# Least Inferable Policies for Markov Decision Processes

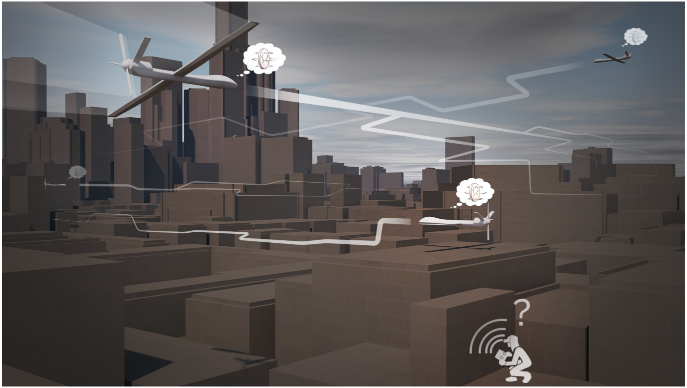
Mustafa O. Karabag, Ufuk Topcu

autonomous SYSTEMS GROUP

M. O. Karabag, M. Ornik, and U. Topcu. "Least Inferable Policies for Markov Decision Processes." In *2019 American Control Conference*, pages 1224-1231, 2019.

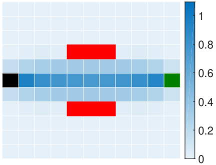
# Motivation

An agent that is performing a task in a stochastic environment while being observed by an adversary, should not have an inferable policy.

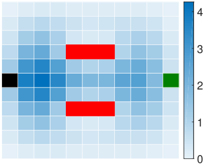


# Motivation

Increasing the entropy does not imply non-inferabilility of the policy.




Least inferable policy




Maximum-entropy policy
(Time limit $= 120$)

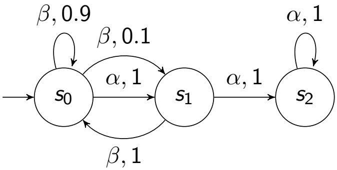
**The idea**

Synthesize a policy that satisfies some task constraints and limits the ability of the observer to infer.

# The model

We model the environment with a Markov decision process (MDP) $\mathcal{M} = (S, \mathcal{A}, \mathcal{P}, s_0)$.

- $S$ is a finite set of states,
- $\mathcal{A}$ is a finite set of actions,
- $\mathcal{P} : S \times \mathcal{A} \times S \to [0, 1]$ is the transition probability function,
- $s_0$ is the initial state.




A policy is a sequence $\pi = \mu_0 \mu_1 \ldots$ where each $\mu_t : S \times \mathcal{A} \to [0, 1]$ is a function such that $\sum_{a \in \mathcal{A}(s)} \mu_t(s, a) = 1$ for every $s \in S$.

# The problem

- The task constraint of the agent is to reach a set $S_{reach}$ of states with high probability.

- The adversary observes the transitions of the agent at a set $W$ of states to estimate the transition probabilities.

- The objective of the agent is to minimize the information leaked to the adversary on the transition probabilities.
  - What is **leaked information**?

# Informativeness of a random variable

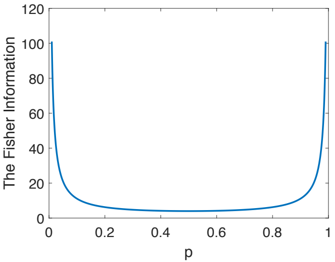
The *Fisher information* of a discrete random variable $X$ on $\theta$ is

$$I_X(\theta) := \mathbb{E}_X\left[\left(\underbrace{\frac{\partial f(X|\theta)}{\partial \theta}}_{\textit{score}}\right)^2 \middle| \theta\right]$$

where $f(X|\theta)$ is the probability mass function.

Example: $Y \sim \textit{Bernoulli}(p)$.
$I_Y(p) = (p(1-p))^{-1}$

If $p = 0$ or 1, the inference is easy, i.e., the estimation error is low.



# Informativeness of a random variable

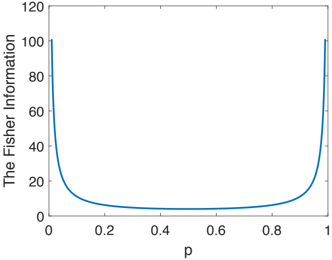
The *Fisher information* of a discrete random variable $X$ on $\theta$ is

$$I_X(\theta) := \mathbb{E}_X\left[\left(\underbrace{\frac{\partial f(X|\theta)}{\partial \theta}}_{\textit{score}}\right)^2 \middle| \theta\right]$$

where $f(X|\theta)$ is the probability mass function.

Example: $Y \sim \textit{Bernoulli}(p)$.
$I_Y(p) = (p(1-p))^{-1}$

If $p = 0$ or $1$, the inference is easy, i.e., the estimation error is low.




$X \leftrightarrow$ successor states, $\theta \leftrightarrow$ transition probabilities

# Lower bound on the estimation error

The *Fisher information* of a discrete random variable $X$ on $\theta$ is

$$I_X(\theta) := \mathbb{E}_X\left[\left(\frac{\partial f(X|\theta)}{\partial\theta}\right)^2 \middle| \theta\right]$$

where $f(X|\theta)$ is the probability mass function.

*The Cramèr-Rao Bound*: Suppose the random variable $X$ is parametrized by $\theta$. The variance of any unbiased estimator $\hat{\theta}$ of $\theta$ is lower bounded by the reciprocal of the Fisher information $I_X(\theta)$:

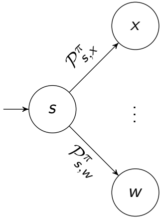
$$Var(\hat{\theta}) \geq \frac{1}{I_X(\theta)}.$$

# Information leaked from a single transition

The *transition information* of a state $s$ is

$$\iota_s^\pi := \frac{1}{\sum_{q \in S} I_Q(\mathcal{P}_{s,q}^\pi)^{-1}}$$

where $Q$ is the random variable denoting the successor state of state $s$.




Analogous to Fisher information: Let $\hat{\mathcal{P}}_s$ be an unbiased estimator of the transition probabilities $\mathcal{P}_s^\pi$ at state $s$. Then,

$$trace(Var(\hat{\mathcal{P}}_s)) \geq \frac{1}{\iota_s^\pi}.$$

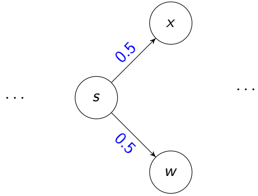
# Information leaked from a path

The *total information* of a path $\xi = s_0 s_1 \ldots$ is

$$\iota^{\pi}_{W,\xi} := \sum_{t=0}^{\infty} \mathbb{1}_W(s_t)\iota^{\pi}_{s_t}.$$

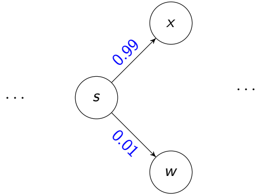
Quantity vs. Informativeness of observations

25 samples from




vs.

1 sample from




Both can be inferred equally well.

# The problem

**Given**

- an MDP $\mathcal{M} = (S, \mathcal{A}, \mathcal{P}, s_0)$,
- a set $S_{reach}$ of states,
- a probability threshold $\nu_{reach}$,
- the set $W$ of observed states,

**compute**

$$\begin{aligned} \min_{\pi} \quad & \overbrace{\mathbb{E}_{\xi}\left[\iota^{\pi}_{W,\xi}\right]}^{\text{Expected total information}} \\ \text{subject to} \quad & \Pr^{\pi}(Reach[S_{reach}]) \geq \nu_{reach}. \end{aligned}$$

# Limiting the policy space

**Assumption**

The policy $\pi$ of the agent is stationary, i.e., $\pi = \mu\mu\ldots$.

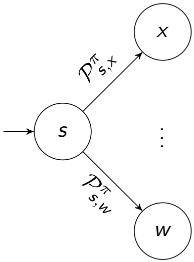
For a stationary policy,

- the expected state residence time at state $s$ is $x_s^\pi = \mathbb{E}[\sum_{t=0}^{\infty} \mathbb{1}_s(s_t)]$,
- the expected state-action residence time at state $s$ and action $a$ is $x_{s,a}^\pi = x_s^\pi \pi_{s,a}$,
- the expected state-state residence time from state $s$ to state $q$ is $y_{s,q}^\pi = \sum_{a \in \mathcal{A}(s)} x_{s,a} \mathcal{P}_{s,a,q}$.

# The transition information under a stationary policy $\pi$

In terms of transition probabilities:

$$\iota_s^\pi = \left( \sum_{q \in S} \mathcal{P}_{s,q}^\pi (1 - \mathcal{P}_{s,q}^\pi) \right)^{-1}.$$




In terms of expected residence times:

$$\iota_s^\pi = \left( \sum_{q \in S} \frac{y_{s,q}^\pi}{x_s^\pi} \left( 1 - \frac{y_{s,q}^\pi}{x_s^\pi} \right) \right)^{-1}.$$

# A minimum-information admissible policy can be synthesized with a convex optimization problem

$$\min_{x^{\pi}_{s,a}} \sum_{w \in W} x^{\pi}_{w} \iota^{\pi}_{w}$$

Expected total information

$$\text{subject to } \iota^{\pi}_{w} = \left( \sum_{q \in S} \frac{y^{\pi}_{w,q}}{x^{\pi}_{w}} \left( 1 - \frac{y^{\pi}_{s,q}}{x^{\pi}_{w}} \right) \right)^{-1}, \quad \forall w \in W$$

$$x^{\pi}_{s,a} \geq 0, \quad \forall s \in S \setminus C, \ \forall a \in \mathcal{A}(s)$$

$$x^{\pi}_{s} = \sum_{a \in \mathcal{A}(s)} x^{\pi}_{s,a}, \quad \forall s \in S \setminus C$$

$$y^{\pi}_{s,q} = \sum_{a \in \mathcal{A}(s)} x^{\pi}_{s,a} \mathcal{P}_{s,a,q}, \quad \forall s \in S \setminus C, \ \forall q \in S$$

$$x^{\pi}_{s} - \sum_{q \in S} y_{q,s} = \mathbb{1}_{s_0}(s), \quad \forall s \in S \setminus C$$

Flow equations to describe feasible policies

$$\sum_{q \in S_{reach}} \sum_{s \in S \setminus C} y^{\pi}_{s,q} + \mathbb{1}_{s_0}(q) \geq \nu_{reach}.$$

The task constraint

$C$ is the set of the end component states

# Lower bound on the estimation error for a state

Let $\sigma_w$ be the mean-squared error of an (**any**) unbiased estimator for the transition probabilities at state $w$.

A random path of the agent is the observed data.

## Proposition

For an MDP $\mathcal{M}$ and a stationary policy $\pi \in \Pi^{St}(\mathcal{M})$,

$$\sigma_w \geq \frac{(\Pr^{\pi}(Reach[w]))^2}{x_w^{\pi} \iota_w^{\pi}}$$

for every state $w \in W$.

The reachability probability to state $w$ under stationary policy $\pi$

The expected leaked information from state $w$

# Lower bound on the total estimation error

## Corallary

For an MDP $\mathcal{M}$ and a stationary policy $\pi \in \Pi^{St}(\mathcal{M})$,

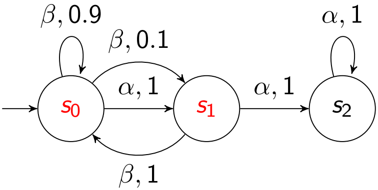
the total MSE $\sum_{w \in W} \sigma_w$ satisfies

The size of set $W$

$$\sum_{w \in W} \sigma_w \geq \frac{\min_{w \in W} \left(\Pr^{\pi}(Reach[w])\right)^2 |W|^2}{\mathbb{E}[\iota^{\pi}_{W,\xi}]}.$$

The expected total information

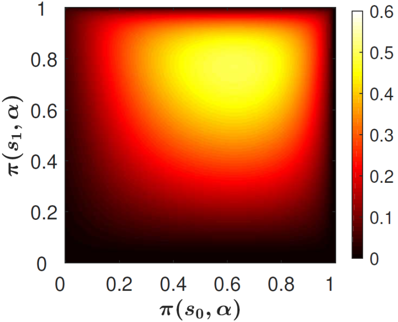
# Examples: Estimation error




$s_0, s_1$: observed

The reachability probability to $s_0$ and $s_1$ is 1 under any policy.

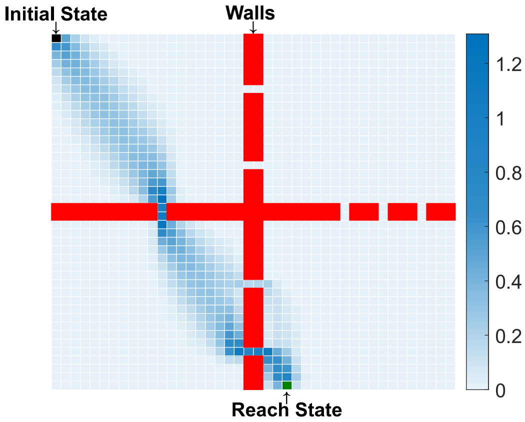
Lower bound on the total mean-squared error.




Reciprocal of the expected total information $\leq$ Total MSE of any unbiased estimator

# Examples: Characteristics of the minimum-information admissible policies

Heat map of the expected residence times:

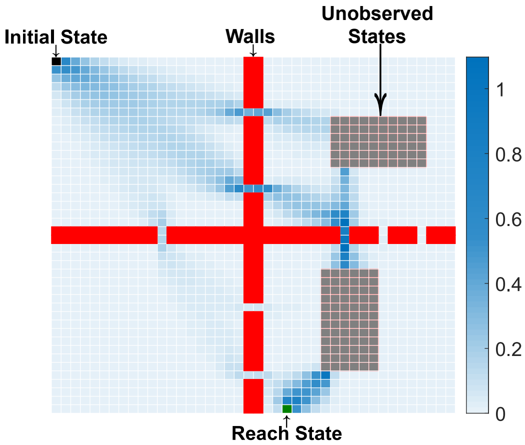



Minimum-information admissible policy yields:

- low number of observations
- less informative observations

# Examples: Characteristics of the minimum-information admissible policies

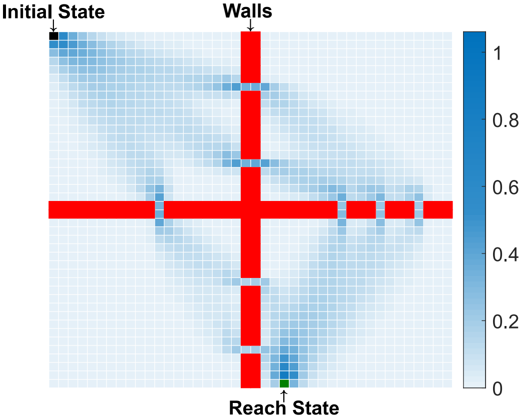
Heat map of the expected residence times:




Minimum-information admissible policy prefers unobserved regions.

# Examples: Characteristics of the minimum-information admissible policies with macro-level transition information

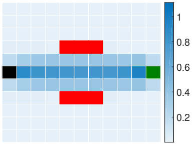
Heat map of the expected residence times:



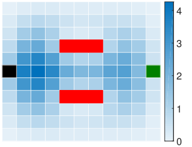

Penalizing the transition information for the gates results in randomization between the gates.

# Examples: Comparison of estimation error to maximum-entropy policies

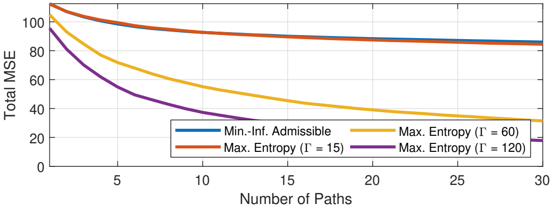
Maximum-entropy policy maximizes the entropy of path distribution given an upper limit Γ on the expected residence times.

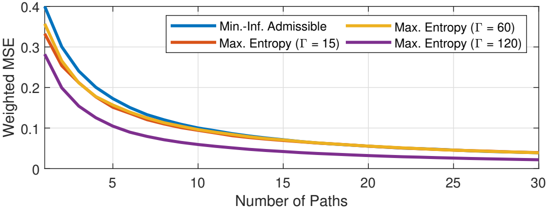



Minimum-information admissible policy



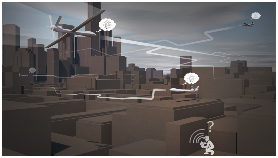

Max. entropy policy (Limit $\Gamma = 120$)

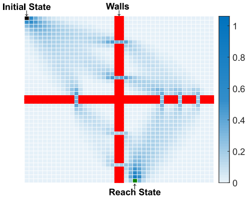


# Summary

- Leaked information can be measured with Fisher information
- Computing a minimum-information admissible policy requires to solve a convex optimization problem.
- Estimation error $\propto 1/$ Expected total information







M. O. Karabag, M. Ornik, and U. Topcu. "Least Inferable Policies for Markov Decision Processes." In *2019 American Control Conference*, pages 1224-1231, 2019.