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

Approx. info. state—(Subramanian and Mahajan)
Many successes of RL in recent years
  ▶ Algorithms based on comprehensive theory

Approx. info. state–(Subramanian and Mahajan)
Many successes of RL in recent years

- Algorithms based on comprehensive theory

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

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

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

Controlled input: $U_t$ \quad Stochastic System \quad Output: $Y_t$

Stochastic input: $W_t$

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

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 $t$. 

Approx. info. state--(Subramanian and Mahajan)
Notion of state in partially observed stochastic dynamical systems

Controlled input: $U_t$ 
Stochastic input: $W_t$ 
Stochastic System 
Output: $Y_t$

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

**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 $\{S_t\}_{t \geq 0}$ for predicting output assuming that the stochastic inputs are observed.

**Step 2** Define a belief state $B_t \in \Delta(S)$:

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


Approx. info. state–(Subramanian and Mahajan)
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.


Approx. info. state—(Subramanian and Mahajan)
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

Controlled input: $U_t$  
Stochastic input: $W_t$  

\[
\text{Output: } Y_t = f_t(U_{1:t}, W_{1:t}).
\]

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

Controlled input: $U_t$ ➔ Stochastic System ➔ Output: $Y_t$

Stochastic input: $W_t$

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 $(U_{t:T}, W_{t:T})$,

$Y_{t:1:T}^{(1)} = Y_{t:1:T}^{(2)}$, a.s.
Notion of state in partially observed stochastic dynamical systems

Controlled input: $U_t$ → Stochastic System → Output: $Y_t$

Stochastic input: $W_t$

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

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 $(U_{t:T}, W_{t:T})$,

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

FORECASTING OUTPUTS IN DISTRIBUTION

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

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


Notion of state in partially observed stochastic dynamical systems

Controlled input: $U_t$  
Stochastic input: $W_t$  
Output: $Y_t$

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

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 $(U_{t:T}, W_{t:T})$,

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

FORECASTING OUTPUTS IN DISTRIBUTION

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

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

Too restrictive ...
Now let’s construct the state space

Now let’s construct the state space.

**FORECASTING OUTPUTS IN DISTRIBUTION**

\[ H_t^{(1)} \sim H_t^{(2)} \text{ if for all future CONTROL inputs } U_{t:T}, \]
\[ \mathbb{P}(Y_{t:T}^{(1)} | H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} | H_t^{(2)}, U_{t:T}) \]

Approx. info. state–(Subramanian and Mahajan)
FORECASTING OUTPUTS IN DISTRIBUTION

\[ H_t^{(1)} \sim H_t^{(2)} \text{ if for all future CONTROL inputs } U_{t:T}, \]
\[ \mathbb{P}(Y_{t:T}^{(1)} | H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} | H_t^{(2)}, U_{t:T}) \]

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:**
  \[ \mathbb{P}(Z_{t+1} | H_t, U_t) = \mathbb{P}(Z_{t+1} | Z_t, U_t). \]
- **SUFFICIENT TO PREDICT OUTPUT:**
  \[ \mathbb{P}(Y_t | H_t, U_t) = \mathbb{P}(Y_t | Z_t, U_t). \]

Now let’s construct the state space
Now let’s construct the state space

**FORECASTING OUTPUTS IN DISTRIBUTION**

\[ H_t^{(1)} \sim H_t^{(2)} \text{ if for all future CONTROL inputs } U_{t:T}, \]
\[ \mathbb{P}(Y_{t:T}^{(1)} | H_t^{(1)}, U_{t:T}) = \mathbb{P}(Y_{t:T}^{(2)} | H_t^{(2)}, U_{t:T}) \]

**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:**
  \[ \mathbb{P}(Z_{t+1} | H_t, U_t) = \mathbb{P}(Z_{t+1} | Z_t, U_t). \]

- **SUFFICIENT TO PREDICT OUTPUT:**
  \[ \mathbb{P}(Y_t | H_t, U_t) = \mathbb{P}(Y_t | Z_t, U_t). \]

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 construct the state space

### Forecasting Outputs in Distribution

\[ H^{(1)}_t \sim H^{(2)}_t \text{ if for all future CONTROL inputs } U_{t:T}, \]
\[ P(Y^{(1)}_{t:T} | H^{(1)}_t, U_{t:T}) = P(Y^{(2)}_{t:T} | H^{(2)}_t, U_{t:T}) \]

### 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**:
  \[ P(Z_{t+1} | H_t, U_t) = P(Z_{t+1} | Z_t, U_t). \]

- **Sufficient to Predict Output**:
  \[ P(Y_t | H_t, U_t) = P(Y_t | Z_t, U_t). \]

### Key Questions

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

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

Approx. info. state–(Subramanian and Mahajan)
An information state for dynamic programming
Stochastic System

Controlled input: $U_t$  
Stochastic input: $W_t$  
Output: $Y_t$  
Reward: $R_t$

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

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

**Stochastic System**

**Controlled input:** $U_t$  
**Stochastic input:** $W_t$  
**Output:** $Y_t$  
**Reward:** $R_t$

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

- **Sufficient to estimate expected reward:**
  $$\mathbb{E}[R_t | H_t, U_t] = \mathbb{E}[R_t | Z_t, U_t].$$
## Dynamic programming using information state

### PROPERTIES OF INFORMATION STATE

<table>
<thead>
<tr>
<th>(SUFFICIENT FOR DYNAMIC PROGRAMMING)</th>
</tr>
</thead>
</table>

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

---

Approx. info. state—(Subramanian and Mahajan)
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 \( U_t = \tilde{g}_t(Z_t) \).

**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:
  \[
P(Z_{t+1} | H_t, U_t) = P(Z_{t+1} | Z_t, U_t).
  \]

- SUFFICIENT TO ESTIMATE EXPECTED REWARD:
  \[
  E[R_t | H_t, U_t] = E[R_t | Z_t, U_t].
  \]

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

Approx. info. state—(Subramanian and Mahajan)
Dynamic programming using information state

PRELIMINARY THEOREM

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

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

2. 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 \( \{\tilde{g}_t\}_{t=1}^{T}, \tilde{g}_t : \mathcal{Z}_t \to \mathcal{U} \), is optimal if it satisfies
\[ \tilde{g}_t(z_t) \in \arg \max_{u_t \in \mathcal{U}} Q_t(z_t, u_t). \]
What about approximations?
INTEGRAL PROBABILITY METRIC (IPM)

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

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

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


Approx. info. state–(Subramanian and Mahajan)
INTEGRAL PROBABILITY METRIC (IPM)

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

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

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

EXAMPLES

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

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

---


Approx. info. state–(Subramanian and Mahajan)
**Approximate information state**

\[(\varepsilon, \delta)\text{-APPROXIMATE INFORMATION STATE (AIS)}\]

Given a function class \(\mathcal{F}\), a compression \(\{Z_t\}_{t \geq 1}\) of history (i.e., \(Z_t = \varphi_t(H_t)\)) is called an \(\{(\varepsilon_t, \delta_t)\}_{t \geq 1}\) 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

\[
\left| \mathbb{E}[R_t | H_t = h_t, U_t = u_t] - \tilde{R}_t(\varphi_t(h_t), u_t) \right| \leq \varepsilon_t
\]

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

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

Then,

\[
d_{\mathcal{F}}(\mu_t, \nu_t(\cdot | \varphi_t(h_t), u_t)) \leq \delta_t.
\]
Approximate dynamic programming using AIS

**MAIN THEOREM**

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

Recursively define the following functions:

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

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

$$\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} \left\{ \tilde{R}_t(z_t, u_t) \right.$$  

$$+ \int V_{t+1}(z_{t+1}) \nu_t(dz_{t+1} \mid z_t, u_t) \bigg\}.$$  

Let $\pi = (\pi_1, \ldots, \pi_T)$ denote the corresponding policy.
Approximate dynamic programming using AIS

**MAIN THEOREM**

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

Recursively define the following functions:

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

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

\[
\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} \left\{ R_t(z_t, u_t) + \int V_{t+1}(z_{t+1}) \nu_t(dz_{t+1} | z_t, u_t) \right\}.
\]

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

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

- for any history $h_t$,
  \[
  |V_t(h_t) - \hat{V}_t(\varphi_t(h_t))| \leq \varepsilon_T + \sum_{s=t}^{T} (\varepsilon_s + K_s \delta_s).
  \]

- for any history $h_t$,
  \[
  |V_t(h_t) - V_{t}^{\pi}(h_t)| \leq 2[\varepsilon_T + \sum_{s=t}^{T} (\varepsilon_s + K_s \delta_s)].
  \]
In the definition of AIS, we can replace
\[ d_\mathfrak{F}(\mathbb{P}(\mu_t, \nu_t(\cdot|Z_t = \varphi_t(h_t), U_t = u_t))) \leq \delta_t \]
by
\[ Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t) \]
\[ d_\mathfrak{F}(\mathbb{P}(Y_t|H_t = h_t, U_t = u_t), \mathbb{P}(Y_t|Z_t = \varphi_t(h_t), U_t = u_t)) \leq \delta_t. \]
In the definition of AIS, we can replace
\[ d_{\mathcal{F}}(\mathbb{P}(\mu_t, \nu_t(\cdot|Z_t = \varphi_t(h_t), U_t = u_t)) \leq \delta_t \]
by

\[ Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t) \]

\[ d_{\mathcal{F}}(\mathbb{P}(Y_t|H_t = h_t, U_t = u_t), \mathbb{P}(Y_t|Z_t = \varphi_t(h_t), U_t = u_t)) \leq \delta_t. \]

The AIS process \( \{Z_t\}_{t \geq 1} \) need not be Markov!!
AIS: Some remarks

In the definition of AIS, we can replace
\[ d_\delta(\mathbb{P}(\mu_t, \nu_t(\cdot|Z_t = \varphi_t(h_t), U_t = u_t)) \leq \delta_t \]
by
\[ Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t) \]
\[ d_\delta(\mathbb{P}(Y_{t+1}|H_t = h_t, U_t = u_t), \mathbb{P}(Y_t|Z_t = \varphi_t(h_t), U_t = u_t)) \leq \delta_t. \]

The AIS process \( \{Z_t\}_{t \geq 1} \) need not be Markov!!

Two ways to interpret the results:
\[ \text{Given the information state space } \mathcal{Z}, \text{ find the best compression } \varphi_t: \mathcal{H}_t \rightarrow \mathcal{Z} \]
\[ \text{Given any compression function } \varphi_t: \mathcal{H}_t \rightarrow \mathcal{Z}_t, \text{ find the approximation error. } \]
In the definition of AIS, we can replace
\[ d_{\mathcal{S}}(\mathbb{P}(\mu_t, \nu_t(\cdot|Z_t = \varphi_t(h_t), U_t = u_t)) \leq \delta_t \]
by
\[ Z_{t+1} = \text{function}(Z_t, Y_{t+1}, U_t) \]
\[ d_{\mathcal{S}}(\mathbb{P}(Y_t|H_t = h_t, U_t = u_t), \mathbb{P}(Y_t|Z_t = \varphi_t(h_t), U_t = u_t)) \leq \delta_t. \]

The AIS process \( \{Z_t\}_{t \geq 1} \) need not be Markov!!

Two ways to interpret the results:
\[ \square \text{ Given the information state space } \mathcal{Z}, \text{ find the best compression } \varphi_t : \mathcal{H}_t \to \mathcal{Z} \]
\[ \square \text{ Given any compression function } \varphi_t : \mathcal{H}_t \to \mathcal{Z}_t, \text{ find the approximation error.} \]

Results naturally extend to infinite horizon
Some examples
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 $\varphi: \mathcal{X} \rightarrow \mathcal{Z}$.

- 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 $\varphi: \mathcal{X} \to \mathcal{Z}$.

- 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 $(\varepsilon, \delta)$ AIs

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

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

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

Example 1: Error bounds on state aggregation


Approx. info. state–(Subramanian and Mahajan)
Example 2: Approximation bounds for using quantized obs.


Approx. info. state–(Subramanian and Mahajan)
Example 2: Approximation bounds for using quantized obs.

- Proposed as a heuristic algorithms
- No performance bounds


Approx. info. state—(Subramanian and Mahajan)
Example 2: Approximation bounds for using quantized obs.

- Proposed as a heuristic algorithms
- No performance bounds

\[
\{Z_t\}_{t \geq 1} \text{ IS AN } (\varepsilon, \delta) \text{ AIS}
\]

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

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


Approx. info. state–(Subramanian and Mahajan)
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 $X_t^i$, control $U_t^i$.

- **Dynamics**
  \[
  \mathbb{P}(X_{t+1}|X_t, U_t) = \prod_{i=1}^n P(X_{t+1}^i|X_t^i, U_t^i, M_t)
  \]

- **Per-step reward**
  \[
  R(X_t, U_t) = \frac{1}{n} \sum_{i=1}^n r(X_t^i, U_t^i, M_t)
  \]

- **Empirical mean-field:**
  \[
  M_t(x) = \frac{1}{n} \sum_{i=1}^n \delta_{X_t^i}(x).
  \]

- **Statistical mean-field:**
  \[
  \tilde{m}_t(x) = \mathbb{P}(X_t^i = x).
  \]
Example 3: Approximation bounds for mean-field teams

$n$ agents: state $X_i^t$, control $U_i^t$.

- **Dynamics**
  \[
  \mathbb{P}(X_{t+1}^i|X_t^i, U_t^i) = \prod_{i=1}^{n} P(X_{t+1}^i|X_t^i, U_t^i, M_t)
  \]

- **Per-step reward**
  \[
  R(X_t^i, U_t^i) = \frac{1}{n} \sum_{i=1}^{n} r(X_t^i, U_t^i, M_t)
  \]

- **Info structure**: $I_t^i = \{X_t^i\}$.

- **Empirical mean-field**:
  \[
  M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i^t}(x).
  \]

- **Statistical mean-field**:
  \[
  \bar{m}_t(x) = \mathbb{P}(X_t^i = x).
  \]
Example 3: Approximation bounds for mean-field teams

- n agents: state $X_t^i$, control $U_t^i$.
- **Dynamics**
  \[
  \mathbb{P}(X_{t+1}|X_t, U_t) = \prod_{i=1}^n P(X_{t+1}^i|X_t^i, U_t^i, M_t)
  \]
- **Per-step reward**
  \[
  R(X_t, U_t) = \frac{1}{n} \sum_{i=1}^n r(X_t^i, U_t^i, M_t)
  \]
- **Info structure**: $I_t^i = \{X_t^i\}$.
- **Expanded info structure**: $\tilde{I}_t^i = \{X_t^i, M_t\}$.

Empirical mean-field:
\[
M_t(x) = \frac{1}{n} \sum_{i=1}^n \delta_{X_t^i}(x).
\]
Statistical mean-field:
\[
\bar{m}_t(x) = \mathbb{P}(X_t^i = x).
\]
Example 3: Approximation bounds for mean-field teams

\[ \mathbb{P}(X_{t+1} | X_t, U_t) = \prod_{i=1}^{n} P(X_{t+1}^i | X_t^i, U_t^i, M_t) \]

![Dynamics](image)

\[ R(X_t, U_t) = \frac{1}{n} \sum_{i=1}^{n} r(X_t^i, U_t^i, M_t) \]

![Per-step reward](image)

\[ M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_t^i}(x). \]

![Empirical mean-field](image)

\[ \bar{m}_t(x) = \mathbb{P}(X_t^i = x). \]

![Statistical mean-field](image)

\[ \mathcal{J}^* \leq \tilde{\mathcal{J}}^* \]

![Info structure](image)

\[ \mathcal{I}_t^i = \{X_t^i\}. \]

![Expanded info structure](image)

\[ \tilde{\mathcal{I}}_t^i = \{X_t^i, M_t\}. \]

\[ \mathcal{J}^* \leq \tilde{\mathcal{J}}^* \]

\[ \mathcal{J}^* \leq \tilde{\mathcal{J}}^* \]

\[ (A) r(x, u, m) \text{ and } P(y|x, u, m) \text{ are Lipschitz in } m. \]

\[ \{\bar{m}_t\}_{t=1}^{\infty} \text{ is an } (\epsilon, \delta) \text{ AIS for expanded info structure, where } \epsilon, \delta \in \Theta(1/\sqrt{n}). \]
Example 3: Approximation bounds for mean-field teams

- **n agents**: state $X^i_t$, control $U^i_t$.

  - **Dynamics**
    
    $\mathbb{P}(X_{t+1} | X_t, U_t) = \prod_{i=1}^n P(X^i_{t+1} | X^i_t, U^i_t, M_t)$

  - **Per-step reward**
    
    $R(X_t, U_t) = \frac{1}{n} \sum_{i=1}^n r(X^i_t, U^i_t, M_t)$

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

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

    - $J^* \leq \tilde{J}^*$, $\tilde{J}^* - J^* \leq K/\sqrt{n}$

    - $\tilde{J}^* \leq J^* \leq \tilde{J}^* + K/\sqrt{n}$.

- **Empirical mean-field**:

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

- **Statistical mean-field**:

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

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

  - $\{\tilde{m}_t\}_{t \geq 1}$ is an $(\epsilon, \delta)$ AIS for expanded info structure, where $\epsilon, \delta \in O(1/\sqrt{n})$. 
Now to reinforcement learning for partially observed systems.
Reinforcement learning setup

State aggregator:

\[ \mathcal{L}_{AIS} = \alpha_t |\tilde{R}_t - R_t| + (1 - \alpha_t) d_\delta(\nu_t, \mu_t) \]

\( \xi \): Parameters of the aggregator

Updated using SGD with LR \( \alpha_k \)

Approx. info. state–(Subramanian and Mahajan)
Reinforcement learning setup

- **State aggregator**:
  \[
  \mathcal{L}_{\text{AIS}} = \alpha_t |\tilde{R}_t - R_t| + (1 - \alpha_t) d_{\tilde{\delta}}(\nu_t, \mu_t)
  \]
  \(\xi\): Parameters of the aggregator
  Updated using SGD with LR \(\alpha_k\)

- **Value approximator**:
  \(\phi\): parameters of \(Q(z, u)\) approximator.
  Updated using TD(0) or TD(\(\lambda\)) with LR \(b_k\).
Reinforcement learning setup

- **State aggregator:**
  \[ \mathcal{L}_{\text{AIS}} = \alpha_t |\tilde{R}_t - R_t| + (1 - \alpha_t) d_{\tilde{\delta}}(\nu_t, \mu_t) \]
  \( \xi \): Parameters of the aggregator
  Updated using SGD with LR \( a_k \)

- **Value approximator:**
  \( \varphi \): parameters of \( Q(z, u) \) approximator.
  Updated using TD(0) or TD(\( \lambda \)) with LR \( b_k \).

- **Policy approximator:**
  \( \theta \): parameters of \( \pi(u | z) \)
  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: $4 \times 4$ Grid Environment
4 × 4 Grid Environment

Approx. info. state–(Subramanian and Mahajan)
Numerical Results: Tiger Environment

Approx. info. state–(Subramanian and Mahajan)
Tiger Environment

Approx. info. state–(Subramanian and Mahajan)
Numerical Results: Cheese Maze Environment
Cheese Maze Environment

Approx. info. state–(Subramanian and Mahajan)
Approx. info. state—(Subramanian and Mahajan)
Now let’s construct 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**:
  $$\mathbb{P}(Y_t \mid H_t, U_t) = \mathbb{P}(Y_t \mid Z_t, U_t).$$

**KEY QUESTIONS**

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

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

\((\varepsilon, \delta)\)-APPROXIMATE INFORMATION STATE (AIS)

Given a function class \(\mathcal{F}\), a compression \(\{Z_t\}_{t \geq 1}\) of history (i.e., \(Z_t = \phi_t(H_t)\)) is called an \(\{(\varepsilon_t, \delta_t)\}_{t \geq 1}\) 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

- \(|E[R_t|H_t = h_t, U_t = u_t] - \tilde{R}_t(\phi_t(h_t), u_t)| \leq \varepsilon_t\)

- For any Borel set \(A\) of \(Z_t\), define
  \[\mu_t(A) = \mathbb{P}(Z_{t+1} \in A | H_t = h_t, U_t = u_t)\]

Then,

\[d_{\mathcal{F}}(\mu_t, \nu_t(\cdot | \phi_t(h_t), u_t)) \leq \delta_t.\]
Summary

New let’s construct the state space

Approximate dynamic programming using AIS

**MAIN THEOREM**

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

Recursively define the following functions:

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

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

$$
\hat{V}_t(z_t) = \max_{u_t \in \mathcal{U}} \left\{ R_t(z_t, u_t) 
+ \int V_{t+1}(z_{t+1}) \nu_t(dz_{t+1} | z_t, u_t) \right\}.
$$

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

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

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

\[|V_t(h_t) - V_t^\pi(h_t)| \leq 2[\varepsilon_T + \sum_{s=t}^{T} (\varepsilon_s + K_s \delta_s)].\]
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 $\varphi: \mathcal{X} \rightarrow \mathcal{Z}$.

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

\[
\{Z_t\}_{t \geq 1} \text{ IS AN } (\epsilon, \delta) \text{ AIS}
\]

\[
\epsilon = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} |r(x,u) - r(\varphi(x), u)| \quad \text{or, equivalently, } r(\cdot, u) \text{ has a } \mathcal{F} \text{-constant } K_r
\]

\[
\delta = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} d(\mathbb{P}(X_+ | X = x, U = u), \mathbb{P}(X_+ | X \in \varphi^{-1}(\varphi(x)), U = u)).
\]

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


Approx. info. state–(Subramanian and Mahajan)
Approx. info. state—(Subramanian and Mahajan)

Example 2: Approximation bounds for using quantized obs.

- Proposed as a heuristic algorithms
- No performance bounds

\[ \{Z_t\}_{t \geq 1} \text{ IS AN } (\varepsilon, \delta) \text{ AIS} \]

\[ \varepsilon = \sup_{h_t, u_t} |\mathbb{E}[R_t|h_t, u_t] - \hat{R}_t(h_t, u_t)| \]

\[ \delta = \sup_{h_t, u_t} d_\mathcal{F}(\mathbb{P}(\hat{Y}_{t+1}|h_t, u_t), \mathbb{P}(\hat{Y}_{t+1}|\varphi(h_t), u_t)) \]


Approx. info. state—(Subramanian and Mahajan)
Now let’s construct the state space.

**Approximate info. state**–(Subramanian and Mahajan)

### Example 3: Approximation bounds for mean-field teams

- **n** agents: state $X_t^i$, control $U_t^i$.
- **Dynamics**
  
  \[ \mathbb{P}(X_{t+1}^i|X_t^i, U_t^i) = \prod_{i=1}^n \mathbb{P}(X_{t+1}^i|X_t^i, U_t^i, M_t) \]
- **Per-step reward**
  \[ R(X_t, U_t) = \frac{1}{n} \sum_{i=1}^n r(X_t^i, U_t^i, M_t) \]
- **Info structure**: $I^i_t = \{X^i_t\}$.
- **Expanded info structure**: $\tilde{I}^i_t = \{X^i_t, M_t\}$.
  \[ J^* \leq \tilde{J}^*, \quad \tilde{J}^* - J^* \leq K/\sqrt{n} \]
  \[ \tilde{J}^* \leq \bar{J}^* \leq J^* + K/\sqrt{n}. \]

- **Empirical mean-field**:
  \[ M_t(x) = \frac{1}{n} \sum_{i=1}^n \delta_{X_t^i}(x). \]
- **Statistical mean-field**:
  \[ \bar{m}_t(x) = \mathbb{P}(X_t^i = x). \]

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

\[ \{\bar{m}_t\}_{t \geq 1} \text{ is an } (\varepsilon, \delta) \text{ AIS for expanded info structure, where } \varepsilon, \delta \in O(1/\sqrt{n}). \]
Approx. info. state—(Subramanian and Mahajan)

Example 1: Error bounds on state aggregation

For any Borel set \( \{ Z \} = \delta = \text{AIS} \)

Suppose \( \text{AIS} \) has a video observation for the control input \( \{ \phi(x), U = u \} \).

Same complexity as identifying the state info structure, where

For any history \( \phi(x) \), \( U = u \).

Memory

Can this be used for dynamic programming using \( \text{AIS} \)?

Approximate dynamic programming using \( \text{AIS} \)

4 × 4 Grid Environment

Approx. info. state—(Subramanian and Mahajan)
Now let's construct the state space.

Approximate dynamic programming using AIS

Example 2: Approximation bounds for using quantized obs. states.

PROPERTIES OF INFORMATION STATE

Example 2: Approximation bounds for using quantized obs. states.

Recursively define the following functions:

Given a function class \( \hat{V} \),

is called an \( \epsilon \)-approximate policy.

Approximate information state (Subramanian and Mahajan)

Approximate information state (Subramanian and Mahajan)

Approximate information state (Subramanian and Mahajan)

Let \( \pi \) be an \( \epsilon \)-optimal policy.

For any Borel set \( Z \),

\[
\sup_{\pi} \mathbb{E}[R(Z, \pi)] - \mathbb{E}[R(Z, \pi)] \leq \epsilon.
\]

(\( \epsilon, \delta \))

\( \mathbb{E}[R(Z, \pi)] \leq \hat{R}(Z) + \epsilon \).

Example 2: Approximation bounds for using quantized obs. states.

\( \delta = \frac{\epsilon}{(1 - \beta) \mathbb{P}(Z) \mathbb{P}(X)} \).

Approximate information state (Subramanian and Mahajan)

Though not performance bounds, expanded info structure has

SUFFICIENT TO PREDICT OUTPUT:

SUFFICIENT TO PREDICT ITSELF:

Tiger Environment


<table>
<thead>
<tr>
<th>Samples ( \times 10^6 )</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>-40</td>
</tr>
<tr>
<td>0.2</td>
<td>-20</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>0.6</td>
<td>10</td>
</tr>
<tr>
<td>0.8</td>
<td>15</td>
</tr>
<tr>
<td>1.0</td>
<td>20</td>
</tr>
<tr>
<td>1.2</td>
<td>19</td>
</tr>
<tr>
<td>1.4</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Samples ( \times 10^6 )</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>-40</td>
</tr>
<tr>
<td>0.2</td>
<td>-20</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>0.6</td>
<td>10</td>
</tr>
<tr>
<td>0.8</td>
<td>15</td>
</tr>
<tr>
<td>1.0</td>
<td>20</td>
</tr>
<tr>
<td>1.2</td>
<td>19</td>
</tr>
<tr>
<td>1.4</td>
<td>18</td>
</tr>
</tbody>
</table>

Planning solution

RPG

AIS

Approximate information state (Subramanian and Mahajan)

Approximate information state (Subramanian and Mahajan)

Approximate information state (Subramanian and Mahajan)
Now let's construct the state space. Approximate dynamic programming using AIS.

Example 1: Approximation bounds for mean field teams.

**Cheese Maze Environment**

![Graph showing performance over samples for Planning solution, RPG, and AIS]

Approx. info. state—(Subramanian and Mahajan)

Approx. info. state—(Subramanian and Mahajan)
AIS provides a conceptually clean framework for approximate DP and online RL in partially observed systems.