IMPERIAL

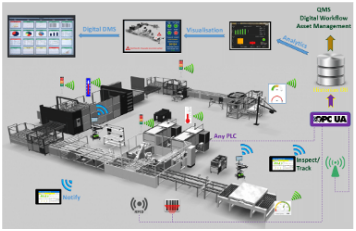
Actions Speak Louder Than Words: Rate-reward Tradeoff in Markov Decision Processes

Haotian Wu, Gongpu Chen, Deniz Gündüz
Imperial College London

ICLR 2025, Singapore

Motivation

Communication plays a vital role in decision-making systems



Smart Factory

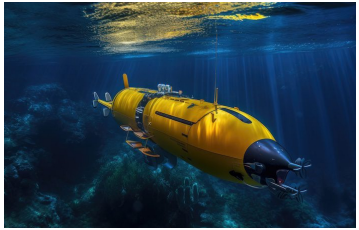
Autonomous Driving

Multi-agent Systems

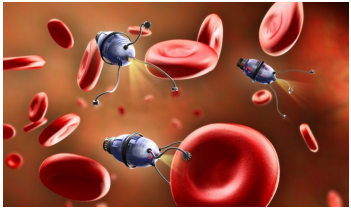
Dedicated channels are needed for explicit communication (5G, WiFi, Bluetooth · · ·)

Motivation

Dedicated communication channels are not available in many scenarios



Underwater environment
Wireless communications in deep water environments are challenging due to signal attenuation



Nano-robots
Electromagnetic communication is not feasible due to size and energy limitations

Battlefield
Wireless communications are highly susceptible to malicious attacks, including eavesdropping and channel jamming.

How can agents communicate without dedicated channels?

Motivation

Implicit communications are abundant in nature

Waggle dance of bees
Honeybees communicate the location, distance, and quality of foods to their hive mates through a series of specific movements.

Starling murmurations
Starlings communicate with nearby birds via their position and movement to adjust their flight path.

Fish schooling
By observing the movements of nearby individuals, fish can adjust their own speed and direction to match the group.

The MDP environment can be viewed as a communication channel

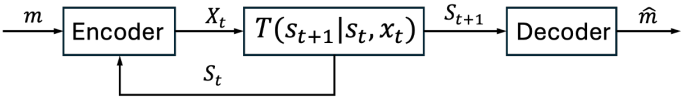
The procedure of MDP control:

- At time t, the environment is in state $s_t \in \mathcal{S}$
- The controller takes an action $x_t \in \mathcal{X}$ according to s_t
- The environment state transitions to s_{t+1} and generates a reward r_t

The MDP environment can be viewed as a communication channel

A control step v.s. A channel use

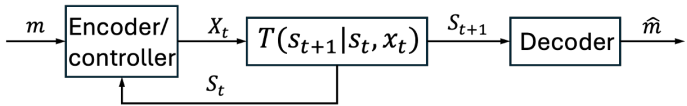
- At time t, the environment is in state $s_t \in \mathcal{S}$
  - s_t corresponds to the channel state at time t
- The controller takes an action $x_t \in \mathcal{X}$ according to s_t
  - x_t corresponds to the channel input at time t
- The environment state transitions to s_{t+1} and generates a reward r_t
  - s_{t+1} corresponds to the channel output at time t
  - the current output is the next channel state



Channel law: $P(s_{t+1}|x^t, s^t) = P_{S^+|X,S}(s_{t+1}|x_t, s_t) = T(s_{t+1}|x_t, s_t)$

How to use the channel?

- **Encoder**: select action via $\mathcal{E}_t : \mathcal{M} \times \mathcal{S}^t \times \mathcal{X}^{t-1} \to \mathcal{X}$
  - encoder and controller represent the two aspects of the same entity (action selection)
  - an encoding policy can be viewed as a history-dependent control policy
- **Decoder**: decode the message via $\mathcal{D} : \mathcal{S}^n \to \mathcal{M}$



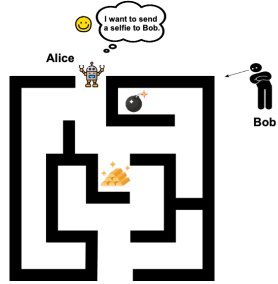
The action-state channel

Messages are encoded into the action sequence and decoded from the state sequence

Message encoding may reduce the control reward

System Model

Integrated control and coding



- an agent interacts with an MDP $(\mathcal{S}, \mathcal{X}, \mathsf{r}, \mathsf{T})$
- a passive receiver that can observe the MDP states
  - the receiver does not affect the MDP control
- the agent has two objectives:
  1. Control: Maximize rewards from the MDP environment
  2. Communication: Communicate messages to the receiver.

Control performance: long-term average reward G

Communication performance: transmission rate R and probability of error $\mathsf{P}_{\mathsf{e}}^{(n)}$

Integrated control and communication:

$$\max_{\mathcal{E},\mathcal{D}} \ \mathsf{R}_{\mathcal{E},\mathcal{D}} \tag{1}$$

$$\text{s.t. } \mathsf{G}_{\mathcal{E}} \geq \mathsf{V} \quad \text{and} \quad \mathsf{P}_{\mathsf{e}}^{(n)} \leq \sigma. \tag{2}$$

What's the fundamental tradeoff between MDP reward and channel capacity?

Reward-capacity Tradeoff

Capacity of the action-state channel without constraint

Theorem

The capacity of the action-state channel without reward constraint is given by

$$C = \max_{\{\pi(x|s):x\in\mathcal{X},s\in\mathcal{S}\}} I(X;S^+|S)$$

where X, S, and S^+ follow a joint distribution given by

$$p(x,s^+,s) = \rho_\pi(s)\pi(x|s)T(s^+|s,x),\ x\in\mathcal{X}, s, s^+\in\mathcal{S}.$$

Notations: $I(X;S^+|S)$ is the conditional mutual information, $\pi(x|s)$ denotes the probability of selection action x in state s, ρ_π denotes the stationary state distribution under policy π.

Insights: the capacity of the action-state channel can be achieved by encoding messages into a stationary randomized policy for the MDP, without relying on historical information.

Reward-capacity Tradeoff

Capacity of the action-state channel with constraint

Theorem

The capacity of the action-state channel with reward constraint V is the optimal value of the following convex optimization problem:

$$C(V) := \max_{w \in \mathcal{W}} I(w, T)$$
$$\text{s.t.} \sum_{s \in \mathcal{S}} \sum_{x \in \mathcal{X}} w(s,x) r(s,x) \geq V$$

where $I(w,T)$ is a concave function of $w \in \mathcal{W}$ defined as

$$I(w,T) \triangleq \sum_{s \in \mathcal{S}} \sum_{x \in \mathcal{X}} w(s,x) \sum_{s' \in \mathcal{S}} T(s'|s,x) \log \frac{T(s'|s,x) \sum_{x''} w(s,x'')}{\sum_{x'} T(s'|s,x') w(s,x')}.$$

In addition, $C(V)$ is a concave function.

The optimal input distribution $\pi^*(x|s)$ can NOT be used as a practical coding policy!

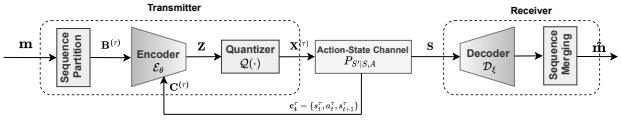
- From the control perspective, π^* is a stationary policy:

$$\pi^*(x|s) = \frac{w(s,x)}{\sum_{x'} w(s,x')}$$

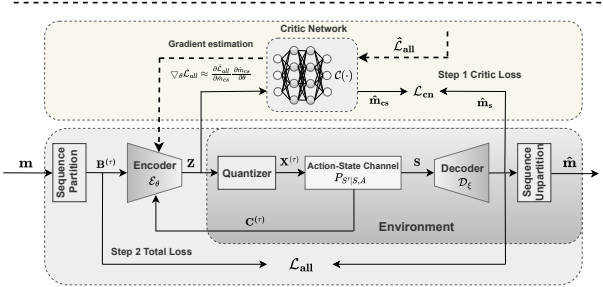
- A coding policy selects actions according to state s and message m
- Need to translate the control policy π^* to a coding policy

A practical coding scheme

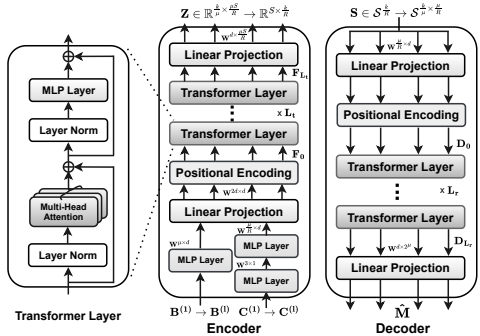
Act2Comm: a transformer-based coding scheme



(a). Model workflow when inference



(b). Two-step training strategy



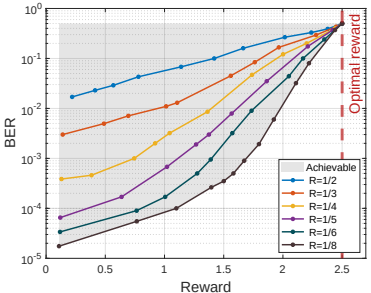
(c). The architecture and parameters of the proposed encoder and decoder.

Experiments

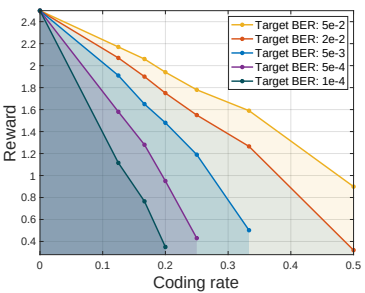
Group 1: Lucky Wheel

MDP environment:

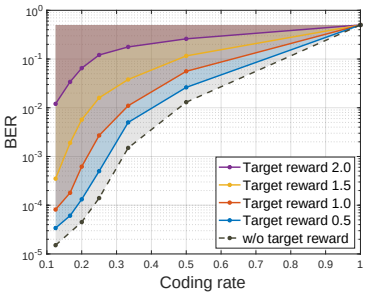
- 3 states
- 2 actions



BER v.s. Reward.



Rate v.s. Reward.



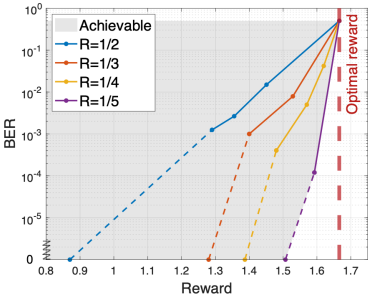
Rate v.s. BER.

Experiments

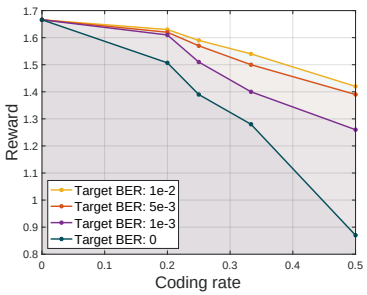
Group 2: Catch the Ball

MDP environment:

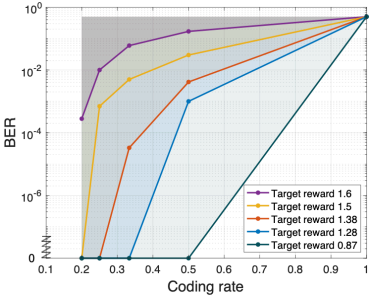
- 27 states
- 3 actions



BER v.s. Reward.



Rate v.s. Reward.



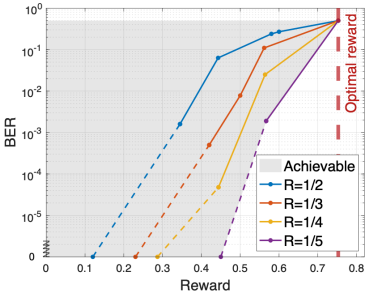
Rate v.s. BER.

Experiments

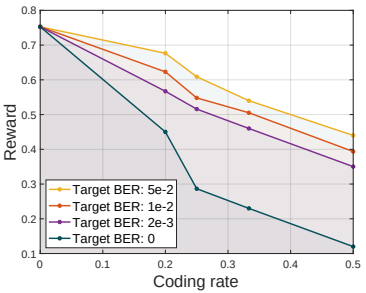
Group 3: Erratic Robot

MDP environment:

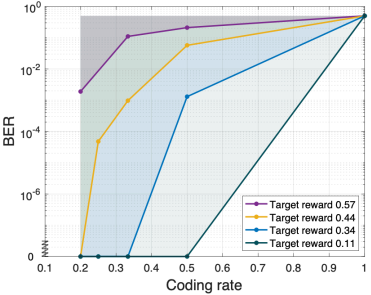
- 16 states
- 5 actions



BER v.s. Reward.



Rate v.s. Reward.



Rate v.s. BER.

IMPERIAL

Thank you

Actions Speak Louder Than Words: Rate-reward Tradeoff in Markov Decision Processes
ICLR 2025, Singapore