Constrained Entropy Maximization for Markov Decision Processes

Yagiz Savas and Ufuk Topcu

u-t-autonomous.info

aUTonomous SYSTEMS GROUP

In collaboration with Murat Cubuktepe, Mustafa Karabag and Melkior Ornik

His only regret—not to have killed de Gaulle

The real Jackal

By Ted Morgan

PARIS. On the dust jacket of my edition of Frederick Forsyth's "The Day of the Jackal" (now a major motion picture starring Edward Fox) is the following summary:

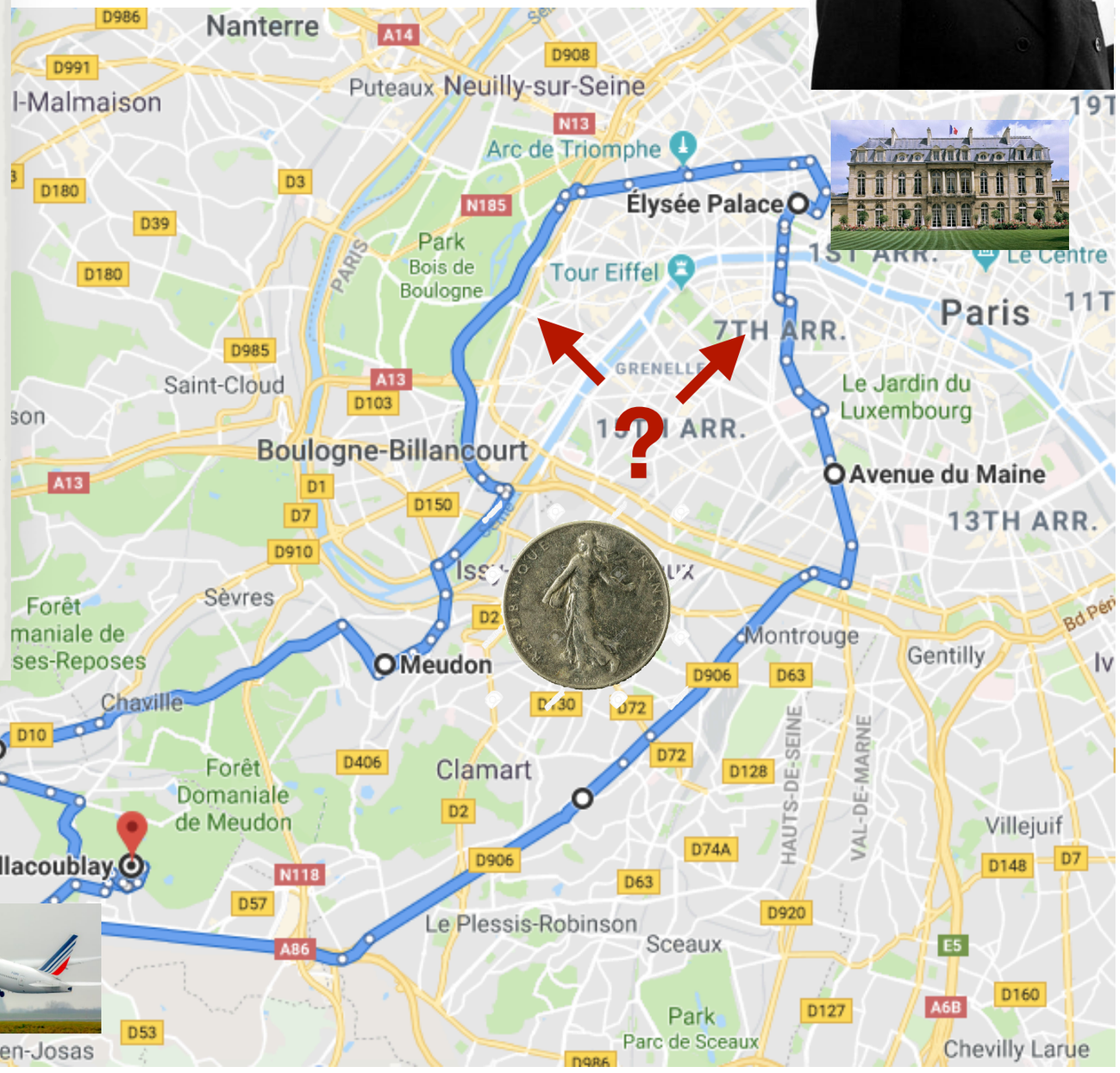
"His mission: kill President Charles de Gaulle.
"His code name: Jackal.
"His price half a million dollars.
"His demand: total secrecy, even from his employers."

The blurb goes on to describe the Jackal as "a tall, blond Englishman with opaque gray eyes —a killer at the top of his grisly profession."

Now it happens that the real "Jackal," that is, the man who really tried to kill de Gaulle, is a short, bespectacled, baldish Frenchman with clear blue eyes and the candid, pink-cheeked face of an aging choirboy. His code name was Max, his price was not a penny, and his demand was to put the army in power and keep Algeria French. His real name is Alain de Bougrenet de la Tocnaye. Like the Jackal's, his assassination attempt failed, but unlike the Jackal, he was not killed. He was tried, sentenced to death, pardoned, then amnestied in 1968. He now lives quietly in a two-room bachelor apartment in the shadow of the Eiffel Tower and operates a small trucking firm. I met him recently through friends. Annoyed at the mixture of fact and fiction in "The Jackal," which describes de la Tocnaye's attempt as a prelude to the hiring of the Jackal, he agreed to tell his story "to set the record straight." The point to remember is that the "Petit Clamart" attempt led by de la Tocnaye and graphically described in the novel really happened, but that after its failure, no further attempts to recruit assassins, domestic or foreign, were made. Since fiction borrows from reality, I have returned the compliment by borrowing some of Mr. Forsyth's chapter headings.

1. Anatomy of a crime

Aug. 22, 1962, was a cool, overcast day in Paris, more like autumn than summer. That morning, General de Gaulle came into the city from his country retreat at Colombey-les-deux-Eglises to preside over a Cabinet meeting. For some time now, ever since a bomb buried in a pile of sand had gone off along his route, without doing any damage, he had been making the trip by car and helicopter: driving from Colombey 40 miles to the airport of Saint-Dizier, flying 150 miles to the military airport of Villacoublay, and driving with a light escort the eight miles from Villacoublay to



Does randomly choosing the route make sense?

Lands tails
with probability p

Lands heads
with probability 1-p

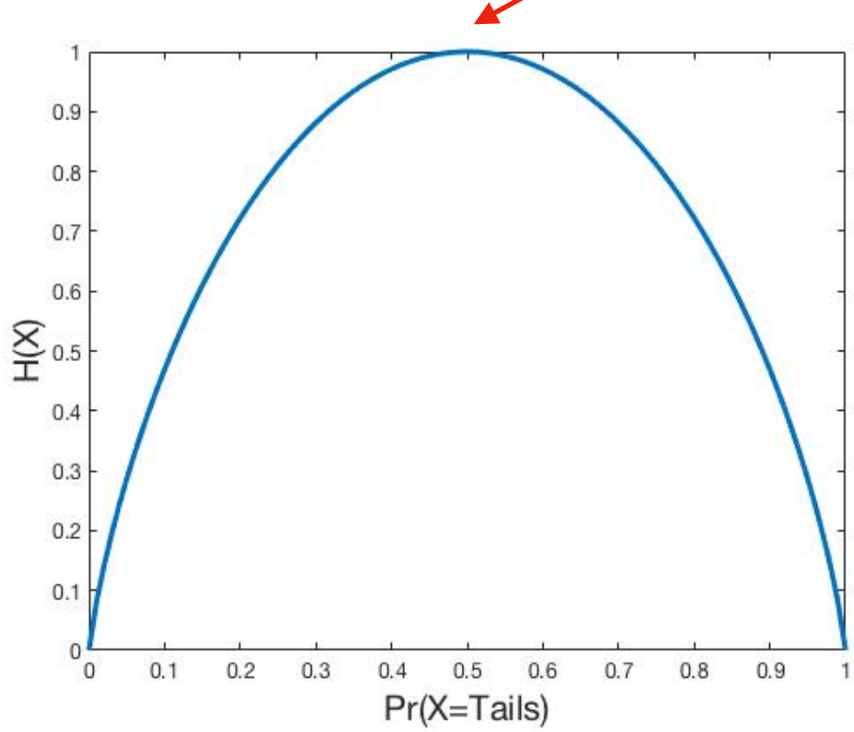
If p=1 or p=0, we know
what is going to happen (almost surely)

If 0<p<1, we can measure how much
we don't know what is going to happen

Entropy of the random variable X

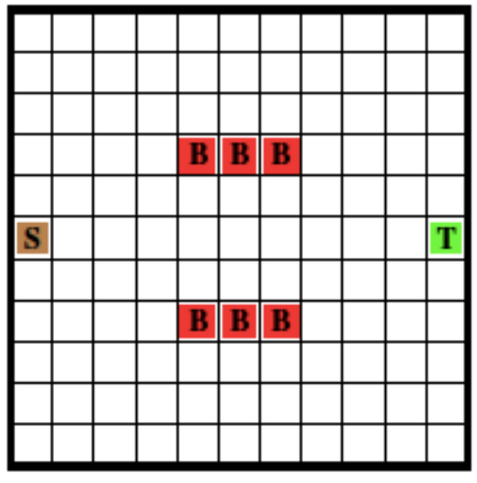
$$H(X)= -p \log(p) -(1-p)\log(1-p)$$

Unbiased coin

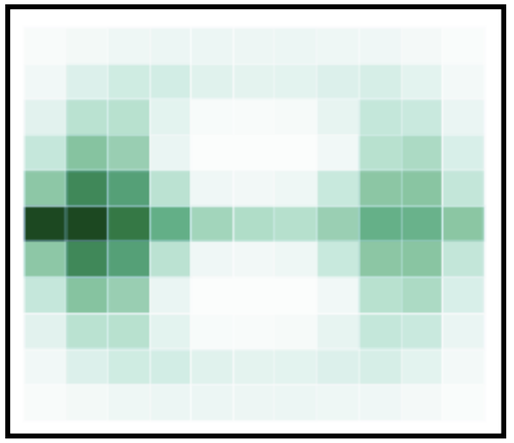


An informal statement of the considered problem

For an agent



- whose behavior in a stochastic environment
  is modeled as a Markov decision process (MDP),



- whose task is expressed by a reward function
  or a temporal logic formula,

Synthesize a policy that maximizes the entropy of the MDP
while guaranteeing the completion of the task with desired probability

Related work

- The maximum entropy of an interval Markov chain (IMC) is used to quantify the maximum information leakage from a deterministic software in [1].
  - Necessary and sufficient conditions for the finiteness of the maximum entropy of an IMC
  - The proposed approach cannot be used for Markov decision processes (MDPs)

- Complexity of computing the maximum entropy of an MDP is established in [2].
  - Maximum entropy can be computed in time polynomial in the size of the MDP
  - The proposed approach cannot be used to synthesize an entropy-maximizing policy

- The synthesis of a policy with maximum entropy is considered in [3].
  - A policy with maximum entropy is not the same with a policy that maximizes the entropy of the MDP
  - The proposed approach was believed to be non-convex

[1] F. Biondi et al. “Maximizing entropy over Markov processes”, Journal of Logical and Algebraic Methods in Programming, vol. 83, no. 5, pp. 384 – 399, 2014

[2] T. Chen and T. Han. “On the complexity of computing maximum entropy for Markovian models”, Conference on Foundation of Software Technology and Theoretical Computer Science, vol. 29, 2014, pp. 571–583.

[3] P. Paruchuri et al. “Security in multiagent systems by policy randomization”, Conference on Autonomous Agents and Multiagent Systems, 2006, pp. 273–280

Contributions

- Necessary and sufficient conditions for the existence of a stationary policy that maximizes the entropy of an MDP

- A polynomial-time algorithm for the synthesis of a stationary policy that maximizes the entropy of an MDP

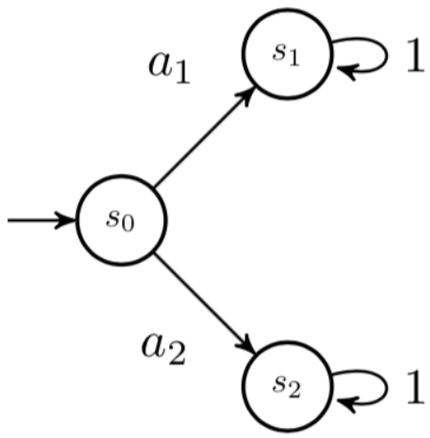
- A polynomial-time algorithm for the synthesis of a stationary policy that maximizes the entropy of an MDP under expected reward *and* temporal logic constraints

[1] Y. Savas et al. “Entropy maximization for constrained Markov decision processes”, Allerton Conference on Communication, Control, and Computing, 2018

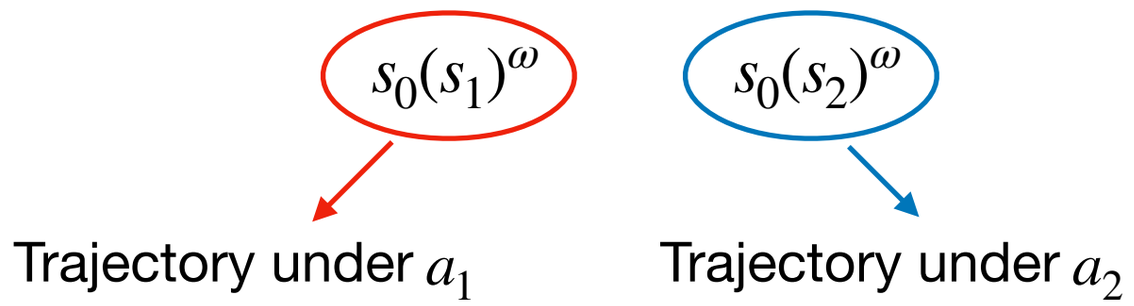
[2] Y. Savas et al. “Entropy maximization for Markov decision processes under temporal logic constraints”, Transactions on Automatic Control, 2019

The synthesis of a policy for a simple system

Let’s consider a simple system



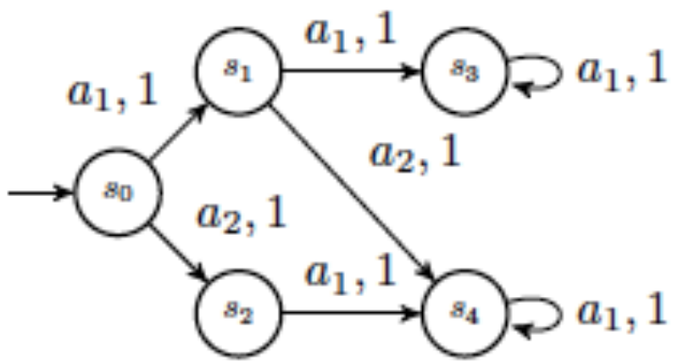
There are two possible trajectories:



If we take both actions with probability 0.5, we follow each trajectory with equal probability.

The synthesis of a policy for a ‘less simple’ system

Consider a bit more complex example

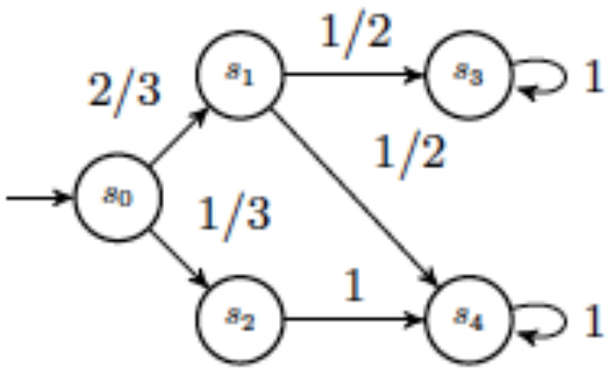


There are three possible trajectories:

$s_0 s_1 (s_3)^\omega$ $s_0 s_1 (s_4)^\omega$ $s_0 s_2 (s_4)^\omega$

Trajectory under $a_1 a_1$ Trajectory under $a_1 a_2$ Trajectory under $a_2 a_1$

The entropy-maximizing policy induces the following Markov chain

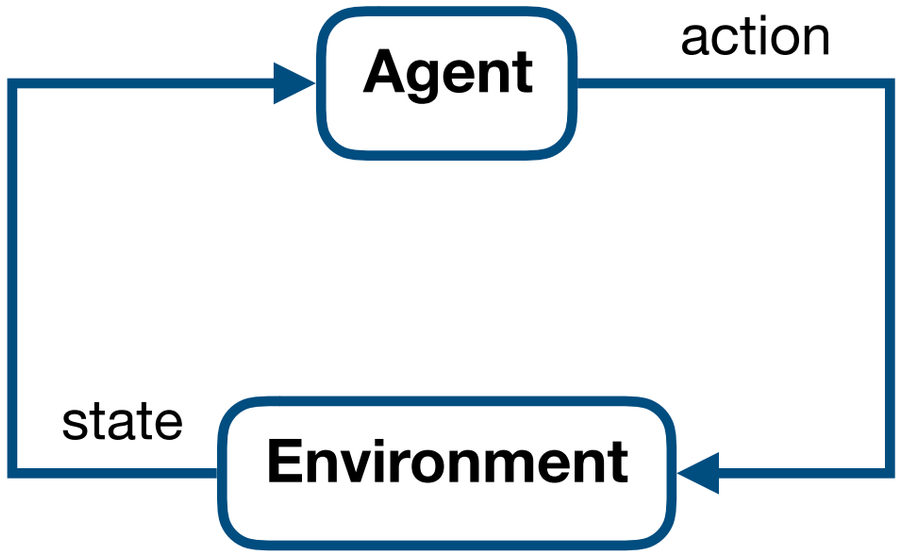


Each trajectory is followed with equal probability

<u>Not all actions are taken with equal probability</u>

Markov decision processes

Markov decision processes model
sequential decision-making under uncertainty



A *Markov decision process (MDP)* is a tuple $\mathscr{M} = (S, s_0, A, \mathbb{P})$

- S is a finite set of states
- s_0 is the initial state
- A is a finite set of actions
- $\mathbb{P}$ is a transition function

A *policy* for an MDP is a sequence $\pi = (\mu_0, \mu_1, \ldots)$ where $\mu_k : S \to \Delta(A)$

The set of all policies is $\Pi(\mathscr{M})$, the set of all stationary policies is $\Pi^S(\mathscr{M})$

The entropy of a stochastic process

Entropy of a discrete random variable X is

$$H(X) := -\sum_{x \in \mathcal{X}} Pr(X = x) \log Pr(X = x)$$

Entropy of a stochastic process $\mathbb{X} = \{X_k \in \mathcal{X} : k \in \mathbb{N}\}$ is

$$H(\mathbb{X}) := \lim_{k \to \infty} H(X_0, X_1, \ldots, X_k)$$

Measures the randomness of admissible realizations

Maximum entropy of an MDP

For an MDP $\mathscr{M}$, a policy π induces a stochastic process $\mathbb{X} = \{X_k \in S, k \in \mathbb{N}\}$.

The entropy of the *induced stochastic process* is denoted by

$$H(\mathscr{M}, \pi) := \lim_{k \to \infty} H(X_1, X_2, \ldots, X_k).$$

The maximum entropy of an MDP is then

$$H(\mathscr{M}) := \sup_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi).$$

Why do we consider only stationary policies?

$$\sup_{\pi \in \Pi(\mathscr{M})} H(\mathscr{M}, \pi) = \sup_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi)$$ [1]

[1] T. Chen and T. Han. “On the complexity of computing maximum entropy for Markovian models”, Conference on Foundation of Software Technology and Theoretical Computer Science, vol. 29, 2014, pp. 571–583.

Properties of the maximum entropy

The maximum entropy of an MDP is

finite if and only if $H(\mathscr{M}) = \max_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi) < \infty$ ← max can be attained and finite

infinite if and only if $H(\mathscr{M}) = \max_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi) = \infty$ ← max can be attained and infinite

unbounded if and only if

$$(i) \quad H(\mathscr{M}) = \sup_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi) = \infty$$

← sup is infinite but cannot be attained

$$(ii) \quad H(\mathscr{M}, \pi) < \infty \quad \forall \pi \in \Pi^S(\mathscr{M})$$

Formal problem statement

- For an MDP M, provide an algorithm to verify whether there exists a policy $\pi^\star \in \Pi^S(\mathscr{M})$ such that

$$H(\mathscr{M}) = H(\mathscr{M}, \pi^\star)\,.$$

We will provide necessary and sufficient conditions
for the existence of an entropy-maximizing stationary policy.

- If there exists an entropy-maximizing stationary policy, provide an algorithm to synthesize it.

  If such a policy does not exist, synthesize a policy $\pi^\star \in \Pi^S(\mathscr{M})$ such that

$$H(\mathscr{M}, \pi^\star) \geq L$$

  for a given constant L.

We will provide a polynomial-time algorithm
to synthesize an entropy-maximizing stationary policy.

Entropy as a sum of immediate rewards

For a state $s \in S$, let the local entropy be $L^{\pi}(s) := -\sum_{s' \in S} \mathbb{P}^{\pi}_{s,s'} \log \mathbb{P}^{\pi}_{s,s'}$,

Entropy gain in one step starting from state s

and the expected residence time be $\xi^{\pi}(s) := \sum_{k=0}^{\infty} Pr^{\pi}(X_k = s \,|\, X_0 = s_0)\,.$

Number of visits to state s

(One step reward) X (Number of visits)

Then, we have $H(\mathscr{M}, \pi) = \sum_{s \in S} L^{\pi}(s)\xi^{\pi}(s)\,.$[1]

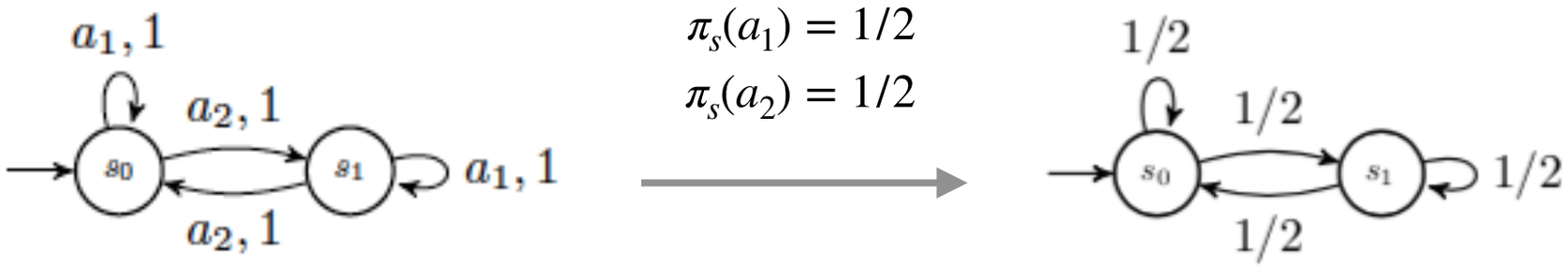
Finite if and only if all recurrent states have zero local entropy

Well, then we also have $H(\mathscr{M}) = \sup_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi) = \sup_{\pi \in \Pi^S(\mathscr{M})} \left[\sum_{s \in S} L^{\pi}(s)\xi^{\pi}(s)\right].$

(1) F. Biondi et al, “Maximizing entropy over Markov processes,” Journal of Logical and Algebraic Methods in Programming, vol. 83, no. 5, pp. 384 – 399, 2014

Verifying the existence of an entropy-maximizing policy: infinite case

Intuitively, maximum entropy is infinite if $L^\pi(s) > 0$ and $\xi^\pi(s) = \infty$ for some $\pi \in \Pi^S(\mathscr{M})$ and state s.



$$L^\pi(s_0) = L^\pi(s_1) = 1$$
$$\xi^\pi(s_0) = \xi^\pi(s_1) = \infty$$

A state s is recurrent under policy π if $\xi^\pi(s) = \infty$.

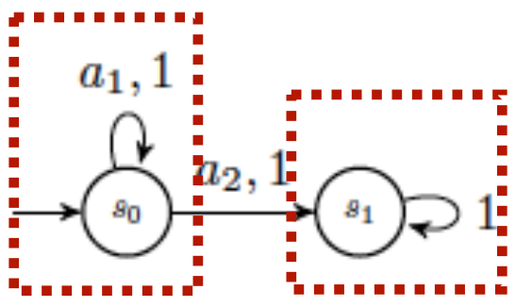
A state s is stochastic under policy π if $L^\pi(s) > 0$.

Maximum entropy is infinite if and only if
there exists a state that is both
recurrent and stochastic under some stationary policy.

Verifying the existence of an entropy-maximizing policy: unbounded case

Maximum entropy is unbounded if we can spend ***almost*** infinite time in a state with nonzero local entropy.

MEC but not bottom strongly connected



Bottom strongly connected MEC

A maximal end component (MEC) of an MDP:

- A pair (C,D) where C is a set of states
  D is a set of actions
- The directed graph (C,D) is strongly connected
- There is no (C',D') such that $(C', D') \neq (C, D),\ \ C \subseteq C',\ \ D \subseteq D'$

Bottom strongly connected MEC:

Under no action, the agent can leave the set C

Maximum entropy is unbounded if and only if
it is not infinite and there exists a maximal end component
that is not bottom strongly connected.

Verifying the existence of an entropy-maximizing policy: finite case

What about finite maximum entropy?

Recall the definition of finite maximum entropy

$$H(\mathscr{M}) = \max_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi) < \infty$$

Maximum entropy is finite if and only if
it is not infinite and not unbounded.

Corollary:

$$\sup_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi) < \infty \implies \sup_{\pi \in \Pi(\mathscr{M})} H(\mathscr{M}, \pi) = \max_{\pi \in \Pi^S(\mathscr{M})} H(\mathscr{M}, \pi)$$

An optimal stationary policy is optimal also in the space of all admissible policies

The synthesis of an entropy-maximizing policy: finite case

$$\max_{x(s,a)} \quad -\sum_{s\in S\setminus C}\sum_{t\in S}\eta(s,t)\log\left(\frac{\eta(s,t)}{\nu(s)}\right)$$

Relative entropy between
the number of times entering a state and
the number of times transitioning to a successor state

Subject to: $$\sum_{a\in A(s)} x(s,a) - \sum_{t\in S}\sum_{a\in A(t)} x(t,a)\mathbb{P}_{t,a,s} = \alpha(s) \qquad \forall s \in S\setminus C$$

Flow constraint:
if you visit a state,
you should also leave that state

$$\nu(s) = \sum_{a\in A(s)} x(s,a) \qquad \forall s \in S\setminus C$$

Expected number of visits to state s

$$\eta(s,t) = \sum_{t\in S} x(s,a)\mathbb{P}_{s,a,t} \qquad \forall s \in S\setminus C \quad \forall t \in S$$

Expected number of transitions
from state s to state t

$$x(s,a) \geq 0 \qquad \forall s \in S\setminus C \quad \forall a \in A$$

Expected number of visits to
state-action pair (s,a)

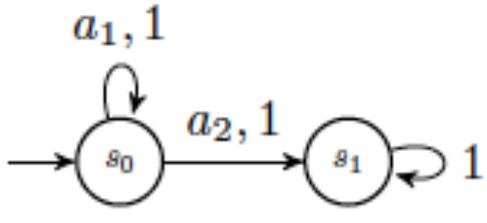
A convex optimization with exponential cone constraints

C : States in MECs

$\alpha(\cdot)$: Initial state distribution

The synthesis of an entropy-maximizing policy: unbounded case

What happens if the MDP has unbounded maximum entropy?



Maximum entropy is unbounded because
we can spend *almost* infinite time to gain entropy

Recall that in the convex optimization problem, we have a variable $x(s, a)$
which denotes the expected number of visits to state-action pair (s,a).

Then, we can bound the time spent until reaching a bottom strongly connected MEC
with the additional constraint

$$\sum_{s \in S \setminus S_B} \sum_{a \in A(s)} x(s, a) \leq B \,.$$

S_B : States in bottom strongly connected MECs

Obtain desired level of entropy
by increasing threshold B

Entropy maximization under expected reward constraints

For an MDP, a reward function $R : S \times A \to \mathbb{R}$ and a reward threshold Γ,

synthesize a policy that solves the following problem:

$$\begin{aligned} &\underset{\pi \in \Pi^S(\mathcal{M})}{\text{maximize}} && H(\mathcal{M}, \pi) \\ &\text{subject to:} && \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \mathcal{R}(s_t, a_t) \right] \geq \Gamma. \end{aligned}$$

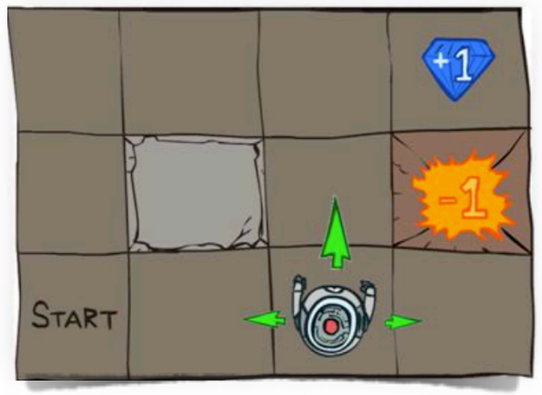
Assuming that $R(s, a) = 0$ for all states in bottom strongly connected MECs,

we can add the following constraint to the convex optimization problem.

$$\sum_{s \in S \setminus S_B} \sum_{a \in A} x(s, a) R(s, a) \geq \Gamma \, .$$

Adding linear temporal logic (LTL) constraints

Does the behavior of the robot satisfy the specification φ ?



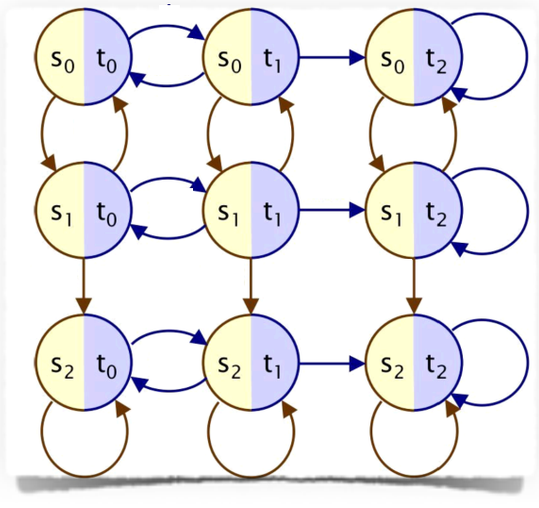
$$\Box\Diamond uav_1 = w_3 \wedge \Box\Diamond uav_1 = w_4 \wedge \Box\Diamond uav_1 = w_5 \wedge$$
$$\Diamond uav_2 = w_1 \wedge \Diamond uav_2 = w_2 \wedge \Diamond uav_2 = w_6$$

$$\Diamond\left(uav = w_1 \wedge \Diamond\left(uav = w_2 \wedge \Diamond uav = w_3\right)\right)$$

$$\Diamond uav = w_1 \wedge \Diamond uav = w_2 \wedge \Diamond uav = w_3 \wedge$$
$$\neg w_2 \,\mathsf{U}\, w_1 \wedge \neg w_3 \,\mathsf{U}\, w_2$$

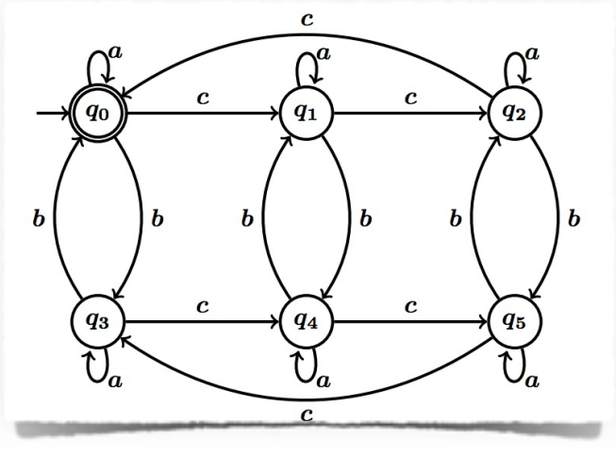
$$\Box(uav = w_1 \rightarrow \bigcirc\Box\neg uav = w_1)$$

MDP *M*



x

“specification” automaton A_φ



= A_φ *x M*

Are certain special states in A_φ *x M reached?*

Entropy maximization under linear temporal logic (LTL) constraints

For an MDP and an LTL formula, synthesize a policy that solves the following problem:

$$\begin{aligned} \underset{\pi \in \Pi^S(\mathcal{M}_p)}{\text{maximize}} \quad & H(\mathcal{M}_p, \pi) \\ \text{subject to:} \quad & \Pr^{\pi}_{\mathcal{M}_p}(s_0 \models \Diamond B) \geq \beta \end{aligned}$$

States in accepting MECs

Desired threshold

Optimization is performed on the product MDP

$\mathscr{M}_p := A_\varphi \times \mathscr{M}$

Then, the policy synthesis procedure is as follows:

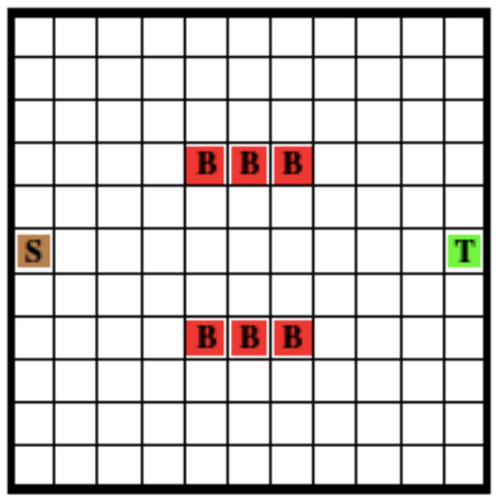
1. Construct the automata corresponding to the given LTL formula
2. Take the product of the given MDP and the automata
3. Make all $s \in B$ absorbing
4. Solve the convex optimization problem with the additional constraints

$$\sum_{s \in B} x(s) \geq \beta$$

$$x(s) - \sum_{t \in S \setminus B} \sum_{a \in A(t)} x(t,a) \mathbb{P}_{t,a,s} = \alpha(s) \quad \forall s \in B$$

Probability of reaching state s

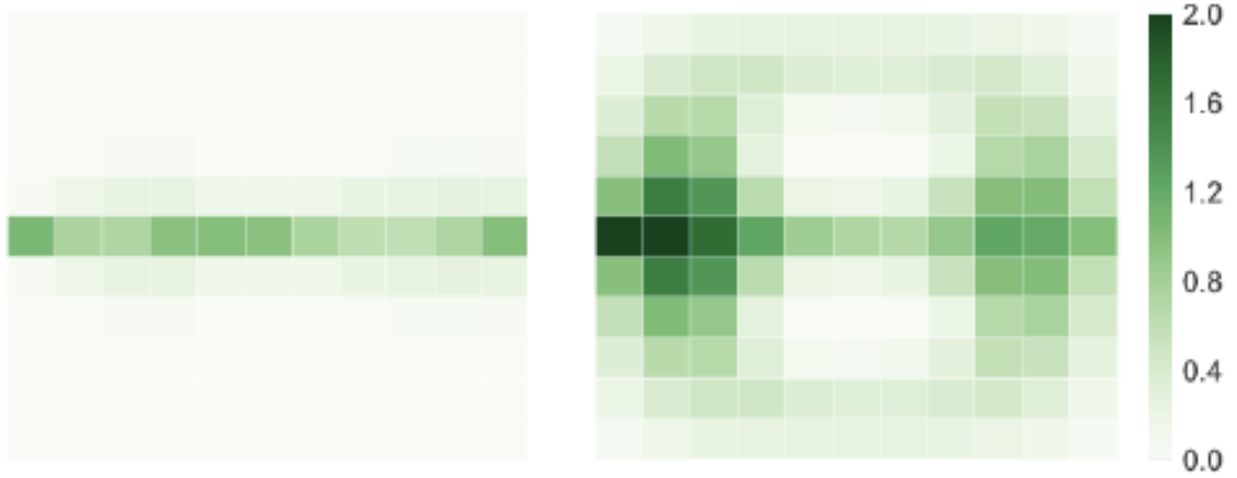
The use of routes for different mission times



Task: Reach the target (green) state within Γ steps
while avoiding red states

Reach the target
as soon as possible

$(\Gamma = 14)$

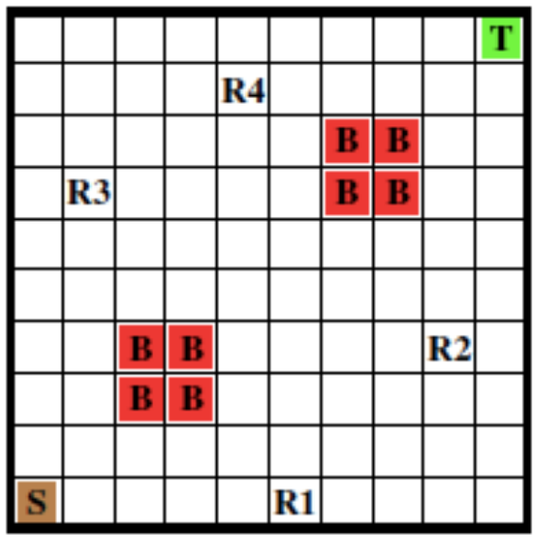


Reach the target
within 60 steps on average

$(\Gamma = 60)$

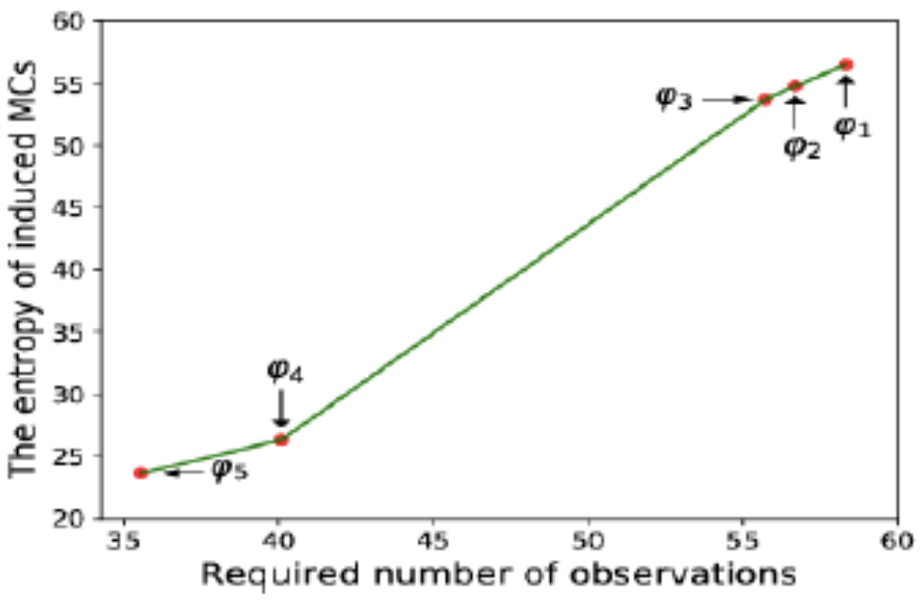
Expected residence time in states $\sum_{a \in A(s)} x(s,a)$

The use of routes for different missions and a fixed mission time



Restrictiveness of the tasks increases

$\varphi_1 = \Box\neg Red \wedge \Diamond\Box T$
$\varphi_2 = \Box\neg Red \wedge \Diamond R4 \wedge \Diamond\Box T$
$\varphi_3 = \Box\neg Red \wedge \Diamond(R3 \wedge \Diamond R4) \wedge \Diamond\Box T$
$\varphi_4 = \Box\neg Red \wedge \Diamond(R2 \wedge \Diamond(R3 \wedge \Diamond R4)) \wedge \Diamond\Box T$
$\varphi_5 = \Box\neg Red \wedge \Diamond(R1 \wedge \Diamond(R2 \wedge \Diamond(R3 \wedge \Diamond R4))) \wedge \Diamond\Box T$



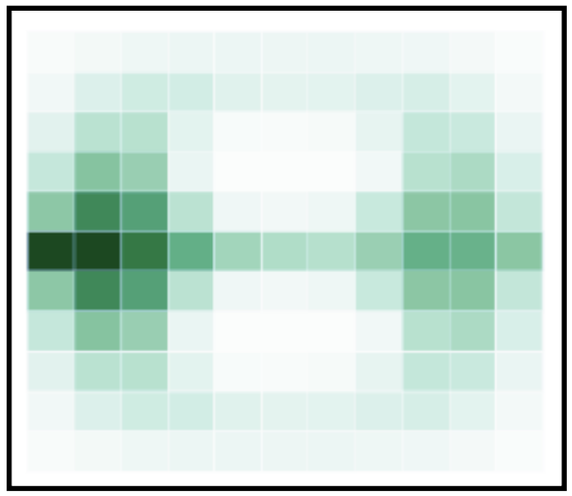
Maximum entropy decreases as the task gets stricter

Number of observations required to infer the trajectory decreases as the task gets stricter

Summary

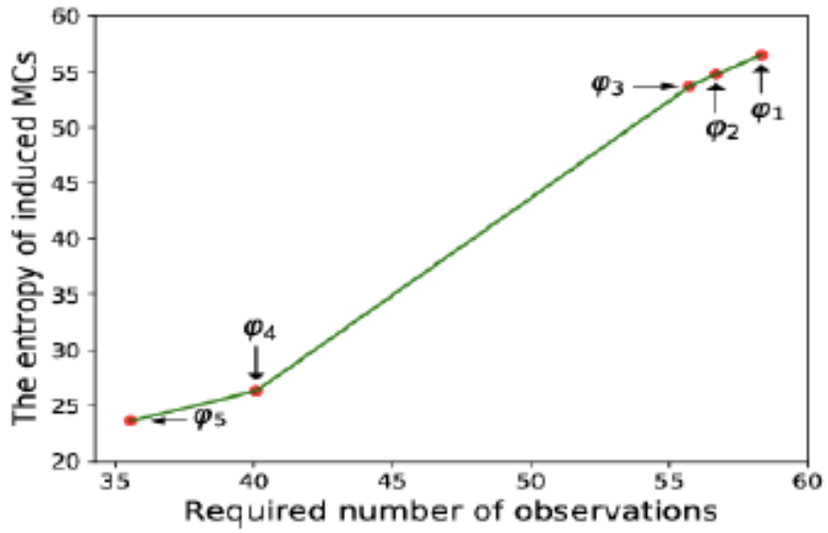
- Existence of an entropy-maximizing policy can be verified efficiently.
- An entropy-maximizing policy can be synthesized efficiently.
- A policy that maximizes entropy while satisfying reward and temporal logic constraints can be synthesized efficiently.

$$\begin{aligned} \underset{\pi \in \Pi^S(\mathcal{M}_p)}{\text{maximize}} \quad & H(\mathcal{M}_p, \pi) \\ \text{subject to:} \quad & \Pr^{\pi}_{\mathcal{M}_p}(s_0 \models \Diamond B) \geq \beta \end{aligned}$$



What is next?

- What happens if the agent has only partial observations?
- What happens if the outside observer is an active player in the game?
- What happens if the outside observer is not interested in the whole trajectory?



[1] M. Hibbard et al. ‘Unpredictable planning under partial observability’, Conference on Decision and Control, 2019

[2] Y. Savas et al. ‘Entropy-regularized stochastic games’, Conference on Decision and Control, 2019