Remote Estimation of Markov Processes over Costly Channels: On Implicit Information Benefits

Edoardo D. Santi[1], Touraj Soleymani[1,2], and Deniz Gündüz[1]
[1] Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom
[2] City St George's School of Science and Technology, University of London, United Kingdom

***Abstract*—In this paper, we study the remote estimation of discrete-state Markov processes over costly point-to-point channels. We formulate this problem as an infinite-horizon optimization problem with two players, i.e., a sensor and a monitor, that have distinct information, and with a reward function that takes into account both the communication cost and the estimation quality. We show that the main challenge in solving this problem is associated with the consideration of implicit information, i.e., information that the monitor can obtain about the source when the sensor is idle. Our main objective is to develop a framework for finding exact or approximate solutions to this problem without neglecting implicit information a priori. To that end, we propose three different algorithms, and discuss their properties. The first one is an alternating optimization algorithm that converges to a Nash equilibrium. The second one optimizes both players' policies jointly, and is guaranteed to find a globally optimal solution. The last one is a heuristic algorithm that can find a near-optimal solution. Finally, we compare the performance of these algorithms through a numerical analysis.**

***Index Terms*—Cyber-physical systems, implicit information, Markov processes, Nash equilibria, Pareto optimality.**

I. Introduction

Cyber-physical systems are distributed dynamical systems that tightly integrate computation, communication, and control [1]. These systems, which can enable capabilities that are far beyond those of today's embedded systems, are envisioned to have various applications in smart cities, smart factories, smart healthcare, and smart transportation. Note that the dynamic and distributed nature of cyber-physical systems necessitates persistent status updating of their components so that changes in the environment can be reflected effectively [2]. This steady influx of real-time data empowers cyber-physical systems to swiftly adapt to evolving conditions, ensuring that control decisions are made based on the most pertinent and up-to-date information [3]–[6].

In this paper, we study the remote estimation of Markov processes over costly channels. In particular, we consider a cyber-physical system composed of a sensor observing a discrete-state Markov source and a remote monitor that needs to be informed about the state of the source. The sensor transmits observed information to the monitor over a costly point-to-point channel, and the monitor estimates the state of the source in real time. The cost of transmission can be a real fixed cost incurred at each transmission, or it can represent an incentive to respect a constraint on the rate of communication, due to, for example, limited transmitter battery capacity, or using a shared communication channel. This problem can be quite challenging when both the communication cost and the estimation quality are taken into account. We show that the main challenge in solving this problem is associated with the consideration of implicit information. Nevertheless, we aim at developing a framework that enables us to solve it either exactly or approximately without neglecting implicit information a priori.

This work received funding from the UKRI for the project AIR (ERC-Consolidator Grant, EP/X030806/1) and the SNS JU project 6G-GOALS under the EU's Horizon program (grant agreement No. 101139232).

A. Related Work

Note that in feedback control, the accuracy of the feedback signal directly depends on the quality of the state estimates. This implies that in order to tackle a remote control problem, one should first address the corresponding remote estimation problem, as the most foundational task. The remote estimation of continuous-state Markov processes over costly channels is addressed in [3], [4], [7]–[10]. These works characterize the optimal policies rigorously, and shed light on the role of implicit information. However, these results cannot be generalized to discrete-state Markov sources and they make restrictive assumptions on the form of the Markov processes. The work in [11] shows the form of the optimal communication policy for discrete Markov sources, without communication costs and constrained to a finite set of communication symbols in the finite horizon. This work is extended in [12]–[14] to the infinite horizon, where an operational solution is provided. The remote estimation of discrete-state Markov processes over costly channels, the problem of interest in the present paper, is addressed in [15]–[23]. Closed-form threshold polices are derived in [15] for sources with a symmetric Toeplitz transition matrix. It is shown in [16] that a piece-wise linear convex decreasing function can represent the trade-off curve between the estimation error and the transmission rate. Under similar conditions, general properties of the optimal policies are discussed in [17]. The effect of channel noise in the context of the remote control problem is studied in [18], where different heuristic policies are compared. The effect of channel noise is also studied in [19], where different policies for a two-state Markov process are proposed, taking into account the importance of these states on the actions to be taken by the

monitor. This work is extended in [20] to N-state Markov processes, where an optimization-based method is proposed for finding the optimal parameter of a randomized stationary policy. Other works, such as [21]–[23], propose solutions for solving variations of Markov decision processes (MDPs) in which the monitor needs to pay a fixed price to either observe the current or the next state. Remote monitoring of two-state Markov sources is studied in a multi-user scenario in [24] in the context of random access protocols. Nevertheless, none of the above works on discrete-state Markov sources take advantage of implicit information.

We should highlight that this body of research falls within the category of pragmatic (a.k.a. goal-oriented) communication [5], [6], [25], [26], where the state/context of the receiver becomes relevant when deciding the communication policy.

II. Problem formulation

Consider a sensor observing the state of a source and a remote monitor that needs to be informed about this state in real time. The source is modelled by a discrete-time finite-state Markov chain. At each time step, the sensor observes the current state of the source, and decides whether to transmit the state value to the monitor, incurring a fixed transmission cost $c_t = 1$ at time t and guaranteeing the correct instantaneous estimation of the state at the monitor; or not to transmit, which incurs a zero cost $c_t = 0$ but leaves the monitor to guess the state of the source. Consequently, a unit common reward $r_t = 1$ is gained at time t if the monitor's guess matches the actual state of the source, otherwise zero reward $r_t = 0$ is gained. The combined reward at time t is therefore given by $R_t = r_t - \lambda \cdot c_t$, where λ is chosen depending on which point of the trade-off curve we would like to operate on.

This remote estimation problem can be formulated formally as a two-player team game denoted by $\hat{M} = (\mathcal{I}, \mathcal{S}, \mathcal{A}, P, \mathcal{Z}, O, R)$, where $\mathcal{I} = \{i_1, i_2\}$ is the set of players, i.e., the sensor and the monitor, where the latter receives observations and acts after the former has already acted; at time t, $s_t \in \mathcal{S}$ is the state of the Markov chain, where $\mathcal{S}$ is the state space; $\mathcal{A} = \mathcal{A}^1 \times \mathcal{A}^2$ is the joint action space, where the action of the sensor is $a_t^1 \in \mathcal{A}^1 = \{0, 1\}$ such that $a_t^1 = 1$ means to transmit and $a_t^1 = 0$ means not to transmit, and the action of the monitor is $a_t^2 \in \mathcal{A}^2 = \mathcal{S}$, which represents the state estimated by the monitor; P is the transition probability matrix such that the element $P_{s,s'}$ represents the probability of transitioning from state s to state s'; $\mathcal{Z} = \mathcal{Z}^1 \times \mathcal{Z}^2$ is the joint observation space, where $z_t^1 \in \mathcal{S} = \mathcal{Z}^1$ and $z_t^2 \in \mathcal{S} \cup \{\epsilon\} = \mathcal{Z}^2$ and ϵ is the empty observation that occurs when no message is sent; $O = \{O^1, O^2\}$ represents the set of the observation functions, where $O^1(z_t^1, s_t) = Pr(z_t^1 | s_t) = \mathbb{1}[z_t^1 = s_t]$ and $O^2(z_t^2, s_t, a_t^1) = Pr(z_t^2 | s_t, a_t^1) = a_t^1 \mathbb{1}[z_t^2 = s_t] + (1 - a_t^1)\mathbb{1}[z_t^2 = \epsilon]$; and finally R is the combined reward, which is equal to $1 - \lambda$ when the state is estimated correctly due to a transmission, 1 when the state is estimated correctly without a transmission, 0 for an incorrect estimation of the state without transmission, and $-\lambda$ for an incorrect estimation despite a transmission. Note that the latter is not possible with a reasonable estimation policy. Given a variable, we use the notation $x_{t_1:t_2}$ to denote the sequence of values that the variable takes between and including the time-steps t_1 and t_2.

We are interested in finding the transmission and estimation policies that jointly maximize the infinite horizon average reward function, i.e., we would like to solve the following optimization problem:

Problem 1:

$$\underset{\pi \in \Pi}{\text{maximize}} \; \mathbb{E}_{\sim b_0, P, \pi} \left[\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} R_t \right], \tag{1}$$

where b_0 is the initial distribution of the source's states, and Π is the set of joint history-dependent stochastic policies for players 1 and 2. Note that Π represents the most comprehensive set of achievable policies. The optimal value of Problem 1 is denoted by J^*.

III. Role of Implicit Information

We refer to the information that the monitor obtains about the state of the source when the sensor is idle as *implicit information*. This information is relevant as the two players are jointly optimized and they are aware of each others' policies. However, jointly optimizing the players in this context is not trivial, as the optimal policy of the sensor depends on that of the monitor, and vice versa. Note that Problem 1 can be simplified if we neglect implicit information. This accordingly leads to a decoupling in the design of the sensor and the monitor. However, this approach, which neglects implicit information *a priori* leads to a suboptimal solution in general as it does not take full advantage of the available information.

In this study, we aim to devise methodologies that can find exact or approximate solutions to this problem without neglecting the implicit information. More specifically, we propose three different algorithms to solve Problem 1. The first one deals with the interdependency of the transmission and estimation policies by optimizing the policy of one player while fixing the other, and then repeats this process until convergence. The second algorithm aims to achieve global optimality by recasting the original two-player problem into a single-player occupancy MDP, and thus optimizes both policies jointly. Finally, we propose a policy which does not require parameter optimization and is the optimal solution whenever perfect reconstruction at the monitor is required.

IV. Alternating Optimization Algorithm

In this section, we propose an algorithm that finds a Nash equilibrium to Problem 1. It is clear that for the sensor, the state $x = (s, s_m, n) \in \mathcal{X} = \mathcal{S}^2 \times \mathbb{Z}^+$ is a sufficient statistic for the purpose of finding an optimal policy, where s is the current state of the source, s_m is the last transmitted state, and $n \triangleq t - \tau$, where t is the current time-step and τ is the time of the last transmission. Similarly, we set the monitor states as $y = (s_m, n) \in \mathcal{Y} = \mathcal{S} \times \mathbb{Z}^+$, which is equivalent to the state representation of the sensor without including the current state, as it is not known by the monitor.

Algorithm 1 summarizes the alternating optimization algorithm. Note that step k of the algorithm indicates a combined optimization of both players. We fix the estimation policy $\pi^2 : \mathcal{Y} \mapsto \mathcal{A}^2$ and initialize it as $\arg\max_i((P^n)^\intercal e_{s_m})_i$, which is the basic monitor policy that disregards implicit information. The Markov chain and the monitor form an MDP denoted by $M_1^k = (\mathcal{X}, \mathcal{A}^1, R_1^k, P_1)$. The state space $\mathcal{X}$ is formed by states $x = (s, s_m, n) \in \mathcal{S}^2 \times \{1, 2, \ldots, n_{max}\}$. The action space $\mathcal{A}^1$ is still $\{0, 1\}$. The reward function depends on the policy of the monitor and is defined as $R_1^k(x, 0) = \mathbb{1}(\pi_2^{k-1}(s_m, n) = s)$ and $R_1^k(x, 1) = 1 - \lambda$, corresponding to the two possible actions. The transition operator is $P_1 \in \{P_1^0, P_1^1\}$, where $P_1^0(x, x') = P(s, s')\mathbb{1}(s'_m = s_m)\mathbb{1}(n' = n + 1)$ and $P_1^1(x, x') = P(s, s')\mathbb{1}(s'_m = s)\mathbb{1}(n' = 0)$, corresponding to the two possible actions. We constrain n to never exceed a maximum value n_{max}. To do so, we modify the reward and transition functions when $n = n_{max}$, so that they give the results corresponding to a state transmission regardless of the sensor's action. The average reward infinite horizon problem for this MDP can be solved using the relative value iteration algorithm, as long we satisfy a sufficient condition for convergence, such as having a state $x \in \mathcal{X}$ that is reachable for every other state in $\mathcal{X}$ under all policies [27]. If these assumptions do not hold, an algorithm for multichain MDPs or standard value iteration with a discount factor $\gamma \approx 1$ can be used instead. We initialize the value function $v_1^k(x) = 0$, $\forall x \in X$. At each step k, the algorithm repeatedly applies the following operator to v_1^k until a stopping condition is satisfied:

$$(\Gamma^k V)(x) = \max_{a \in \mathcal{A}^1} R_1^k(x, a) + \mathbb{E}_{x' \sim P_1^a}[V(x')] - V(x_{\text{ref}}), \quad (2)$$

where x_{ref} is a fixed state in $\mathcal{X}$. From the resulting function v_1^k, we extract the optimal transmission policy π_1^k that maximizes the average reward, given the current policy π_2^{k-1}.

Afterwards, the transmission policy is fixed and the optimal estimation policy is found. From the point of view of the monitor, we have an MDP denoted by $M_2^k = (\mathcal{Y}, \mathcal{A}^2, R_2^k, P_1^k)$. The state space $\mathcal{Y}$ is formed by the set of states $y = (s_m, n) \in \mathcal{S} \times \{0, 1, ..., n_{max} - 1\}$. The action space $\mathcal{A}^2 = \mathcal{S}$. The reward function outputs an expectation given the belief over states

$$R_2^k((s_m, n), a) = \begin{cases} b^k(s_m, n)_a - \lambda, & n = 0, \\ b^k(s_m, n)_a, & n > 0, \end{cases} \quad (3)$$

where the belief of the monitor, using the implicit information, is $b^k(s_m, 0) = e_{s_m}$ and $b^k(s_m, n) \propto P^\intercal b^k(s_m, n-1) \circ (1 - \pi_1^k(\cdot, s_m, n))$, for $n \geq 1$, where "$\circ$" represents element-wise multiplication. The transition function is P_2^k and is independent of the action taken, but it depends on the transmission policy and it is defined as

$$P_2^k((s_m, n), (s_m, n+1)) = \\ = (P^\intercal b^k(s_m, n))^\intercal (\mathbf{1} - \pi_k^1(\cdot, s_m, n+1)), \quad (4)$$

$$P_2^k((s_m, n), (s'_m, 0)) = \\ = e_{s'_m}^\intercal P^\intercal b^k(s_m, n) \cdot \pi_k^1(s'_m, s_m, n+1), \quad (5)$$

Algorithm 1 Alternating Optimization Algorithm

1: $\pi_0^2(s_m, n) \leftarrow (P^n)^\intercal e_{s_m}$ ▷ initialize monitor policy
2: $k \leftarrow 0$
3: **while** $J_k \neq J_{k-1}$ **do** ▷ loop until convergence
4: $k \leftarrow k + 1$
5: $\pi_k^1 \leftarrow RVI(\Gamma, \pi_{k-1}^2)$ ▷ improve sensor policy
6: $\pi_k^2(s_m, n) = \arg\max_{a \in \mathcal{A}^2}(b^k(s_m, n))_a$ ▷ improve monitor policy
7: **end while**

with all other transitions having probability zero. The independence of the transition probabilities from the action taken confirms that, to maximize this MDP's value function, it is sufficient to maximize the immediate reward at each time step. The relation in Eq. (3) suggests that we can do this by setting $\pi_2^k(s_m, n) = \arg\max_{a \in \mathcal{A}^2}(b^k(s_m, n))_a$.

The next theorem provides a convergence result for the alternating optimization algorithm.

Theorem 1: *The infinite horizon average reward of the system at the kth sequential policy improvement step $J^k = J(\pi_1^k, \pi_2^k)$ convergences to a fixed point as $k \to \infty$, and the converged policies form a Nash equilibrium.*

Proof: Fixing the monitor policy and applying relative value iteration to the sensor policy finds the best response to the fixed monitor policy. Fixing the sensor policy and setting the monitor policy as above finds the best response to the fixed sensor policy. Both lead to a non-negative change in the expected average reward of the system. Combining this with the fact that the rewards are bounded we get that the algorithm will converge. As applying each policy improvement step finds the best response to the other policy, once two local improvement steps have made no changes, the two policies must be already the best responses to each other, thus a Nash equilibrium is reached. The detailed proof is provided in Appendix A of the extended version [28]. ■

Remark 1: *Note that Problem 1 might possess multiple Nash equilibria, and Theorem 1 guarantees the convergence to one of these equilibria. A question that arises in relation to such a Nash equilibrium is whether it is globally optimal. Answering this question requires developing further techniques such as the one proposed in the next section. The current algorithm's complexity is given by iteratively solving discrete state MDPs and it has the advantage of being less computationally complex than the following.*

It can also be shown that this algorithm can be adapted to achieve a Nash equilibrium for any stationary memoryless channel without receiver feedback.

V. Joint Optimization Algorithm

In this section, we propose an algorithm derived using the notion of occupancy state [29] in decentralized partially observable MDPs (dec-POMDPs) that finds a globally optimal solution to Problem 1. Let an occupancy-state MDP be denoted by $\tilde{M} = (\tilde{\mathcal{S}}, \tilde{\mathcal{A}}, \tilde{P}, \tilde{R})$, where $\tilde{\mathcal{S}} \in \Delta(\mathcal{S})$ is the state space such that each state is a belief over states of the Markov

Algorithm 2 Joint Optimization Algorithm

$\tilde{\pi} \leftarrow RVI\ Q-learning(Problem2)$ ▷ solve Problem 2 to obtain the policy
$b \leftarrow b_0$; ▷ initialize belief using initial state distribution
loop
  $s \leftarrow sample(P^\intercal e_s)$ ▷ sensor observes the Markov state
  $\pi \leftarrow \tilde{\pi}(b)$ ▷ obtain the decision rule from the policy
  if $\pi(s) = 1$ **then** ▷ if the sensor transmits
    $y = s$ ▷ the monitor observes the state
    $b \leftarrow e_y$ ▷ the monitor updates its belief
  else
    $y = \epsilon$ ▷ the monitor does not receive a message
    $b \leftarrow normalize(b \circ (1 - \pi(s)))$ ▷ bayesian update of the monitor belief
  end if
  $a^2 = \arg\max_i b_i$ ▷ monitor estimates state
  $b \leftarrow P^\intercal b$ ▷ account for new time-step transition
end loop

chain, given the monitor's knowledge; $\tilde{\mathcal{A}} : \mathcal{S} \mapsto \{0, 1\}$ is the action space, where each action represents a decision rule of the sensor, mapping from the observed state of the sensor to the action taken; such a decision rule is represented by a vector of dimensions $|\mathcal{S}| \times 1$, where the nth element of this vector is the action taken by the sensor when the Markov state is n; $\tilde{P} \in \tilde{\mathcal{S}} \times \tilde{\mathcal{A}} \times \tilde{\mathcal{S}} \mapsto [0, 1]$ is the state transition probability function given by $\tilde{P}(\tilde{s}, \tilde{a}, \tilde{s}') = \sum_{s \in \mathcal{S}} \tilde{s}_s \mathbb{1}(P^\intercal b'_{\tilde{s},\tilde{a}}(s) = \tilde{s}')$ and $b'_{\tilde{s},\tilde{a}}(s) = \tilde{a}_s e_s + (1-\tilde{a}_s)\frac{(1-\tilde{a}) \circ \tilde{s}}{|(1-\tilde{a}) \circ \tilde{s}|_1}$ is the post-transmission belief of being in state s given the pre-transmission belief $\tilde{s}$ and a known decision rule $\tilde{a}$; and finally $\tilde{R} \in \tilde{\mathcal{S}} \times \tilde{\mathcal{A}} \mapsto \mathbb{R}$ is the reward function, mapping beliefs and sensor decision rules to rewards, i.e., $\tilde{R}(\tilde{s}, \tilde{a}) = \sum_{s \in \mathcal{S}} \tilde{s}_s [\tilde{a}_s(1 - \lambda) + (1 - \tilde{a}_s)\mathbb{1}(\arg\max_i (b'_{\tilde{s},\tilde{a}}(i)) = s)]$, where the $\arg\max$ is the optimal monitor action as shown in Section IV.

We can use this formulation to define the following problem, the solution of which provides us with a solution to the original Problem 1. We define

Problem 2:

$$\underset{\tilde{\pi} \in \tilde{\Pi}}{\text{maximize}}\ \mathbb{E}_{\sim b_0, \tilde{P}, \tilde{\pi}} \left[\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{R}(\tilde{s}_t, \tilde{\pi}(\tilde{s}_t)) \right], \quad (6)$$

where $\tilde{\Pi}$ is the set of stationary deterministic policies for $\tilde{M}$. During each time-step at execution, the sensor obtains the current decision rule by taking $\pi_t = \tilde{\pi}(\tilde{s}_t)$, and then acts according to $a_t^1 = (\pi_t)_s$. Algorithm 2 summarizes the joint optimization algorithm.

The next theorem shows that a global optimal solution can be found by the joint optimization algorithm.

Theorem 2: *The transformation of the original two-player problem to the occupancy-state single-player problem is without loss of optimality, and the generated policies are also optimal for Problem 1.*

Proof: The optimal monitor intuitively always estimates the state corresponding to the highest belief. As a result, the environment perceived by the sensor can be modelled by an MDP where the state is a combination of the source's state and the monitor's belief and as the sensor knows what the monitor knows, the problem can be rewritten as a single agent belief MDP problem. The detailed proof is provided in Appendix B of the extended version, which starts from the concept of occupancy state in dec-POMDPs to derive the single agent belief state formulation. ■

Remark 2: *As globally optimal solutions express a stronger solution concept than Nash equilibria, the performance of the solution obtained by Algorithm 2 is better than or equal to that of the solution obtained by Algorithm 1. Nevertheless, Algorithm 2 can be computationally expensive, as it involves approximately solving a belief MDP which has a continuous state space. The next section focuses on a policy that exploits implicit information and can find a near-optimal solution without requiring any parameter optimization.*

It can be shown that this algorithm maintains optimality for a stationary memoryless channel, as long as the information structure is also maintained, i.e. the sensor knows the information state of the monitor, requiring perfect feedback from the monitor if the channel is imperfect.

VI. Perfect Estimation Policy

In this section, we propose a joint communication and estimation policy, which minimizes the average communication frequency while guaranteeing no estimation error, and does not require any parameter optimization. In this policy, at each time-step, the sensor sends a message to the monitor if and only if the state of the source is not the one with the highest pre-transmission probability in the monitor's belief. This is feasible as the sensor knows the monitor's belief. When a message containing the state is sent, the monitor trivially guesses correctly, otherwise, the monitor can exploit the implicit information. This eliminates any ambiguity and allows the monitor to always estimate the correct state. Algorithm 3 summarizes this policy. In this algorithm, $b_0 \in \Delta(\mathcal{S})$ represents the initial distribution over states of the Markov chain, s is the state of the Markov chain, b is the pre-transmission belief, y is the message sent, and a^2 is the monitor's action.

The next theorem shows that this policy is optimal if perfect reconstruction at the monitor is required.

Theorem 3: *The proposed perfect estimation policy obtains an optimal solution that minimizes the average communication frequency subject to the perfect reconstruction constraint.*

Proof: Perfect reconstruction requires that a transmission occurs when there would otherwise be an error. As the optimal monitor always estimates the state with highest belief, we deduce that the optimal sensor transmits whenever the state does not correspond to the highest pre-transmission (and thus post-transmission) belief. The detailed proof is provided in Appendix C of the extended version. ■

Remark 3: *Note that in many safety critical applications it is desired to have a perfect reconstruction at the monitor. The*

Algorithm 3 Perfect Estimation Policy

$b \leftarrow b_0$ ▷ initialize belief using initial state distribution
loop
  $s \leftarrow sample(P^\intercal e_s)$ ▷ sensor observes the Markov state
  if $s \neq \arg\max_i b_i$ **then** ▷ transmission only occurs if the monitor would guess incorrectly otherwise
    $y = s$ ▷ the monitor observes the state
    $b \leftarrow e_y$ ▷ the monitor updates its belief
  else
    $y = \epsilon$ ▷ the monitor does not receive a message
    $b \leftarrow e_{\arg\max_i b_i}$ ▷ the monitor updates its belief to the natural vector corresponding to the only state that would not result in transmission
  end if
  $a^2 = \arg\max_i b_i$ ▷ monitor estimates state
  $b \leftarrow P^\intercal b$ ▷ account for new time-step transition
end loop

proposed perfect estimation policy achieves this optimally by directly exploiting implicit information without requiring any optimization of parameters. In the next section, we show that this policy can be adapted for different communication rates, achieving performance very close to the globally optimal one.

VII. Numerical results

We compare our policies (called `alternate`, `joint` and `perfect est.`) with two other policies adopted in [20], i.e., a uniform policy in which the sensor transmits a message every $u \in \mathbb{Z}^+$ time steps (called `uniform`) and a randomized stationary policy (called `randomized`), where at each time-step, a transmission occurs with probability p_{tx}, independently from the evolution of the system. Note that [20] deals with Markov chains with transition matrices in the form $P = qI - p(J - I)$, where I is the identity matrix, J is an all-ones matrix and $p, q \in [0, 1]$. In their setting, the monitor keeps guessing the last state it received until it receives a new one. The other policies in [20] are only coherent with this assumption. We consider general Markov chains without any restrictions, so we modify their policies so that the monitor takes actions $a^2(s_m, n) = \arg\max_i((P^n)^\intercal e_{s_m})_i$. These policies do not exploit implicit information, as the transmissions are not based on the state of the source. We also compare our policies with a modified perfect estimation policy (called `heuristic`) that neglects the implicit information, and forces the monitor to take the same actions as above: $a^2(s_m, n) = \arg\max_i((P^n)^\intercal e_{s_m})_i)$.

Fig. 1 shows the trade-off between the average correct reconstruction probability and the average rate of communication. Each line represents a different algorithm and each point a specific solution, obtained by a different λ. `randomized` performs the worst for all average channel utilization values. `uniform` is slightly better at medium average channel utilization values. The only way to obtain a point with perfect reconstruction with average channel utilization < 1 is for the sensor to always transmit when the monitor would have otherwise been wrong, as `perfect est.` and `heuristic` do, while `uniform` and `randomized` do not reason about the monitor's state to decide whether to transmit. Then, the average channel utilization at perfect reconstruction depends on the quality of the monitor's beliefs, where a more accurate belief reduces the average channel utilization needed. `heuristic` achieves a point that is not the most leftwards in the plot as its monitor calculates the beliefs sub-optimally, neglecting the implicit information. `perfect est.` adopts the optimal belief, leading to the most leftwards point. These two algorithms only have one point (policy) each in the graph as they do not have adjustable parameters (in the figure we connect those points to the leftmost zero-transmission point as the other points on the connecting line can be achieved through time-sharing). `alternating` matches `heuristic` in performance for perfect reconstruction. We can show that such point is a Nash equilibrium of the player's policies, but it is not the globally optimal solution. At lower average channel utilization, this algorithm performs better than `heuristic` as the dynamic programming-based policy of the sensor schedules samples optimally, given the estimation policy and the transmission cost parameter λ. `joint` achieves the best trade-off boundary. `joint` and `perfect est.`'s performance at perfect reconstruction is the same, which is compatible with our theoretical result. `joint` provides slightly better performance at lower channel utilization, which is attributed to a more intelligent scheduling.

The algorithms can be grouped into 3 classes in terms of performance. The first one includes `randomized` and `uniform`, for which the sensor does not reason about the behavior of the monitor. Then, `heuristic` and `alternating` transmit information more intelligently, being more likely to transmit when this adds more information to the monitor, though the implicit information is not used or used sub-optimally as `alternating` is not guaranteed to converge to a global optimum. Lastly, `perfect est.` and `joint` use the implicit information optimally and achieve the best performance. Note that some algorithms are only able to find policies corresponding to very few points on the trade-off curve. This is because the objective function is the average reward minus transmission cost. We are not directly targeting the trade-off between reward and communication in a way that would allow us to find a continuous Pareto optimal boundary. In our formulation, different values of transmission cost can lead to the same policy and so the same point in Fig. 1. However, we can achieve any point on the line joining any two points on the graph via randomized time sharing. It should be noted that when dealing with different Markov processes, we can obtain a different number of points for each curve. Also, the distance between `joint`'s and `perfect est.`'s curves varies and the point of `alternating` at average correct reconstruction =1 can be anywhere between that of `perfect est.` and `uniform`, depending on which local maximum the algorithm converges to.

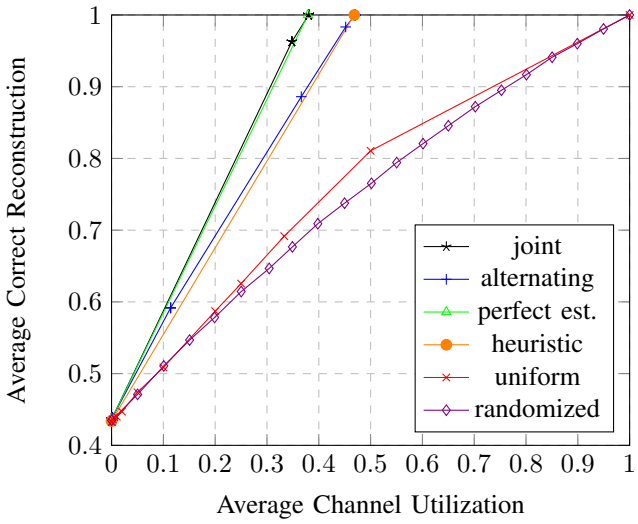

Fig. 1. Trade-off curves between estimation quality and communication cost.

VIII. Conclusion

We developed a framework for finding solutions to the problem of remote estimation of Markov processes over costly channels without neglecting implicit information a priori. First, we proposed an algorithm that alternates between optimizing the transmission and estimation policies, and guarantees a Nash equilibrium. We showed that this algorithm performs vastly better than other policies previously explored in the literature. Then, we proposed the occupancy state formulation, which transforms the original two-player problem into a single-player MDP, and guarantees a globally optimal solution. Lastly, we proposed the perfect estimation policy, in which the messages are sent when the monitor would have otherwise guessed incorrectly, which is optimal in the sense of minimizing communication cost subject to the perfect reconstruction constraint. In future work, we will study the scenario in which the monitor takes actions based on its knowledge about the source state to maximize the total reward it accumulate over time, where the instantaneous reward depends on the current state of the source process and the action taken by the monitor.

References

[1] K.-D. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1287–1308, May 2012.

[2] E. A. Lee, R. Akella, S. Bateni, S. Lin, M. Lohstroh, and C. Menard, "Consistency vs. availability in distributed cyber-physical systems," *ACM Trans. on Embedded Computing Systems*, vol. 22, no. 5, Sep. 2023.

[3] T. Soleymani, J. S. Baras, and S. Hirche, "Value of information in feedback control: Quantification," *IEEE Transactions on Automatic Control*, vol. 67, no. 7, pp. 3730–3737, 2022.

[4] T. Soleymani, J. S. Baras, S. Hirche, and K. H. Johansson, "Value of information in feedback control: Global optimality," *IEEE Transactions on Automatic Control*, vol. 68, no. 6, pp. 3641–3647, 2023.

[5] E. Uysal *et al.*, "Semantic communications in networked systems: A data significance perspective," *IEEE Network*, vol. 36, no. 4, pp. 233–240, 2022.

[6] D. Gündüz *et al.*, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2023.

[7] T. Soleymani, J. S. Baras, S. Hirche, and K. H. Johansson, "Feedback control over noisy channels: Characterization of a general equilibrium," *IEEE Trans. on Automatic Control*, vol. 67, no. 7, pp. 3396–3409, 2021.

[8] T. Soleymani, J. S. Baras, and K. H. Johansson, "State estimation over delayed and lossy channels: An encoder–decoder synthesis," *IEEE Transactions on Automatic Control*, vol. 69, no. 3, pp. 1568–1583, 2024.

[9] A. Molin and S. Hirche, "Event-triggered state estimation: An iterative algorithm and optimality properties," *IEEE Trans. on Automatic Control*, vol. 62, no. 11, pp. 5939–5946, 2017.

[10] G. M. Lipsa and N. C. Martins, "Remote state estimation with communication costs for first-order LTI systems," *IEEE Transactions on Automatic Control*, vol. 56, no. 9, pp. 2013–2025, Sep. 2011.

[11] J. Walrand and P. Varaiya, "Optimal causal coding - decoding problems," *IEEE Trans. on Information Theory*, vol. 29, no. 6, pp. 814–820, 1983.

[12] R. G. Wood, T. Linder, and S. Yüksel, "Optimality of Walrand-Varaiya type policies and approximation results for zero delay coding of Markov sources," in *IEEE Int'l Symp. on Info. Theory (ISIT)*, 2015.

[13] ——, "Optimal zero delay coding of Markov sources: Stationary and finite memory codes," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5968–5980, 2017.

[14] L. Cregg, F. Alajaji, and S. Yüksel, "Reinforcement learning for zero-delay coding over a noisy channel with feedback," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 3939–3944.

[15] J. Chakravorty and A. Mahajan, "On the optimal thresholds in remote state estimation with communication costs," in *53rd IEEE Conference on Decision and Control*, Dec. 2014, pp. 1041–1046.

[16] ——, "Distortion-transmission trade-off in real-time transmission of Markov sources," in *IEEE Info. Theory Workshop (ITW)*, Apr. 2015.

[17] ——, "Fundamental limits of remote estimation of autoregressive Markov processes under communication constraints," *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1109–1124, Jun. 2016.

[18] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *IEEE International Conference on Autonomous Systems (ICAS)*, Aug. 2021, pp. 1–5.

[19] M. Salimnejad, M. Kountouris, and N. Pappas, "State-aware real-time tracking and remote reconstruction of a Markov source," 2023. [Online]. Available: https://arxiv.org/abs/2309.11950

[20] ——, "Real-time remote reconstruction of a Markov source and actuation over wireless," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2023, pp. 1386–1391.

[21] M. Krale, T. D. Simão, and N. Jansen, "Act-then-measure: Reinforcement learning for partially observable environments with active measuring," *Int'l Conf. on Automated Planning and Scheduling*, vol. 33, no. 1, pp. 212–220, Jul. 2023.

[22] C. Bellinger, R. Coles, M. Crowley, and I. Tamblyn, "Active measure reinforcement learning for observation cost minimization," in *Canadian AI*, Jun. 2021.

[23] H. A. Nam, S. Fleming, and E. Brunskill, "Reinforcement learning with state observation costs in action-contingent noiselessly observable Markov decision processes," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[24] G. Cocco, A. Munari, and G. Liva, "Remote monitoring of two-state markov sources via random access channels: An information freshness vs. state estimation entropy perspective," *IEEE Journal on Selected Areas in Information Theory*, vol. 4, pp. 651–666, 2023.

[25] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gündüz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2590–2603, 2021.

[26] D. Gündüz, F. Chiariotti, K. Huang, A. E. Kalør, S. Kobus, and P. Popovski, "Timely and massive communication in 6G: Pragmatics, learning, and inference," *IEEE BITS the Information Theory Magazine*, vol. 3, no. 1, pp. 27–40, 2023.

[27] S. Mahadevan, "Average reward reinforcement learning: Foundations, algorithms, and empirical results," *Machine learning*, vol. 22, pp. 159–195, 1996.

[28] E. D. Santi, T. Soleymani, and D. Gunduz, "Remote estimation of Markov processes over costly channels: On the benefits of implicit information," 2024. [Online]. Available: https://arxiv.org/abs/2401.17999

[29] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet, "Optimally solving Dec-POMDPs as continuous-state MDPs," *Journal of Artificial Intelligence Research*, vol. 55, pp. 443–497, 2016.