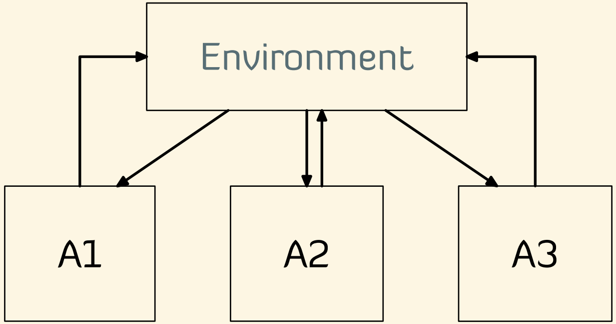
Information state (and its approximations) for stochastic control

Aditya Mahajan
McGill University and GERAD

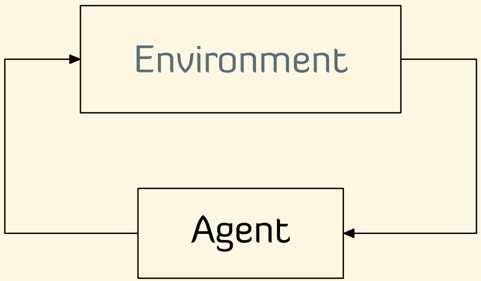
Joint work with Jayakumar Subramanian (McGill University)

BIRS-CMO Workshop
Multi-Stage Stochastic Optimization for Clean Energy Transition
26 September 2019

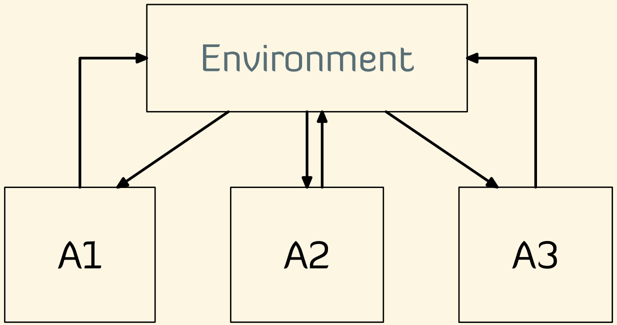


Online reinforcement learning for decentralized multi-agent systems

Approx. info state–(Mahajan)

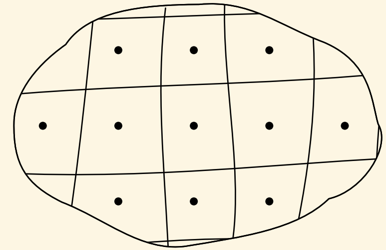


Online reinforcement learning for POMDPs

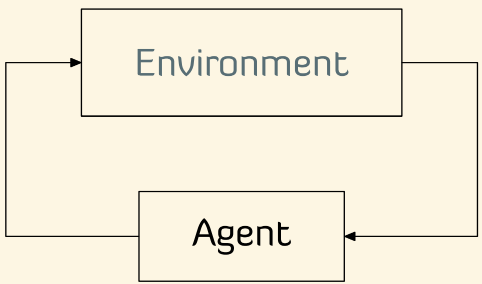


Online reinforcement learning for decentralized multi-agent systems

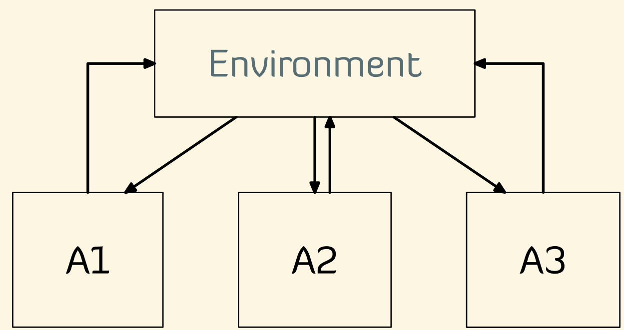
Approx. info state–(Mahajan)



Deriving approximation bounds for MDPs and POMDPs

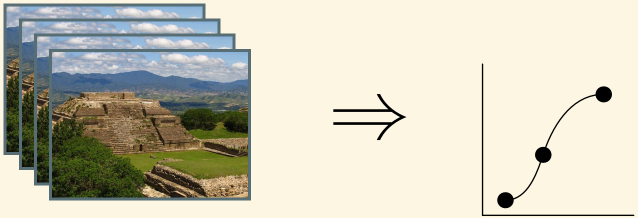


Online reinforcement learning for POMDPs

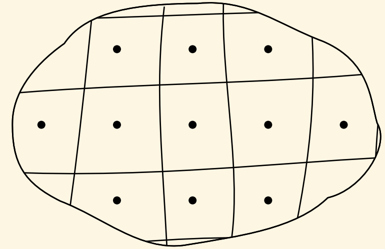


Online reinforcement learning for decentralized multi-agent systems

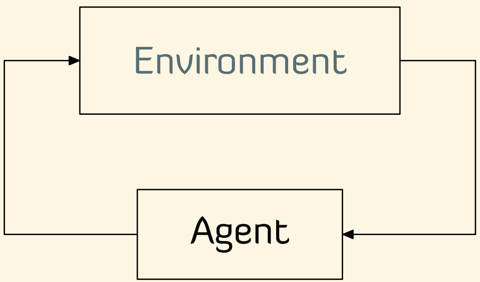
Approx. info state–(Mahajan)



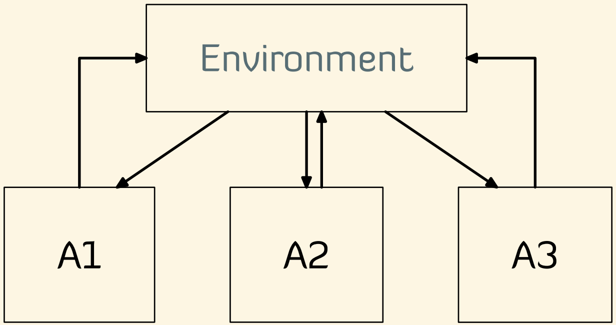
Discovering latent space representtation
for MDPs with high dimensional inputs



Deriving approximation bounds
for MDPs and POMDPs



Online reinforcement
learning for POMDPs



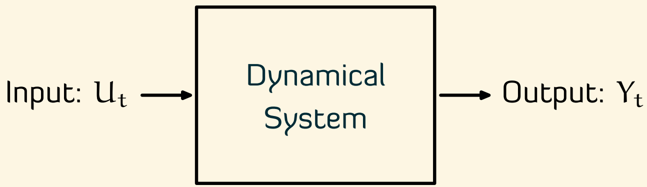
Online reinforcement learning for
decentralized multi-agent systems

Approx. info state-(Mahajan)

. . . .or how stochastic programmers can stop worrying and use state space models.

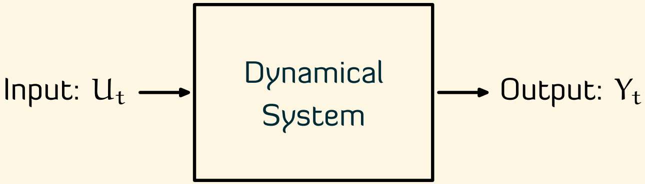
Let’s revisit the notion of state
in stochastic dynamical systems

Notion of state in deterministic dynamical systems

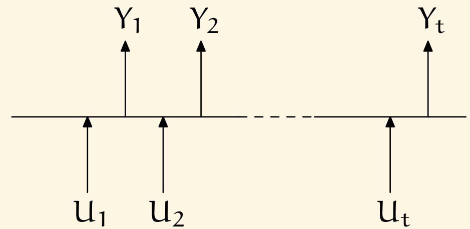


$$Y_t = f_t(U_{1:t}).$$

Notion of state in deterministic dynamical systems

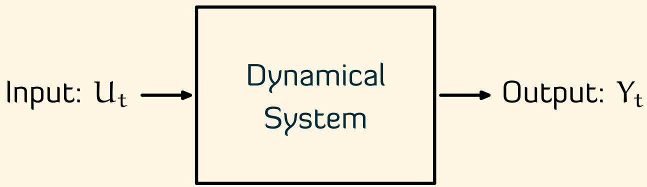


$$Y_t = f_t(U_{1:t}).$$

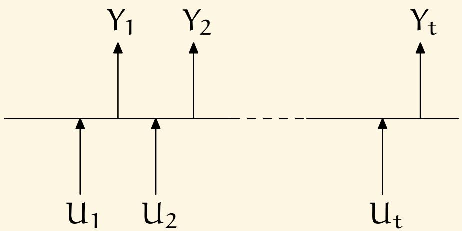


Timing diagram

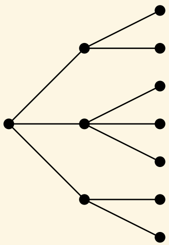
Notion of state in deterministic dynamical systems



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

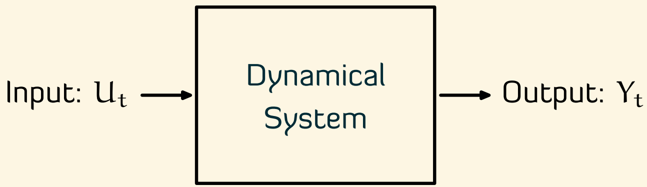


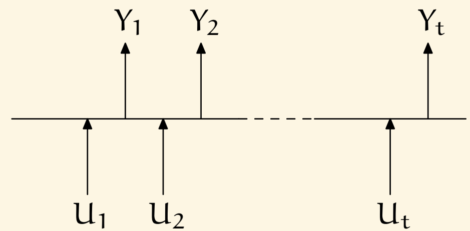
Timing diagram



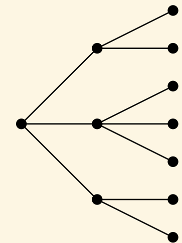
Decision tree

Notion of state in deterministic dynamical systems





Timing diagram



Decision tree

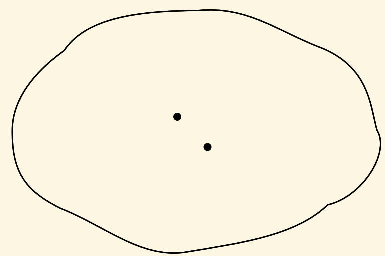
$$Y_t = f_t(U_{1:t}).$$

EQUIVALENCE RELATIONSHIP

Let $H_t = U_{1:t-1}$ denote the history of inputs until time t.

$H_t^{(1)} \sim H_t^{(2)}$ if for all future inputs $U_{t:T}$, the future outputs $Y_{t:T}^{(1)}$ and $Y_{t:T}^{(2)}$ are the same:

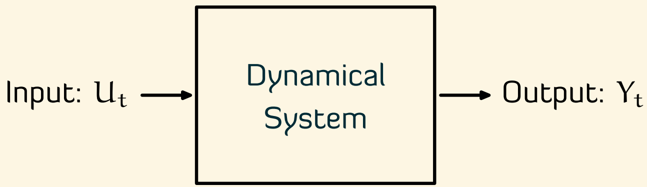
$$f_{t:T}(H_t^{(1)}, U_{t:T}) = f_{t:T}(H_t^{(2)}, U_{t:T})$$



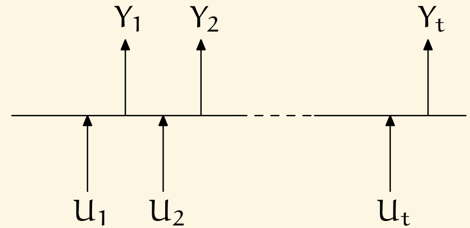
▷ Nerode, "Linear Automaton Transformation", 1958. ▷ Minsky, "Computation: Finite and Infinite Machines", 1967.
▷ Witsenhausen, "Some remarks on the concept of state", 1976.

Approx. info state–(Mahajan)

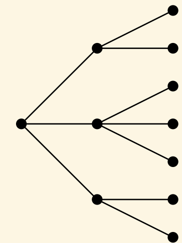
Notion of state in deterministic dynamical systems



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



Timing diagram

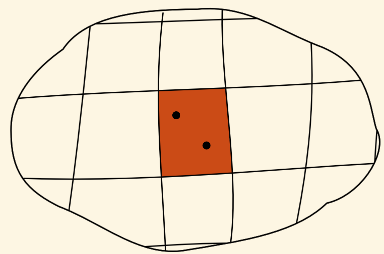

Decision tree

EQUIVALENCE RELATIONSHIP

Let $H_t = U_{1:t-1}$ denote the history of inputs until time t.

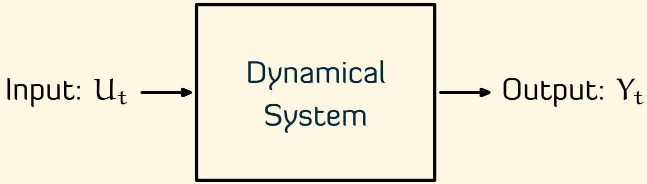
$H_t^{(1)} \sim H_t^{(2)}$ if for all future inputs $U_{t:T}$, the future outputs $Y_{t:T}^{(1)}$ and $Y_{t:T}^{(2)}$ are the same:

$$f_{t:T}(H_t^{(1)}, U_{t:T}) = f_{t:T}(H_t^{(2)}, U_{t:T})$$



▷ Nerode, "Linear Automaton Transformation", 1958. ▷ Minsky, "Computation: Finite and Infinite Machines", 1967.
▷ Witsenhausen, "Some remarks on the concept of state", 1976.

Notion of state in deterministic dynamical systems



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

EQUIVALENCE RELATIONSHIP

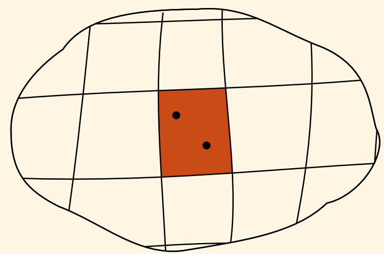
Let $H_t = U_{1:t-1}$ denote the history of inputs until time t.

$H_t^{(1)} \sim H_t^{(2)}$ if for all future inputs $U_{t:T}$, the future outputs $Y_{t:T}^{(1)}$ and $Y_{t:T}^{(2)}$ are the same:

$$f_{t:T}(H_t^{(1)}, U_{t:T}) = f_{t:T}(H_t^{(2)}, U_{t:T})$$

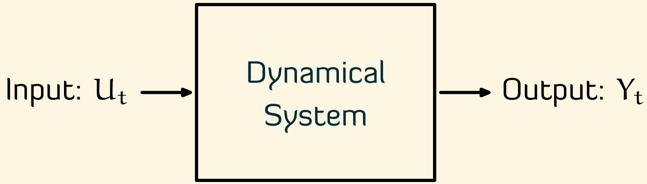
STATE SUFFICIENT FOR I/O MAPPING

Let $\mathcal{H}_t$ denote the space of all histories at time t. Then, the state space at time t is the quotient space $\mathcal{H}_t/\sim$.



- Nerode, "Linear Automaton Transformation", 1958.
- Minsky, "Computation: Finite and Infinite Machines", 1967.
- Witsenhausen, "Some remarks on the concept of state", 1976.

Notion of state in deterministic dynamical systems



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

EQUIVALENCE RELATIONSHIP

Let $H_t = U_{1:t-1}$ denote the history of inputs until time t.

$H_t^{(1)} \sim H_t^{(2)}$ if for all future inputs $U_{t:T}$, the future outputs $Y_{t:T}^{(1)}$ and $Y_{t:T}^{(2)}$ are the same:

$$f_{t:T}(H_t^{(1)}, U_{t:T}) = f_{t:T}(H_t^{(2)}, U_{t:T})$$

STATE SUFFICIENT FOR I/O MAPPING

Let $\mathcal{H}_t$ denote the space of all histories at time t. Then, the state space at time t is the quotient space $\mathcal{H}_t/\sim$.

PROPERTIES OF STATE

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

- UPDATES IN A RECURSIVE MANNER:

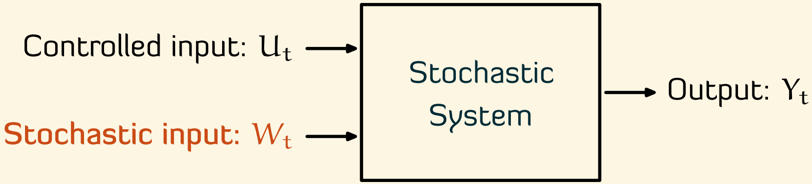
$$X_{t+1} = \text{function}(X_t, U_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$Y_t = \text{function}(X_t, U_t).$$

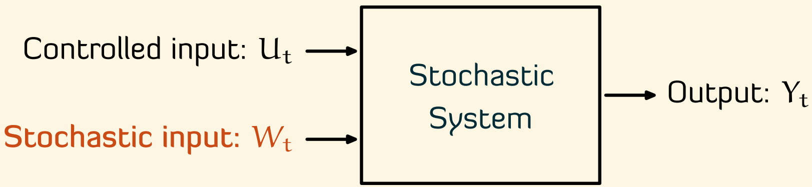
(Ignore: measurability and minimality)

Notion of state in stochastic dynamical systems



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

Notion of state in stochastic dynamical systems



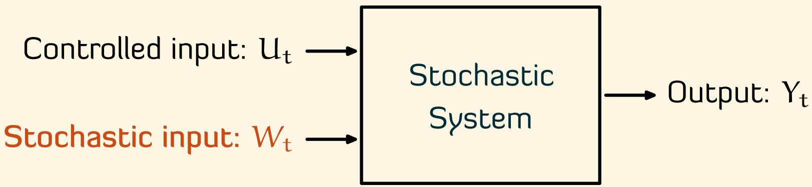
$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

STOCHASTIC INPUT IS OBSERVED

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

There are two ways to define state:

Notion of state in stochastic dynamical systems



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

PREDICTING OUTPUTS ALMOST SURELY

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

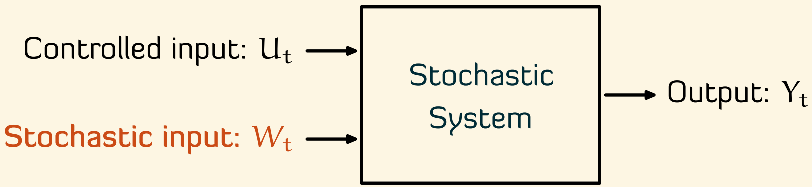
$$Y_{t:T}^{(1)} = Y_{t:T}^{(2)}, \quad a.s.$$

STOCHASTIC INPUT IS OBSERVED

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

There are two ways to define state:

Notion of state in stochastic dynamical systems



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

STOCHASTIC INPUT IS OBSERVED

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

There are two ways to define state:

PREDICTING OUTPUTS ALMOST SURELY

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

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

FORECASTING OUTPUTS IN DISTRIBUTION

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

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

Now let's construct the state space

PREDICTING OUTPUTS ALMOST SURELY

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

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

Now let's construct the state space

PREDICTING OUTPUTS ALMOST SURELY

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

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

PROPERTIES OF STATE

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

- UPDATES IN A RECURSIVE MANNER:

$$X_{t+1} = \text{function}(X_t, U_t, W_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$Y_t = \text{function}(X_t, U_t, W_t).$$

- Kalman, "Mathematical description of linear dynamical systems", 1963.
- Balakrishnan, "Foundations of state-space theory of cts systems", 1967.
- Willems, "The generation of Lyapunov functions for I/O stable systems", 1971.

Now let's construct the state space

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 $U_{t:T}$,

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

PROPERTIES OF STATE

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

- UPDATES IN A RECURSIVE MANNER:

$$X_{t+1} = \text{function}(X_t, U_t, W_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$Y_t = \text{function}(X_t, U_t, W_t).$$

Now let's construct the state space

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 $U_{t:T}$,

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

PROPERTIES OF STATE

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

- UPDATES IN A RECURSIVE MANNER:

$$X_{t+1} = \text{function}(X_t, U_t, W_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$Y_t = \text{function}(X_t, U_t, W_t).$$

PROPERTIES OF STATE

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

- SUFFICIENT TO PREDICT ITSELF:

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

- SUFFICIENT TO PREDICT OUTPUT:

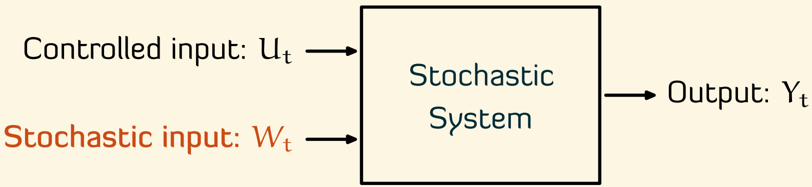
$$\mathbb{P}(Y_t \mid H_t, U_t) = \mathbb{P}(Y_t \mid X_t, U_t).$$

We recover the two basic models
of Markov decision processes!

We recover the two basic models
of Markov decision processes!

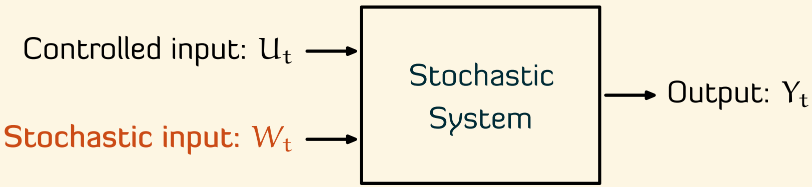
What happens when the
stochastic input is not observered?

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

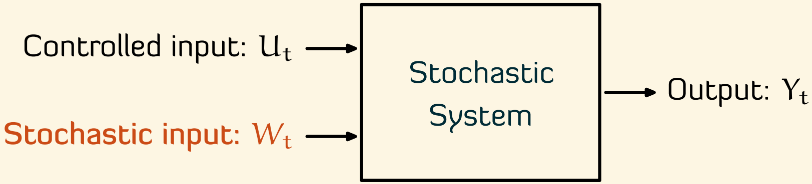


$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

STOCHASTIC INPUT IS NOT OBSERVED

Let $H_t = (U_{1:t-1}, Y_{1:t-1})$ 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 $H_t = (U_{1:t-1}, Y_{1:t-1})$ 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 $B_t \in \Delta(\mathcal{S})$:

$$B_t(s) = \mathbb{P}(S_t = s \mid Y_{1:t-1} = y_{1:t-1}, U_{1:t-1} = u_{1:t-1}), \quad s \in \mathcal{S}.$$

▷ 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.

Approx. info state–(Mahajan)

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.

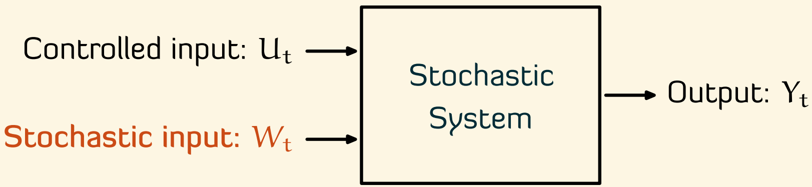
When the state space model is not known analytically (as is the case for black-box models and simulators as well as some real world application such as healthcare), belief states are difficult to construct and difficult to approximate from data.

- 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.

Are there other 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

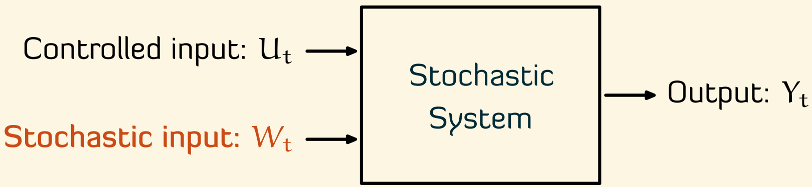


$$Y_t = f_t(U_{1:t}, W_{1:t}).$$

WHEN THE STOCHASTIC INPUT IS NOT OBSERVED

Let $H_t = (U_{1:t-1}, Y_{1:t-1})$ 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

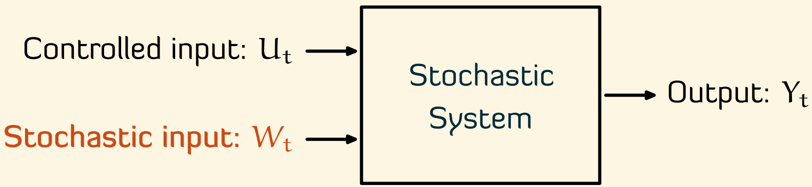
Let $H_t = (U_{1:t-1}, Y_{1:t-1})$ denote the history of inputs and OUTPUTS until time t.

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

PREDICTING OUTPUTS ALMOST SURELY

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

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

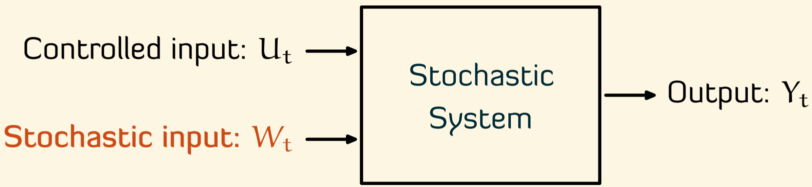
FORECASTING OUTPUTS IN DISTRIBUTION

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

$$\mathbb{P}(Y^{(1)}_{t:T} \mid H^{(1)}_t, U_{t:T}) = \mathbb{P}(Y^{(2)}_{t:T} \mid H^{(2)}_t, U_{t:T})$$

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

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

~~PREDICTING OUTPUTS ALMOST SURELY~~

$$\cancel{H_t^{(1)} \sim H_t^{(2)} \text{ if for all future inputs } (U_{t:T}, W_{t:T}), \quad Y_{t:T}^{(1)} = Y_{t:T}^{(2)}, \quad \text{a.s.}}$$

FORECASTING OUTPUTS IN DISTRIBUTION

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

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

Too restrictive . . .

▷ 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:

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

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:

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

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:

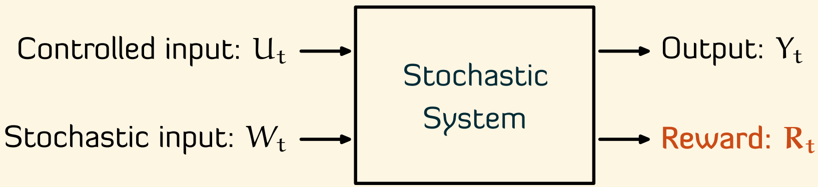
$$\mathbb{P}(Y_t \mid H_t, U_t) = \mathbb{P}(Y_t \mid Z_t, U_t).$$

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



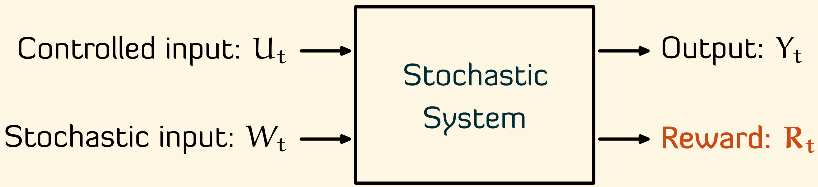
$$Y_t = f_t(U_{1:t}, W_{1:t}),$$

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

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

- There is a hint about this result in Kumar and Varaiya, "Stochastic Systems: estimation, identification, and adaptive control," 1986.

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
  $$U_t = \tilde{g}_t(Z_t).$$
- Let $\{V_t\}_{t=1}^{T+1}$ denote the solution to the following dynamic program: $V_{T+1}(z_{T+1}) = 0$ and for $t \in \{T, \dots, 1\}$,
  $$Q_t(z_t, u_t) = \mathbb{E}[R_t + V_{t+1}(Z_{t+1}) \mid Z_t = z_t, U_t = u_t],$$
  $$V_t(z_t) = \max_{u_t \in \mathcal{U}} Q_t(z_t, u_t).$$
  A policy $\{\tilde{g}_t\}_{t=1}^{T}$, $\tilde{g}_t : \mathcal{Z}_t \to \mathcal{U}$, 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:
  $$\mathbb{E}[R_t \mid H_t, U_t] = \mathbb{E}[R_t \mid Z_t, U_t].$$

- There is a hint about this result in Kumar and Varaiya, "Stochastic Systems: estimation, identification, and adaptive control," 1986.

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.
- ...

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 it satisfies:

- $$\Big| \mathbb{E}[R_t | H_t = h_t, U_t = u_t] - \mathbb{E}[R_t | Z_t = \varphi_t(h_t), U_t = u_t] \Big| < \varepsilon_t$$
- For any Borel set A of $\mathcal{Z}_t$, define
$$\mu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid H_t = h_t, U_t = u_t)$$
and
$$\nu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid Z_t = \varphi_t(h_t), U_t = u_t).$$
Then,
$$d_{\mathfrak{F}}(\mu_t, \nu_t) \leqslant \delta_t.$$

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 it satisfies:

- $$\Big| \mathbb{E}[R_t | H_t = h_t, U_t = u_t] - \mathbb{E}[R_t | Z_t = \varphi_t(h_t), U_t = u_t] \Big| < \varepsilon_t$$

- For any Borel set A of $\mathcal{Z}_t$, define

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

  and

  $$\nu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid Z_t = \varphi_t(h_t), U_t \quad)$$

  Then,

  $$d_{\mathfrak{F}}(\mu_t, \nu_t) \leqslant \delta_t.$$

Approximate information state

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 $t \in \{T, \dots, 1\}$,

$$\hat{Q}_t(z_t, u_t) = \mathbb{E}[R_t + V_{t+1}(Z_{t+1}) \mid Z_t = z_t, U_t = u_t],$$

$$\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} Q_t(z_t, u_t).$$

(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 it satisfies:

- $\Big| \mathbb{E}[R_t | H_t = h_t, U_t = u_t]$
  $- \mathbb{E}[R_t | Z_t = \varphi_t(h_t), U_t = u_t] \Big| < \varepsilon_t$
- For any Borel set A of $\mathcal{Z}_t$, define
  $\mu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid H_t = h_t, U_t = u$
  and
  $\nu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid Z_t = \varphi_t(h_t), U_t$ $)$
  Then,
  $d_{\mathfrak{F}}(\mu_t, \nu_t) \leqslant \delta_t.$

Approximate information state

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 $t \in \{T, \dots, 1\}$,

$$\hat{Q}_t(z_t, u_t) = \mathbb{E}[R_t + V_{t+1}(Z_{t+1}) \mid Z_t = z_t, U_t = u_t],$$

$$\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} Q_t(z_t, u_t).$$

Then, if there exist positive constants $\{K_t\}_{t\geqslant 1}$ such that $\hat{V}_t/K_t \in \mathfrak{F}$, then for any history h_t,

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

(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 it satisfies:

- $\Big|\mathbb{E}[R_t | H_t = h_t, U_t = u_t]$
  $- \mathbb{E}[R_t | Z_t = \varphi_t(h_t), U_t = u_t]\Big| < \varepsilon_t$
- For any Borel set A of $\mathcal{Z}_t$, define
  $\mu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid H_t = h_t, U_t = u$
  and
  $\nu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid Z_t = \varphi_t(h_t), U_t$ $)$
  Then,
  $d_{\mathfrak{F}}(\mu_t, \nu_t) \leqslant \delta_t.$

AIS: Some remarks

In the definition of AIS, we can replace

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

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

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

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

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

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.

AIS: Some remarks

In the definition of AIS, we can replace

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

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 !!

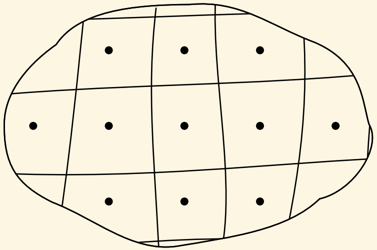
Two ways to interpret the results:

- 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

Analytic example: Error bounds on state discretization

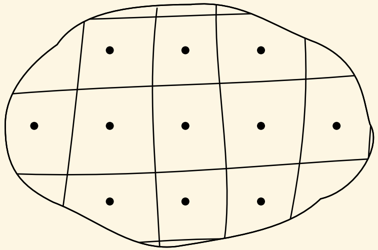


Consider an MDP with state space $\mathcal{X}$ and per-step reward $R_t = r(X_t, U_t)$.

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.

Analytic example: Error bounds on state discretization



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 $\varphi\colon \mathcal{X} \to \mathcal{Z}$.

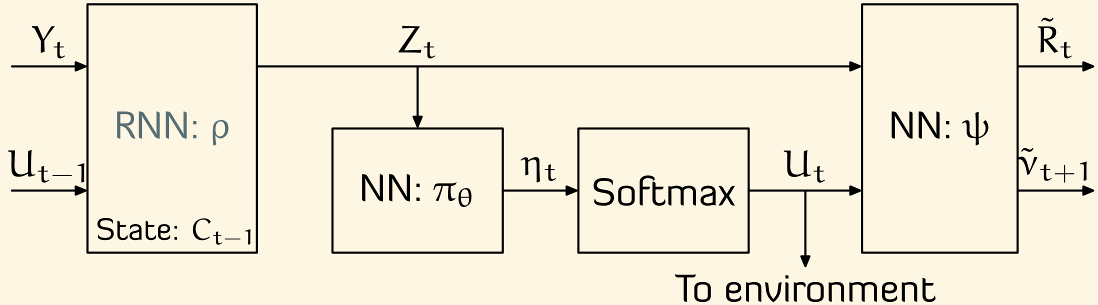
- Let $z = \varphi(x)$ denote the label for x.
- Then $\varphi^{-1}(z)$ denote all states which have label z.

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

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

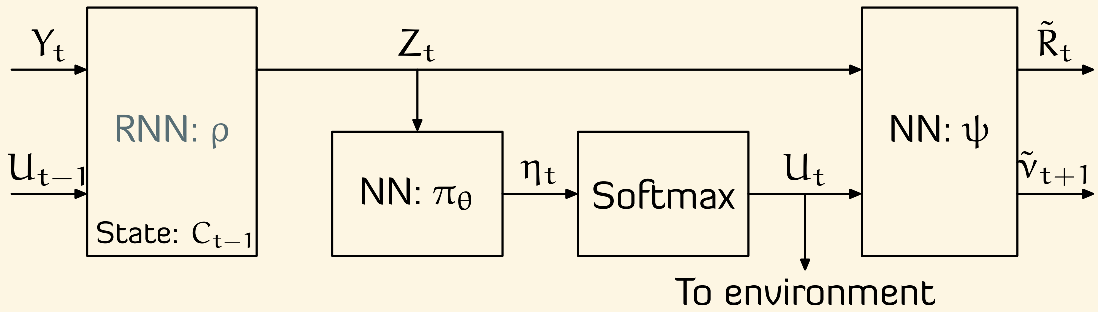
$$\delta = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} d_{\mathfrak{F}}(\mathbb{P}(Z_+ \mid X = x, U = u), \mathbb{P}(Z_+ \mid X \in \varphi^{-1}(\varphi(x)), U = u)).$$

Numerical example: Reinforcement learning for POMDPs

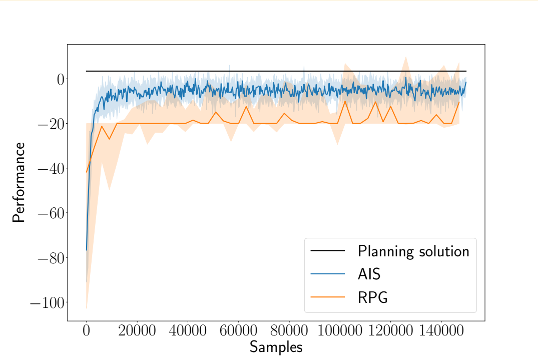


Develop a three time-scale AIS-based actor-critic algorithm for RL in POMDPs.

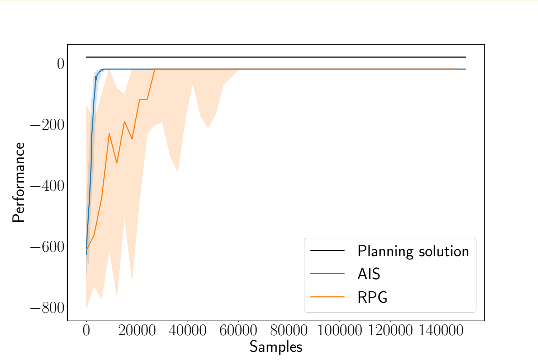
Numerical example: Reinforcement learning for POMDPs



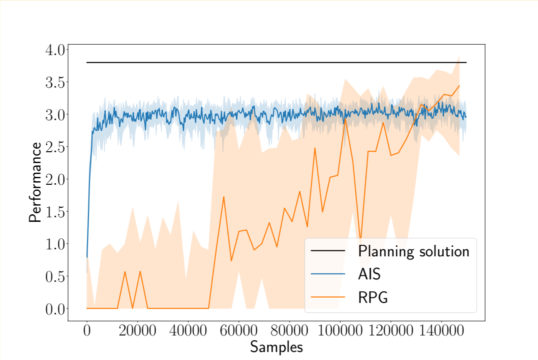
Develop a three time-scale AIS-based actor-critic algorithm for RL in POMDPs.



Voicemail problem



Tiger problem



4 × 4 grid problem

Summary

Summary

Now let's construct the state space

PREDICTING OUTPUTS ALMOST SURELY

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

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

FORECASTING OUTPUTS IN DISTRIBUTION

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

$$\mathbb{P}(Y^{(1)}_{t:T} \mid H^{(1)}_t, U_{t:T}) = \mathbb{P}(Y^{(2)}_{t:T} \mid H^{(2)}_t, U_{t:T})$$

PROPERTIES OF STATE

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

- UPDATES IN A RECURSIVE MANNER:

$$X_{t+1} = \text{function}(X_t, U_t, W_t).$$

- SUFFICIENT TO PREDICT OUTPUT:

$$Y_t = \text{function}(X_t, U_t, W_t).$$

PROPERTIES OF STATE

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

- SUFFICIENT TO PREDICT ITSELF:

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

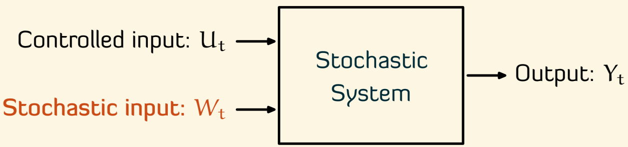
- SUFFICIENT TO PREDICT OUTPUT:

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

Summary

Now let's construct the state space

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 $H_t = (U_{1:t-1}, Y_{1:t-1})$ 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 $B_t \in \Delta(\mathcal{S})$:

$$B_t(s) = \mathbb{P}(S_t = s \mid Y_{1:t-1} = y_{1:t-1}, U_{1:t-1} = u_{1:t-1}), \quad s \in \mathcal{S}.$$

▷ 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.

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

Notion of state in partially observed stochastic dynamical systems

Now let's consturct 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 it satisfies:

- $$\Big| \mathbb{E}[R_t | H_t = h_t, U_t = u_t] - \mathbb{E}[R_t | Z_t = \varphi_t(h_t), U_t = u_t] \Big| < \varepsilon_t$$
- For any Borel set A of $\mathcal{Z}_t$, define
  $$\mu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid H_t = h_t, U_t = u_t)$$
  and
  $$\nu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid Z_t = \varphi_t(h_t), U_t = u_t).$$
  Then,
  $$d_{\mathfrak{F}}(\mu_t, \nu_t) \leqslant \delta_t.$$

Summary

Approximate information state

Approximate information state

MAIN THEOREM

Given a function class $\mathfrak{F}$, let $\{Z_t\}_{t\geq 1}$, 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 $t \in \{T, \dots, 1\}$,

$$\hat{Q}_t(z_t, u_t) = \mathbb{E}[R_t + V_{t+1}(Z_{t+1}) \mid Z_t = z_t, U_t = u_t],$$

$$\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} Q_t(z_t, u_t).$$

Then, if there exist positive constants $\{K_t\}_{t\geqslant 1}$ such that $\hat{V}_t / K_t \in \mathfrak{F}$, then for any history h_t,

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

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

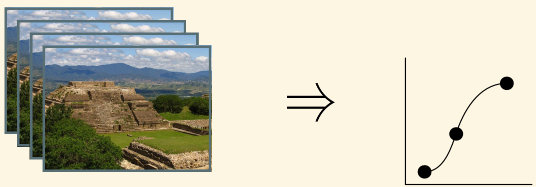
Given a function class $\mathfrak{F}$, a compression $\{Z_t\}_{t\geqslant 1}$ of history (i.e., $Z_t = \varphi_t(H_t)$) is called an $\{(\varepsilon_t, \delta_t)\}_{t\geqslant 1}$ AIS if it satisfies:

- $\left|\mathbb{E}[R_t | H_t = h_t, U_t = u_t] - \mathbb{E}[R_t | Z_t = \varphi_t(h_t), U_t = u_t]\right| < \varepsilon_t$
- For any Borel set A of $\mathcal{Z}_t$, define
  $$\mu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid H_t = h_t, U_t = u_t)$$
  and
  $$\nu_t(A) = \mathbb{P}(Z_{t+1} \in A \mid Z_t = \varphi_t(h_t), U_t = u_t)$$
  Then,
  $$d_{\mathfrak{F}}(\mu_t, \nu_t) \leqslant \delta_t.$$

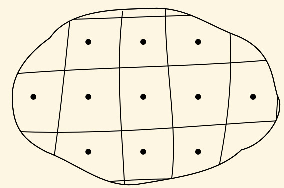
Summary

Conclusion

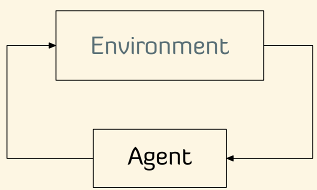
In my biased opinion, the notions of information state and approximate information state provide a conceptually clean framework to think about approximations (and online reinforcement learning) in sequential decision making.



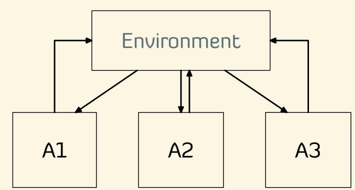
Discovering latent space models



State aggregation



Online RL in POMDPs



Online Multi-agent RL