (s_0, a_0, s_1, a_1, \dots, s)$$

Who cares about non-Markovian policies?

Partial observability

[1](Proposition 4.4.3, Puterman, 2014)

Motivation

Markovian

$\pi(a|s)$

$s_0 \to a_1$
$s_1 \to a_3$
$s_2 \to a_0$
$\dots$

existence of an optimal
deterministic Markovian policy[1]

Non-Markovian

$\pi(a|h)$

$h = (s_0, a_0, s_1, a_1, \dots, s)$

Who cares about non-Markovian policies?

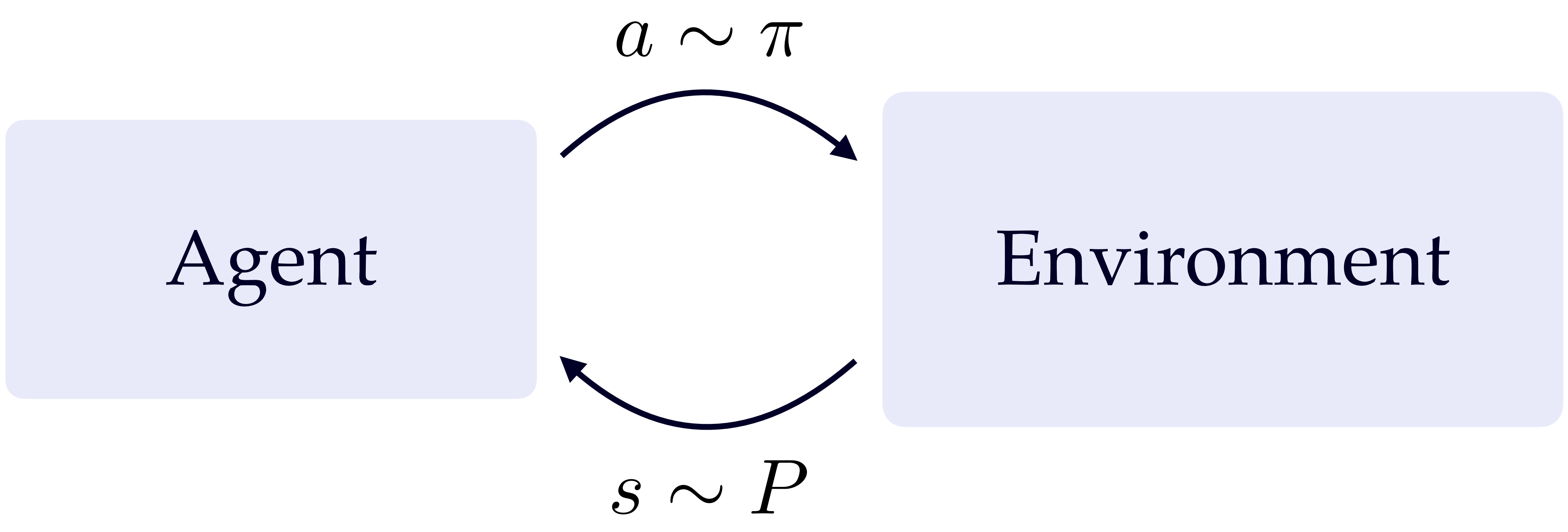
Partial observability
Imitation learning, risk-aversion, pure exploration, ... ?

[1](Proposition 4.4.3, Puterman, 2014)

Problem Setting

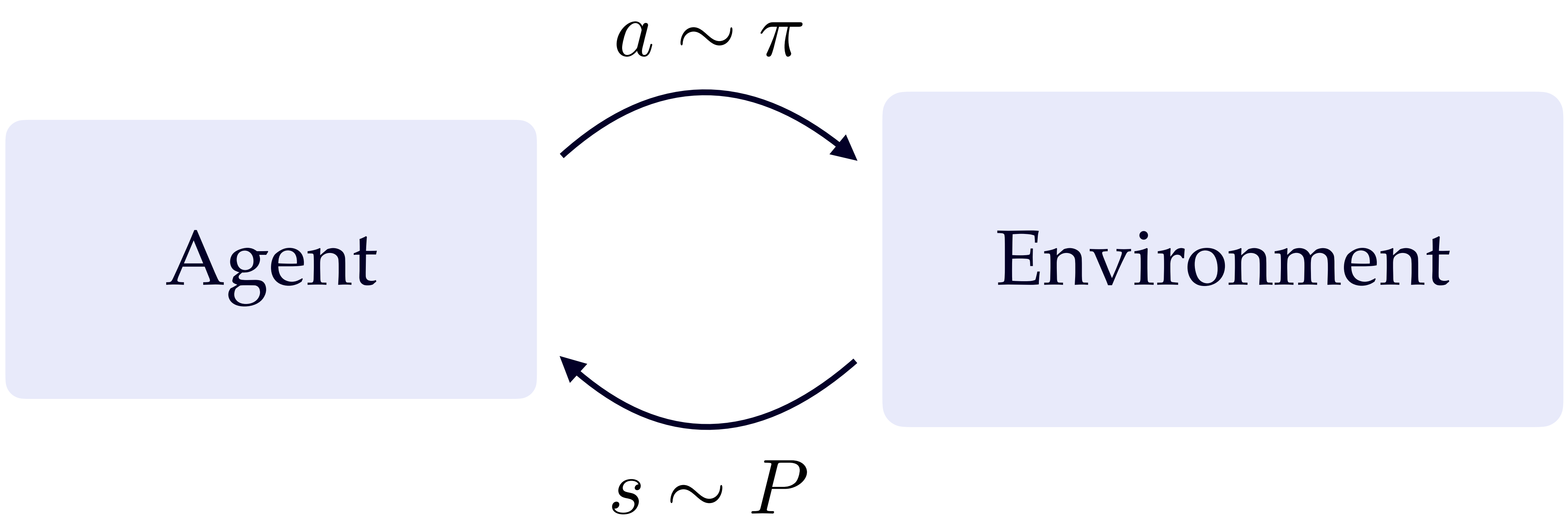
Problem Setting

Controlled Markov Process (CMP)



Problem Setting

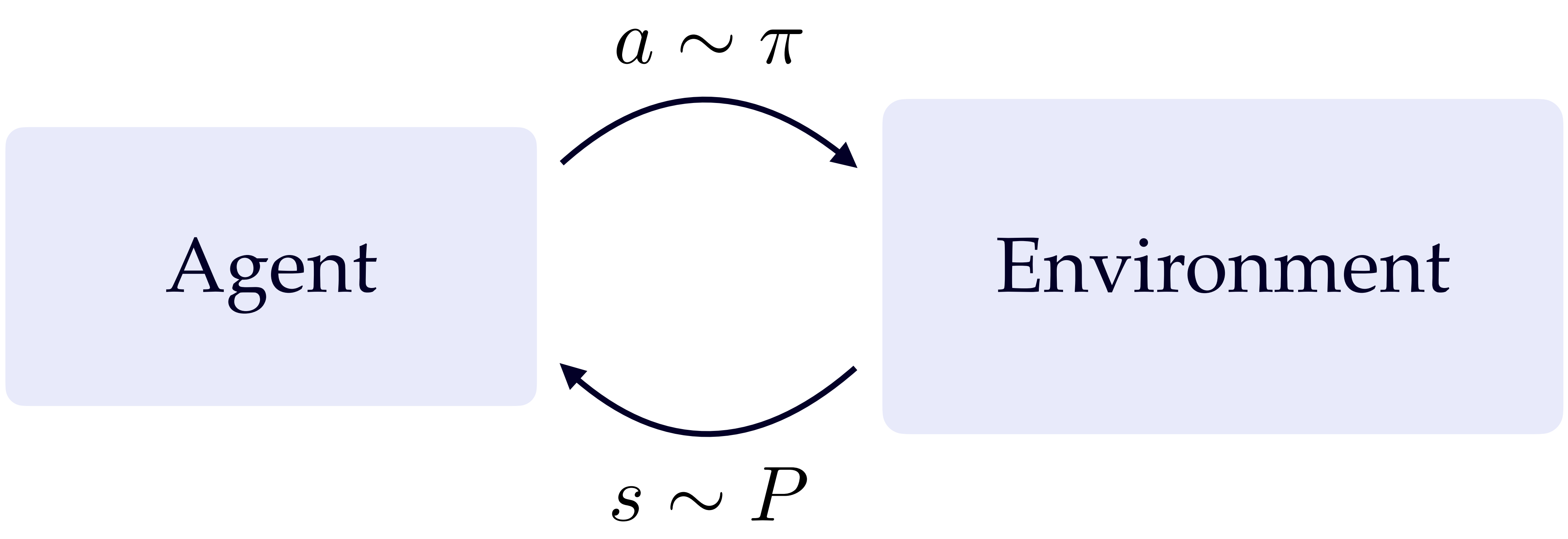
Controlled Markov Process (CMP)



$\mathcal{S}$	discrete set of states
$\mathcal{A}$	discrete set of actions
P	transition matrix
μ	initial state distribution
T	episode horizon

Problem Setting

Controlled Markov Process (CMP)



Π_{NM} set of **non-Markovian** policies

$$\pi : \mathcal{H} \rightarrow \Delta(\mathcal{A})$$

$$\mathcal{H} := \mathcal{S} \times \mathcal{S} \times \ldots$$

Π_{M} set of **Markovian** policies

$$\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$$

$\mathcal{S}$ discrete set of states

$\mathcal{A}$ discrete set of actions

P transition matrix

μ initial state distribution

T episode horizon

Problem Setting

policy π + CMP

$\downarrow$

marginal state distribution

$$d^{\pi}(s) = \frac{1}{T} \sum_{t \in [T]} Pr(s_t = s)$$

Problem Setting

policy π + CMP

$\downarrow$

marginal state distribution

$$d^\pi(s) = \frac{1}{T} \sum_{t \in [T]} Pr(s_t = s)$$

Reinforcement Learning (RL)

$$\mathcal{J}(\pi) = d^\pi \cdot R$$

Problem Setting

policy π + CMP

$\downarrow$

marginal state distribution

$$d^\pi(s) = \frac{1}{T}\sum_{t\in[T]} Pr(s_t = s)$$

Reinforcement Learning (RL)

$$\mathcal{J}(\pi) = d^\pi \cdot R$$

Convex Reinforcement Learning (CRL)[1,2]

$$\mathcal{J}(\pi) = \mathcal{F}(d^\pi)$$

$\mathcal{F}$ is a convex/concave function

[1](Zhang et al., 2020), [2](Zahavy et al., 2021)

Problem Setting

policy π + CMP

↓

marginal state distribution

$$d^\pi(s) = \frac{1}{T} \sum_{t \in [T]} Pr(s_t = s)$$

CFOL Workshop @ ICML

Mutti et al. **"Challenging Common Assumptions in Convex Reinforcement Learning"**. 2022.

←

Convex Reinforcement Learning (CRL)[1,2]

$$\mathcal{J}(\pi) = \mathcal{F}(d^\pi)$$

$\mathcal{F}$ is a convex/concave function

[1](Zhang et al., 2020), [2](Zahavy et al., 2021)

Problem Setting

policy π + CMP

↓

marginal state distribution

$$d^\pi(s) = \frac{1}{T} \sum_{t \in [T]} Pr(s_t = s)$$

this paper

Maximum State Entropy (MSE)[3]

$$\mathcal{E}(\pi) = H(d^\pi) = d^\pi \log d^\pi$$

↑

CFOL Workshop @ ICML

Mutti et al. **"Challenging Common Assumptions in Convex Reinforcement Learning"**. 2022.

←

Convex Reinforcement Learning (CRL)[1,2]

$$\mathcal{J}(\pi) = \mathcal{F}(d^\pi)$$

$\mathcal{F}$ is a convex/concave function

[1](Zhang et al., 2020), [2](Zahavy et al., 2021), [3](Hazan et al., 2019)

Does Non-Markovianity Matter?

$$\mathcal{E}(\pi) := H(d^{\pi})$$

↓

Markovian policies
are sufficient[1]

[1][Puterman, 2014]

Does Non-Markovianity Matter?

state visitation frequency $d(s) = \frac{1}{T} \sum_{t \in [T]} \mathbb{1}(s_t = s)$

$$\mathcal{E}_{\infty}(\pi) := H(d^\pi) = H\Big(\mathop{\mathbb{E}}_{d \sim p^\pi} [d] \Big)$$

Markovian policies are sufficient[1]

[1][Puterman, 2014]

Does Non-Markovianity Matter?

$$\mathcal{E}_{\infty}(\pi) := H(d^{\pi}) = H\Big(\mathop{\mathbb{E}}_{d\sim p^{\pi}}[d]\Big) \geq \mathop{\mathbb{E}}_{d\sim p^{\pi}}\Big[H(d)\Big]$$

↓

Markovian policies are sufficient[1]

(through Jensen's)

[1][Puterman, 2014]

Does Non-Markovianity Matter?

$$\mathcal{E}_{\infty}(\pi) := H(d^{\pi}) = H\Big(\mathop{\mathbb{E}}_{d\sim p^{\pi}}[d]\Big) \geq \mathop{\mathbb{E}}_{d\sim p^{\pi}}\Big[H(d)\Big] =: \mathcal{E}_1(\pi)$$

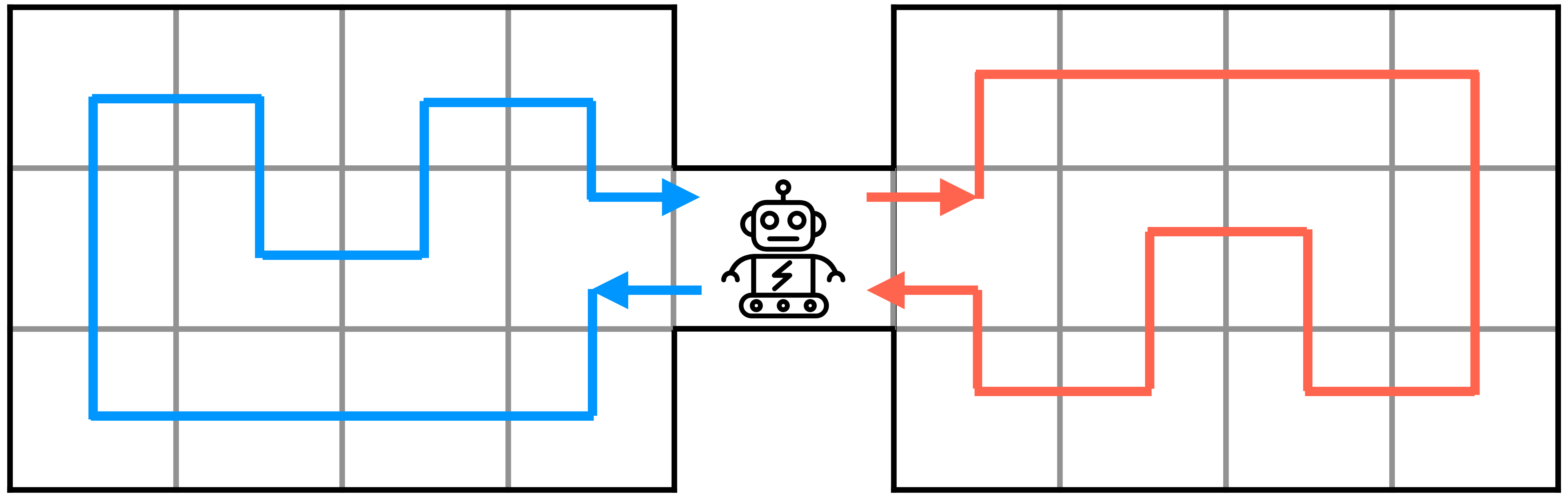
(through Jensen's)

Markovian policies are sufficient[1]

non-Markovianity matters!

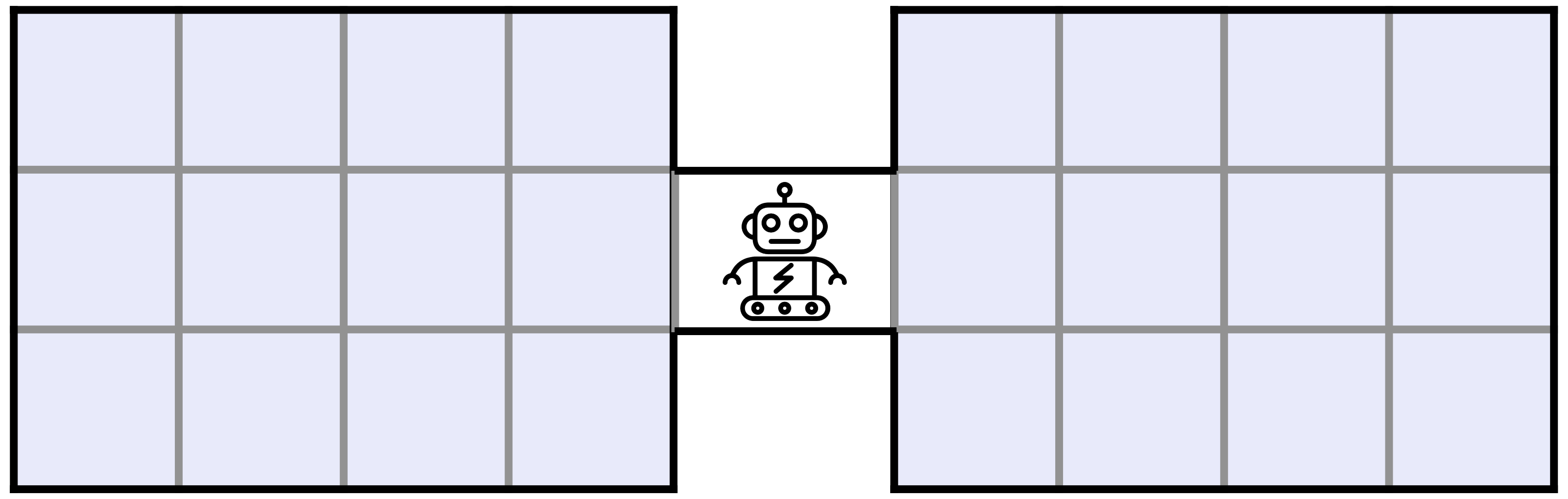
[1][Puterman, 2014]

Illustrative Example

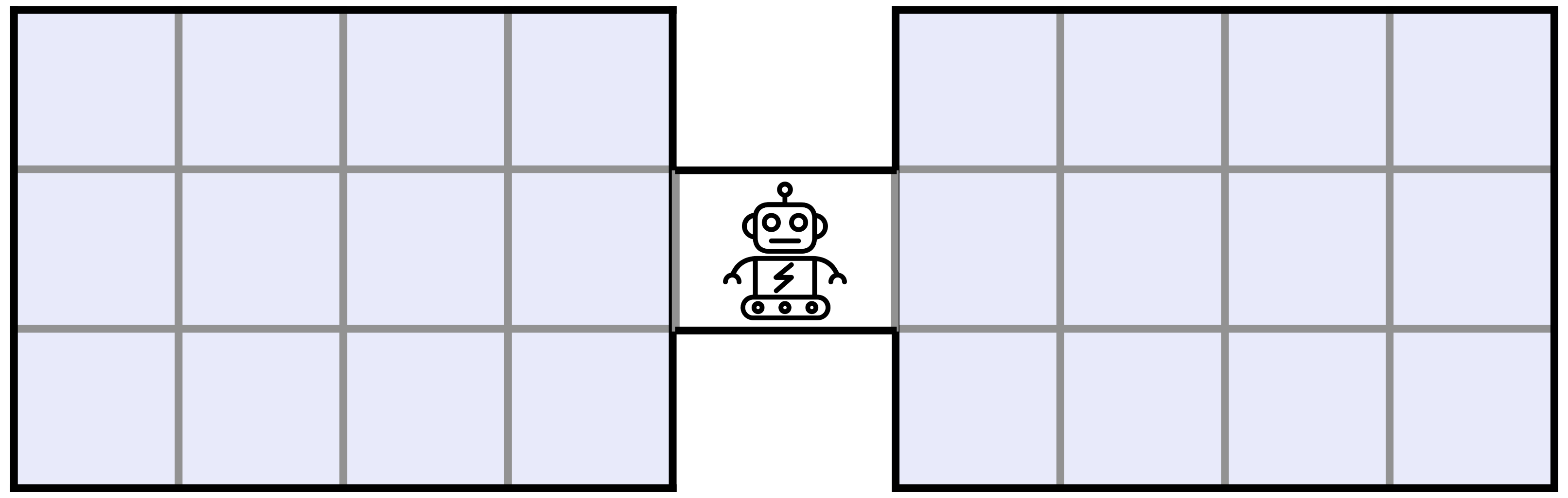


Illustrative Example

Best Markovian Policy

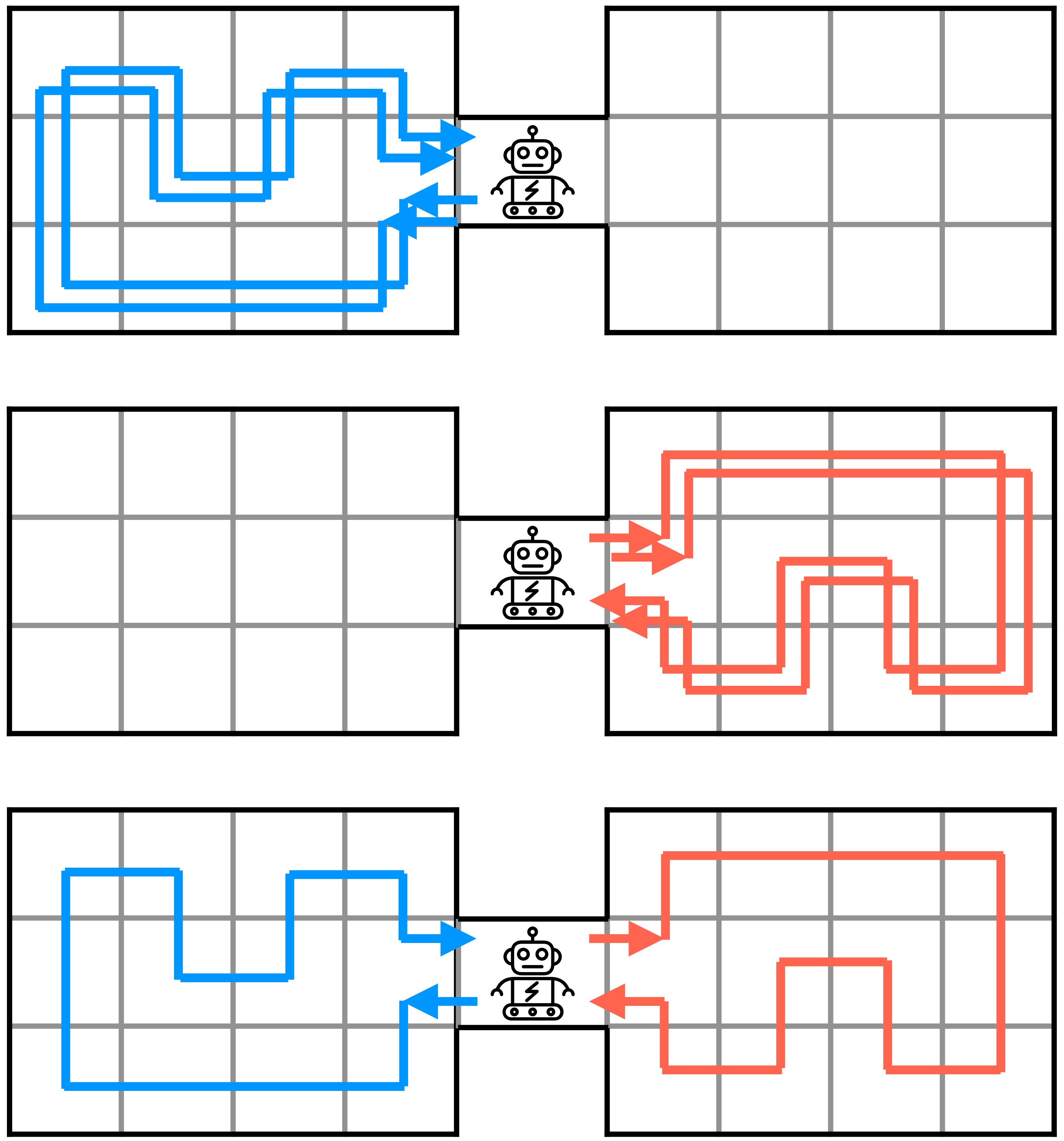


Best Non-Markovian Policy



Illustrative Example

Best Markovian Policy



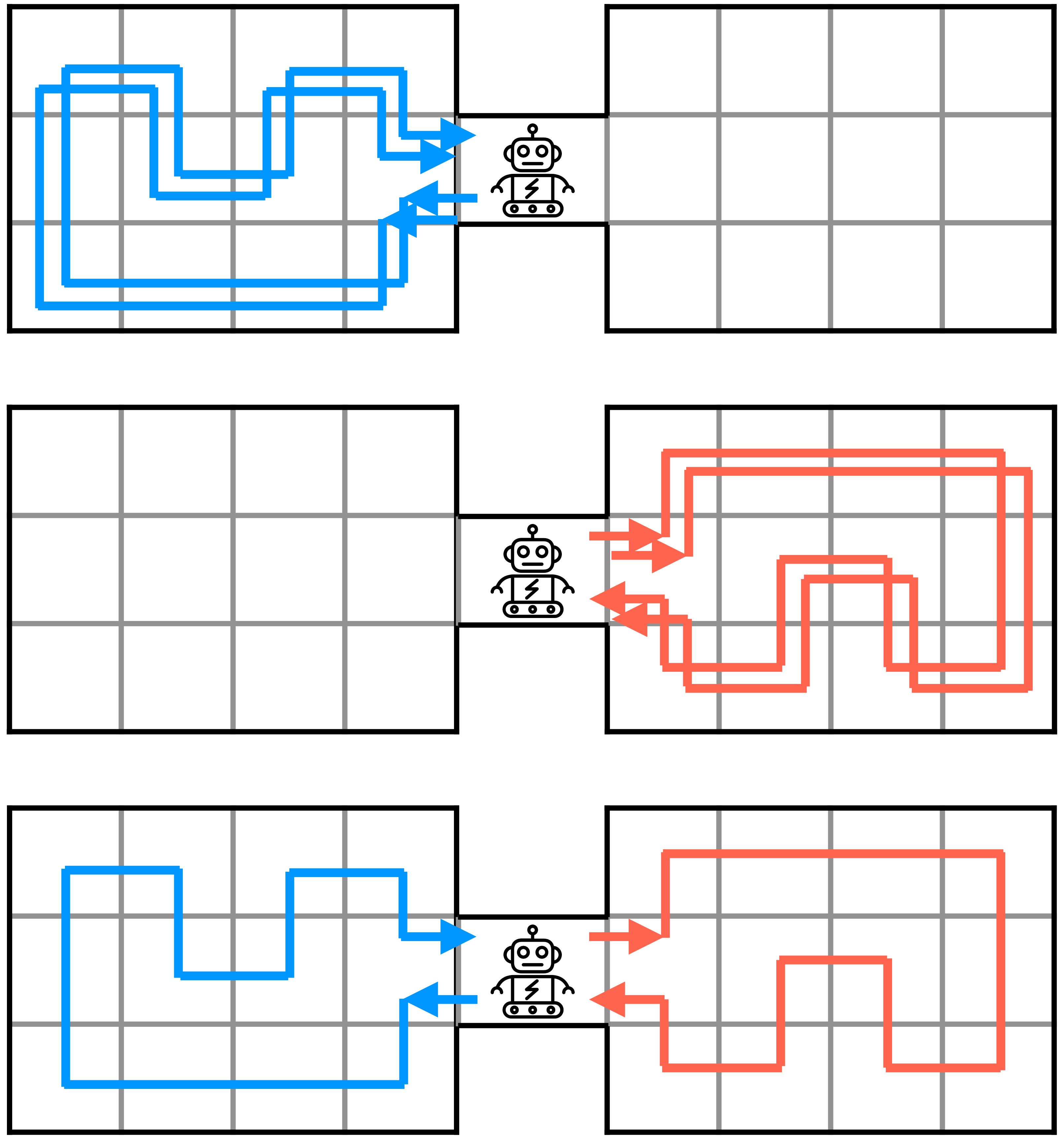
$1/3$

$1/3$

$1/3$

Illustrative Example

Best Markovian Policy



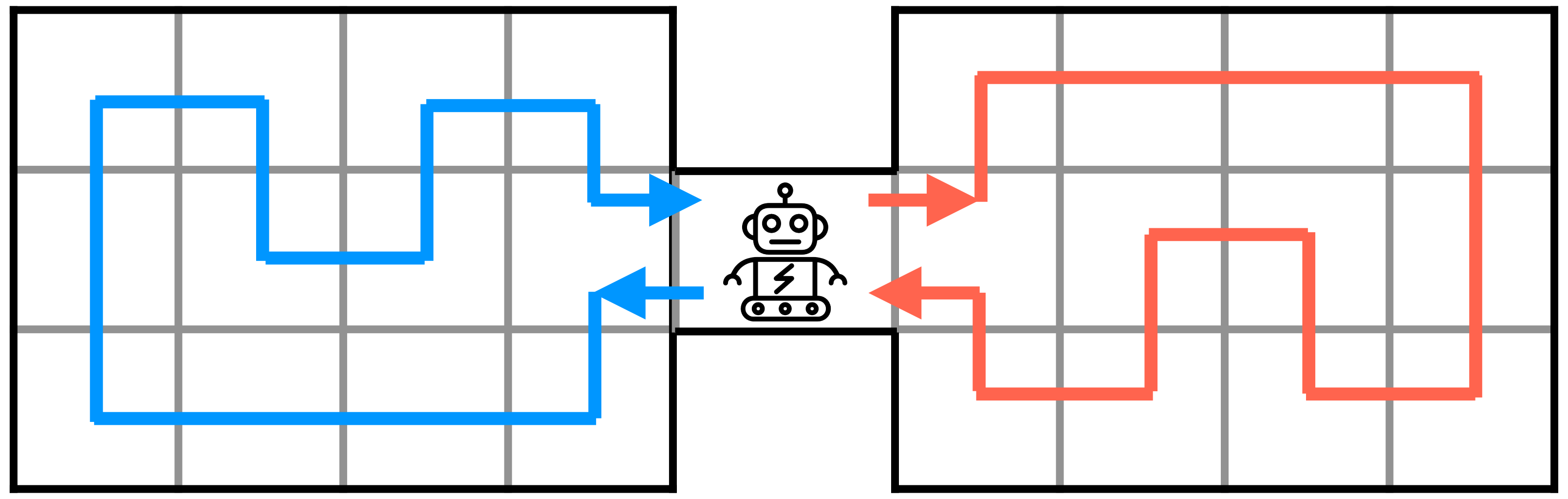
$1/3$

$1/3$

$1/3$

Best Non-Markovian Policy

1



The Importance of Non-Markovianity

Finite-Sample Maximum State Entropy

$$\mathcal{E}_1(\pi) = \mathop{\mathbb{E}}_{d \sim p^\pi} \Big[H(d) \Big]$$

The Importance of Non-Markovianity

Finite-Sample Maximum State Entropy

$$\mathcal{E}_1(\pi) = \mathop{\mathbb{E}}_{d \sim p^\pi} \Big[H(d) \Big]$$

A tool to compare Markovian and non-Markovian policies?

The Importance of Non-Markovianity

Finite-Sample Maximum State Entropy

$$\mathcal{E}_1(\pi) = \mathop{\mathbb{E}}_{d \sim p^{\pi}} \Big[H(d) \Big]$$

A tool to compare Markovian and non-Markovian policies?

The entropy is non-additive, standard regret cannot be used

Regret

Definition (Expected Regret-to-go). *Let* π *be a policy interacting with an MDP over* $T - t$ *steps starting from trajectory* h_t. *We define the expected regret-to-go as*

$$R_{T-t}(\pi, h_t) = H^* - \mathop{\mathbb{E}}_{h_{T-t} \sim p^\pi} \left[H(d_{h_t \oplus h_{T-t}}) \right]$$

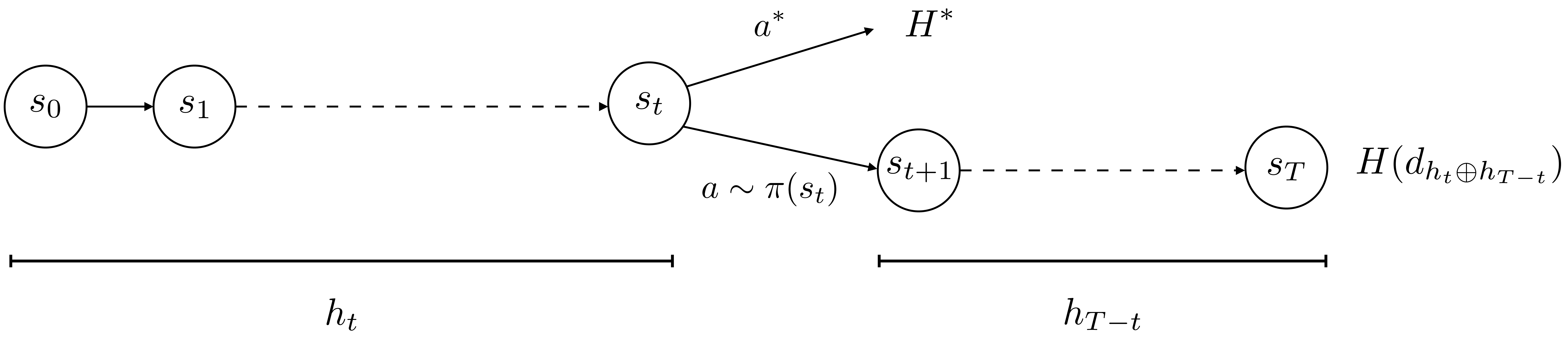
where H^* *is the maximum entropy that can be achieved by any policy starting from* h_t.

Regret

Definition (Expected Regret-to-go). *Let π be a policy interacting with an MDP over $T-t$ steps starting from trajectory h_t. We define the expected regret-to-go as*

$$R_{T-t}(\pi, h_t) = H^* - \mathop{\mathbb{E}}_{h_{T-t} \sim p^\pi} \left[H(d_{h_t \oplus h_{T-t}}) \right]$$

where H^* is the maximum entropy that can be achieved by any policy starting from h_t.



Regret Bounds

Definition (Expected Regret-to-go). *Let* π *be a policy interacting with an MDP over* $T-t$ *steps starting from trajectory* h_t*. We define the expected regret-to-go as*

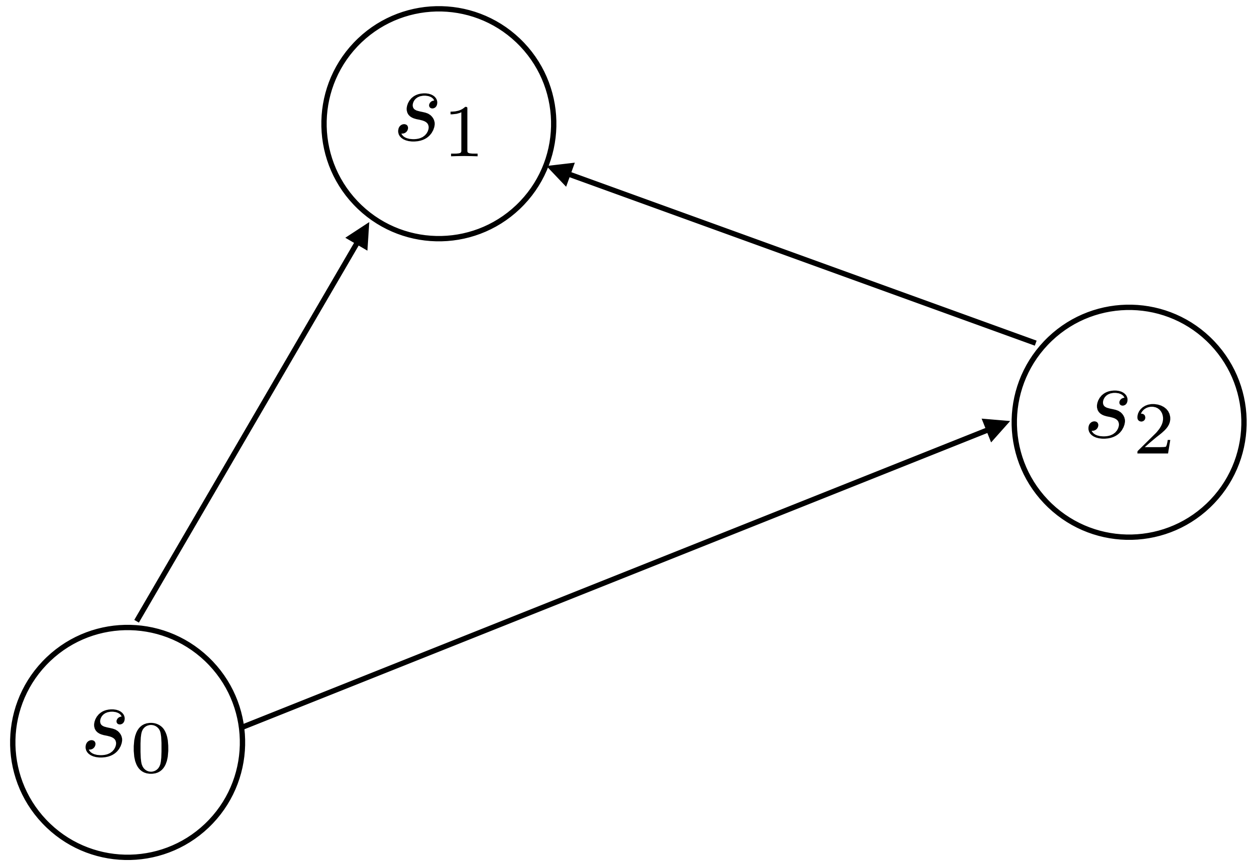
$$R_{T-t}(\pi, h_t) = H^* - \mathop{\mathbb{E}}_{h_{T-t} \sim p^{\pi}} \left[H(d_{h_t \oplus h_{T-t}}) \right]$$

where H^* *is the maximum entropy that can be achieved by any policy starting from* h_t*.*

First step. There exists a deterministic optimal non-Markovian policy π_{NM}

Regret Bounds

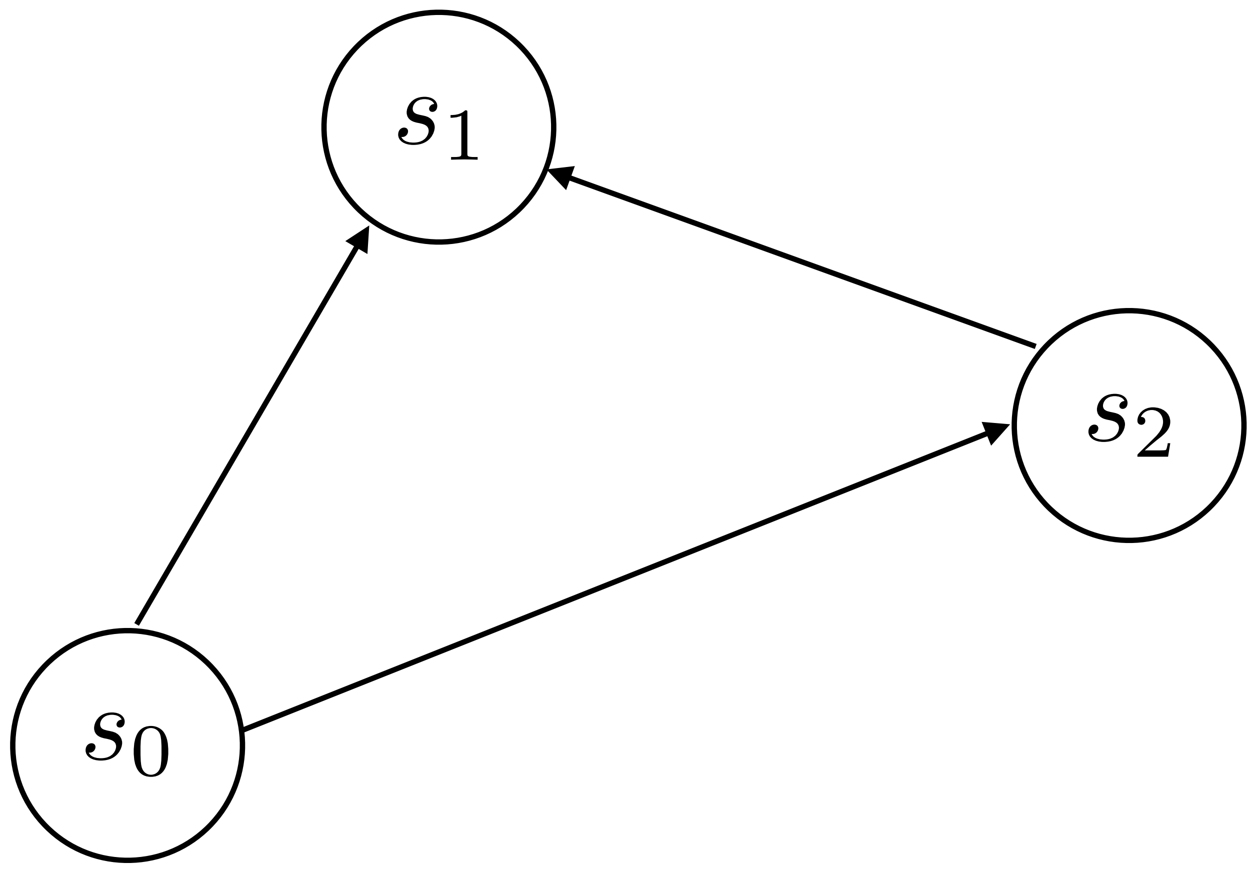
First step. There exists a deterministic optimal non-Markovian policy π_{NM}



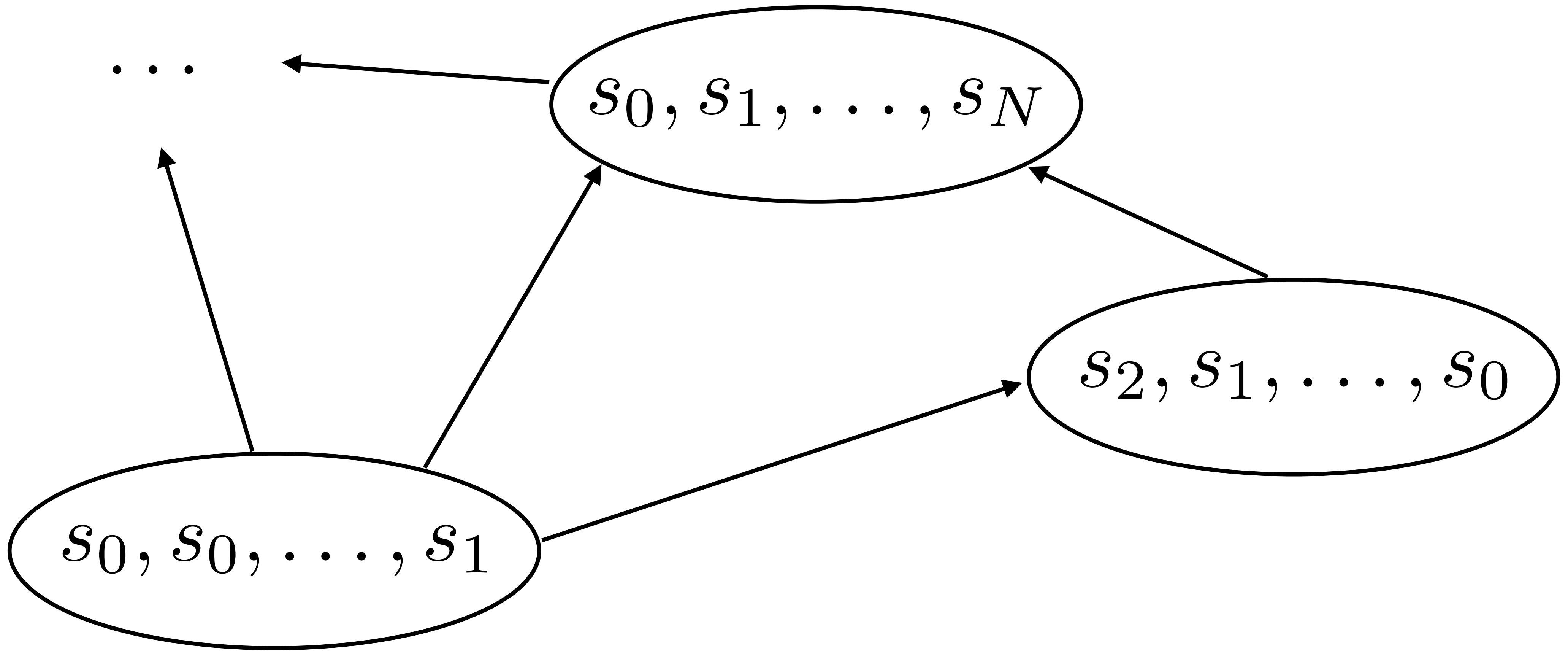
CMP with **MSE** objective

Regret Bounds

First step. There exists a deterministic optimal non-Markovian policy π_{NM}



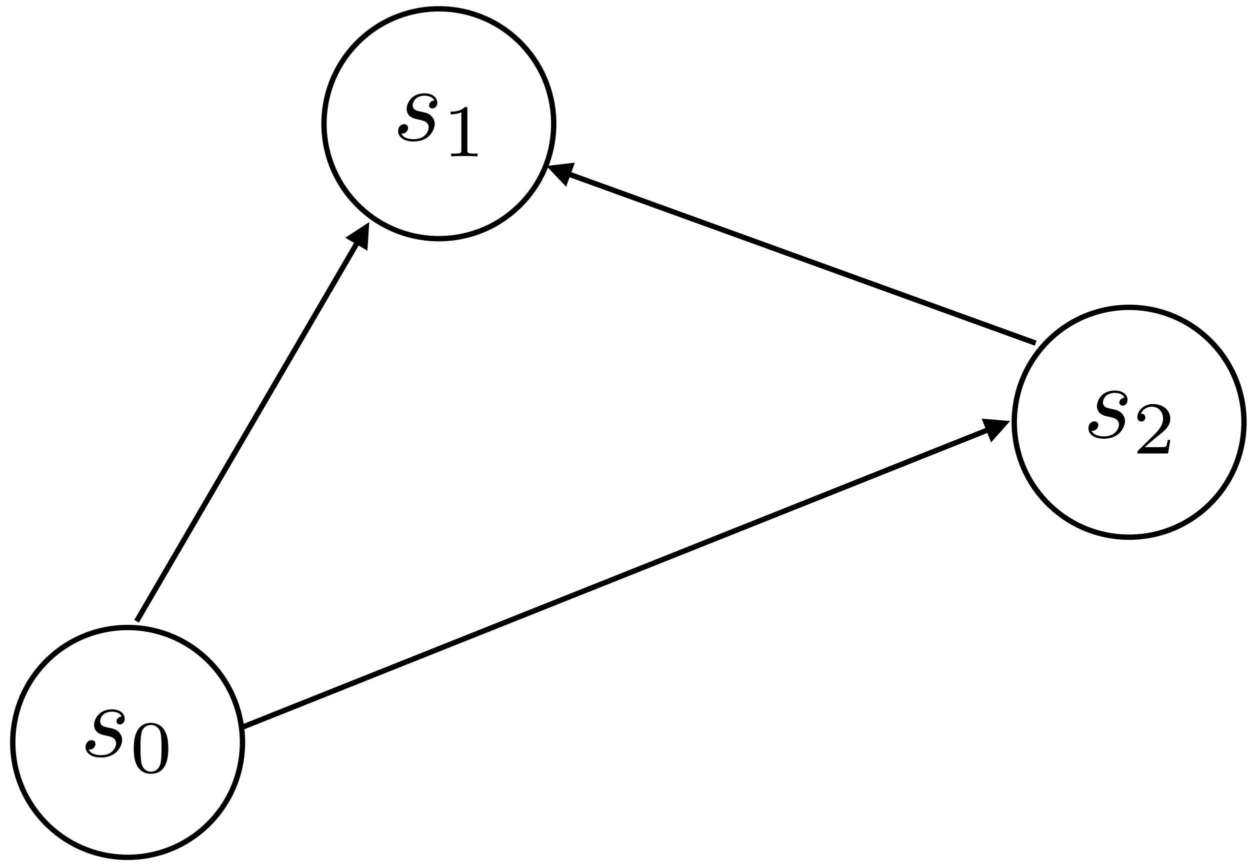
CMP with **MSE** objective



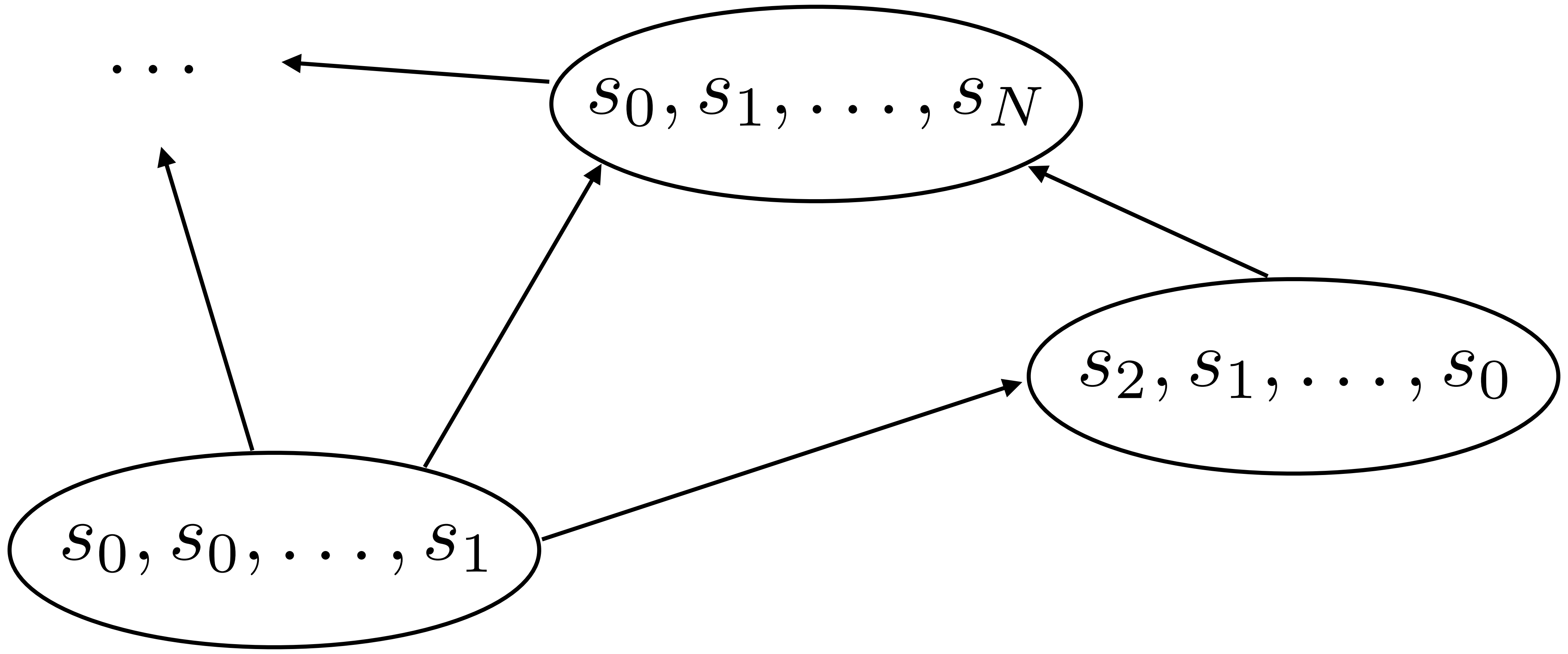
Extended CMP with **reward** $R(h_T) = H(d_{h_T})$

Regret Bounds

First step. There exists a deterministic optimal non-Markovian policy π_{NM}



CMP with **MSE** objective



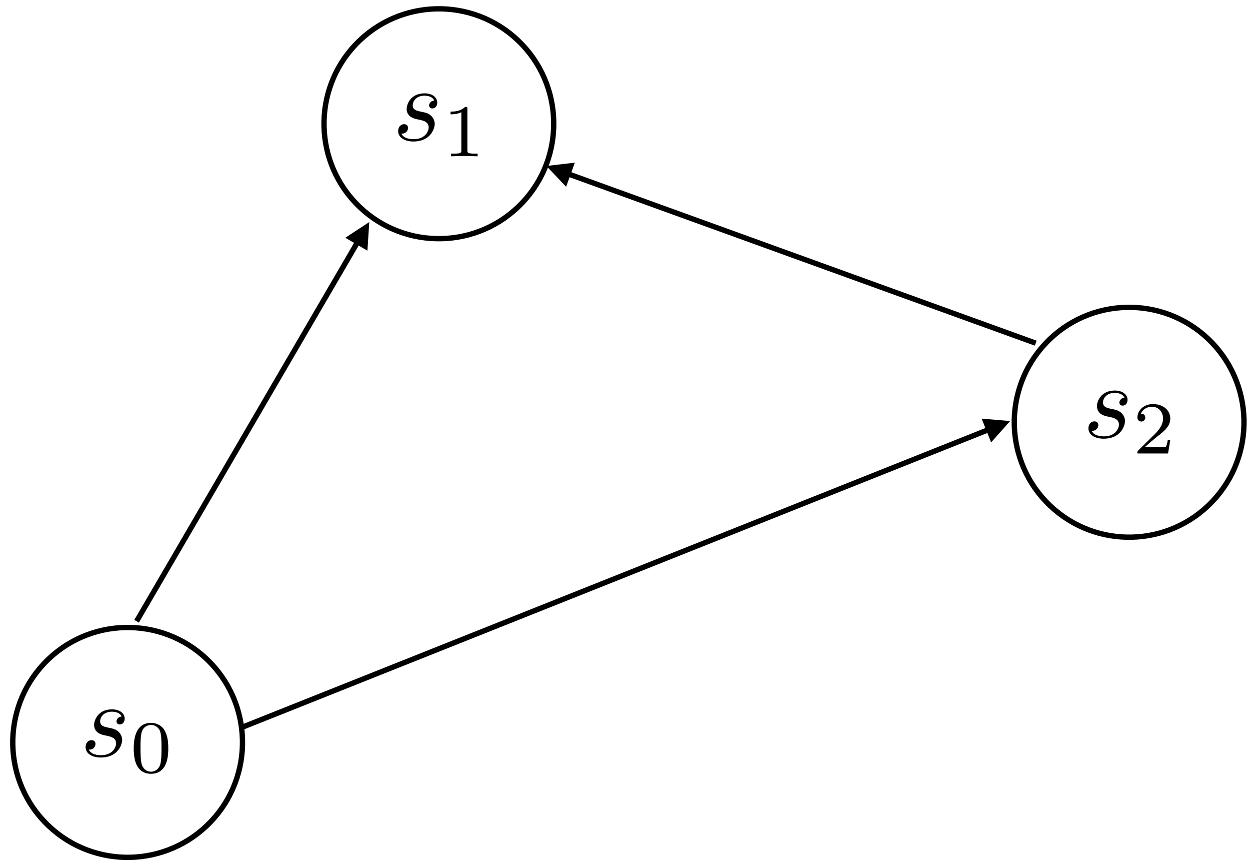
Extended CMP with **reward** $R(h_T) = H(d_{h_T})$

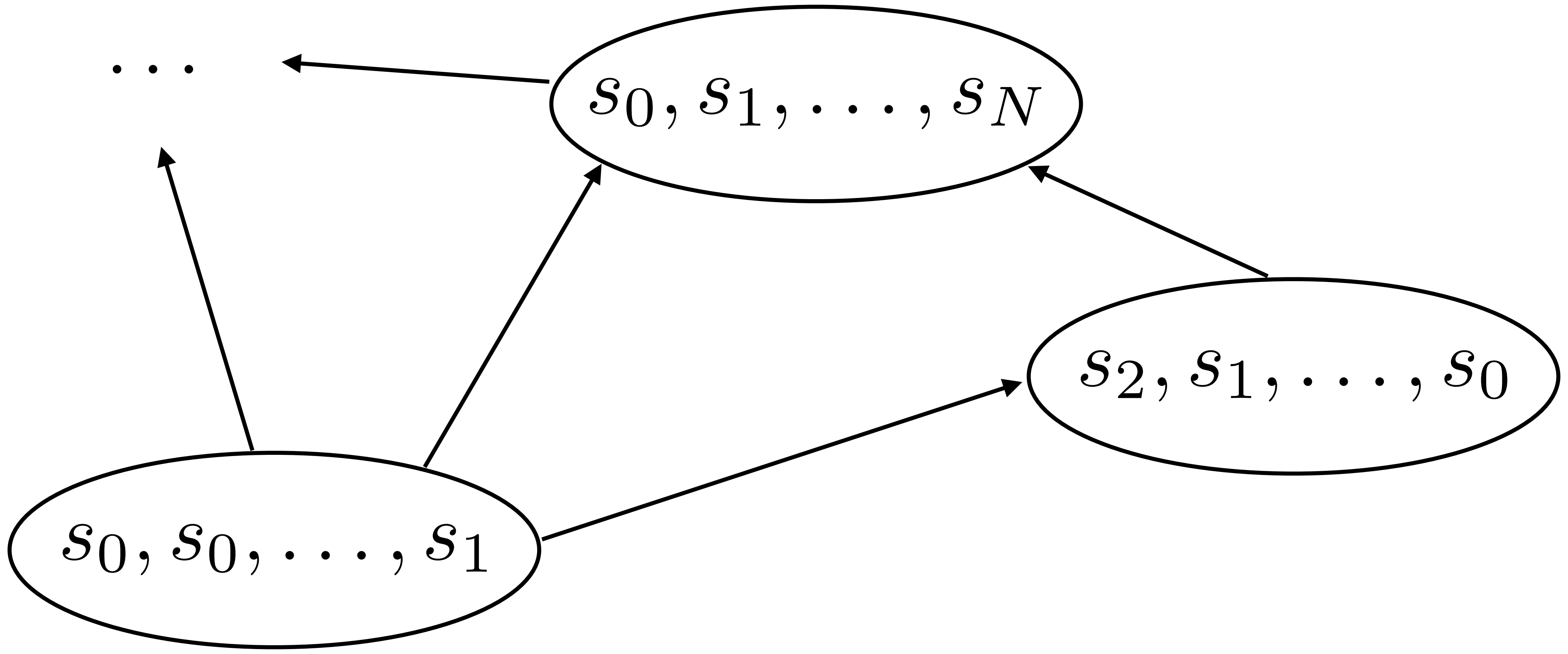
optimal deterministic Markovian policy[1]

[1](Proposition 4.4.3, Puterman, 2014)

Regret Bounds

First step. There exists a deterministic optimal non-Markovian policy π_{NM}





CMP with **MSE** objective

Extended CMP with **reward** $R(h_T) = H(d_{h_T})$

optimal deterministic non-Markovian policy $\Longleftarrow$ optimal deterministic Markovian policy[1]

[1](Proposition 4.4.3, Puterman, 2014)

Regret Bounds

Definition (Expected Regret-to-go). *Let* π *be a policy interacting with an MDP over* $T-t$ *steps starting from trajectory* h_t. *We define the expected regret-to-go as*

$$R_{T-t}(\pi, h_t) = H^* - \mathop{\mathbb{E}}_{h_{T-t} \sim p^\pi} \left[H(d_{h_t \oplus h_{T-t}}) \right]$$

where H^* *is the maximum entropy that can be achieved by any policy starting from* h_t.

First step. There exists a <u>deterministic</u> optimal non-Markovian policy π_{NM}

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

Regret Bounds

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] =$$

Regret Bounds

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

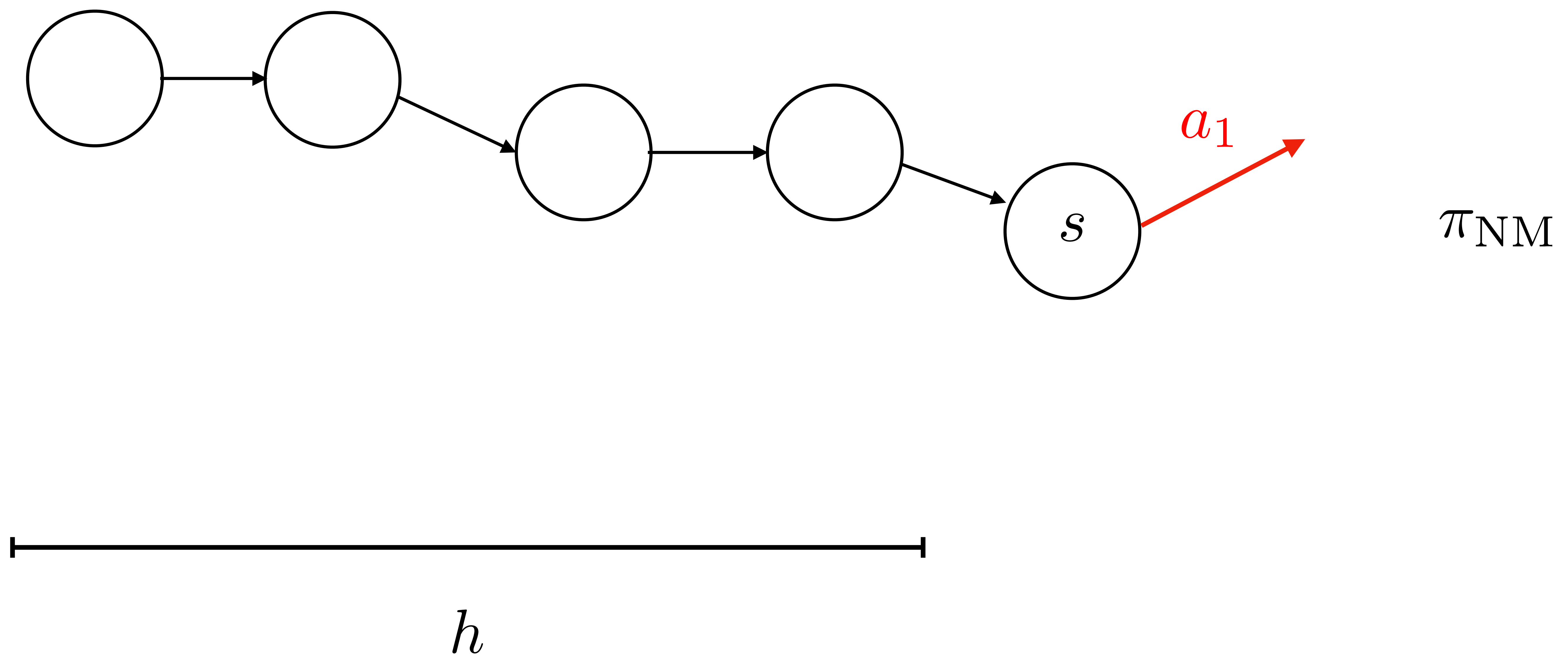
$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] = \underset{hs\sim p_t^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}}\left[\mathbb{E}\left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs))\right]\right]$$

(through the Law of Total Variance and the determinism of π_{NM})

Regret Bounds

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

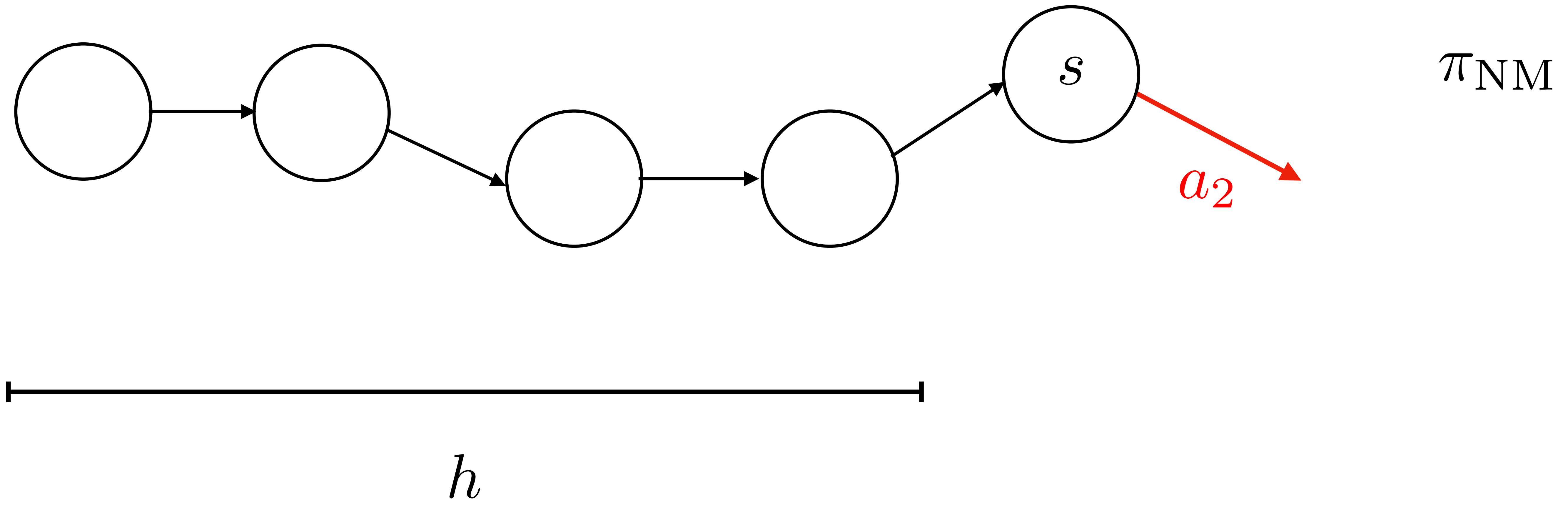
$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] = \mathop{\mathbb{V}\mathrm{ar}}_{hs\sim p_t^{\pi_{\mathrm{NM}}}}\left[\mathbb{E}\left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs))\right]\right]$$



Regret Bounds

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

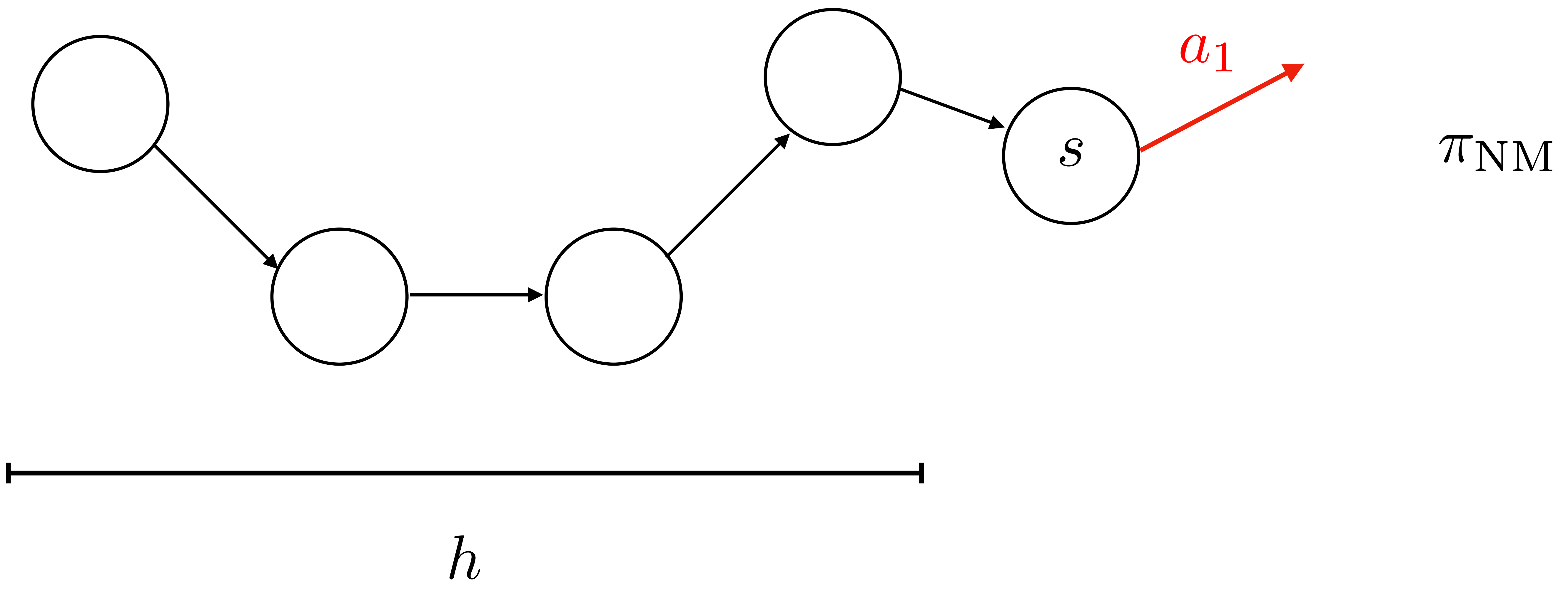
$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] = \underset{hs\sim p_t^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}}\left[\mathbb{E}\left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs))\right]\right]$$



Regret Bounds

Second step. The optimal Markovian policy π_{M} is randomized

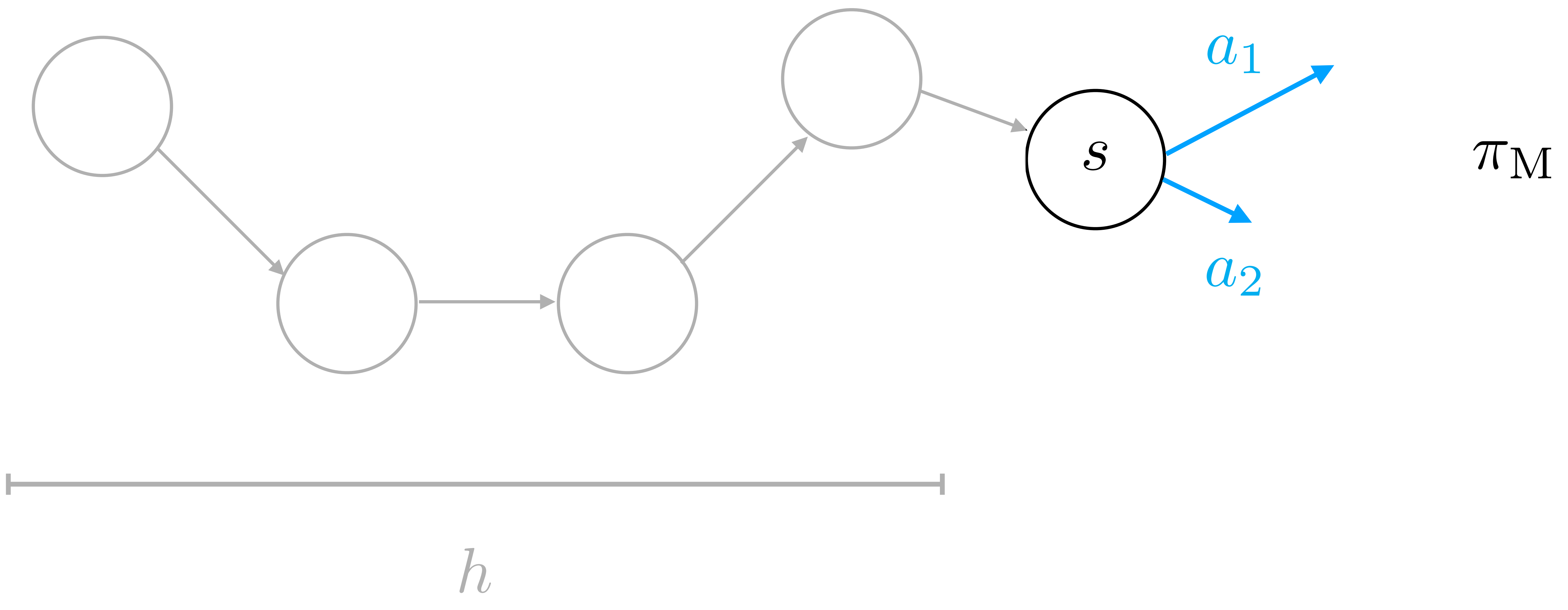
$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] = \mathop{\mathbb{V}\mathrm{ar}}_{hs\sim p_t^{\pi_{\mathrm{NM}}}}\left[\mathbb{E}\left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs))\right]\right]$$



Regret Bounds

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

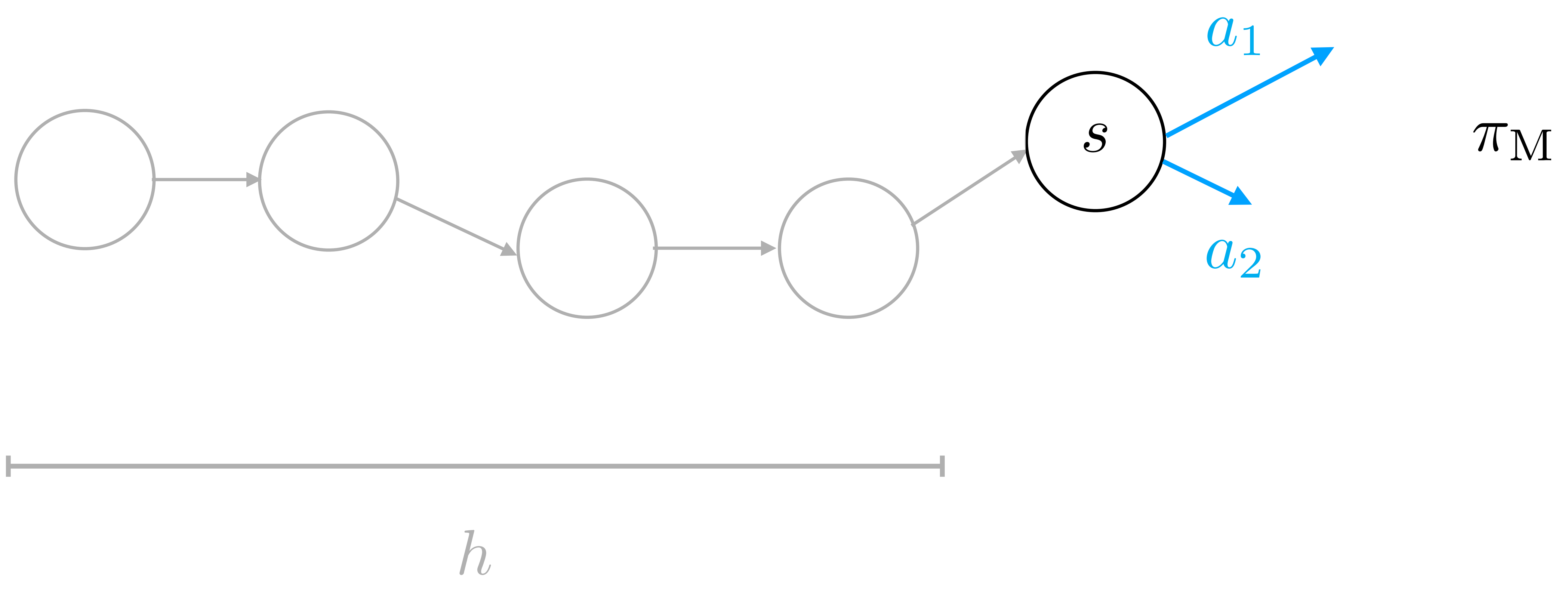
$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] = \underset{hs\sim p_t^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}}\left[\mathbb{E}\left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs))\right]\right]$$



Regret Bounds

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

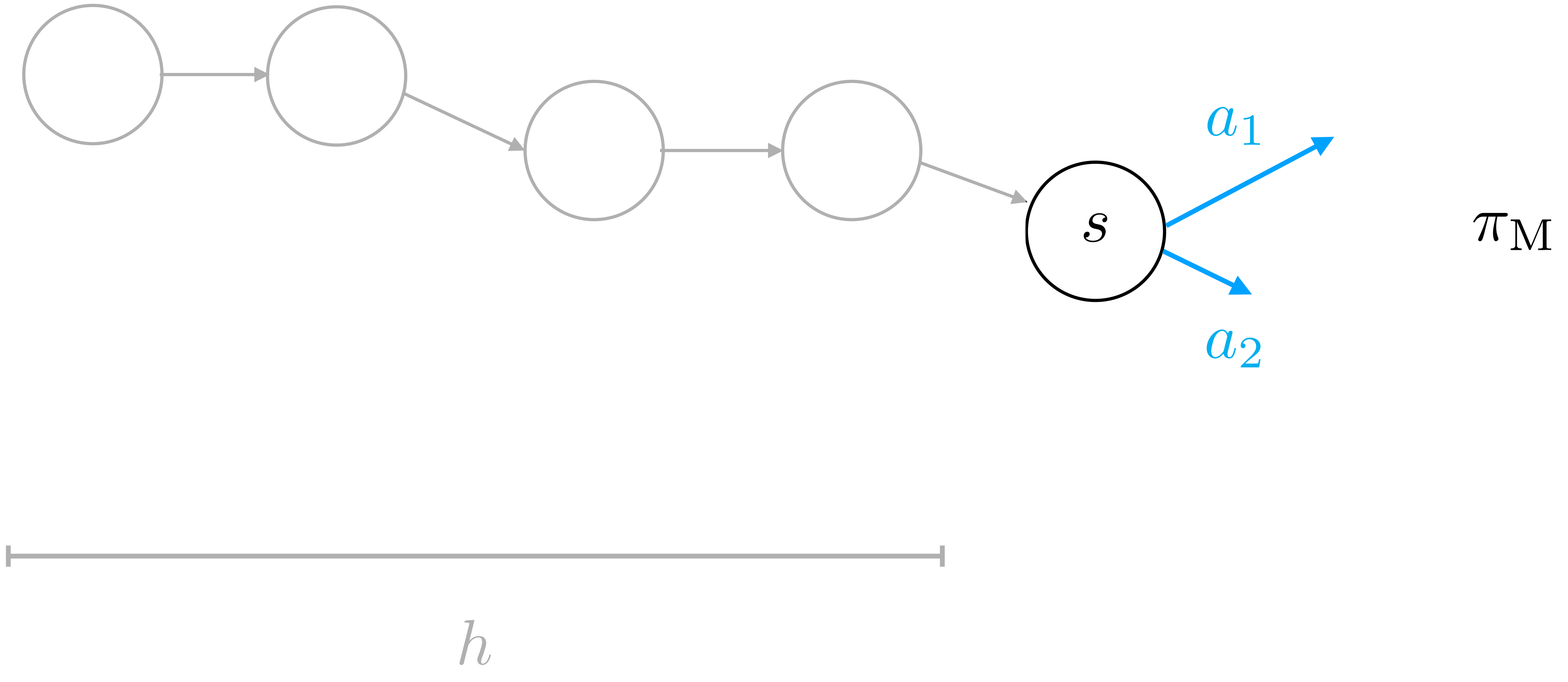
$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] = \underset{hs\sim p_t^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}}\left[\mathbb{E}\left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs))\right]\right]$$



Regret Bounds

Second step. The optimal Markovian policy π_{M} is <u>randomized</u>

$$\mathbb{V}\mathrm{ar}\left[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s,t))\right] = \mathop{\mathbb{V}\mathrm{ar}}_{hs\sim p_t^{\pi_{\mathrm{NM}}}}\left[\mathbb{E}\left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs))\right]\right]$$



Regret Bounds

Definition (Expected Regret-to-go). *Let π be a policy interacting with an MDP over $T-t$ steps starting from trajectory h_t. We define the expected regret-to-go as*

$$R_{T-t}(\pi, h_t) = H^* - \mathop{\mathbb{E}}_{h_{T-t} \sim p^\pi} \left[H(d_{h_t \oplus h_{T-t}}) \right]$$

where H^* is the maximum entropy that can be achieved by any policy starting from h_t.

First step. There exists a deterministic optimal non-Markovian policy π_{NM}

Second step. The optimal Markovian policy π_{M} is randomized

Result. The optimal Markovian policy suffers positive regret

Regret Bounds

Result. The optimal Markovian policy suffers positive regret

$$\mathcal{R}_{T-t}(\pi_{\mathrm{M}}, h_t) \propto \mathbb{V}\mathrm{ar}[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s))]$$

Regret Bounds

Result. The optimal Markovian policy suffers positive regret

$$\mathcal{R}_{T-t}(\pi_{\mathrm{M}}, h_t) \propto \mathbb{V}\mathrm{ar}[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s))]$$

$$\propto \underset{hs \sim p_t^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}} \left[\mathbb{E} \left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs)) \right] \right]$$

Regret Bounds

Result. The optimal Markovian policy suffers positive regret

$$\mathcal{R}_{T-t}(\pi_{\mathrm{M}}, h_t) \propto \mathbb{V}\mathrm{ar}[\mathrm{Ber}(\pi_{\mathrm{M}}(a^*|s))]$$

$$\propto \underset{hs \sim p_t^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}} \left[\mathbb{E} \left[\mathrm{Ber}(\pi_{\mathrm{NM}}(a^*|hs)) \right] \right]$$

$\Longrightarrow$ **non-Markovianity** matters in **finite-sample** MSE

Computational Tractability

Learning the optimal Markovian policy for MSE is known to be provably efficient[1,2]

[1](Hazan et al., 2019), [2](Zhang et al., 2020)

Computational Tractability

Learning the optimal Markovian policy for MSE is known to be provably efficient[1,2]

Is computing the optimal non-Markovian policy for the finite-sample MSE even tractable?

[1](Hazan et al., 2019), [2](Zhang et al., 2020)

Computational Tractability

Learning the optimal Markovian policy for MSE is known to be provably efficient[1,2]

Is computing the optimal non-Markovian policy for the finite-sample MSE even tractable?

Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is* ***NP-hard****.*

[1](Hazan et al., 2019), [2](Zhang et al., 2020)

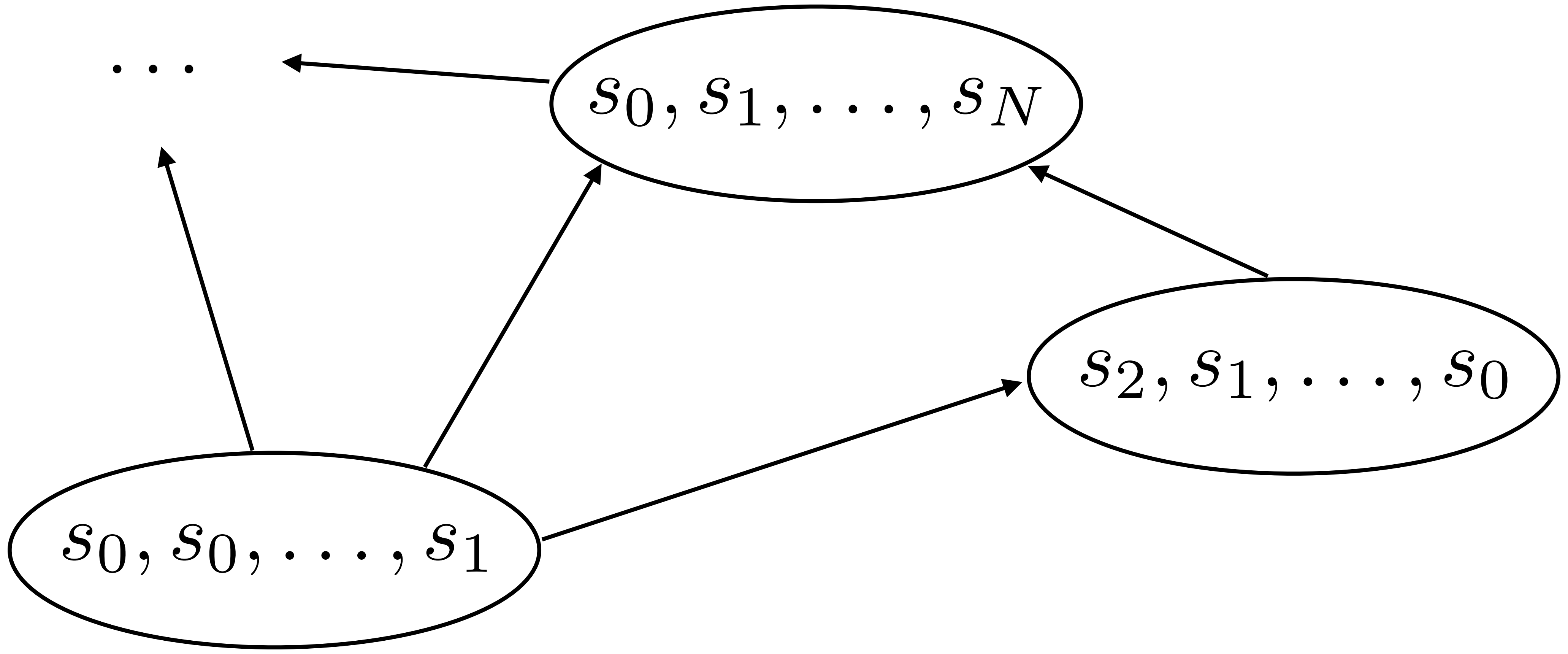
Computational Tractability

Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is* ***NP-hard***.

→ through reduction to POMDP

Computational Tractability

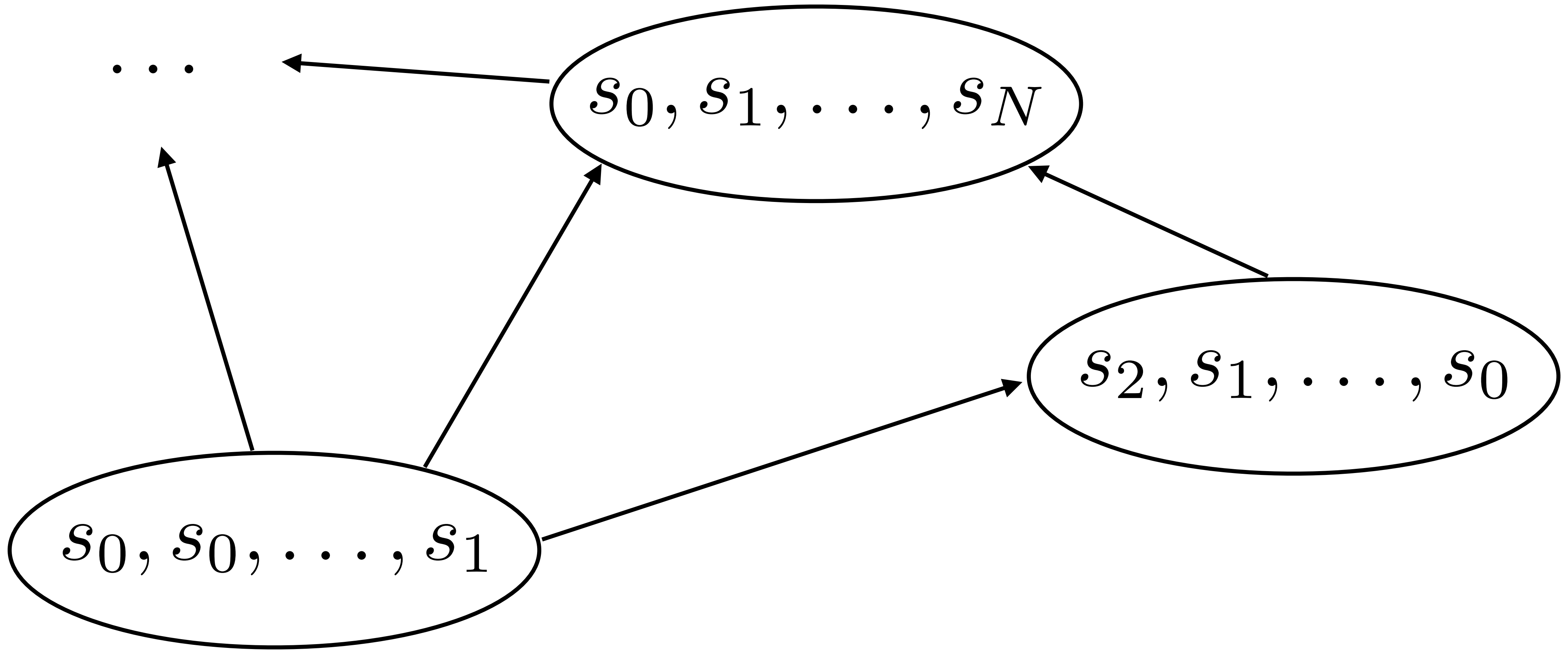
Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is* ***NP-hard****.*



Extended CMP with **reward** $R(h_T) = H(d_{h_T})$

Computational Tractability

Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is* ***NP-hard****.*

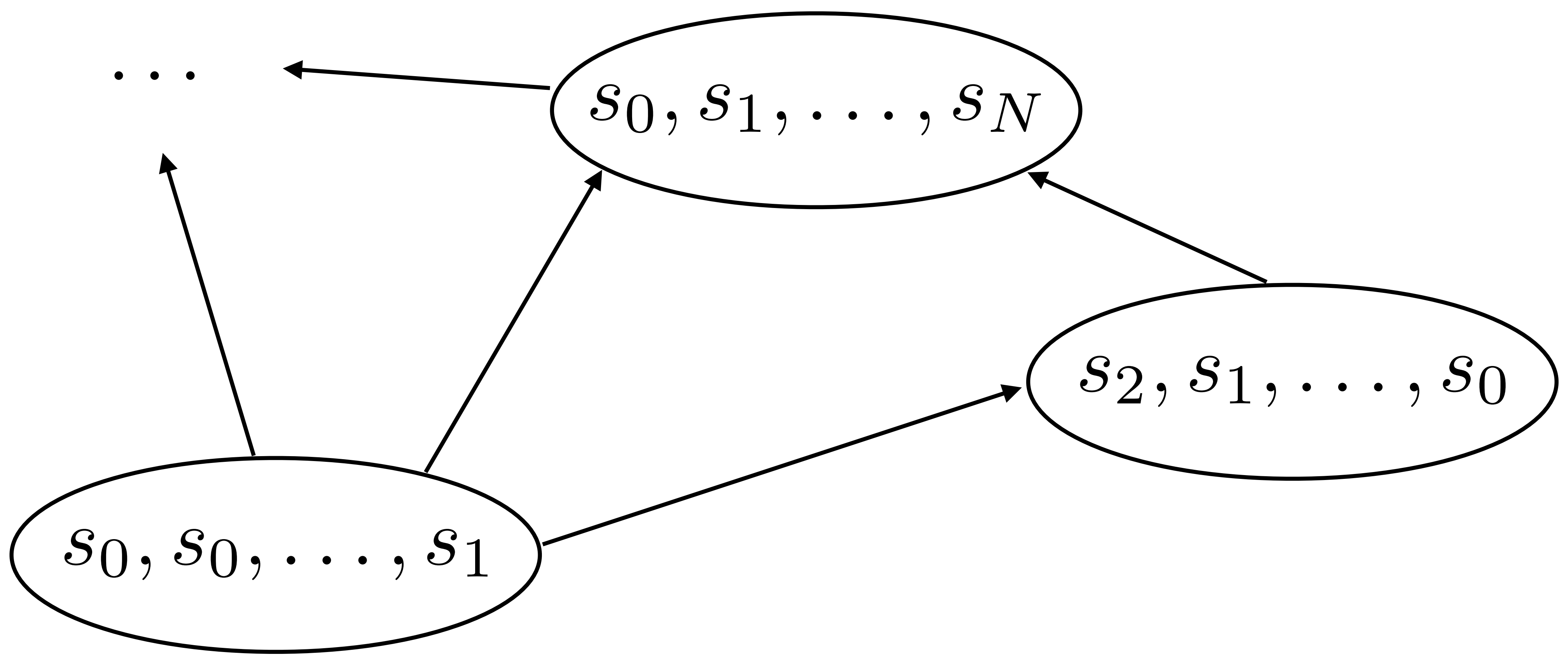


Extended CMP with **reward** $R(h_T) = H(d_{h_T})$

exponential blowup with the horizon

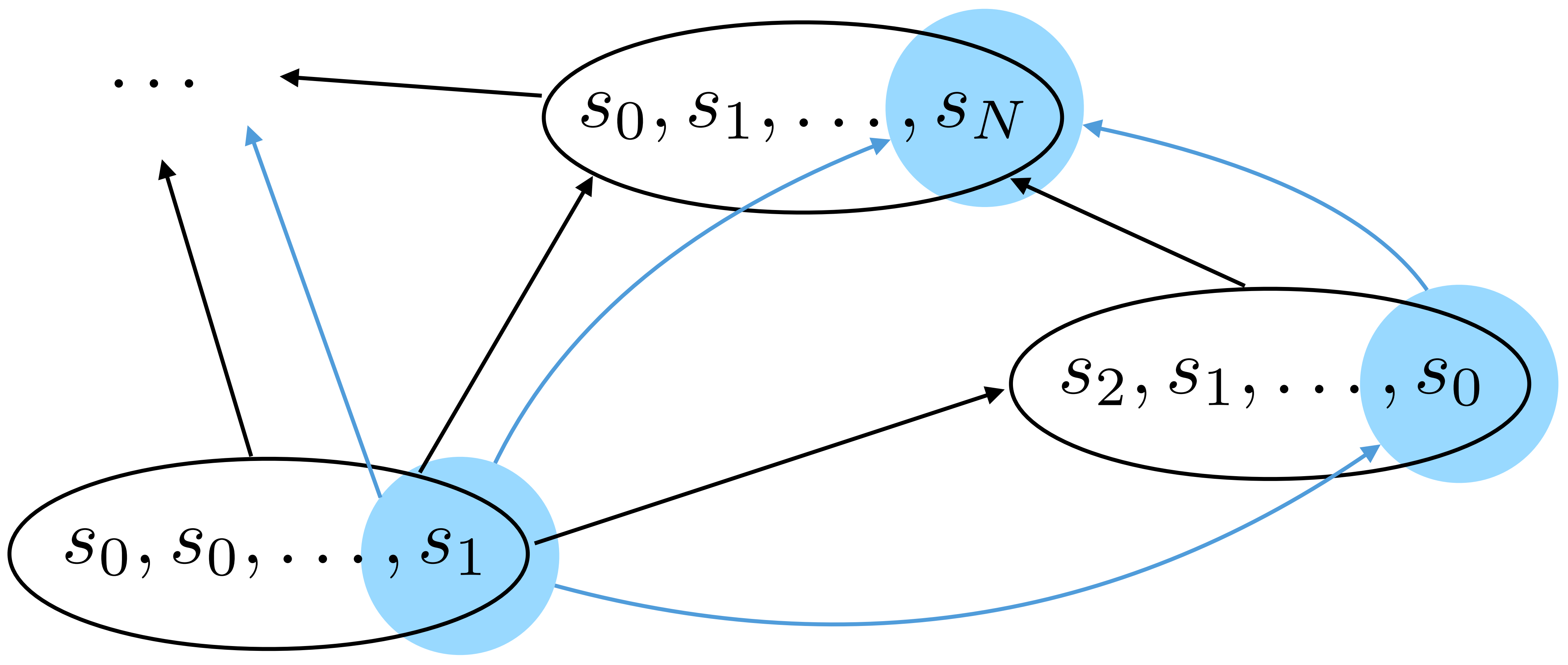
Computational Tractability

Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is **NP-hard**.*



Extended CMP with **reward** $R(h_T) = H(d_{h_T})$

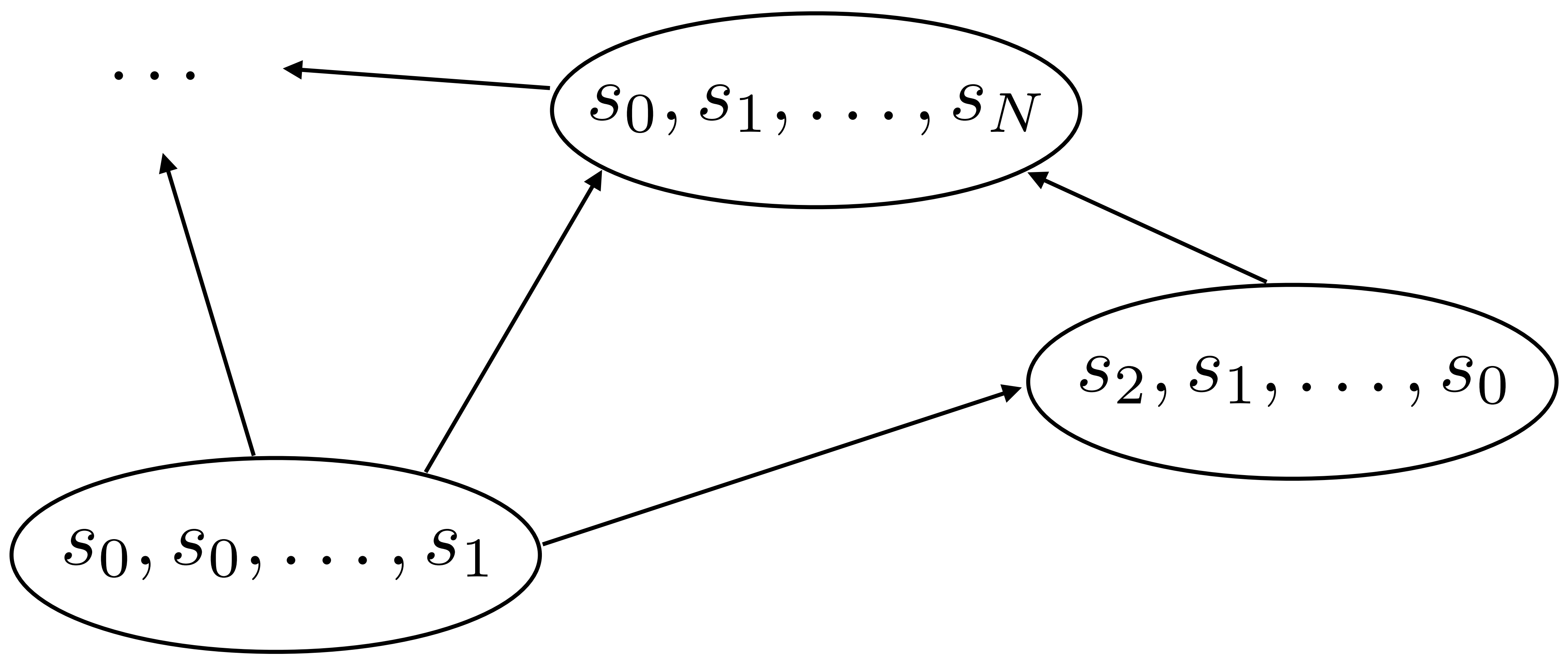
exponential blowup with the horizon

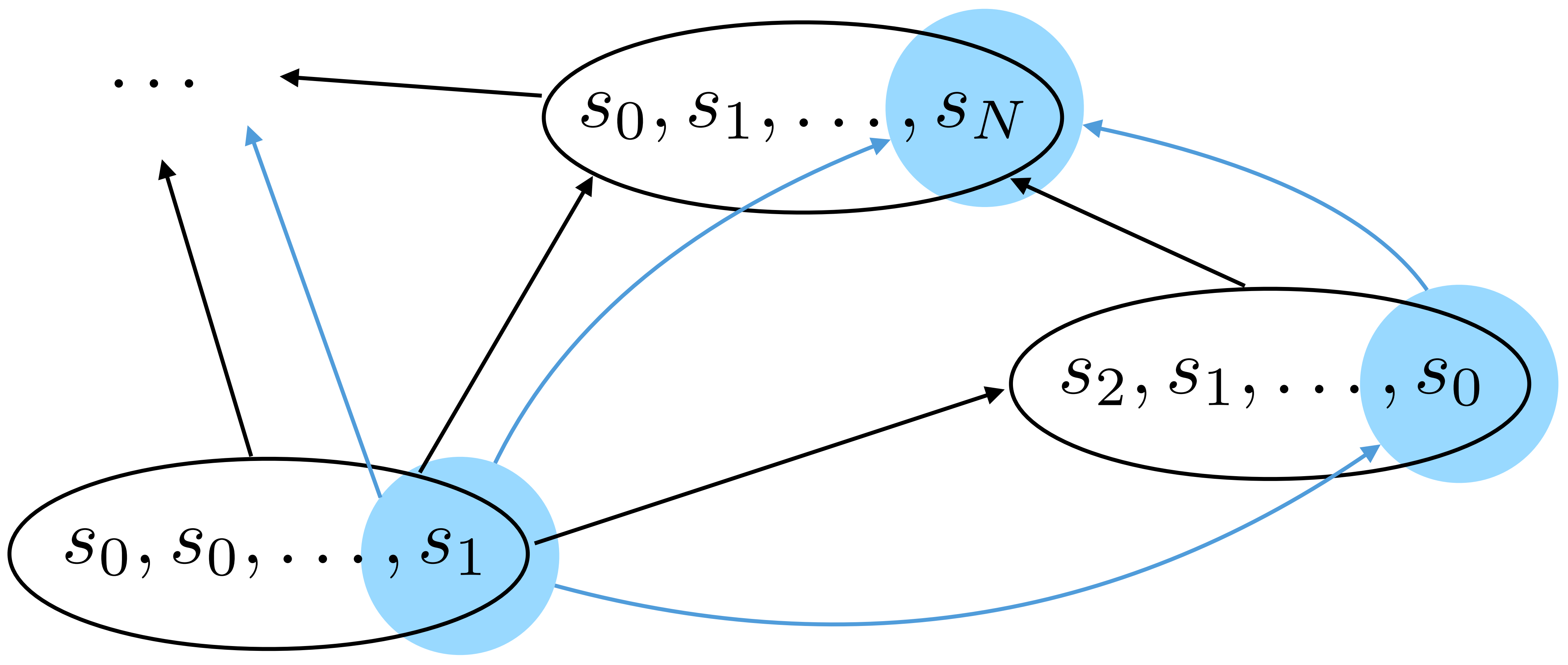


Reduction to a class of **POMDPs**

Computational Tractability

Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is* ***NP-hard****.*





Extended CMP with **reward** $R(h_T) = H(d_{h_T})$

exponential blowup with the horizon

Reduction to a class of **POMDPs** $\geq_p$ **3SAT**

NP-hard problem[1]

[1](Mundhenk et al., 2000)

Computational Tractability

Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is* ***NP-hard****.*

Can we solve the problem with function approximation?

Computational Tractability

Theorem (Computational Complexity). *Optimizing the finite-sample MSE within the space of non-Markovian policies is* ***NP-hard****.*

Can we solve the problem with function approximation?

DARL Workshop @ ICML & Pre-Training Workshop @ ICML

Mutti et al. **"Non-Markovian Policies for Unsupervised Reinforcement Learning in Multiple Environments"**. 2022.

Take Home

Non-Markovian policies are better for **finite-sample** convex objectives

Optimizing non-Markovian policies exactly is often **intractable**

Take Home

Non-Markovian policies are better for **finite-sample** convex objectives

Optimizing non-Markovian policies exactly is often **intractable**

What Is Next?

Approximate methods to optimize non-Markovian policies for convex objectives

Applications: When is it critical to consider a finite-sample objective?

References

(Williams & Zipser, 1989) A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1989.
(Hochreiter & Schmidhuber, 1997) Long short-term memory. Neural computation, 1997.
(Puterman, 2014) Markov decision processes: Discrete stochastic dynamic programming. 2014.
(Zhang et al., 2020) Variational policy gradient method for reinforcement learning with general utilities. NeurIPS, 2020.
(Zahavy et al., 2021) Reward is enough for convex MDPs. NeurIPS, 2021.
(Hazan et al., 2019) Provably efficient maximum entropy exploration. ICML 2019.
(Mundhenk et al., 2000) Complexity of finite-horizon Markov decision process problems. Journal of the ACM, 2020.

Questions?

Poster

Hall E #838

Paper