Sub-optimality bounds for Certainty Equivalence in POMDPs

Aditya Mahajan
McGill University

Joint work with Berk Bozkurt, Ashutosh Nayyar, and Yi Ouyang

CDC Workshop on Information Decentralization
December 2025

- email: aditya.mahajan@mcgill.ca
- web: https://adityam.github.io

Using POMDPs in real-world applications

POMDPs model many real-world applications

- Model applications where the decision maker does not have access to the complete state.
- Examples: Robotics, autonomous systems, finance, healthcare, and other domains

Using POMDPs in real-world applications

POMDPs model many real-world applications

- Model applications where the decision maker does not have access to the complete state.
- **Examples:** Robotics, autonomous systems, finance, healthcare, and other domains

Computational challenges

- **Standard approach:** translate POMDPs to belief-state MDPs

Astrom, "Optimal control of Markov processes with incomplete information," JMAA 1965.
Smallwood, "Optimal control of partially observable Markov processes over a finite horizon," OR 1973.
Papadimitriou and Tsitsiklis, "The complexity of Markov decision processes," MOR 1987.

Using POMDPs in real-world applications

POMDPs model many real-world applications

- Model applications where the decision maker does not have access to the complete state.
- **Examples:** Robotics, autonomous systems, finance, healthcare, and other domains

Computational challenges

- **Standard approach:** translate POMDPs to belief-state MDPs

- Finding optimal policy is PSPACE-hard
- Exact algorithms have exponential worst-case complexity
- Finding approximately optimal policies is also PSPACE-hard
- Heuristic approaches can be efficient but lack provable performance guarantees

---

Astrom, "Optimal control of Markov processes with incomplete information," JMAA 1965.
Smallwood, "Optimal control of partially observable Markov processes over a finite horizon," OR 1973.
Papadimitriou and Tsitsiklis, "The complexity of Markov decision processes," MOR 1987.

Trading off computational tractability and performance

Structured agent-state based policies

- Balance computational tractability and good performance guarantees
- **Examples:** Finite window policies (frame stacking in RL), RNN-based policies
- Agent-state: recursively updatable function of past observations and actions

Trading off computational tractability and performance

Structured agent-state based policies

- Balance computational tractability and good performance guarantees
- Examples: Finite window policies (frame stacking in RL), RNN-based policies
- Agent-state: recursively updatable function of past observations and actions

Sufficient conditions for good performance

- Approximate information state [Subramanian et al., 2022]
- Filter stability [Kara Yüksel 2022; McDonald Yüksel 2022; Golowich et al., 2023.]
- Weakly revealing observations [Liu et al, 2022]
- Low covering numbers [Lee, Long, Hsu 2007]
- Low-rank structure [Guo et al, 2023]
- Revealing observation models [Belly et al, 2025]

- Structured policies can be approximately optimal for specific sub-classes of POMDPs

This talk: Revisit a classical class of structured policies.

Special class of policies: Certainty Equivalence

Classical Certainty Equivalence Principle (LQG)

- In LQG systems, the optimal policy has a special structure:
- **Standard interpretation:**
  - Optimal action is linear function of the MMSE estimate
  - Feedback gain equals to that of the deterministic system (obtained replacing random variables by their means)

Simon, "Dynamic programming under uncertainty with a quadratic criterion function," Econometrica 1956.

Special class of policies: Certainty Equivalence

Classical Certainty Equivalence Principle (LQG)

- In LQG systems, the optimal policy has a special structure:
- **Standard interpretation:**
  - Optimal action is linear function of the MMSE estimate
  - Feedback gain equals to that of the deterministic system (obtained replacing random variables by their means)
- **Alternative interpretation:**
  - Feedback gain equals to that of the stochastic perfectly observed system.

Simon, "Dynamic programming under uncertainty with a quadratic criterion function," Econometrica 1956.

Special class of policies: Certainty Equivalence

Classical Certainty Equivalence Principle (LQG)

- In LQG systems, the optimal policy has a special structure:
- **Standard interpretation:**
  - Optimal action is linear function of the MMSE estimate
  - Feedback gain equals to that of the deterministic system (obtained replacing random variables by their means)
- **Alternative interpretation:**
  - Feedback gain equals to that of the stochastic perfectly observed system.

What if model is not LQG?

- CE remains optimal when there is dual effect [Bar-Shalom Tse 1974; Derpich Yuksel 2022]
- Also optimal for some risk sensitive objectives [Whittle 1986]

Simon, "Dynamic programming under uncertainty with a quadratic criterion function," Econometrica 1956.

Certainty equivalence for general POMDPs

POMDP $\mathcal{P}$

- Finite horizon T; State space $\mathcal{S}$, action space $\mathcal{A}$, observation space $\mathcal{Y}$.
- Dynamics P_t, given by $P_t(ds_{t+1}, dy_t \mid s_t, a_t)$
- Per-step cost $c_t \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\|c_t\|_\infty < \infty$.

Certainty equivalence for general POMDPs

POMDP $\mathcal{P}$

- Finite horizon T; State space $\mathcal{S}$, action space $\mathcal{A}$, observation space $\mathcal{Y}$.
- Dynamics P_t, given by $P_t(ds_{t+1}, dy_t \mid s_t, a_t)$
- Per-step cost $c_t \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\|c_t\|_\infty < \infty$.

Auxiliary Fully Observable MDP $\mathcal{M}$

- $\mathcal{M}$ uses the same dynamics and costs as $\mathcal{P}$ but assumes the controller observes S_t
- $\pi^{\mathcal{M}}$: optimal state-feedback policy for $\mathcal{M}$.

Certainty equivalent (CE) Policy

- Uses an arbitrary state estimation function $\mathcal{E}_t \colon \mathcal{H}_t \to \mathcal{S}$
- CE policy: $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}_t(h_t))$

Technical Assumptions

Assumption 1: Measurable Selection

MDP $\mathcal{M}$ satisfies a measurable selection condition which ensures existance of optimal policy $\pi^{\mathcal{M}}$

Technical Assumptions

Assumption 1: Measurable Selection

MDP $\mathcal{M}$ satisfies a measurable selection condition which ensures existance of optimal policy $\pi^{\mathcal{M}}$

Assumption 2: Smoothness

There exist a sequence of concave and non-decreasing functions $F^P_t, F^c_t \colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$, $t \in \{1, \dots, T\}$, such that for any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$:

- **Dynamics:** $d_{\text{Was}}(P_{S,t}(\cdot|s, a), P_{S,t}(\cdot|s', a)) \leqslant F^P_t(d_{\mathcal{S}}(s, s'))$
- **Cost:** $\left|c_t(s, a) - c_t(s', a)\right| \leqslant F^c_t(d_{\mathcal{S}}(s, s'))$

Special case: When F^P_t and F^c_t are linear, this reduces to standard Lipschitz continuity.

Sub-optimality bounds

Quality of estimator

Worst-case conditional expected estimation error η_t:

$$\eta_t := \sup_{h_t} \mathbb{E}[d_S(S_t, \mathcal{E}_t(h_t)) \mid h_t]$$

We assume η_t is bounded.

Sub-optimality bounds

Quality of estimator

Worst-case conditional expected estimation error η_t:

$$\eta_t := \sup_{h_t} \mathbb{E}[d_S(S_t, \mathcal{E}_t(h_t)) \mid h_t]$$

We assume η_t is bounded.

Theorem 1

Define $\varepsilon_t = F^c_t(\eta_t)$ and $\delta_t = F^P_t(\eta_t) + \eta_{t+1}$. Under our assumptions, the CE policy satisfies:

$$W^{\mathcal{P}, \mu^{\mathcal{E}}}_t(h_t) - W^{\mathcal{P}}_t(h_t) \leqslant 2\alpha_t$$

where

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \big[\delta_\tau \mathrm{Lip}(V^{\mathcal{M}}_{\tau+1}) + \varepsilon_{\tau+1}\big], \qquad \text{where } V^{\mathcal{M}}_{\tau+1} \text{ is the opt. value fn. for MDP } \mathcal{M}.$$

Certainty equivalence using state abstraction

State abstraction

- Abstract state space $\tilde{\mathcal{S}}$ with metric $d_{\tilde{\mathcal{S}}}$
- Abstraction function $\phi\colon \mathcal{S} \to \tilde{\mathcal{S}}$ and stochastic kernels $\lambda^P, \lambda^c\colon \tilde{\mathcal{S}} \to \Delta(\mathcal{S})$
- Construct abstract MDP $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{S}}, \mathcal{A}, \{\tilde{P}_t\}_{t=1}^{T-1}, \{\tilde{c}_t\}_{t=1}^{T}, T \rangle$:
  - Dynamics: $\tilde{P}_t(\tilde{S}_{t+1} \in M_{\tilde{\mathcal{S}}} | \tilde{s}_t, a_t) = \int_{\phi^{-1}(\tilde{s}_t)} P_{\mathcal{S},t}\big(\phi(S_{t+1}) \in M_{\tilde{\mathcal{S}}} | s_t, a_t\big) \lambda^P(ds_t|\tilde{s}_t)$
  - Cost: $\tilde{c}_t(\tilde{s}_t, a_t) = \int_{\phi^{-1}(\tilde{s}_t)} c_t(s_t, a_t) \lambda^c(ds_t|\tilde{s}_t)$

- Cost function is a weighted averaging over all states in $\phi^{-1}(\tilde{s}_t)$;

similar interpretation for the dynamics

Certainty equivalence using state abstraction

State abstraction

- Abstract state space $\tilde{\mathcal{S}}$ with metric $d_{\tilde{\mathcal{S}}}$
- Abstraction function $\phi \colon \mathcal{S} \to \tilde{\mathcal{S}}$ and stochastic kernels $\lambda^P, \lambda^c \colon \tilde{\mathcal{S}} \to \Delta(\mathcal{S})$
- Construct abstract MDP $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{S}}, \mathcal{A}, \{\tilde{P}_t\}_{t=1}^{T-1}, \{\tilde{c}_t\}_{t=1}^{T}, T \rangle$:
  - Dynamics: $\tilde{P}_t(\tilde{S}_{t+1} \in M_{\tilde{\mathcal{S}}} | \tilde{s}_t, a_t) = \int_{\phi^{-1}(\tilde{s}_t)} P_{\mathcal{S},t}\big(\phi(S_{t+1}) \in M_{\tilde{\mathcal{S}}} | s_t, a_t\big)\, \lambda^P(ds_t | \tilde{s}_t)$
  - Cost: $\tilde{c}_t(\tilde{s}_t, a_t) = \int_{\phi^{-1}(\tilde{s}_t)} c_t(s_t, a_t) \lambda^c(ds_t | \tilde{s}_t)$

- Cost function is a weighted averaging over all states in $\phi^{-1}(\tilde{s}_t)$;

  similar interpretation for the dynamics

Assumptions

- The model $\tilde{\mathcal{M}}$ satisfies measurable selection
- The model $\tilde{\mathcal{M}}$ is smooth

Sub-optimality bounds for state abstraction

Quality of estimator

Worst-case conditional expected estimation error η_t:

$$\tilde{\eta}_t := \sup_{h_t} \mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t)) \mid h_t]$$

We assume $\tilde{\eta}_t$ is bounded.

Sub-optimality bounds for state abstraction

Quality of estimator

Worst-case conditional expected estimation error η_t:

$$\tilde{\eta}_t := \sup_{h_t} \mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t)) \mid h_t]$$

We assume $\tilde{\eta}_t$ is bounded.

Theorem 2

Define $\tilde{\varepsilon}_t = F_t^c(\tilde{\eta}_t)$ and $\tilde{\delta}_t = F_t^P(\tilde{\eta}_t) + \tilde{\eta}_{t+1}$. Under our assumptions, the CE policy satisfies:

$$W_t^{\mathcal{P}, \mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \leqslant 2\tilde{\alpha}_t$$

where

$$\tilde{\alpha}_t = \tilde{\varepsilon}_t + \sum_{\tau=t}^{T-1} \big[\tilde{\delta}_\tau \mathrm{Lip}(V_{\tau+1}^{\tilde{\mathcal{M}}}) + \tilde{\varepsilon}_{\tau+1}\big],$$ where $V_{\tau+1}^{\tilde{\mathcal{M}}}$ is the opt. value fn. for MDP $\tilde{\mathcal{M}}$.

Some Examples

Example 1: Bounded Observation Noise

System Model

- $\mathcal{Y} = \mathcal{S}$ and $d_{\mathcal{S}}(Y_t, S_t) \leqslant r$.
- $\mathcal{M}$ satisfies measurable selection.
- Dynamics and cost are Lipschitz continuous with Lipschitz constants L_t^P and L_t^c.

Example 1: Bounded Observation Noise

System Model

- $\mathcal{Y} = \mathcal{S}$ and $d_{\mathcal{S}}(Y_t, S_t) \leqslant r$.
- $\mathcal{M}$ satisfies measurable selection.
- Dynamics and cost are Lipschitz continuous with Lipschitz constants L^{P}_{t} and L^{c}_{t}.

Certainty equivalent policy

- $\mathcal{E}_t(h_t) = y_t$
- $\mu^{\mathcal{E}}_t(h_t) = \pi^{\mathcal{M}}_t(y_t)$

Example 1: Bounded Observation Noise

System Model

- $\mathcal{Y} = \mathcal{S}$ and $d_{\mathcal{S}}(Y_t, S_t) \leqslant r$.
- $\mathcal{M}$ satisfies measurable selection.
- Dynamics and cost are Lipschitz continuous with Lipschitz constants L^P_t and L^c_t.

Certainty equivalent policy

- $\mathcal{E}_t(h_t) = y_t$
- $\mu^{\mathcal{E}}_t(h_t) = \pi^{\mathcal{M}}_t(y_t)$

Sub-optimality bound

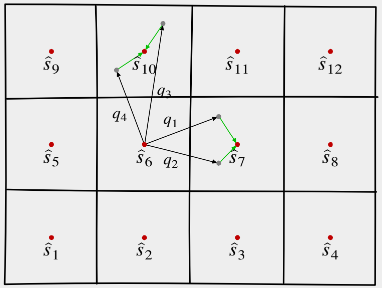
- $\mathbb{E}[d_{\mathcal{S}}(S_t, Y_t) \mid h_t] \leqslant r$. Thus, $\eta_t \leqslant r$.
- $\varepsilon_t \leqslant rL^c_t$ and $\delta_t \leqslant r(1 + L^P_t)$.
- Hence, $W^{\mathcal{P}, \mu^{\mathcal{E}}}_t(h_t) - W^{\mathcal{P}}_t(h_t) \leqslant 2rL_T$ where

$$L_T = \Bigl[L^c_t + \sum_{\tau=t}^{T-1} \bigl[(1 + L^P_\tau)\,\mathrm{Lip}(V^{\mathcal{M}}_{\tau+1}) + L^c_{\tau+1}\bigr]\Bigr]$$

Example 2: Bounded obs noise with state quantization

System Model

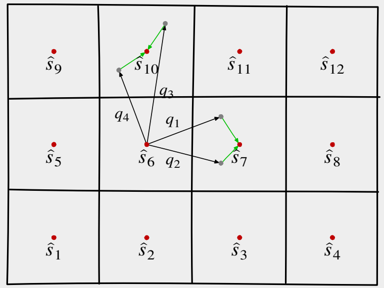
- Same as previous model but $\mathcal{S}$ is so large that we cannot solve MDP $\mathcal{M}$.
- Quantize state space $\mathcal{S}$ into K bins. Quantized state $\tilde{\mathcal{S}} = \{1, \dots, K\}$.
- Quantization function $\phi \colon \mathcal{S} \to \tilde{\mathcal{S}}$.



Example 2: Bounded obs noise with state quantization

System Model

- Same as previous model but $\mathcal{S}$ is so large that we cannot solve MDP $\mathcal{M}$.
- Quantize state space $\mathcal{S}$ into K bins. Quantized state $\tilde{\mathcal{S}} = \{1, \dots, K\}$.
- Quantization function $\phi \colon \mathcal{S} \to \tilde{\mathcal{S}}$.



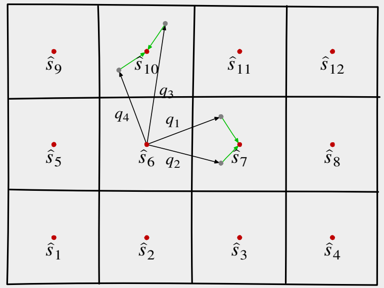
Certainty equivalent policy

- Solve quantized MDP $\tilde{\mathcal{M}}$.
- $\mathcal{E}_t(h_t) = \phi(Y_t)$.
- $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\tilde{\mathcal{M}}}(\phi(Y_t))$

Example 2: Bounded obs noise with state quantization

System Model

- Same as previous model but $\mathcal{S}$ is so large that we cannot solve MDP $\mathcal{M}$.
- Quantize state space $\mathcal{S}$ into K bins. Quantized state $\tilde{\mathcal{S}} = \{1, \dots, K\}$.
- Quantization function $\phi \colon \mathcal{S} \to \tilde{\mathcal{S}}$.



Certainty equivalent policy

- Solve quantized MDP $\tilde{\mathcal{M}}$.
- $\mathcal{E}_t(h_t) = \phi(Y_t)$.
- $\mu^{\mathcal{E}}_t(h_t) = \pi^{\tilde{\mathcal{M}}}_t(\phi(Y_t))$

Sub-optimality bound

- $\tilde{\eta}_t \leqslant \bar{r} := r + 2D$

  where D is diameter of quantization cell.
- Thus,

$$\tilde{\varepsilon}_t \leqslant L^c_t(\bar{r}) \quad \text{and} \quad \tilde{\delta}_t \leqslant L^P_t(\bar{r}) + \bar{r}$$

Example 3: Intermittently degraded observation

System Model

- $\mathcal{Y} = \mathcal{S}$ and $\mathcal{M}$ satisfies measurable selection.
- Observation is either bad (with prob. p) or good.
- **Good obs:** $d_{\mathcal{S}}(Y_t, S_t) \leqslant r$.
- **Bad obs:** $d_{\mathcal{S}}(Y_t, S_t) \leqslant R$, where $R > r$.
- Dynamics and cost are Lipschitz continuous

Example 3: Intermittently degraded observation

System Model

- $\mathcal{Y} = \mathcal{S}$ and $\mathcal{M}$ satisfies measurable selection.
- Observation is either bad (with prob. p) or good.
- **Good obs:** $d_{\mathcal{S}}(Y_t, S_t) \leqslant r$.
- **Bad obs:** $d_{\mathcal{S}}(Y_t, S_t) \leqslant R$, where $R > r$.
- Dynamics and cost are Lipschitz continuous

Certainty equivalent policy

- $\mathcal{E}_t(h_t) = y_t$
- $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t)$

Example 3: Intermittently degraded observation

System Model

- $\mathcal{Y} = \mathcal{S}$ and $\mathcal{M}$ satisfies measurable selection.
- Observation is either bad (with prob. p) or good.
- **Good obs:** $d_{\mathcal{S}}(Y_t, S_t) \leqslant r$.
- **Bad obs:** $d_{\mathcal{S}}(Y_t, S_t) \leqslant R$, where $R > r$.
- Dynamics and cost are Lipschitz continuous

Certainty equivalent policy

- $\mathcal{E}_t(h_t) = y_t$
- $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t)$

Sub-optimality bound

- $\mathbb{E}[d_{\mathcal{S}}(S_t, Y_t) \mid h_t] \leqslant (1-p)r + pR$. Thus, $\eta_t \leqslant (1-p)r + pR$.
- $\varepsilon_t \leqslant [(1-p)r + pR]\, L_t^c$ and $\delta_t \leqslant [(1-p)r + pR](1 + L_t^P)$.
- Hence, $W_t^{\mathcal{P}, \mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \leqslant 2[(1-p)r + pR]\, L_T$

Example 4: Certainty equivalence in adaptive control

System Model

- Parameterized MDP $\mathcal{M}_X(\theta)$, $\theta \in \Theta$, with state space $\mathcal{X}$, action space $\mathcal{A}$.
- Dynamics $P_{X,\theta}$ and per-step cost ℓ_θ. Assumed to be Lipschitz continuous.

- POMDP with state (X_t, θ), observation $(X_t, \ell_\theta(X_{t-1}, A_{t-1}))$
- Corresponding MDP $\mathcal{M} = \mathcal{M}_X(\theta)$.

Example 4: Certainty equivalence in adaptive control

System Model

- Parameterized MDP $\mathcal{M}_X(\theta)$, $\theta \in \Theta$, with state space $\mathcal{X}$, action space $\mathcal{A}$.
- Dynamics $P_{X,\theta}$ and per-step cost ℓ_θ. Assumed to be Lipschitz continuous.

- POMDP with state (X_t, θ), observation $(X_t, \ell_\theta(X_{t-1}, A_{t-1}))$
- Corresponding MDP $\mathcal{M} = \mathcal{M}_X(\theta)$.

Certainty equivalent policy

- Let $\hat{\theta}_t$ be any estimator of θ.
- $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(x_t, \hat{\theta}_t) = \pi_t^{\mathcal{M}_X(\hat{\theta}_t)}(x_t)$

Example 4: Certainty equivalence in adaptive control

System Model

- Parameterized MDP $\mathcal{M}_X(\theta)$, $\theta \in \Theta$, with state space $\mathcal{X}$, action space $\mathcal{A}$.
- Dynamics $P_{X,\theta}$ and per-step cost ℓ_θ. Assumed to be Lipschitz continuous.

- POMDP with state (X_t, θ), observation $(X_t, \ell_\theta(X_{t-1}, A_{t-1}))$
- Corresponding MDP $\mathcal{M} = \mathcal{M}_X(\theta)$.

Certainty equivalent policy

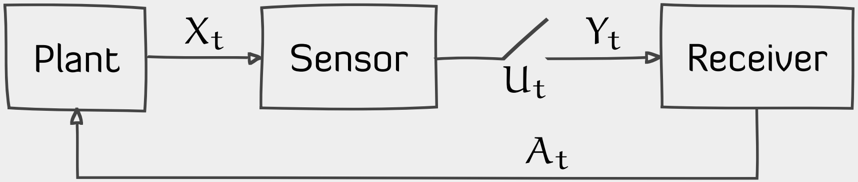
- Let $\hat{\theta}_t$ be any estimator of θ.
- $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(x_t, \hat{\theta}_t) = \pi_t^{\mathcal{M}_X(\hat{\theta}_t)}(x_t)$

Sub-optimality bound

- $\eta_t = \sup_{h_t} \mathbb{E}[d_\Theta(\theta, \hat{\theta}_t) \mid h_t]$.
- Thus, $\varepsilon_t \leqslant L^c \eta_t$ and $\delta_t \leqslant L^P \eta_t + \eta_{t+1}$.

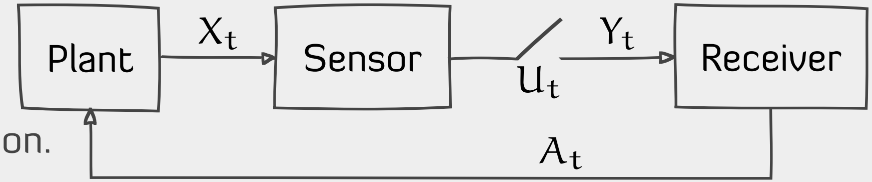
- If η_t decays sufficiently fast, we can obtain upper bounds on performance loss even as $T \to \infty$.

Example 5: Remote estimation with event-triggered comm



Example 5: Remote estimation with event-triggered comm

Event-triggered communication



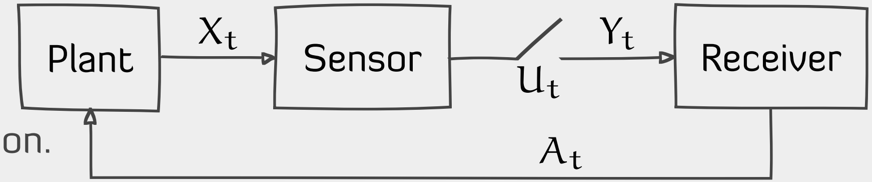
- Let $g \colon \mathcal{X} \times \mathcal{A} \to \mathcal{X}$ is a pre-specified function.
- The remote controller generates an estimate

$$\hat{X}_{t|t-1} = g(X_{t-1|t-1}, A_{t-1}) \quad \text{and} \quad \hat{X}_{t|t} = \begin{cases} Y_t & \text{if } Y_t = \mathfrak{E} \\ \hat{X}_{t|t-1} & \text{otherwise} \end{cases}$$

- **Event-triggered communication:** Communicate if $d_{\mathcal{X}}(X_t, \hat{X}_{t|t-1}) > r$.

Example 5: Remote estimation with event-triggered comm

Event-triggered communication



- Let $g \colon \mathcal{X} \times \mathcal{A} \to \mathcal{X}$ is a pre-specified function.
- The remote controller generates an estimate

$$\hat{X}_{t|t-1} = g(X_{t-1|t-1}, A_{t-1}) \quad \text{and} \quad \hat{X}_{t|t} = \begin{cases} Y_t & \text{if } Y_t = \mathfrak{E} \\ \hat{X}_{t|t-1} & \text{otherwise} \end{cases}$$

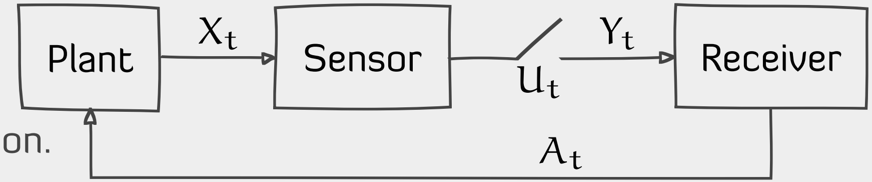
- **Event-triggered communication:** Communicate if $d_{\mathcal{X}}(X_t, \hat{X}_{t|t-1}) > r$.

Certainty equivalent policy

- POMDP with $S_t = (X_t, \hat{X}_{t|t-1})$ and obs. Y_t.
- State estimate $\mathcal{E}_t(h_t) = (\hat{x}_{t|t}, \hat{x}_{t|t-1})$.
- $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(\hat{x}_{t|t}, \hat{x}_{t|t-1}) = \pi_t^{\mathcal{M}X}(\hat{x}_{t|t})$.

Example 5: Remote estimation with event-triggered comm

Event-triggered communication



- Let $g\colon \mathcal{X} \times \mathcal{A} \to \mathcal{X}$ is a pre-specified function.
- The remote controller generates an estimate

$$\hat{X}_{t|t-1} = g(X_{t-1|t-1}, A_{t-1}) \quad \text{and} \quad \hat{X}_{t|t} = \begin{cases} Y_t & \text{if } Y_t = \mathfrak{E} \\ \hat{X}_{t|t-1} & \text{otherwise} \end{cases}$$

- **Event-triggered communication:** Communicate if $d_{\mathcal{X}}(X_t, \hat{X}_{t|t-1}) > r$.

Certainty equivalent policy

- POMDP with $S_t = (X_t, \hat{X}_{t|t-1})$ and obs. Y_t.
- State estimate $\mathcal{E}_t(h_t) = (\hat{x}_{t|t}, \hat{x}_{t|t-1})$.
- $\mu^{\mathcal{E}}_t(h_t) = \pi^{\mathcal{M}}_t(\hat{x}_{t|t}, \hat{x}_{t|t-1}) = \pi^{\mathcal{M}_X}_t(\hat{x}_{t|t})$.

Sub-optimality bound

- $\mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(h_t)) \mid h_t] \leqslant r$. Thus, $\eta_t \leqslant r$.
- Hence,

$$\varepsilon_t \leqslant F^c_t(r) \quad \text{and} \quad \delta_t \leqslant F^P_t(r) + r$$

Example 6: Non-homogeneous multi-particle systems

System Model

- n particles, state of particle $X^i_t \in \mathcal{X}$
- Global state $X_t = (X^1_t, \dots, X^n_t)$
- Global observation $Y_t = X_t + N_t$
- Weighted mean-field: $M_t = \sum_{i=1}^{n} \alpha^i X^i_t$
- Dynamics: $X^i_{t+1} = \bar{f}(M_t, A_t, W_t) + f^i(X_t, A_t, W_t)$
- Cost: $c(X_t, A_t) = \bar{\ell}(M_t, A_t) + \sum_{i=1}^{n} \alpha^i \ell^i(X_t, A_t)$

Example 6: Non-homogeneous multi-particle systems

System Model

- n particles, state of particle $X^i_t \in \mathcal{X}$
- Global state $X_t = (X^1_t, \dots, X^n_t)$
- Global observation $Y_t = X_t + N_t$

- Weighted mean-field: $M_t = \sum_{i=1}^n \alpha^i X^i_t$

- Dynamics: $X^i_{t+1} = \bar{f}(M_t, A_t, W_t) + f^i(X_t, A_t, W_t)$
- Cost: $c(X_t, A_t) = \bar{\ell}(M_t, A_t) + \sum_{i=1}^n \alpha^i \ell^i(X_t, A_t)$

Assumptions

- $\bar{f}$ and $\bar{\ell}$ are Lipschitz continuous.
- f^i and ℓ^i are small (in sup-norm)

Example 6: Non-homogeneous multi-particle systems

System Model

- n particles, state of particle $X^i_t \in \mathcal{X}$
- Global state $X_t = (X^1_t, \dots, X^n_t)$
- Global observation $Y_t = X_t + N_t$

- Weighted mean-field: $M_t = \sum_{i=1}^n \alpha^i X^i_t$

- Dynamics: $X^i_{t+1} = \bar{f}(M_t, A_t, W_t) + f^i(X_t, A_t, W_t)$
- Cost: $c(X_t, A_t) = \bar{\ell}(M_t, A_t) + \sum_{i=1}^n \alpha^i \ell^i(X_t, A_t)$

Assumptions

- $\bar{f}$ and $\bar{\ell}$ are Lipschitz continuous.
- f^i and ℓ^i are small (in sup-norm)

Certainty equivalent policy and sub-optimality bounds

- $\mathcal{E}_t(h_t) = \sum_{i=1}^n \alpha^i Y^i_t$
- $\mu^{\mathcal{E}}_t(h_t) = \pi^{\mathcal{M}}_t(\mathcal{E}_t(h_t))$

- $\tilde{\varepsilon}_t \leqslant L^{\bar{\ell}} \sum^{i=1}{}^n \alpha^i r^i + 2\beta$ and $\tilde{\delta}_t \leqslant L^{\bar{f}} \sum^{i=1}{}^n \alpha^i r^i + 2 \sum_{i=1}^n \alpha^i \gamma^i$

Proof Outline

Approximate Information State

Given a sequence $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)$ and $\delta = (\delta_1, \dots, \delta_T)$, a process $\{Z_t\}_{t=1}^{T}$ is an (ε,δ)**-approximate information state (AIS)** if there exists

- History compression functions $\sigma_t^{\mathsf{AIS}} \colon \mathcal{H}_t \to \mathcal{Z}$
- Cost approximation functions $c_t^{\mathsf{AIS}} \colon \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$
- Dynamics approximation functions $P_t^{\mathsf{AIS}} \colon \mathcal{Z} \times \mathcal{A} \to \mathcal{Z}$

Subramanian, Sinha, Seraj, Mahajan, "Approximate Information State for approximate planning and learning . . .", JMLR 2022.

Approximate Information State

Given a sequence $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)$ and $\delta = (\delta_1, \dots, \delta_T)$, a process $\{Z_t\}_{t=1}^{T}$ is an **(ε,δ)-approximate information state (AIS)** if there exists

- History compression functions $\sigma_t^{\text{AIS}} \colon \mathcal{H}_t \to \mathcal{Z}$
- Cost approximation functions $c_t^{\text{AIS}} \colon \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$
- Dynamics approximation functions $P_t^{\text{AIS}} \colon \mathcal{Z} \times \mathcal{A} \to \mathcal{Z}$

such that

- $\left| \mathbb{E}[c_t(S_t, a_t) \mid h_t] - c_t^{\text{AIS}}(\sigma_t^{\text{AIS}}(h_t), a_t) \right| \leqslant \varepsilon_t$
- $d_{\text{Was}}(\nu_t(\cdot \mid h_t, a_t), P_t^{\text{AIS}}(\cdot \mid z_t, a_t)) \leqslant \delta_t$, where $\nu_t(M_Z \mid h_t, a_t) = \mathbb{P}(Z_{t+1} \in M_Z \mid h_t, a_t)$.
- MDP $\langle \mathcal{Z}, \mathcal{A}, P^{\text{AIS}}, c^{\text{AIS}} \rangle$ satisfies measurable selection.

Subramanian, Sinha, Seraj, Mahajan, "Approximate Information State for approximate planning and learning . . .", JMLR 2022.

AIS Approximation Bound

AIS Dynamic Program

Let $\{Z_t\}_{t=1}^{T}$ be an (ε, δ)-AIS. Define:

$$V_t^{\mathsf{AIS}}(z_t) = \min_{a \in \mathcal{A}} \left\{ c_t^{\mathsf{AIS}}(z_t, a) + \int_{\mathcal{Z}} P_t^{\mathsf{AIS}}(dz' | z_t, a) V_{t+1}^{\mathsf{AIS}}(z') \right\}.$$

Let μ_t^{AIS} be the corresponding arg min policy.

Subramanian, Sinha, Seraj, Mahajan, "Approximate Information State for approximate planning and learning . . .", JMLR 2022.

AIS Approximation Bound

AIS Dynamic Program

Let $\{Z_t\}_{t=1}^{T}$ be an (ε, δ)-AIS. Define:

$$V_t^{\mathsf{AIS}}(z_t) = \min_{a \in \mathcal{A}} \left\{ c_t^{\mathsf{AIS}}(z_t, a) + \int_{\mathcal{Z}} P_t^{\mathsf{AIS}}(dz' | z_t, a) \, V_{t+1}^{\mathsf{AIS}}(z') \right\}.$$

Let μ_t^{AIS} be the corresponding arg min policy.

AIS Policy

Define the policy μ^{AIS} for POMDP $\mathcal{P}$ as

$$\mu_t^{\mathsf{AIS}}(h_t) = \pi_t^{\mathsf{AIS}}(\sigma_t^{\mathsf{AIS}}(h_t))$$

Subramanian, Sinha, Seraj, Mahajan, "Approximate Information State for approximate planning and learning ...", JMLR 2022.

AIS Approximation Bound

AIS Dynamic Program

Let $\{Z_t\}_{t=1}^{T}$ be an (ε, δ)-AIS. Define:

$$V_t^{\mathsf{AIS}}(z_t) = \min_{a \in \mathcal{A}} \left\{ c_t^{\mathsf{AIS}}(z_t, a) + \int_{\mathcal{Z}} P_t^{\mathsf{AIS}}(dz' | z_t, a) \, V_{t+1}^{\mathsf{AIS}}(z') \right\}.$$

Let μ_t^{AIS} be the corresponding arg min policy.

AIS Policy

Define the policy μ^{AIS} for POMDP $\mathcal{P}$ as

$$\mu_t^{\mathsf{AIS}}(h_t) = \pi_t^{\mathsf{AIS}}(\sigma_t^{\mathsf{AIS}}(h_t))$$

AIS Approximation Bound

$$W_t^{\mathcal{P}, \mu^{\mathsf{AIS}}}(h_t) - W_t^{\mathcal{P}}(h_t) \leqslant 2\alpha$$

where

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \left[\delta_\tau \mathrm{Lip}(V_{\tau+1}^{\mathsf{AIS}}) + \varepsilon_{\tau+1} \right]$$

Subramanian, Sinha, Seraj, Mahajan, "Approximate Information State for approximate planning and learning . . .", JMLR 2022.

Proof Outline

Show that CE policy is an AIS

Under smoothness assumptions:

- $\Big|\mathbb{E}[c_t(S_t, a_t)|h_t] - \tilde{c}_t(\mathcal{E}_t(h_t), a_t)\Big| \leqslant F^c_t(\eta_t).$
- $d_{\text{Was}}(\hat{\psi}_t(\cdot|h_t, a_t), \tilde{P}_t(\cdot|\mathcal{E}_t(h_t), a_t)) \leqslant F^P_t(\eta_t) + \eta_{t+1}$

  where

$$\hat{\psi}_t(M_{\tilde{S}}|h_t, a_t) = \mathbb{P}(\mathcal{E}_{t+1}(H_{t+1}) \in M_{\tilde{S}}|h_t, a_t)$$

$$\tilde{P}_t(M_{\tilde{S}}|\tilde{s}_t, a_t) = \mathbb{P}(\tilde{S}_{t+1} \in M_{\tilde{S}}|\tilde{s}_t, a_t)$$

- The main result (Theorem 2) follows from the AIS bounds.

Conclusion

- CE policies are practical and attractive for non-LQG settings.
- Results agree with engineering intuition: the sub-optimality of CE policies depends on the quality of the estimator and smoothness of the model.
- The approximation bounds are based on AIS theory.
- CE policies are not appropriate for all models: for instance, if the agent has an option to pay a cost to sense the true state of the MDP, a CE policy will never choose the sensing action.

- email: aditya.mahajan@mcgill.ca
- web: https://adityam.github.io

Thank you