Approximate information state for partially observed systems

Jayakumar Subramanian and Aditya Mahajan
McGill University

Thanks to Amit Sinha and Raihan Seraj for simulation results

IEEE Conference on Decision and Control
11 December 2019

Many successes of RL in recent years

- Algorithms based on comprehensive theory

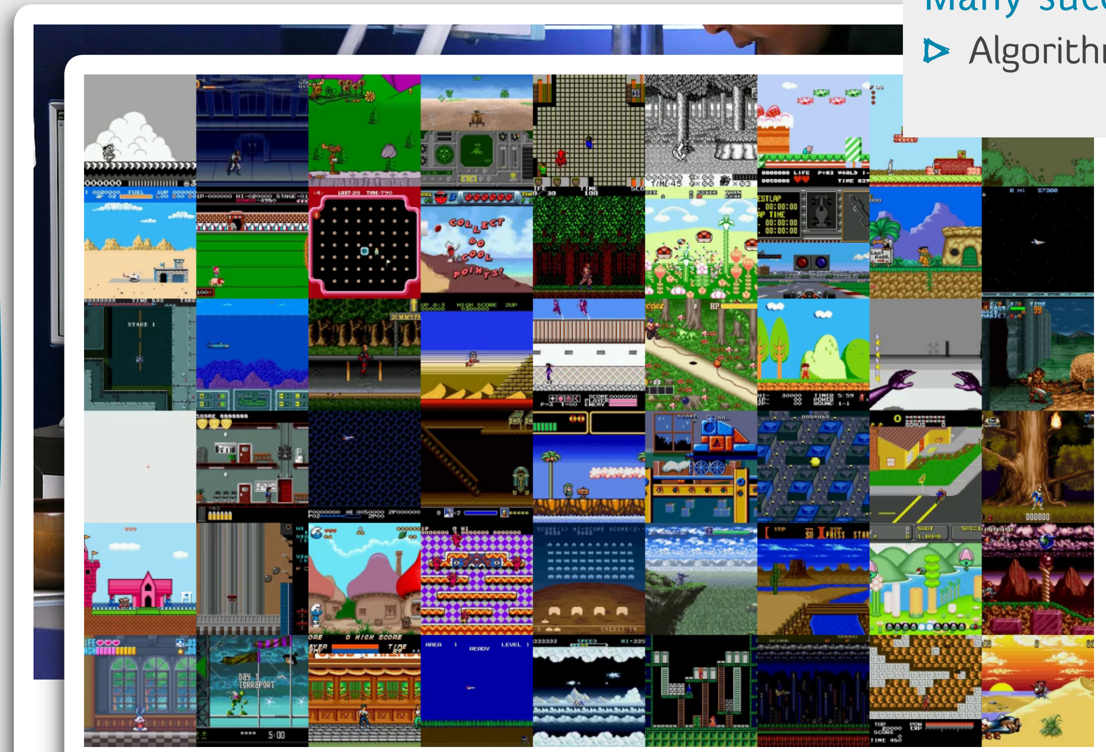
Many successes of RL in recent years

- Algorithms based on comprehensive theory

Alpha Go

Many successes of RL in recent years

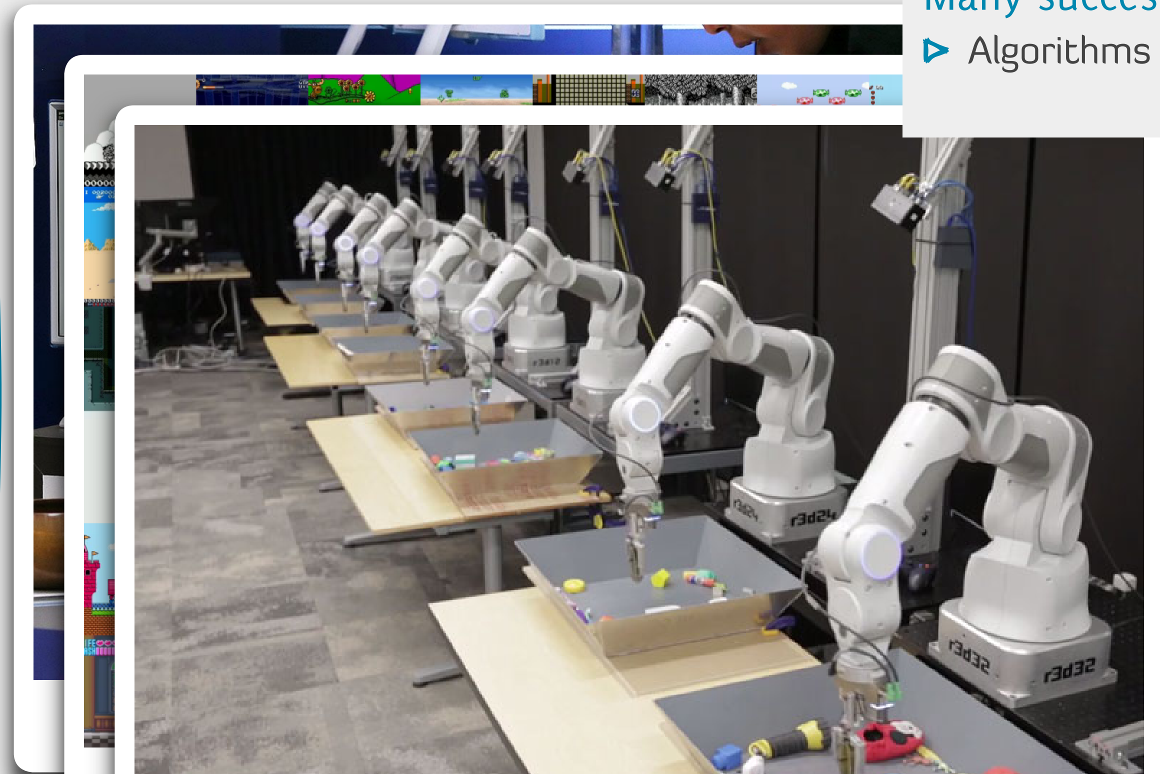
- Algorithms based on comprehensive theory



Arcade games

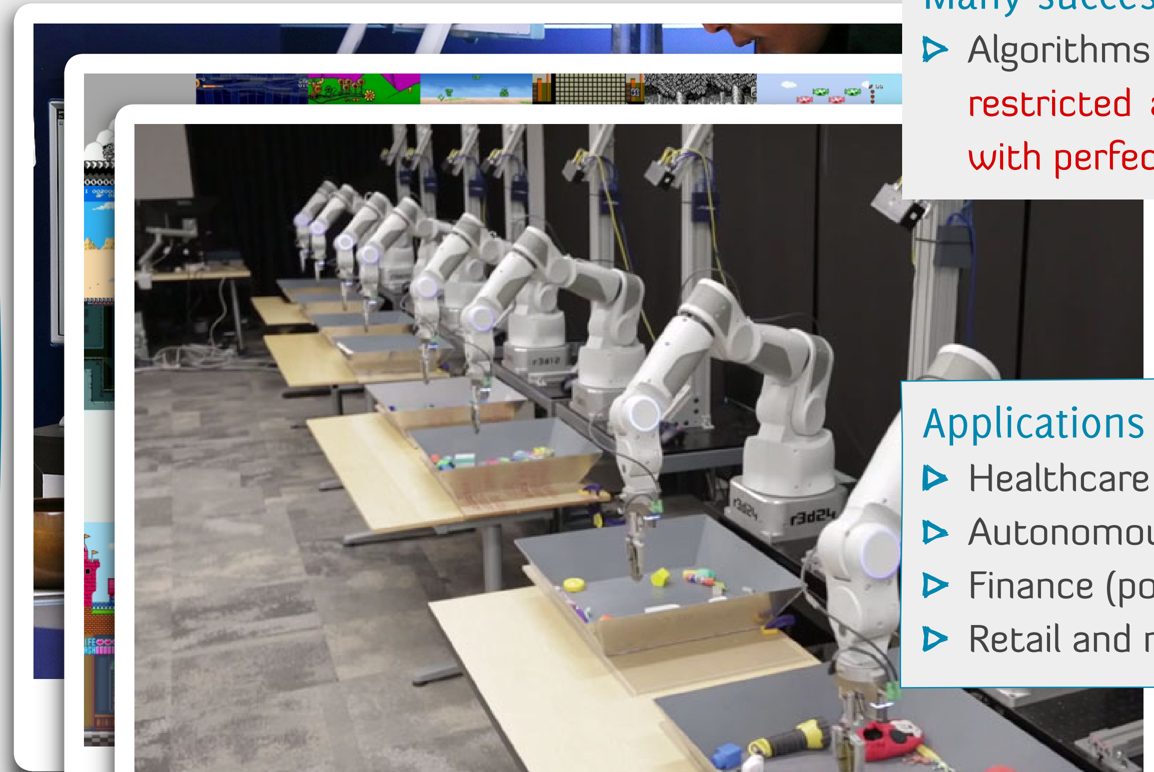
Many successes of RL in recent years

- Algorithms based on comprehensive theory



Robotics

Approx. info. state–(Subramanian and Mahajan)



Robotics

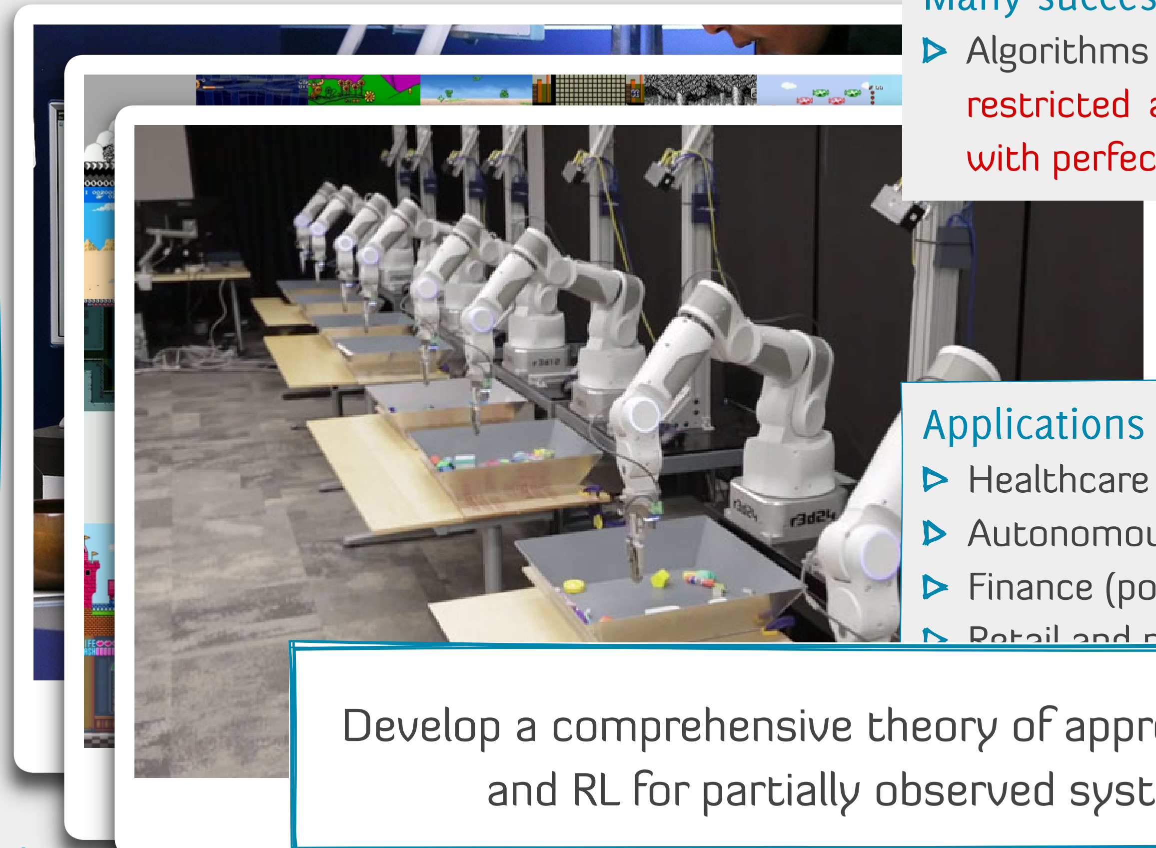
Many successes of RL in recent years

- Algorithms based on comprehensive theory restricted almost exclusively to systems with perfect state observations.

Applications with partially observed state

- Healthcare
- Autonomous driving
- Finance (portfolio management)
- Retail and marketing

Approx. Info. state–(Subramanian and Mahajan)



Many successes of RL in recent years

- Algorithms based on comprehensive theory **restricted almost exclusively to systems with perfect state observations.**

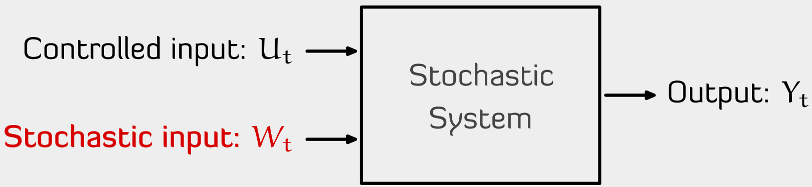
Applications with partially observed state

- Healthcare
- Autonomous driving
- Finance (portfolio management)
- Retail and marketing

Develop a comprehensive theory of approximate DP and RL for partially observed systems

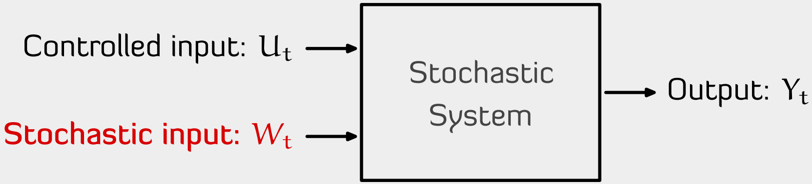
Notion of information state for partially observed systems

Notion of state in partially observed stochastic dynamical systems



$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

Notion of state in partially observed stochastic dynamical systems

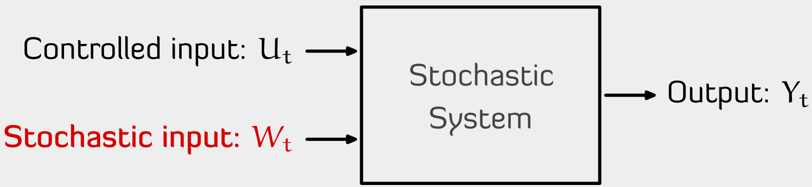


$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

STOCHASTIC INPUT IS NOT OBSERVED

Let $H_t = (Y_{1:t-1}, U_{1:t-1})$ denote the history of inputs and OUTPUTS until time t.

Notion of state in partially observed stochastic dynamical systems



$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

STOCHASTIC INPUT IS NOT OBSERVED

Let $H_t = (Y_{1:t-1}, U_{1:t-1})$ denote the history of inputs and OUTPUTS until time t.

TRADITIONAL SOLUTION: BELIEF STATES

Step 1 Identify a state $\{S_t\}_{t \geqslant 0}$ for predicting output assuming that the stochastic inputs are observed.

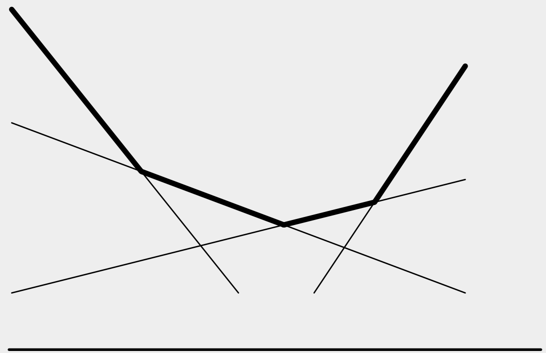
Step 2 Define a BELIEF STATE $B_t \in \Delta(\mathcal{S})$:

$$B_t(s) = \mathbb{P}(S_t = s \mid H_t = h_t), \quad s \in \mathcal{S}.$$

▷ Astrom, "Optimal control of Markov decision processes with incomplete state information," 1965. ▷ Striebel, "Sufficient statistics in the optimal control of stochastic systems," 1965. ▷ Baum and Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," 1966. ▷ Stratonovich, "Conditional Markov processes," 1960.

Partially observed Markov decision processes (POMDPs): Pros and Cons of belief state representation

Value function is piecewise linear and convex.

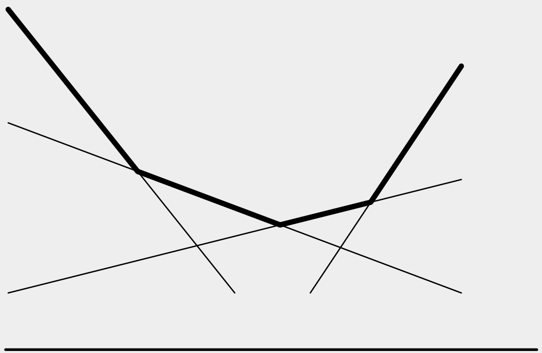


Is exploited by various efficient algorithms.

- Smallwood and Sondik, "The optimal control of partially observable Markov process over a finite horizon," 1973.
- Chen, "Algorithms for partially observable Markov decision processes," 1988.
- Kaelbling, Littmam, Cassandra, "Planning and acting in partially observable stochastic domains," 1998.
- Pineau, Gordon, Thrun, "Point-based value iteration: an anytime algorithm for POMDPs," 2003.

Partially observed Markov decision processes (POMDPs): Pros and Cons of belief state representation

Value function is piecewise linear and convex.



Is exploited by various efficient algorithms.

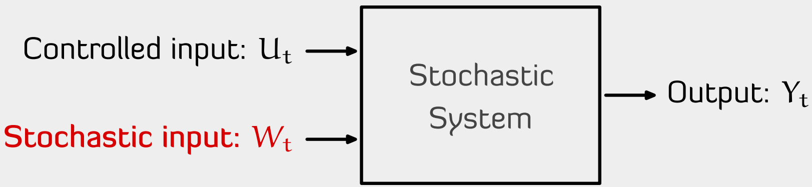
When the state space model is not known analytically (as is the case for black-box models and simulators as well as some real world application such as healthcare), belief states are difficult to construct and difficult to approximate from data.

- Smallwood and Sondik, "The optimal control of partially observable Markov process over a finite horizon," 1973.
- Chen, "Algorithms for partially observable Markov decision processes," 1988.
- Kaelbling, Littmam, Cassandra, "Planning and acting in partially observable stochastic domains," 1998.
- Pineau, Gordon, Thrun, "Point-based value iteration: an anytime algorithm for POMDPs," 2003.

Is there another ways to model partially observed systems which is more amenable to approximations?

Let's go back to first principles.

Notion of state in partially observed stochastic dynamical systems

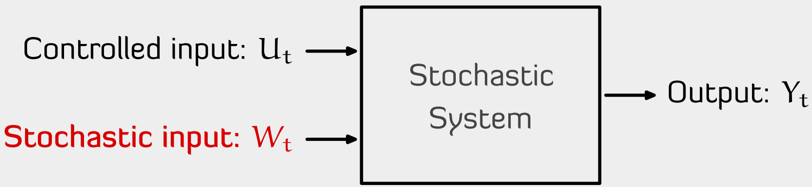


$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

WHEN THE STOCHASTIC INPUT IS NOT OBSERVED

Let $H_t = (Y_{1:t-1}, U_{1:t-1})$ denote the history of inputs and OUTPUTS until time t.

Notion of state in **partially observed** stochastic dynamical systems



$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

WHEN THE STOCHASTIC INPUT IS NOT OBSERVED

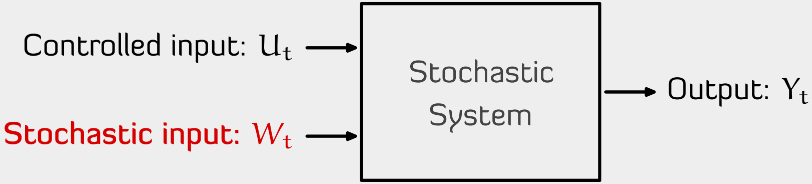
Let $H_t = (Y_{1:t-1}, U_{1:t-1})$ denote the history of inputs and OUTPUTS until time t.

PREDICTING OUTPUTS ALMOST SURELY

$H_t^{(1)} \sim H_t^{(2)}$ if for all future inputs $(U_{t:T}, W_{t:T})$,

$$Y_{t:T}^{(1)} = Y_{t:T}^{(2)}, \quad a.s.$$

Notion of state in partially observed stochastic dynamical systems



$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

WHEN THE STOCHASTIC INPUT IS NOT OBSERVED

Let $H_t = (Y_{1:t-1}, U_{1:t-1})$ denote the history of inputs and OUTPUTS until time t.

PREDICTING OUTPUTS ALMOST SURELY

$H_t^{(1)} \sim H_t^{(2)}$ if for all future inputs $(U_{t:T}, W_{t:T})$,

$$Y_{t:T}^{(1)} = Y_{t:T}^{(2)}, \quad a.s.$$

FORECASTING OUTPUTS IN DISTRIBUTION

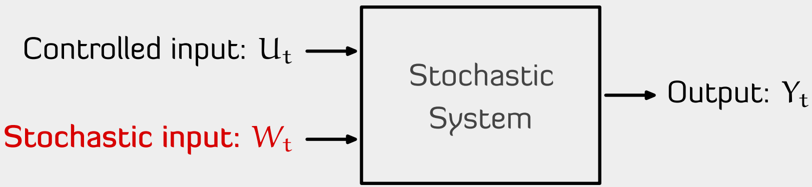
$H_t^{(1)} \sim H_t^{(2)}$ if for all future CONTROL inputs $U_{t:T}$,

$$\mathbb{P}(Y_{t:T}^{(1)} \mid H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} \mid H_t^{(2)}, U_{t:T})$$

- Grassberger, "Complexity and forecasting in dynamical systems," 1988.
- Cruthfield and Young, "Inferring statistical complexity," 1989.

Approx. info. state–(Subramanian and Mahajan)

Notion of state in partially observed stochastic dynamical systems



$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

WHEN THE STOCHASTIC INPUT IS NOT OBSERVED

Let $H_t = (Y_{1:t-1}, U_{1:t-1})$ denote the history of inputs and OUTPUTS until time t.

PREDICTING OUTPUTS ALMOST SURELY

$H_t^{(1)} \sim H_t^{(2)}$ if for all future inputs $(U_{t:T}, W_{t:T})$,

$$Y_{t:T}^{(1)} = Y_{t:T}^{(2)}, \quad a.s.$$

Too restrictive . . .

FORECASTING OUTPUTS IN DISTRIBUTION

$H_t^{(1)} \sim H_t^{(2)}$ if for all future CONTROL inputs $U_{t:T}$,

$$\mathbb{P}(Y_{t:T}^{(1)} \mid H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} \mid H_t^{(2)}, U_{t:T})$$

- Grassberger, "Complexity and forecasting in dynamical systems," 1988.
- Cruthfield and Young, "Inferring statistical complexity," 1989.

Now let's consturct the state space

FORECASTING OUTPUTS IN DISTRIBUTION

$H_t^{(1)} \sim H_t^{(2)}$ if for all future CONTROL inputs $U_{t:T}$,

$$\mathbb{P}(Y_{t:T}^{(1)} \mid H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} \mid H_t^{(2)}, U_{t:T})$$

Now let's consturct the state space

FORECASTING OUTPUTS IN DISTRIBUTION

$H_t^{(1)} \sim H_t^{(2)}$ if for all future CONTROL inputs $U_{t:T}$,

$$\mathbb{P}(Y_{t:T}^{(1)} \mid H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} \mid H_t^{(2)}, U_{t:T})$$

PROPERTIES OF INFORMATION STATE

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:

$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$\mathbb{P}(Y_t \mid H_t, U_t) = \mathbb{P}(Y_t \mid Z_t, U_t).$$

Approx. info. state–(Subramanian and Mahajan)

Now let's consturct the state space

FORECASTING OUTPUTS IN DISTRIBUTION

$H_t^{(1)} \sim H_t^{(2)}$ if for all future CONTROL inputs $U_{t:T}$,

$$\mathbb{P}(Y_{t:T}^{(1)} \mid H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} \mid H_t^{(2)}, U_{t:T})$$

PROPERTIES OF INFORMATION STATE

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:

$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$\mathbb{P}(Y_t \mid H_t, U_t) = \mathbb{P}(Y_t \mid Z_t, U_t).$$

Same complexity as identifying the state sufficient for forecasting outputs for the case of perfect observations (which was Step 1 for belief state formulations)

Now let's consturct the state space

FORECASTING OUTPUTS IN DISTRIBUTION

$H_t^{(1)} \sim H_t^{(2)}$ if for all future CONTROL inputs $U_{t:T}$,

$$\mathbb{P}(Y_{t:T}^{(1)} \mid H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} \mid H_t^{(2)}, U_{t:T})$$

Same complexity as identifying the state sufficient for forecasting outputs for the case of perfect observations (which was Step 1 for belief state formulations)

PROPERTIES OF INFORMATION STATE

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:

$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

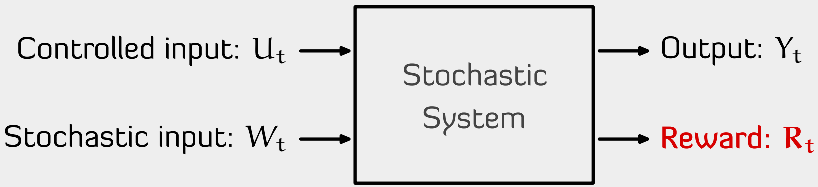
$$\mathbb{P}(Y_t \mid H_t, U_t) = \mathbb{P}(Y_t \mid Z_t, U_t).$$

KEY QUESTIONS

- Can this be used for dynamic programming?
- What is the right notion of approximations in this framework?

An information state for dynamic programming

Predicting output vs optimizing expected rewards over time



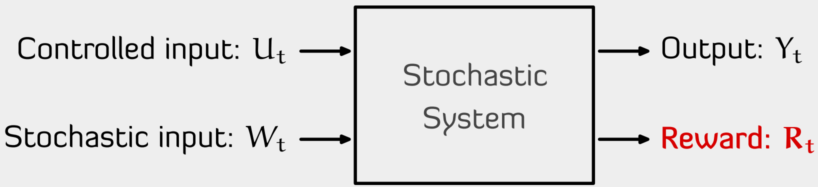
$$Y_t = f_t(U_{1:t}, W_{1:t}),$$

$$R_t = r_t(U_{1:t}, W_{1:t}).$$

Choose $U_t = g_t(Y_{1:t-1}, U_{1:t-1})$ to

$$\max \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

Predicting output vs optimizing expected rewards over time



$$Y_t = f_t(U_{1:t}, W_{1:t}),$$

$$R_t = r_t(U_{1:t}, W_{1:t}).$$

Choose $U_t = g_t(Y_{1:t-1}, U_{1:t-1})$ to

$$\max \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

PROPERTIES OF INFORMATION STATE (SUFFICIENT FOR DYNAMIC PROGRAMMING)

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:

$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$

- SUFFICIENT TO ESTIMATE EXPECTED REWARD:

$$\mathbb{E}[R_t \mid H_t, U_t] = \mathbb{E}[R_t \mid Z_t, U_t].$$

Dynamic programming using information state

PROPERTIES OF INFORMATION STATE

(SUFFICIENT FOR DYNAMIC PROGRAMMING)

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:

$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$

- SUFFICIENT TO ESTIMATE EXPECTED REWARD:

$$\mathbb{E}[R_t \mid H_t, U_t] = \mathbb{E}[R_t \mid Z_t, U_t].$$

Dynamic programming using information state

PRELIMINARY THEOREM

If $\{Z_t\}_{t \geqslant 1}$ is any information state process. Then:

- There is no loss of optimality in restricting attention to policies of the form
$$U_t = \tilde{g}_t(Z_t).$$

PROPERTIES OF INFORMATION STATE

(SUFFICIENT FOR DYNAMIC PROGRAMMING)

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:
$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$
- SUFFICIENT TO ESTIMATE EXPECTED REWARD:
$$\mathbb{E}[R_t \mid H_t, U_t] = \mathbb{E}[R_t \mid Z_t, U_t].$$

- Bohlin (1970)
- David and Varaiya (1972)
- Kumar and Varaiya (1984).

Dynamic programming using information state

PRELIMINARY THEOREM

If $\{Z_t\}_{t \geq 1}$ is any information state process. Then:

- There is no loss of optimality in restricting attention to policies of the form
$$U_t = \tilde{g}_t(Z_t).$$
- Let $\{V_t\}_{t=1}^{T+1}$ denote the solution to the following dynamic program: $V_{T+1}(z_{T+1}) = 0$ and for $t \in \{T, \ldots, 1\}$,
$$Q_t(z_t, u_t) = \mathbb{E}[R_t + V_{t+1}(Z_{t+1}) \mid Z_t = z_t, U_t = u_t],$$
$$V_t(z_t) = \max_{u_t \in \mathcal{U}} Q_t(z_t, u_t).$$
A policy $\{\tilde{g}_t\}_{t=1}^{T}$, $\tilde{g}_t \colon \mathcal{Z}_t \to \mathcal{U}$, is optimal if it satisfies
$$\tilde{g}_t(z_t) \in \arg\max_{u_t \in \mathcal{U}} Q_t(z_t, u_t).$$

PROPERTIES OF INFORMATION STATE

(SUFFICIENT FOR DYNAMIC PROGRAMMING)

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:
$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$
- SUFFICIENT TO ESTIMATE EXPECTED REWARD:
$$\mathbb{E}[R_t \mid H_t, U_t] = \mathbb{E}[R_t \mid Z_t, U_t].$$

- Bohlin (1970)
- David and Varaiya (1972)
- Kumar and Varaiya (1984).

What about approximations?

Preliminary: A family of pseudometrics on probability distribution

INTEGRAL PROBABILITY METRIC (IPM)

Let $\mathcal{P}$ denote the set of probability measures on a measurable space $(\mathcal{X}, \mathfrak{G})$.

Given a class $\mathfrak{F}$ of real-valued bounded measureable functions on $(\mathcal{X}, \mathfrak{G})$, the integral probability metric (IPM) between two probability distributions $\mu, \nu \in \mathcal{P}$ is given by:

$$d_{\mathfrak{F}}(\mu, \nu) = \sup_{f \in \mathfrak{F}} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right|.$$

▷ Müller, "Integral probability metrics and their generating classes of functions," 1997.

Preliminary: A family of pseudometrics on probability distribution

INTEGRAL PROBABILITY METRIC (IPM)

Let $\mathcal{P}$ denote the set of probability measures on a measurable space $(\mathcal{X}, \mathfrak{G})$.

Given a class $\mathfrak{F}$ of real-valued bounded measureable functions on $(\mathcal{X}, \mathfrak{G})$, the integral probability metric (IPM) between two probability distributions $\mu, \nu \in \mathcal{P}$ is given by:

$$d_{\mathfrak{F}}(\mu, \nu) = \sup_{f \in \mathfrak{F}} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right|.$$

EXAMPLES

- If $\mathfrak{F} = \{f : \|f\|_\infty \leqslant 1\}$,
  $d_{\mathfrak{F}} =$ Total variation distance.
- If $\mathfrak{F} = \{f : |f|_L \leqslant 1\}$,
  $d_{\mathfrak{F}} =$ Wasserstein distance.
- If $\mathfrak{F} = \{f : \|f\|_\infty + |f|_L \leqslant 1\}$,
  $d_{\mathfrak{F}} =$ Dudley metric.
- ...

We say a function f has a $\mathfrak{F}$-constant K if $f/K \in \mathfrak{F}$.

Müller, "Integral probability metrics and their generating classes of functions," 1997.

Approximate information state

(ε, δ)-APPROXIMATE INFORMATION STATE (AIS)

Given a function class $\mathfrak{F}$, a compression $\{Z_t\}_{t \geqslant 1}$ of history (i.e., $Z_t = \varphi_t(H_t)$) is called an $\{(\varepsilon_t, \delta_t)\}_{t \geqslant 1}$ AIS if there exist:

- a function $\tilde{R}_t(Z_t, U_t)$, and
- a stochastic kernel $\nu_t(Z_{t+1} | Z_t, U_t)$

such that

- $\left| \mathbb{E}[R_t | H_t = h_t, U_t = u_t] - \tilde{R}_t(\varphi_t(h_t), u_t) \right| \leqslant \varepsilon_t$
- For any Borel set A of $\mathcal{Z}_t$, define

  $$\mu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid H_t = h_t, U_t = u_t)$$

  Then,

  $$d_{\mathfrak{F}}(\mu_t, \nu_t(\cdot | \varphi_t(h_t), u_t)) \leqslant \delta_t.$$

Approximate dynamic programming using AIS

MAIN THEOREM

Given a function class $\mathfrak{F}$, let $\{Z_t\}_{t \geqslant 1}$, where $Z_t = \varphi_t(H_t)$, be an $\{(\varepsilon_t, \delta_t)\}_{t \geqslant 1}$ AIS.

Recursively define the following functions:

$$\hat{V}_{T+1}(z_{T+1}) = 0$$

and for $t \in \{T, \dots, 1\}$:

$$\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} \Big\{ \tilde{R}_t(z_t, u_t) + \int V_{t+1}(z_{t+1}) \nu_t(dz_{t+1} \mid z_t, u_t) \Big\}.$$

Let $\pi = (\pi_1, \dots, \pi_T)$ denote the corresponding policy.

Approximate dynamic programming using AIS

MAIN THEOREM

Given a function class $\mathfrak{F}$, let $\{Z_t\}_{t \geqslant 1}$, where $Z_t = \varphi_t(H_t)$, be an $\{(\varepsilon_t, \delta_t)\}_{t \geqslant 1}$ AIS.

Recursively define the following functions:

$$\hat{V}_{T+1}(z_{T+1}) = 0$$

and for $t \in \{T, \dots, 1\}$:

$$\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} \Big\{ \tilde{R}_t(z_t, u_t) + \int V_{t+1}(z_{t+1}) \nu_t(dz_{t+1} \mid z_t, u_t) \Big\}.$$

Let $\pi = (\pi_1, \dots, \pi_T)$ denote the corresponding policy.

Then, if the value function $\hat{V}_t$ has $\mathfrak{F}$-constant K_t, then

- for any history h_t,

$$\Big| V_t(h_t) - \hat{V}_t(\varphi_t(h_t)) \Big| \leqslant \varepsilon_T + \sum_{s=t}^{T} (\varepsilon_s + K_s \delta_s).$$

- for any history h_t,

$$\Big| V_t(h_t) - V_t^{\pi}(h_t) \Big| \leqslant 2 \Big[\varepsilon_T + \sum_{s=t}^{T} (\varepsilon_s + K_s \delta_s) \Big].$$

AIS: Some remarks

In the definition of AIS, we can replace

$$d_{\mathfrak{F}}(\mathbb{P}(\mu_t, \nu_t(\cdot|Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t$$

by

- $Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t)$
- $d_{\mathfrak{F}}(\mathbb{P}(Y_t|H_t = h_t, U_t = u_t), \mathbb{P}(Y_t|Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t$.

AIS: Some remarks

In the definition of AIS, we can replace

$$d_{\mathfrak{F}}(\mathbb{P}(\mu_t, \nu_t(\cdot|Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t$$

by

- $Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t)$
- $d_{\mathfrak{F}}(\mathbb{P}(Y_t|H_t = h_t, U_t = u_t), \mathbb{P}(Y_t|Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t.$

The AIS process $\{Z_t\}_{t \geqslant 1}$ need not be Markov !!

AIS: Some remarks

In the definition of AIS, we can replace

$$d_{\mathfrak{F}}(\mathbb{P}(\mu_t, \nu_t(\cdot|Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t$$

by

- $Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t)$
- $d_{\mathfrak{F}}(\mathbb{P}(Y_t|H_t = h_t, U_t = u_t), \mathbb{P}(Y_t|Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t$.

The AIS process $\{Z_t\}_{t \geqslant 1}$ need not be Markov !!

Two ways to interpret the results:

- Given the information state space $\mathcal{Z}$, find the best compression $\varphi_t : \mathcal{H}_t \to \mathcal{Z}$
- Given any compression function $\varphi_t : \mathcal{H}_t \to \mathcal{Z}_t$, find the approximation error.

AIS: Some remarks

In the definition of AIS, we can replace

$$d_{\mathfrak{F}}(\mathbb{P}(\mu_t, \nu_t(\cdot | Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t$$

by

- $Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t)$
- $d_{\mathfrak{F}}(\mathbb{P}(Y_t | H_t = h_t, U_t = u_t), \mathbb{P}(Y_t | Z_t = \varphi_t(h_t), U_t = u_t)) \leqslant \delta_t$.

The AIS process $\{Z_t\}_{t \geqslant 1}$ need not be Markov !!

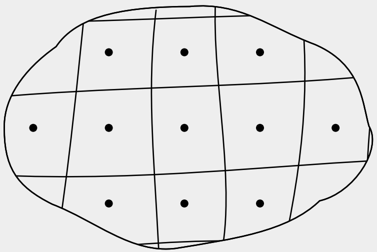
Two ways to interpret the results:

- Given the information state space $\mathcal{Z}$, find the best compression $\varphi_t \colon \mathcal{H}_t \to \mathcal{Z}$
- Given any compression function $\varphi_t \colon \mathcal{H}_t \to \mathcal{Z}_t$, find the approximation error.

Results naturally extend to infinite horizon

Some examples

Example 1: Error bounds on state aggregation

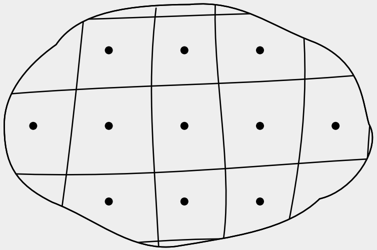


Consider an MDP with state space $\mathcal{X}$ and per-step reward $R_t = r(X_t, U_t)$.

Suppose $\mathcal{X}$ is quantized to a discrete set $\mathcal{Z}$ using $\varphi\colon \mathcal{X} \to \mathcal{Z}$.

- Let $z = \varphi(x)$ denote the label for x.
- Then $\varphi^{-1}(z)$ denote all states which have label z.

Example 1: Error bounds on state aggregation



Consider an MDP with state space $\mathcal{X}$ and per-step reward $R_t = r(X_t, U_t)$.

Suppose $\mathcal{X}$ is quantized to a discrete set $\mathcal{Z}$ using $\varphi\colon \mathcal{X} \to \mathcal{Z}$.

- Let $z = \varphi(x)$ denote the label for x.
- Then $\varphi^{-1}(z)$ denote all states which have label z.

$\{Z_t\}_{t \geqslant 1}$ IS AN (ε, δ) AIS

$$\varepsilon = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} \left| r(x,u) - r(\varphi(x), u) \right|$$

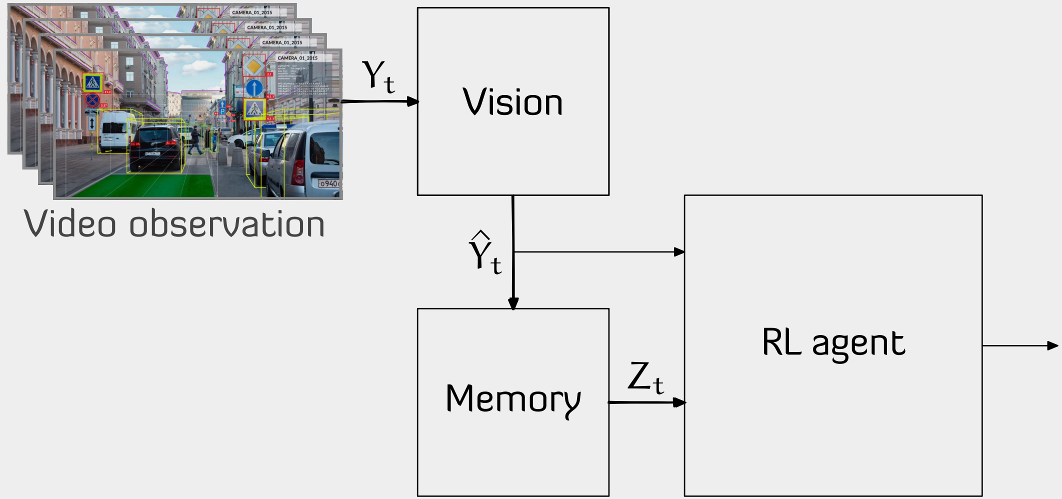
or, equivalently, $r(\cdot, u)$ has a $\mathfrak{F}$-cosntant K_r

$$\delta = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} d_{\mathfrak{F}}(\mathbb{P}(X_+ \mid X = x, U = u), \mathbb{P}(X_+ \mid X \in \varphi^{-1}(\varphi(x)), U = u)).$$

or, equivalently, $\mathbb{P}(X_+ | X = \cdot, U = u)$ has a $\mathfrak{F}$-constant of K_d.

- Bertsekas, "Convergence of discretization procedures in dynamic programming," 1975.

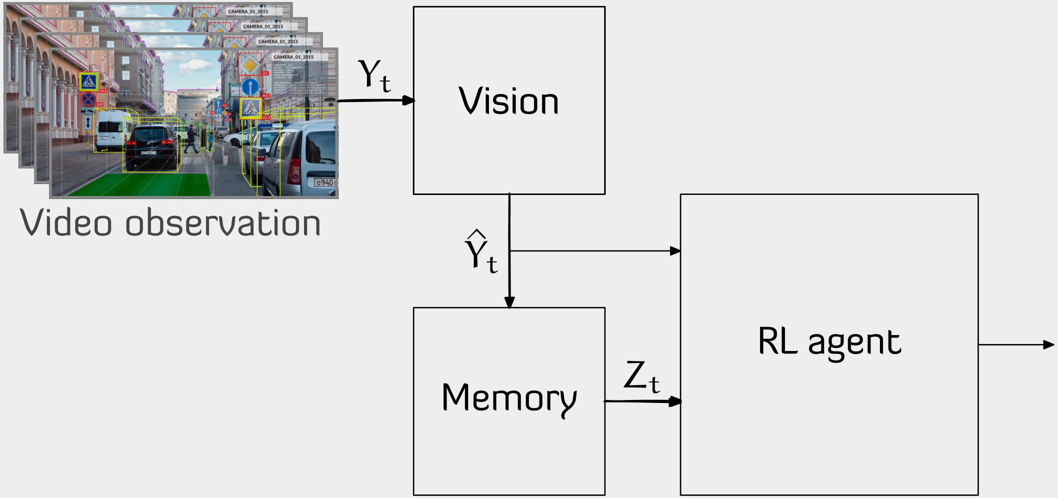
Example 2: Approximation bounds for using quantized obs.



▷ Ha, Schmidhuber, "World Models", 2018.

Example 2: Approximation bounds for using quantized obs.

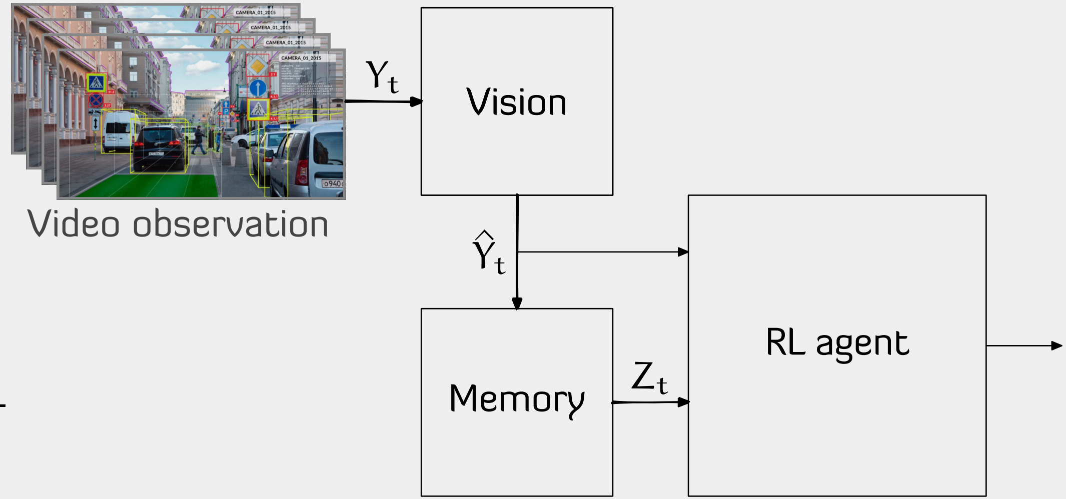
- Proposed as a heuristic algorithms
- No performance bounds



- Ha, Schmidhuber, "World Models", 2018.

Example 2: Approximation bounds for using quantized obs.

- Proposed as a heuristic algorithms
- No performance bounds



$\{Z_t\}_{t \geq 1}$ IS AN (ε, δ) AIS

$$\varepsilon = \sup_{h_t, u_t} \left| \mathbb{E}[R_t | h_t, u_t] - \tilde{R}_t(\varphi_t(h_t), u_t) \right|$$

$$\delta = \sup_{h_t, u_t} d_{\mathfrak{F}}(\mathbb{P}(\hat{Y}_{t+1} | h_t, u_t), \mathbb{P}(\hat{Y}_{t+1} | \varphi_t(h_t), u_t))$$

Ha, Schmidhuber, "World Models", 2018.

Example 3: Approximation bounds for mean-field teams

n agents: state X^i_t, control U^i_t.

- Empirical mean-field:

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i_t}(x).$$

- Statistical mean-field:

$$\bar{m}_t(x) = \mathbb{P}(X^i_t = x).$$

Example 3: Approximation bounds for mean-field teams

n agents: state X^i_t, control U^i_t.

- Dynamics

$$\mathbb{P}(\mathbf{X}_{t+1} | \mathbf{X}_t, \mathbf{U}_t) = \prod_{i=1}^{n} P(X^i_{t+1} | X^i_t, U^i_t, M_t)$$

- Per-step reward

$$R(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n} \sum_{i=1}^{n} r(X^i_t, U^i_t, M_t)$$

- Empirical mean-field:

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i_t}(x).$$

- Statistical mean-field:

$$\bar{m}_t(x) = \mathbb{P}(X^i_t = x).$$

Example 3: Approximation bounds for mean-field teams

n agents: state X^i_t, control U^i_t.

- Dynamics

$$\mathbb{P}(\mathbf{X}_{t+1} | \mathbf{X}_t, \mathbf{U}_t) = \prod_{i=1}^{n} P(X^i_{t+1} | X^i_t, U^i_t, M_t)$$

- Per-step reward

$$R(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n} \sum_{i=1}^{n} r(X^i_t, U^i_t, M_t)$$

- Empirical mean-field:

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i_t}(x).$$

- Statistical mean-field:

$$\bar{m}_t(x) = \mathbb{P}(X^i_t = x).$$

- **Info structure**: $I^i_t = \{X^i_t\}$.

Example 3: Approximation bounds for mean-field teams

n agents: state X^i_t, control U^i_t.

- Dynamics

$$\mathbb{P}(\mathbf{X}_{t+1} | \mathbf{X}_t, \mathbf{U}_t) = \prod_{i=1}^{n} P(X^i_{t+1} | X^i_t, U^i_t, M_t)$$

- Per-step reward

$$R(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n} \sum_{i=1}^{n} r(X^i_t, U^i_t, M_t)$$

- Empirical mean-field:

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i_t}(x).$$

- Statistical mean-field:

$$\bar{m}_t(x) = \mathbb{P}(X^i_t = x).$$

- **Info structure**: $I^i_t = \{X^i_t\}$.
- **Expanded info structure**: $\tilde{I}^i_t = \{X^i_t, M_t\}$.

$$\mathcal{J}^* \leqslant \tilde{\mathcal{J}}^*$$

Approx. info. state–(Subramanian and Mahajan)

Example 3: Approximation bounds for mean-field teams

n agents: state X^i_t, control U^i_t.

- Dynamics

$$\mathbb{P}(\mathbf{X}_{t+1} | \mathbf{X}_t, \mathbf{U}_t) = \prod_{i=1}^{n} P(X^i_{t+1} | X^i_t, U^i_t, M_t)$$

- Per-step reward

$$R(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n} \sum_{i=1}^{n} r(X^i_t, U^i_t, M_t)$$

- Empirical mean-field:

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i_t}(x).$$

- Statistical mean-field:

$$\bar{m}_t(x) = \mathbb{P}(X^i_t = x).$$

- **Info structure**: $I^i_t = \{X^i_t\}$.
- **Expanded info structure**: $\tilde{I}^i_t = \{X^i_t, M_t\}$.

$$\mathcal{J}^* \leqslant \tilde{\mathcal{J}}^*$$

(A) $r(x, u, m)$ and $P(y | x, u, m)$ are Lipschitz in m.

$\{\bar{m}_t\}_{t \geqslant 1}$ is an (ε, δ) AIS for expanded info structure, where $\varepsilon, \delta \in \mathcal{O}(1/\sqrt{n})$.

Example 3: Approximation bounds for mean-field teams

n agents: state X^i_t, control U^i_t.

- Dynamics

$$\mathbb{P}(\mathbf{X}_{t+1} | \mathbf{X}_t, \mathbf{U}_t) = \prod_{i=1}^{n} P(X^i_{t+1} | X^i_t, U^i_t, M_t)$$

- Per-step reward

$$R(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n} \sum_{i=1}^{n} r(X^i_t, U^i_t, M_t)$$

- Empirical mean-field:

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i_t}(x).$$

- Statistical mean-field:

$$\bar{m}_t(x) = \mathbb{P}(X^i_t = x).$$

- **Info structure**: $I^i_t = \{X^i_t\}$.
- **Expanded info structure**: $\tilde{I}^i_t = \{X^i_t, M_t\}$.

$$\mathcal{J}^* \leqslant \tilde{\mathcal{J}}^*, \qquad \tilde{\mathcal{J}}^* - \bar{\mathcal{J}}^* \leqslant K/\sqrt{n}$$

$$\bar{\mathcal{J}}^* \leqslant \mathcal{J}^* \leqslant \bar{\mathcal{J}}^* + K/\sqrt{n}.$$

(A) $r(x, u, m)$ and $P(y|x, u, m)$ are Lipschitz in m.

$\{\bar{m}_t\}_{t \geqslant 1}$ is an (ε, δ) AIS for expanded info structure, where $\varepsilon, \delta \in \mathcal{O}(1/\sqrt{n})$.

Now to reinforcement learning for partially observed systems.

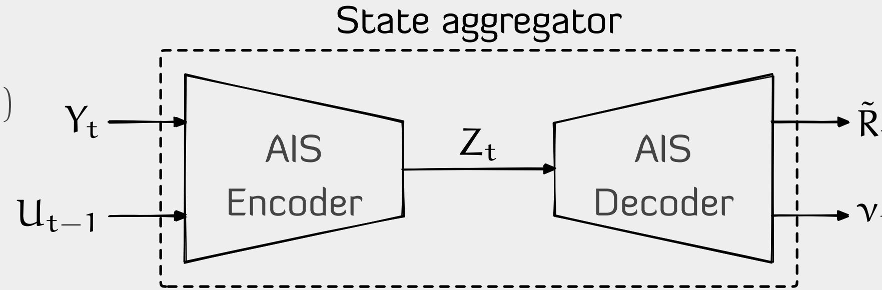
Reinforcement learning setup

- **State aggregator**:

  $\mathcal{L}_{\text{AIS}} = \alpha_t|\tilde{R}_t - R_t| + (1-\alpha_t)d_{\mathfrak{F}}(\nu_t, \mu_t)$

  ξ: Parameters of the aggregator

  Updated using SGD with LR a_k



Reinforcement learning setup

- **State aggregator**:

  $\mathcal{L}_{\text{AIS}} = \alpha_t |\tilde{R}_t - R_t| + (1 - \alpha_t) d_{\mathfrak{F}}(\nu_t, \mu_t)$

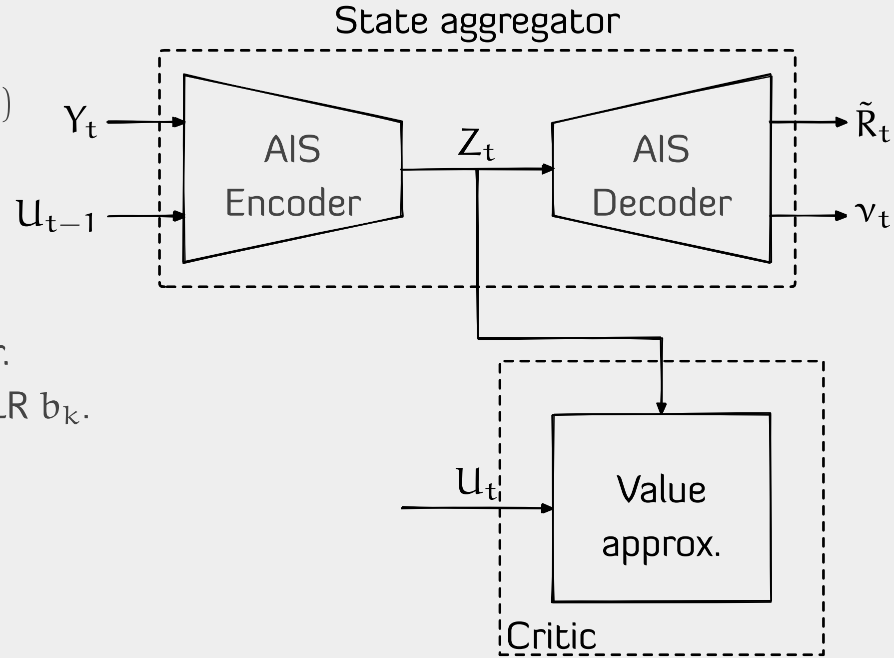
  ξ: Parameters of the aggregator

  Updated using SGD with LR a_k

- **Value approximator**:

  φ: parameters of $Q(z, u)$ approximator.

  Updated using TD(0) or TD(λ) with LR b_k.



Reinforcement learning setup

- **State aggregator**:

  $\mathcal{L}_{AIS} = \alpha_t|\tilde{R}_t - R_t| + (1-\alpha_t)d_{\mathfrak{F}}(\nu_t, \mu_t)$

  ξ: Parameters of the aggregator

  Updated using SGD with LR a_k

- **Value approximator**:

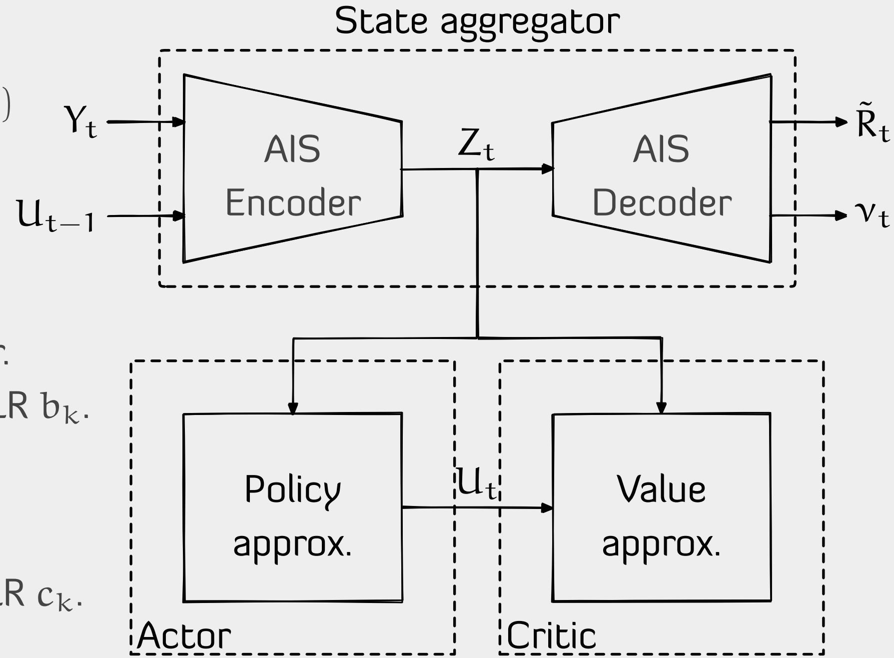
  φ: parameters of $Q(z,u)$ approximator.

  Updated using TD(0) or TD(λ) with LR b_k.

- **Policy approximator**:

  θ: parameters of $\pi(u \mid z)$

  Updated using policy gradient with LR c_k.



Reinforcement learning setup

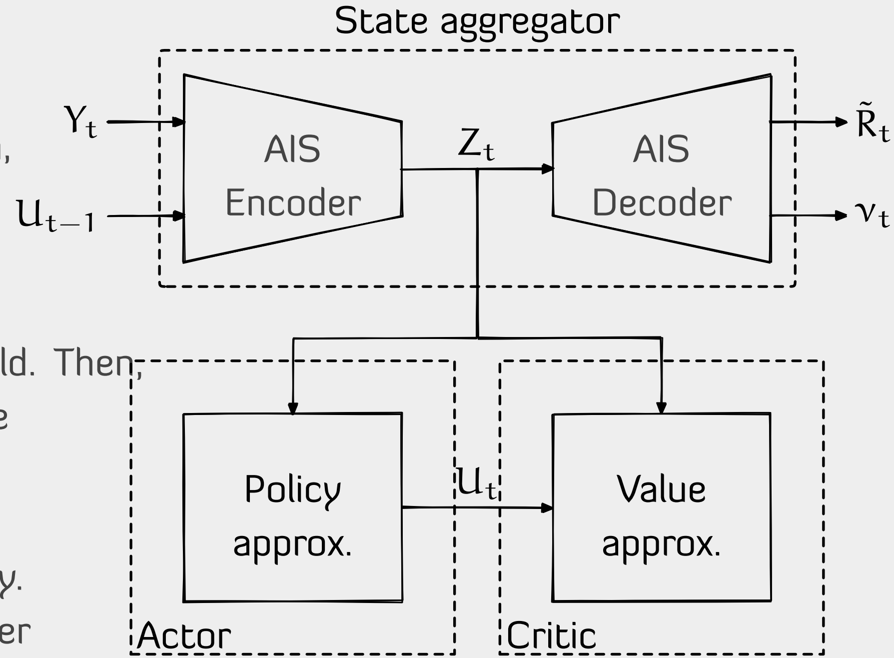
CONVERGENCE RESULT

If the learning rates satisfy conditions for three time-scale stochastic approximation, the compatibility condition

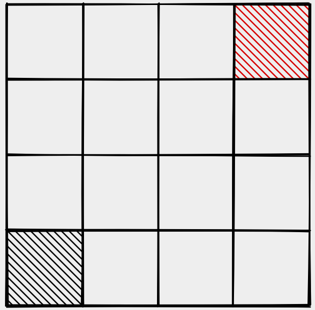
$$\frac{\partial Q(z,u)}{\partial \varphi} = \frac{1}{\pi(u|z)} \frac{\partial \pi(u|z)}{\partial \theta}$$

and additional mild technical conditions hold. Then,

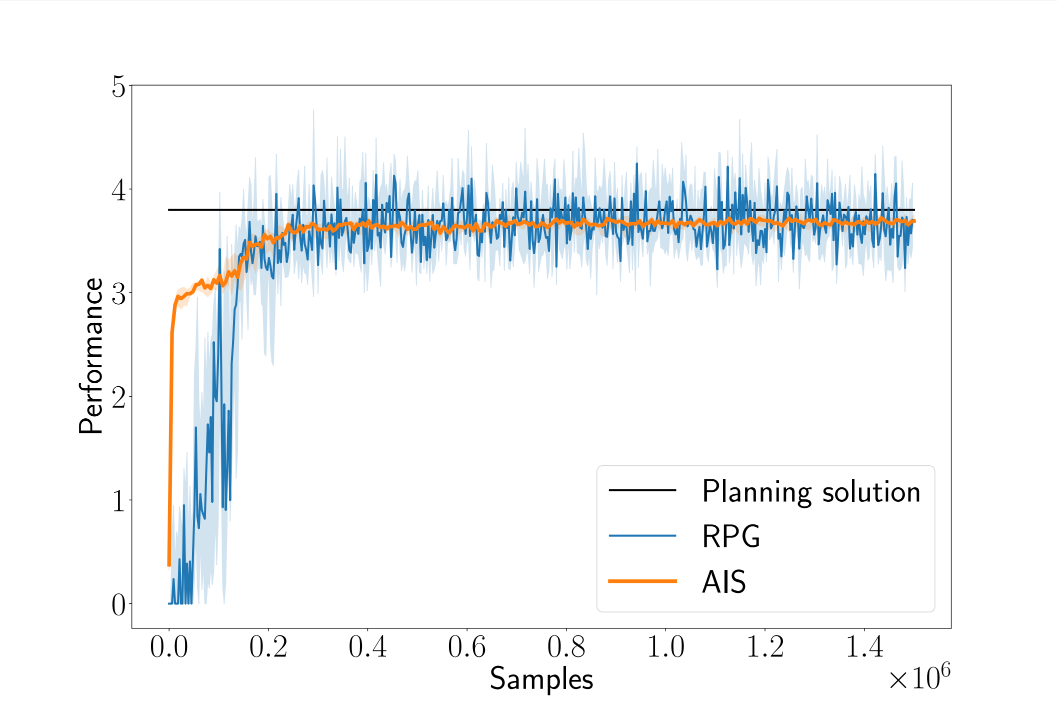
- State aggregator converges (with some approximation error)
- The critic converges to the best approximator within the specified family.
- The actor converges to a local maximizer within the family of policy approximators.

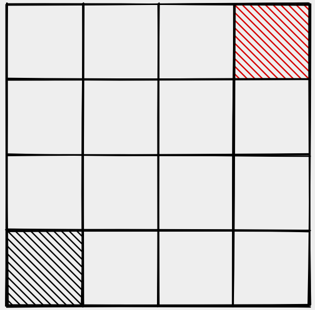


Numerical Results: 4×4 Grid Environment



4 × 4 Grid Environment

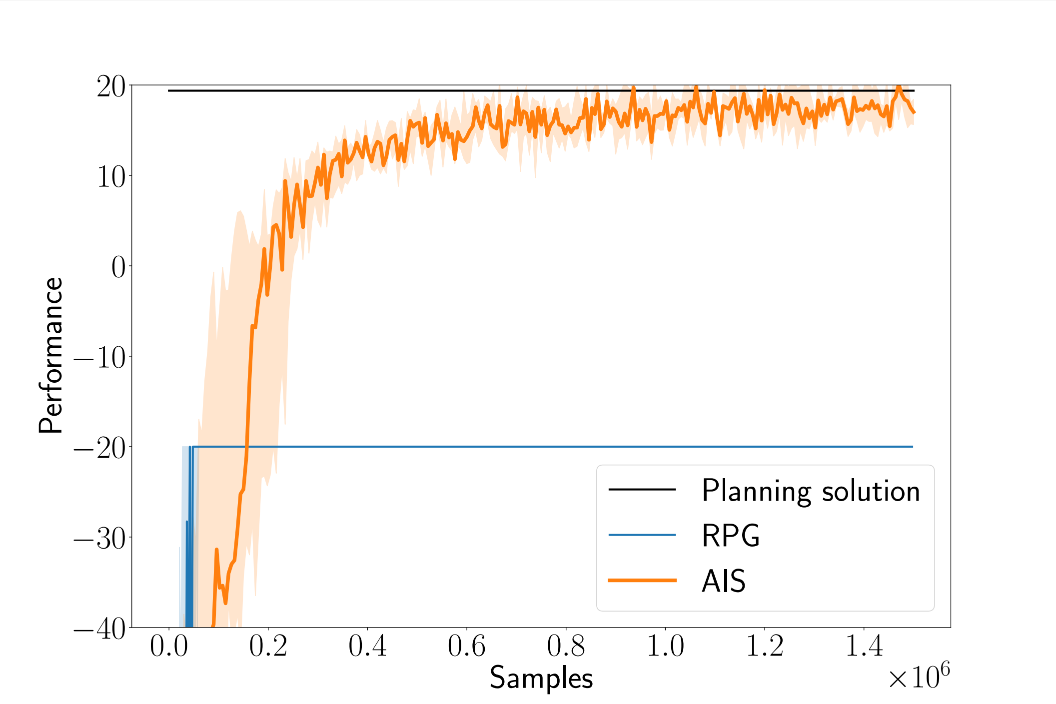




Approx. info. state–(Subramanian and Mahajan)

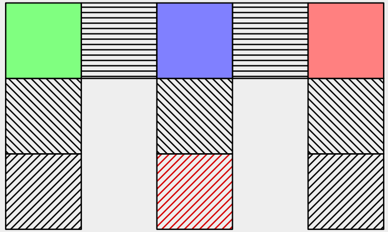
Numerical Results: Tiger Environment

Tiger Environment

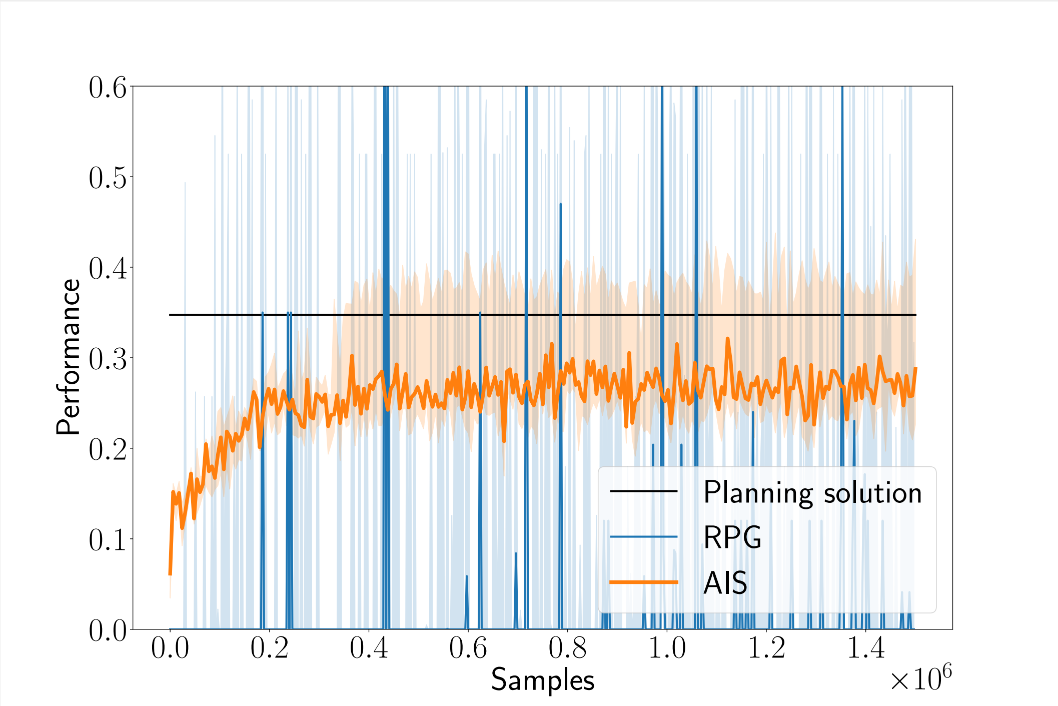


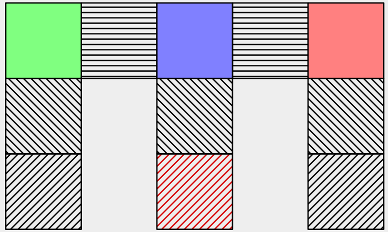
Approx. info. state–(Subramanian and Mahajan)

Numerical Results: Cheese Maze Environment



Cheese Maze Environment





Approx. info. state–(Subramanian and Mahajan)

Summary

Summary

Now let's consturct the state space

FORECASTING OUTPUTS IN DISTRIBUTION

$H_t^{(1)} \sim H_t^{(2)}$ if for all future CONTROL inputs $U_{t:T}$,

$$\mathbb{P}(Y_{t:T}^{(1)} \mid H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} \mid H_t^{(2)}, U_{t:T})$$

Same complexity as identifying the state sufficient for forecasting outputs for the case of perfect observations (which was Step 1 for belief state formulations)

PROPERTIES OF INFORMATION STATE

The info state Z_t at time t is a "compression" of past inputs that satisfies the following:

- SUFFICIENT TO PREDICT ITSELF:

$$\mathbb{P}(Z_{t+1} \mid H_t, U_t) = \mathbb{P}(Z_{t+1} \mid Z_t, U_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$\mathbb{P}(Y_t \mid H_t, U_t) = \mathbb{P}(Y_t \mid Z_t, U_t).$$

KEY QUESTIONS

- Can this be used for dynamic programming?
- What is the right notion of approximations in this framework?

Summary

Now let's construct the state space

Approximate information state

(ε, δ)-APPROXIMATE INFORMATION STATE (AIS)

Given a function class $\mathfrak{F}$, a compression $\{Z_t\}_{t \geqslant 1}$ of history (i.e., $Z_t = \varphi_t(H_t)$) is called an $\{(\varepsilon_t, \delta_t)\}_{t \geqslant 1}$ AIS if there exist:

- a function $\tilde{R}_t(Z_t, U_t)$, and
- a stochastic kernel $\nu_t(Z_{t+1} | Z_t, U_t)$

such that

- $\left| \mathbb{E}[R_t | H_t = h_t, U_t = u_t] - \tilde{R}_t(\varphi_t(h_t), u_t) \right| \leqslant \varepsilon_t$
- For any Borel set A of $\mathcal{Z}_t$, define

  $$\mu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid H_t = h_t, U_t = u_t)$$

  Then,

  $$d_{\mathfrak{F}}(\mu_t, \nu_t(\cdot | \varphi_t(h_t), u_t)) \leqslant \delta_t.$$

Summary

Now let's construct the state space

Approximate dynamic programming using AIS

MAIN THEOREM

Given a function class $\mathfrak{F}$, let $\{Z_t\}_{t \geqslant 1}$, where $Z_t = \varphi_t(H_t)$, be an $\{(\varepsilon_t, \delta_t)\}_{t \geqslant 1}$ AIS.

Recursively define the following functions:

$$\hat{V}_{T+1}(z_{T+1}) = 0$$

and for $t \in \{T, \dots, 1\}$:

$$\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} \Big\{ \tilde{R}_t(z_t, u_t) + \int V_{t+1}(z_{t+1}) \nu_t(dz_{t+1} \mid z_t, u_t) \Big\}.$$

Let $\pi = (\pi_1, \dots, \pi_T)$ denote the corresponding policy.

Then, if the value function $\hat{V}_t$ has $\mathfrak{F}$-constant K_t, then

- for any history h_t,

$$\big| V_t(h_t) - \hat{V}_t(\varphi_t(h_t)) \big| \leqslant \varepsilon_T + \sum_{s=t}^{T} (\varepsilon_s + K_s \delta_s).$$

- for any history h_t,

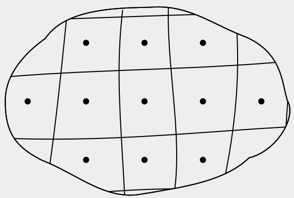
$$\big| V_t(h_t) - V_t^{\pi}(h_t) \big| \leqslant 2 \Big[\varepsilon_T + \sum_{s=t}^{T} (\varepsilon_s + K_s \delta_s) \Big].$$

Summary

Now let's consturct the state space

Approximate dynamic programming using AIS

Example 1: Error bounds on state aggregation



Consider an MDP with state space $\mathcal{X}$ and per-step reward $R_t = r(X_t, U_t)$.

Suppose $\mathcal{X}$ is quantized to a discrete set $\mathcal{Z}$ using $\varphi\colon \mathcal{X} \to \mathcal{Z}$.

- Let $z = \varphi(x)$ denote the label for x.
- Then $\varphi^{-1}(z)$ denote all states which have label z.

$\{Z_t\}_{t \geq 1}$ IS AN (ε, δ) AIS

$$\varepsilon = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} \big| r(x,u) - r(\varphi(x), u) \big| \qquad \text{or, equivalently, } r(\cdot, u) \text{ has a } \mathfrak{F}\text{-cosntant } K_r$$

$$\delta = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} d_{\mathfrak{F}}(\mathbb{P}(X_+ \mid X = x, U = u), \mathbb{P}(X_+ \mid X \in \varphi^{-1}(\varphi(x)), U = u)).$$

or, equivalently, $\mathbb{P}(X_+ | X = \cdot, U = u)$ has a $\mathfrak{F}$-constant of K_d.

- Bertsekas, "Convergence of discretization procedures in dynamic programming," 1975.

Approx. info. state–(Subramanian and Mahajan)

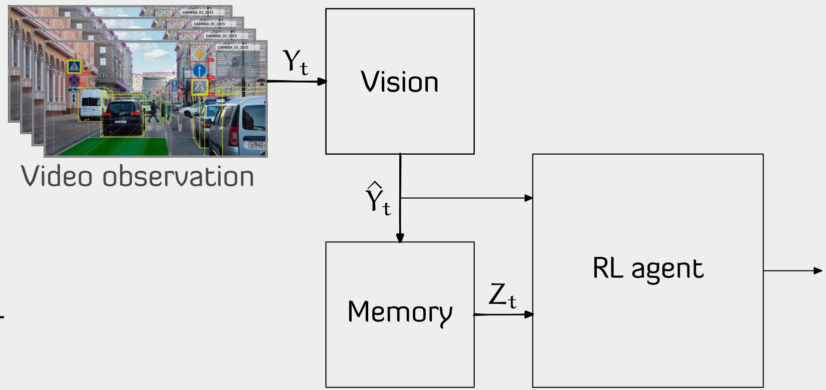
Summary

Now let's construct the state space

Approximate dynamic programming using AIS

Example 2: Approximation bounds for using quantized obs.

- Proposed as a heuristic algorithms
- No performance bounds



$\{Z_t\}_{t \geqslant 1}$ IS AN (ε, δ) AIS

$$\varepsilon = \sup_{h_t, u_t} \left| \mathbb{E}[R_t | h_t, u_t] - \tilde{R}_t(\varphi_t(h_t), u_t) \right|$$

$$\delta = \sup_{h_t, u_t} d_{\mathfrak{F}}(\mathbb{P}(\hat{Y}_{t+1} | h_t, u_t), \mathbb{P}(\hat{Y}_{t+1} | \varphi_t(h_t), u_t))$$

- Ha, Schmidhuber, "World Models", 2018.

Summary

Now let's construct the state space

Approximate dynamic programming using AIS

Example 3: Approximation bounds for mean-field teams

n agents: state X^i_t, control U^i_t.

- Dynamics

$$\mathbb{P}(\mathbf{X}_{t+1} | \mathbf{X}_t, \mathbf{U}_t) = \prod_{i=1}^{n} P(X^i_{t+1} | X^i_t, U^i_t, M_t)$$

- Per-step reward

$$R(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n} \sum_{i=1}^{n} r(X^i_t, U^i_t, M_t)$$

- Empirical mean-field:

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i_t}(x).$$

- Statistical mean-field:

$$\bar{m}_t(x) = \mathbb{P}(X^i_t = x).$$

- **Info structure**: $I^i_t = \{X^i_t\}$.
- **Expanded info structure**: $\tilde{I}^i_t = \{X^i_t, M_t\}$.

$$\mathcal{J}^* \leqslant \tilde{\mathcal{J}}^*, \qquad \tilde{\mathcal{J}}^* - \bar{\mathcal{J}}^* \leqslant K/\sqrt{n}$$

$$\bar{\mathcal{J}}^* \leqslant \mathcal{J}^* \leqslant \bar{\mathcal{J}}^* + K/\sqrt{n}.$$

(A) $r(x, u, m)$ and $P(y|x, u, m)$ are Lipschitz in m.

$\{\bar{m}_t\}_{t \geqslant 1}$ is an (ε, δ) AIS for expanded info structure, where $\varepsilon, \delta \in \mathcal{O}(1/\sqrt{n})$.

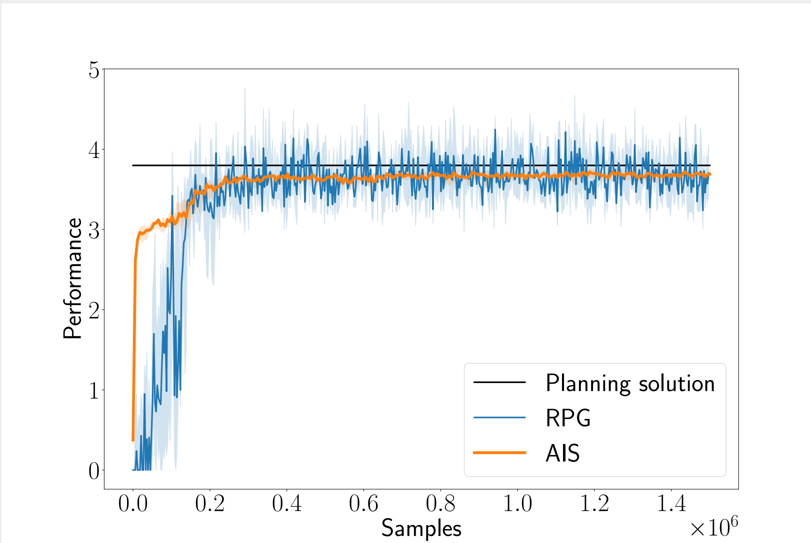
Summary

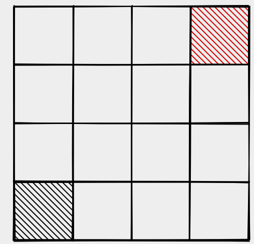
Now let's consturct the state space

Approximate dynamic programming using AIS

Example 2: Approximation bounds for mean field teams

4×4 Grid Environment





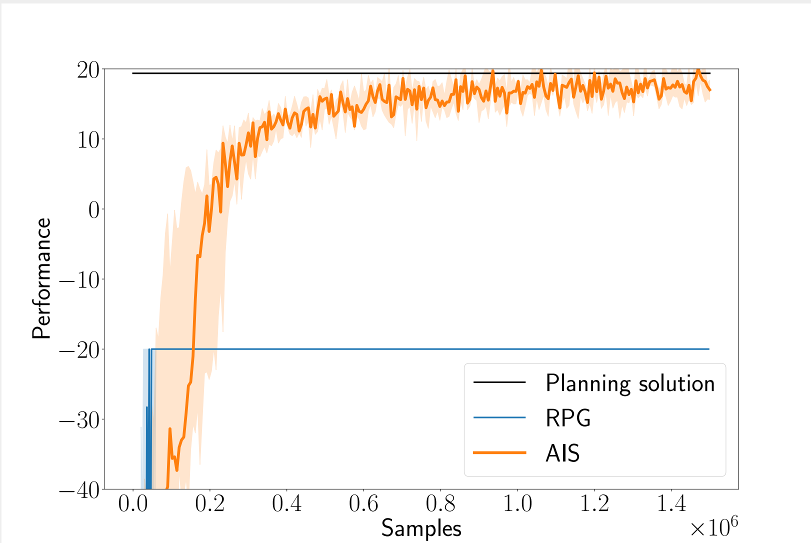
Summary

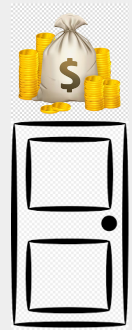
Now let's consturct the state space

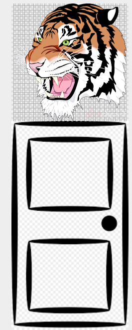
Approximate dynamic programming using AIS

Example a: Approximation bounds for mean field teams

Tiger Environment







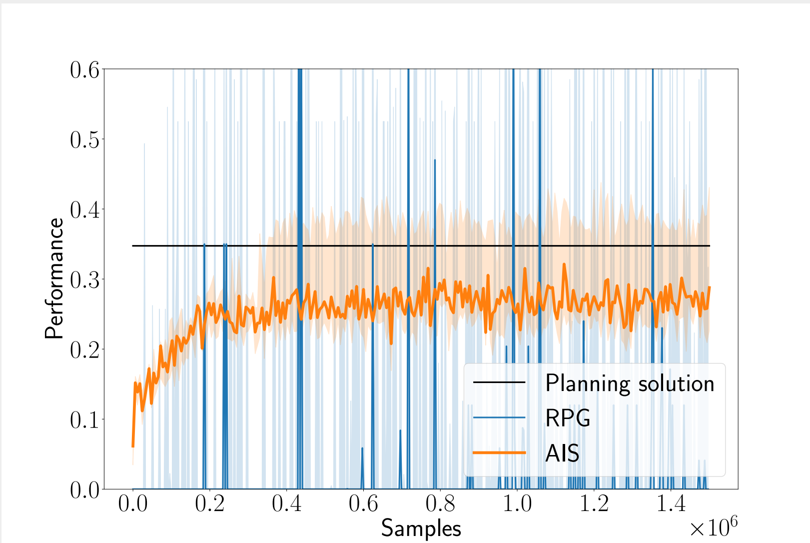
Summary

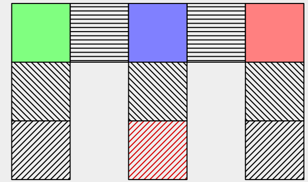
Now let's consturct the state space

Approximate dynamic programming using AIS

Example 2: Approximation bounds for mean field teams

Cheese Maze Environment





Summary

Now let's consturct the state space

Approximate dynamic programming using AIS

Example 2: Approximation bounds for mean field teams

Choose Maze Environment

AIS provides a conceptually clean framework for approximate DP and online RL in partially observed systems