A Memory-Based Reinforcement Learning Approach to Integrated Sensing and Communication

Homa Nikbakht[1], Michèle Wigger[2], Shlomo Shamai (Shitz)[3], and H. Vincent Poor[1]
[1]Princeton University, [2]LTCI, Télécom Paris, IP Paris, [3]Technion
{homa, poor}@princeton.edu, michele.wigger@telecom-paris.fr, sshlomo@ee.technion.ac.il

2024 58th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-5405-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEEECONF60004.2024.10942618

***Abstract*—In this paper, we consider a point-to-point integrated sensing and communication (ISAC) system, where a transmitter conveys a message to a receiver over a channel with memory and simultaneously estimates the state of the channel through the backscattered signals from the emitted waveform. Using Massey's concept of directed information for channels with memory, we formulate the capacity-distortion tradeoff for the ISAC problem when sensing is performed in an online fashion. Optimizing the transmit waveform for this system to simultaneously achieve good communication and sensing performance is a complicated task, and thus we propose a deep reinforcement learning (RL) approach to find a solution. The proposed approach enables the agent to optimize the ISAC performance by learning a reward that reflects the difference between the communication gain and the sensing loss. Since the state-space in our RL model is à priori unbounded, we employ deep deterministic policy gradient algorithm (DDPG). Our numerical results suggest a significant performance improvement when one considers unbounded state-space as opposed to a simpler RL problem with reduced state-space. In the extreme case of degenerate state-space only memoryless signaling strategies are possible. Our results thus emphasize the necessity of well exploiting the memory inherent in ISAC systems.**

I. Introduction

Integrating sensing and communication (ISAC) into a single system is motivated by reducing hardware costs, bandwidth usage, and power consumption. It is enabled by several features anticipated for 6G communication systems: higher frequency bands (from mmWave up to THz), wider bandwidths and denser distributions of massive antenna arrays [1]–[5].

Recently, deep learning technology has demonstrated its capability in various wireless communication applications such as channel estimation, signal detection, and resource allocation [6], [7]. Motivated by this, some recent studies have focused on enhancing the performance of an ISAC system using deep reinforcement learning approaches in different model-based and model-free settings and for a wide range of applications [8]–[10].

In this paper, we propose a reinforcement learning (RL) approach to study fundamental limits of ISAC systems with memory by adopting a deep deterministic policy gradient (DDPG) algorithm [11] where an agent simultaneously takes sensing and communication actions to optimize the ISAC performance. More specifically, we use Massey's concept of directed information for channels with memory [12] and formulate the capacity-distortion trade-off under an online-sensing framework. This formulation includes an optimization problem where the objective is to optimize the transmit waveform so as to simultaneously achieve good communication and sensing performance. We simplify this optimization problem for the class of *unifilar channels* [13], and reformulate the problem in this special case as a Markov decision process (MDP). Finally, using a DDPG algorithm, we numerically evaluate the capacity-distortion trade-off for a specific example. Our numerical results suggest a significant performance improvement if our RL approach is applied to the full model with unbounded state-space as opposed to a restricted and simplified model with limited state-space. In the extreme case of degenerate state-space, the RL formulation only allows for memoryless signaling strategies and is highly suboptimal.

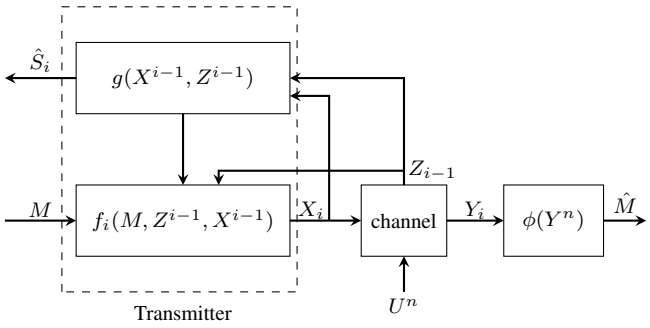

Fig. 1: system model

The capacity-distortion trade-off of ISAC systems has been studied both in the asymptotic [14]–[16] and finite blocklength [17] regimes. However, while most works focus on memoryless ISAC channels, [16] considered a very general model with memory similar to the one in this work. The difference between our work and [16] lies in our focus on online-estimators that sense the targets in an online manner and not just at the end of the communication. Moreover, we managed to simplify the general complicated expression for the class of unifilar channels.

II. Problem Setup

Consider the point-to-point setup in Figure 1 where a dual function ISAC transmitter (Tx) wishes to communicate the message $M \in [1 : 2^{n\mathsf{R}}]$ to a receiver (Rx), where R denotes the rate of communication and n the blocklength of transmission. At the same time, the Tx also collects backscattering signals to estimate sensing parameters of the system.

More specifically, at each discrete-time $i \in \{1, \dots, n\}$, the Tx emits a channel input to the system, and the environment creates two outputs: the receive signal Y_i observed at the Rx and the backscattered signal Z_i observed at the Tx. Both signals depend

on an internal state sequence $U_1, \ldots, U_n$ of the environment. In practical contexts, this state sequence can model obstacles, fading phenomena, or also velocities or directions of vehicles in the neighborhood of the Tx or the Rx. We formalize a very general model with channel transition law

$$P_{Z_i Y_i | X^i U^i Z^{i-1} Y^{i-1}}, \quad i \in \{1, \ldots, n\}, \tag{1}$$

in function of the current and past inputs and states and the past outputs and backscatterers. Both the channel from the Tx to the Rx (also called communication channel) as well as the channel from the Tx to itself (also called sensing or radar channel) can thus have arbitrary memory and arbitrarily depend on the past state symbols U^i. The two channels can either be dependent or independent, depending on the specific choice of channel laws $\{P_{Z_i Y_i | X^i U^i Z^{i-1} Y^{i-1}}\}$ that one considers.

The Tx can compute its channel inputs in an interactive way, depending on the previously observed backscatterers. Thus, at a given time $i \in \{1, \ldots, n\}$, the Tx produces the input X_i as

$$X_i = f_i(M, Z^{i-1}, X^{i-1}) \tag{2}$$

using some encoding function f_i on appropriate domains.

In terms of sensing, the Tx is interested in estimating a given target S_i which can depend on the channel's internal state U_i and inputs/outputs. The Tx produces a time-i state estimate $\hat{S}_i$ in an online manner based on its previous observations:

$$\hat{S}_i = g_i(X^i, Z^i). \tag{3}$$

Sensing performance is measured by average block distortion:

$$\Delta^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[d(S_i, \hat{S}_i)], \tag{4}$$

where $d(\cdot, \cdot)$ is a given bounded per-symbol distortion function.

The optimal estimator $g_i^*(\cdot, \cdot)$ is easily obtained from the target distribution $P_{S_i | X^i, Z^i}$ implicitly defined by the channel. We wish to set:

$$\hat{S}_i := \operatorname*{argmin}_{\hat{s}} \sum_{s} P_{S_i | X^i Z^i}(s_i | x^i, z^i) d(s_i, \hat{s}). \tag{5}$$

The receiver waits until it observes all its n channel outputs Y^n and then decodes message M as

$$\hat{M} = \psi(Y^n), \tag{6}$$

using a well-chosen decoding function ϕ that acts on appropriate domains. (The Rx thus does not have explicit knowledge of the state à priori, only what it learns from its observed outputs.)

In an information-theoretic tradition, communication performance is measured by the rate R that allows to drive the Rx's error probability $\epsilon^{(n)} := \mathbb{P}[\hat{M} \neq M]$ to 0 asymptotically. This is formalized in the following section.

III. Capacity-Distortion Tradeoff

Definition 1: A rate-distortion pair (R, D) is said to be achievable if there exists a sequence (in n) of $(2^{n\mathsf{R}}, n)$ codes and encoding, estimation and decoding functions $f_1, \ldots, f_n, g_1, \ldots . g_n, \psi$ that simultaneously satisfy

$$\lim_{n \to \infty} \epsilon^{(n)} = 0, \tag{7}$$

$$\lim_{n \to \infty} \Delta^{(n)} \leq \mathsf{D}. \tag{8}$$

The capacity-distortion trade-off $\mathsf{C}(\mathsf{D})$ is the largest rate R such that the rate-distortion tuple (R, D) is achievable.

The capacity-distortion trade-off $\mathsf{C}(\mathsf{D})$ of our model can be obtained following similar steps to [16], see the following Proposition 1. The only difference between the model here and the one in [16] lies in the way the transmitter estimates the state sequence. While in [16], the state estimation is performed *at the very end* of the communication, here we impose *online estimators* where the i-th state symbol has to be estimated at the same time as producing the i-th channel inputs.

Proposition 1: The capacity-distortion trade-off $\mathsf{C}(\mathsf{D})$ is

$$\mathsf{C}(\mathsf{D}) = \lim_{n \to \infty} \frac{1}{n} \sup_{\{P_{X_i | X^{i-1} Z^{i-1}}\}_{i=1}^n} \sum_{i=1}^{n} I(X^i; Y_i | Z^{i-1}), \tag{9}$$

$$\text{subject to} \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[d(S_i, g_i^*(X^i, Z^i))] \leq \mathsf{D}, \tag{10}$$

where $g_i^*(\cdot, \cdot)$ is the argmin-estimator in (5).

Above formula for the capacity-distortion trade-off is difficult to evaluate due to the limit $n \to \infty$ and the supremum over the conditional laws. It simplifies for certain classes of channels, such as obviously memoryless channels or the set of *unifilar channels* [13] on which we shall focus in this article.

Definition 2: Consider perfect feedback, i.e., $Y = Z$. A state-dependent channel is called *unifilar* if

$$P(y_i | x^i, y^{i-1}, u^i) = P(y_i | x_i, u_i) \tag{11a}$$

and

$$u_i = \phi(x_i, u_{i-1}, y_i) \tag{11b}$$

for a given state-transition function $\phi(\cdot)$ on appropriate domains.

Theorem 1: Given D the capacity-distortion trade-off of a connected unifilar channel, where the initial state s_0 is available to both the encoder and decoder, is given by the following optimization problem:

$$\mathsf{C}(\mathsf{D}) = \lim_{n \to \infty} \max_{\{P_{X_i | U_{i-1} Y^{i-1}}\}_{i=1}^n} \frac{1}{n} \sum_{i=1}^{n} I(X_i, U_{i-1}; Y_i | Y^{i-1}) \tag{12}$$

subject to

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{y^i, x^i} \min_{\hat{s}} \sum_{s} P_{S_i | X^i Y^i}(s_i | x^i, y^i) d(s_i, \hat{s}) \leq \mathsf{D}. \tag{13}$$

Proof: The term in (12) is equivalent to the capacity of a unifilar channel and has been proved in [18, Theorem 1]. The condition in (5) stems from the optimal estimator in (5). ∎

IV. Reinforcement Learning Approach to ISAC

To evaluate the capacity-distortion trade-off of the proposed ISAC system, we require to solve a complex multi-letter optimization problem, see Theorem 1. Our approach is to first present a Markov decision process (MDP) formulation of this optimization problem. We then employ RL where we model the Tx with an agent performing both sensing and communication tasks. For this purpose, we use the DDPG algorithm [11] which is an actor-critic model-free RL algorithm that operates over continuous action spaces and is of deterministic gradient policy.

TABLE I: MDP formulation of the capacity-distortion trade-off

state δ_{i-1}	$P_{U_{i-1}\|Y^{i-1}}(\cdot\|y^{i-1})$
action a_i	$P_{X_i\|U_{i-1}Y^{i-1}}(\cdot\|\cdot, y^{i-1})$
reward r_i	$I(X_i, U_{i-1}; Y_i\|Y^{i-1}) - \beta\mathbb{E}[d(S_i, \hat{S}_i)]$.
disturbance	y_i

A. MDP formulation of the capacity-distortion trade-off

To formulate the capacity-distortion trade-off as an MDP, we require to determine the triple (δ_{i-1}, a_i, r_i) where δ_{i-1} is the state at time i, a_i is the action at time i and r_i is the corresponding reward at time i. For some fixed $\beta \in [0,1]$, we determine this triple as in Table I. A new state δ_i is then generated according to Equation (14) shown on the next page.

B. DDPG Algorithm

The training procedure of this algorithm consists of K episodes each of T sequential steps [11]. In each step, we perform the following two operations:

1) Collecting Experience at the Agent: Given the current state δ_{i-1}, the agent takes an action $a_i = A_\mu(\delta_{i-1})$ according to the ϵ-greedy policy as follows:

$$a_i = \begin{cases} \underset{\tilde{a}_1}{\text{argmax}}\, Q_\pi(\delta_{i-1}, \tilde{a}_1), & \text{w.p. } 1-\epsilon, \\ \text{A random action} & \text{w.p. } \epsilon, \end{cases} \tag{15}$$

with $\epsilon \in [0,1]$. After taking the action a_i, the agent observes the incurred reward r_i and the new state s_i. The tuple $\{\delta_{i-1}, a_i, r_i, \delta_i\}$ is then stored in a *replay buffer* denoted by B.

2) Improving agents and environment networks: Draw N samples randomly from B. For each transition $j \in \{1, \ldots, N\}$, we compute the sampled estimate for future rewards denoted by b_j. We then minimize the following objective function:

$$L(w) = \frac{1}{N}\sum_{j=1}^{N} \left(Q_w(\delta_{j-1}, A_\mu(\delta_{j-1})) - b_j\right)^2 \tag{16}$$

over the parameters of the environment network w. To maximize the estimate of future cumulative rewards, the parameter μ is updated as follows

$$\mu \to \mu + \frac{\eta}{N}\sum_{j=1}^{N} \boldsymbol{\nabla}_A Q_w\left(\delta_{j-1}, A\right)|_{A=A_\mu(\delta_{j-1})} \boldsymbol{\nabla}_\mu A_\mu(\delta_{j-1}), \tag{17}$$

where η is the learning rate at the agent.

V. Example: Binary Channel with Multiplicative Bernoulli State

Consider the channel

$$Y_i = S_i X_i, \quad i \in \{1, \ldots, n\} \tag{18}$$

with binary alphabets $\mathcal{X} = \mathcal{S} = \mathcal{Y} \in \{0,1\}$. Assume that the feedback is perfect, i.e., $Z_i = Y_i$, and that the target sequence satisfies

$$S_i = S_{i-1} \oplus \tilde{S}_i, \qquad i = 1, 2, \ldots, \tag{19}$$

with $\{\tilde{S}_i\}$ being i.i.d Bernoulli(p) and $S_0 = 0$ deterministically. We consider the Hamming distortion measure $d(s, \hat{s}) = s \oplus \hat{s}$.

For each time $i = 1, 2, \ldots$, and depending on the past inputs $X_1, \ldots, X_{i-1}$, let L_i denote the time evolved since the input was 1 for the last time. This is, L_i is such that

$$X_{i-1} = \cdots = X_{i-L_i+1} = 0 \quad \text{and} \quad X_{i-L_i} = 1. \tag{20}$$

Then, define the time-i auxiliary random variable

$$U_i := (L_i, S_{i-L_i}). \tag{21}$$

Notice that

$$\begin{aligned} U_i &= \phi(X_i, Y_i, U_{i-1}) = \phi(X_i, Y_i, (L_{i-1}, S_{i-1-L_{i-1}})) \\ &= \begin{cases} (L_{i-1}+1, S_{i-L_{i-1}}) & \text{if } X_i = 0 \\ (0, Y_i/X_i) & \text{else.} \end{cases} \end{aligned} \tag{22}$$

In other words, the Tx can calculate the auxiliary sequence $\{U_i\}$ using an online procedure.

We next determine the optimal estimator to estimate the target S_i based on X^i and Y^i. When $X_i = 1$, the Tx should obviously set $\hat{S}_i = Y_i$ resulting in distortion $d(S_i, \hat{S}_i) = 0$. To understand how to estimate S_i when $X_i = 0$, notice that by (19):

$$S_i = S_{i-L_i} \oplus \tilde{S}_{i-L_i+1} \oplus \cdots \oplus \tilde{S}_i. \tag{23}$$

Conditioned on S_{i-L_i}, it is thus trivially independent of inputs, outputs, and states prior to time $i - L_i + 1$. Moreover, since by definition of L_i we have $X_{i-L_i+1} = \cdots = X_{i-1} = 0$, it is also independent of inputs and outputs after time $i - L_i + 1$. The optimal estimator is thus the maximum likelihood estimator based on $U_i = (L_i, S_{i-L_i})$, which sets

$$\hat{S}_i = S_{i-L_i} \oplus \mathbb{1}\{p_{L_i} > 1 - p_{L_i}\} \tag{24}$$

where we define (notice that the right-hand side in the following expression does not depend on ℓ):

$$p_\ell := \mathbb{P}[\tilde{S}_{i-\ell+1} \oplus \cdots \oplus \tilde{S}_i = 1]. \tag{25}$$

Conditioned on the value of $L_i = \ell$ and given that $X_i = 0$, the distortion for the time-i symbol is thus

$$d(S_i, \hat{S}_i) = \min\{p_\ell, 1 - p_\ell\}. \tag{26}$$

Following a similar reasoning as in the derivation of the distortion under $X_i = 0$, we can conclude that given (U_i, X_i) the channel output and the feedback signal are conditionally independent of the previous inputs, outputs, and auxiliaries:

$$(X^{i-1}, Y^{i-1}, U^{i-1}) \to (X_i, U_i) \to Y_i, \quad i = 1, 2, \ldots \tag{27}$$

Combined with (22), this establishes that channel in this example can be seen as a unifilar channel according to Definition 2, where the auxiliaries U_i acts as the time-i state and the channel transition law is:

$$\mathbb{P}[Y_i = y | X_i = x, U_i = (\ell, s)] = \begin{cases} 1 & y = x = 0 \\ 0 & y = 1, x = 0 \\ p_\ell & y \neq s, x = 1 \\ 1 - p_\ell & y = s, x = 1. \end{cases}$$

Notice that by solving the linear recursion $p_\ell = (1-p)p_{\ell-1} + p(1-p_\ell)$, one obtains that

$$p_\ell = \frac{1}{2} - \frac{1}{2}(1-2p)^\ell. \tag{28}$$

We evaluate Theorem 1 for our unifilar channel.

Theorem 2: Given D, the capacity-distortion trade-off of the binary channel with multiplicative state in (18) is given by the following optimization problem:

$$\mathsf{C}(\mathsf{D}) = \lim_{n\to\infty} \max_{\mathcal{P}} \frac{1}{n}\sum_{i=1}^{n} H_b\left(\sum_{\ell=1}^{i-1}\alpha_i(\ell)\right) - \sum_{\ell=1}^{i-1}\kappa_i(\ell)H_b(p_\ell), \tag{29a}$$

$$\text{s.t.:} \sum_{i=1}^{n}\sum_{\ell=1}^{i-1}\mathbb{P}[X_i=0,\ell_i=\ell]\min\{p_\ell, 1-p_\ell\} \le n\mathsf{D}. \tag{29b}$$

where $H_b(\cdot)$ is the binary entropy function and

$$\mathcal{P} := \{P_{X_i|U_{i-1}}(x_i|u_{i-1})\}_{i=1}^{n}, \tag{30}$$

$$\kappa_i(\ell) := \mathbb{P}[X_i=1, L_{i-1}=\ell], \tag{31}$$

$$\begin{aligned}\alpha_i(\ell) := {}& \mathbb{P}[X_i=0, L_{i-1}=\ell] \\ & + \mathbb{P}[X_i=1, L_{i-1}=\ell, S_{i-L_i}=1]\cdot p_\ell \\ & + \mathbb{P}[X_i=1, L_{i-1}=\ell, S_{i-L_i}=0]\cdot(1-p_\ell).\end{aligned} \tag{32}$$

Proof: We have for the i-th term:

$$\begin{aligned} & I(X_i, U_{i-1}; Y_i|Y^{i-1}) \\ &= H(Y_i|Y^{i-1}) - H(Y_i|X_i, U_{i-1}, Y^{i-1}) && (33)\\ &= H_b(\mathbb{P}[Y_i=0|Y^{i-1}=y^{i-1}]) \\ &\quad - \sum_{\ell=1}^{i-1}\mathbb{P}[X_i=1, L_i=\ell]H_b(p_\ell) && (34)\\ &= H_b\left(\sum_{\ell=1}^{i-1}\alpha_i(\ell)\right) - \sum_{\ell=1}^{i-1}\kappa_i(\ell)H_b(p_\ell), && (35)\end{aligned}$$

where κ_i and α_i are defined in (31) and (32), respectively. Note that the first term in (35) is due to the fact that

$$\begin{aligned} & \mathbb{P}[Y_i=0|Y^{i-1}=y^{i-1}] \\ &= \sum_{\ell=1}^{i-1}\sum_{s}\sum_{x}\mathbb{P}[X_i=x, U_{i-1}=(\ell,s)] \\ &\qquad \mathbb{P}[Y_i=0|Y^{i-1}=y^{i-1}, X_i=x, U_{i-1}=(\ell,s)] && (36)\\ &= \sum_{\ell=1}^{i-1}\mathbb{P}[X_i=0, L_i=\ell] \\ &+ \sum_{\ell=1}^{i-1} p_\ell\cdot\mathbb{P}[X_i=1, L_i=\ell, S_{i-L_i}=1] && (37)\\ &+ \sum_{\ell=1}^{i-1}(1-p_\ell)\cdot\mathbb{P}[X_i=1, L_i=\ell, S_{i-L_i}=0]. && (38)\end{aligned}$$

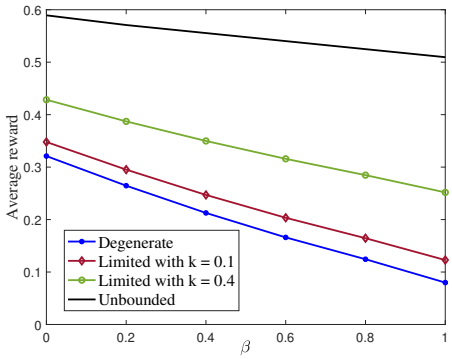

Fig. 2: Average reward as a function of β with $n = 5000$ and $p = 0.1$ for degenerate, limited with $k \in \{0.1, 0.4\}$ and unbounded state spaces.

The second term follows by the fact that conditioning on $L_i = \ell$, if $X_i = 0$ then $H(Y_i|X_i, U_{i-1}, Y^{i-1}) = 0$, and if $X_i = 1$ then $H(Y_i|X_i, U_{i-1}, Y^{i-1}) = H_b(p_\ell)$.

The distortion calculation follows immediately from (26) and because when $X_i = 1$ then the i-th distortion term is zero. ■

VI. Numerical Analysis

In our implementation, we trained the agent for 500 episodes, each containing 100 consecutive blocks. The Monte Carlo evaluation length for the average reward was chosen to be 1000. We solve the optimization problem in (29) using an RL approach with full (unbounded), limited and degenerate sate spaces. In the limited case, we restrict the algorithm state space to a fraction $k \in [0,1]$ of the full state space δ_{i-1}. In the degenerate case, the optimization problem in (29) is solved for memoryless strategies, i.e., for X_i independent of U_{i-1}, or equivalently using an RL approach with constant state-space.

Fig. 2 illustrates the average reward versus the parameter β for $n = 5000$ and $p = 0.1$ for unbounded, limited with $k \in \{0.1, 0.4\}$ and degenerate state spaces. In a similar way, Fig. 3 illustrates this average reward as a function of both β and p. We observe that the average reward is high for very large (i.e., $p > 0.8$) and very small (i.e., $p < 0.2$) values of p, and generally whenever $p \neq 0.5$, the average reward is strictly larger with unbounded state space compared to the limited and degenerate cases. For $p = 0.5$, the situation is somehow degenerate and average reward is the same for the cases with degenerate, limited and unbounded state spaces.

Fig. 4 illustrates the communication gain versus D for the cases with unbounded and degenerate state spaces when $p = 0.3$. As can be seen from this figure, enlarging the state space in our RL framework significantly improves sensing and communication performances.

$$\delta_i(u_i) = \frac{\sum_{x_i,u_{i-1}}\delta_{i-1}(u_{i-1})a_i(x_i,u_{i-1})P(y_i|x_i,u_{i-1})\mathbb{1}\{u_i = f(x_i,u_{i-1},y_i\}}{\sum_{x_i,u_{i-1},\tilde{u}_i}\delta_{i-1}(u_{i-1})a_i(x_i,u_{i-1})P(y_i|x_i,u_{i-1})\mathbb{1}\{\tilde{u}_i = f(x_i,u_{i-1},y_i\}}. \tag{14}$$

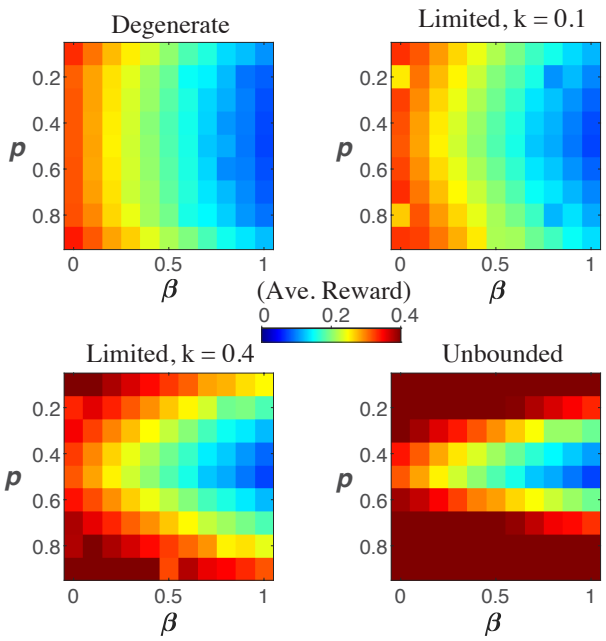

Fig. 3: Average reward as a function of p and β for $n = 5000$ for degenerate, limited with $k \in \{0.1, 0.4\}$ and unbounded state spaces.

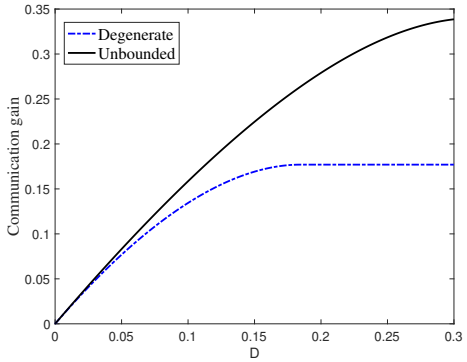

Fig. 4: Communication gain versus the maximum distortion D with $p = 0.3$ for degenerate and unbounded state space cases.

VII. Conclusions

We have considered an ISAC system where a transmitter sends a message to a receiver over a channel with memory and simultaneously estimates given targets by analyzing the backscattered signals from the emitted waveform. Estimation of the targets was performed in an online matter. We have used Massey's concept of directed information to derive the capacity-distortion trade-off for this ISAC setup and simplified the expression for the class of unifilar channels. We then presented an MDP formulation of the resulting waveform optimization problem and solved it by employing the DDPG algorithm. Our numerical results have shown a significant performance improvement when the RL approach can take advantage of the full (unbounded) state space as compared to models with limited (or even degenerate) state spaces.

Aknowledgment

The work of H. V. Poor has been supported by the U.S National Science Foundation under Grant ECCS-2335876. The work of S. Shamai was supported by the German Research Foundation (DFG) via the German-Israeli Project Cooperation (DIP), under Project SH 1937/1-1, and by the Ollendorff Minerva Center of the Technion. The work of M. Wigger has been supported by the European Research Council under Grant Agreement 101125691.

References

[1] V. Koivunen, M. F. Keskin, H. Wymeersch, M. Valkama, and N. González-Prelcic, "Multicarrier ISAC: Advances in waveform design, signal processing and learning under non-idealities,"arXiv:2406.18476, June 2024.

[2] H. Luo, F. Gao, H. Lin, S. Ma, and H. V. Poor, "YOLO: An efficient terahertz band integrated sensing and communications scheme with beam squint," *IEEE Trans. on Wireless Comm.*, Feb. 2024.

[3] F. Liu et al., "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE J. on Sel. Areas in Comm.*, vol. 40, no. 6, pp. 1728–1767, June 2022.

[4] M. J. Ahmadi, R. F. Schaefer, H. V. Poor, "Integrated sensing and communications for unsourced random access: fundamental limits," arXiv:2404.19431, 2024.

[5] S. K. Mohammed, R. Hadani, A. Chockalingam and R. Calderbank, "OTFS—predictability in the delay-Doppler domain and its value to communication and radar sensing," *IEEE BITS the Information Theory Magazine*, vol. 3, no. 2, pp. 7-31, June 2023.

[6] Y. Eldar, A. Goldsmith, D. Gündüz, and H. V. Poor, *Machine learning and wireless communication*, Cambridge University Press, 2022.

[7] M. Ade Krisna Respati and B. M. Lee, "A survey on machine learning enhanced integrated sensing and communication systems: Architectures, algorithms, and applications," *IEEE Access*, vol. 12, pp. 170946–170964, 2024.

[8] P. Pulkkinen and V. Koivunen, "Model-based online learning for active ISAC waveform optimization," *IEEE J. of Sel. Topics in Signal Proc.,* 2024.

[9] P. Pulkkinen and V. Koivunen, "Model-free online learning for waveform optimization In integrated sensing and communications," in *Proc. IEEE ICASSP 2023*, Rhodes Island, Greece, pp. 1–5, 04–10 June, 2023.

[10] N. T. Nguyen et al., "Joint communications and sensing hybrid beamforming design via deep unfolding," *IEEE J. of Sel. Topics in Signal Proc.,* 2024.

[11] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," Online: arXiv:1509.02971, July 2019.

[12] J. L. Massey, "Causality, feedback and directed information," in *Proc. of the IEEE ISITA 1990*, pp. 303-–305, Waikiki, Hawaii, Nov. 27-30, 1990.

[13] R. G. Gallager, *Information Theory and Reliable Communication*, New York, NY, USA: Wiley, 1968.

[14] M. Kobayashi, H. Hamad, G. Kramer, and G. Caire, "Joint state sensing and communication over memoryless multiple access channels," in *Proc. IEEE ISIT 2019*, pp. 270–274, Paris, France, 07–12 July, 2019.

[15] M. Ahmadipour, M. Kobayashi, M. Wigger, and G. Caire, "An information-theoretic approach to joint sensing and communication," *IEEE Trans. on Info. Theory*, vol. 70, pp. 1124 – 1146, 2022.

[16] Y. Chen, T. Oechtering, M. Skoglund, and Y. Luo, "On general capacity-distortion formulas of integrated sensing and communication," Online: arXiv:2310.11080, Oct. 2023.

[17] H. Nikbakht, M. Wigger, S. Shamai, and H. V. Poor, "Integrated sensing and communication in the finite blocklength regime," in *Proc. IEEE ISIT 2024*, pp. 2790–2795, Athens, Greece, 07–12 July, 2024.

[18] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the Trapdoor channel with feedback," *IEEE Trans. on Info. Theory*, vol. 54, no. 7, pp. 3150–3165, July 2008.