

Rollout-Based Approximate Dynamic Programming for MDPs with Information-Theoretic Constraints

Zixuan He¹, Charalambos D. Charalambous², and Photios A. Stavrou¹

Abstract—This paper studies a finite-horizon Markov decision problem with information-theoretic constraints, where the goal is to minimize directed information from the controlled source process to the control process, subject to stage-wise cost constraints, aiming for an optimal control policy. We propose a new way of approximating a solution for this problem, which is known to be formulated as an unconstrained MDP with a continuous information-state using Q-factors. To avoid the computational complexity of discretizing the continuous information-state space, we propose a truncated rollout-based backward-forward approximate dynamic programming (ADP) framework. Our approach consists of two phases: an offline base policy approximation over a shorter time horizon, followed by an online rollout lookahead minimization, both supported by provable convergence guarantees. We supplement our theoretical results with a numerical example where we demonstrate the cost improvement of the rollout method compared to a previously proposed policy approximation method, and the computational complexity observed in executing the offline and online phases for the two methods.

I. INTRODUCTION

Sequential decision-making under uncertainty is a fundamental problem that attracts extensive attention across control and communication systems, paving the way for developing approximate dynamic programming (DP) or model-based reinforcement learning methods [2]. Markov decision processes (MDPs) provide a natural framework for modeling and analyzing sequential decision-making problems through the lens of approximate DP (ADP) problems [3].

Directed information (DI) is an information-theoretic measure that quantifies the causal information flow from one process to another in the presence of feedback information [4]. It has been widely used by control and communication communities to quantify fundamental limits in feedback control systems under information-theoretic constraints [5]–[8]. By incorporating DI into the control objective, the controller is encouraged to be “information-frugal”, making decisions based on only the most relevant and necessary state information, thus enabling real-time systems to balance the performance tradeoffs between control and communication.

*The work of Z. He was supported by the Huawei France-EURECOM Chair on Future Wireless Networks. The work of P. A. Stavrou was supported in part by the SNS JU project 6G-GOALS [1] under the EU’s Horizon programme (Grant Agreement No. 101139232) and by the Huawei France-EURECOM Chair on Future Wireless Networks.

¹Z. He and P. A. Stavrou are with the Foundation and Algorithm Group, Communication Systems Department, EURECOM, France. email: {zixuan.he, fotios.stavrou}@eurecom.fr

²C. D. Charalambous is with the Department of Electrical and Computer Engineering, University of Cyprus, Cyprus. email: chadcha@ucy.ac.cy

DI and other causal information-theoretic measures (eg. entropy, transfer-entropy [9]) have been efficiently combined with MDPs to study sequential stochastic models using DP. For instance, [10], [11] considered the entropy as the reward in MDP and partially observable MDP (POMDP) to synthesize the control policies by maximizing the entropy. Other works integrate the information-theoretic quantities as regularized terms. In [12], an entropy-regularized POMDP is used for active state estimation. Perhaps the most relevant work to ours is [13], which introduced the transfer-entropy-regularized MDP (TERMDP) for feedback control problems. Their approach seeks to minimize the control cost subject to a soft constraint, expressed through transfer entropy, a variant of DI, capturing the uncertainty between the state and control processes. Unfortunately, constructing the transfer-entropy regularizer with a standard state-dependent cost results in a nonconvex optimization problem, which makes the proposed dynamic forward-backward Blahut-Arimoto algorithm (BAA) [14] challenging to converge even to a local optimum, while incurring very high computational complexity.

Recently, we proposed a DP approach based on the information state, or belief, to study a lower bound for a zero-delay lossy compression problem by minimizing DI subject to a single-letter distortion function [15]. To address the well-known “curse of dimensionality” arising in the resulting DP recursion, we employed an ADP methodology grounded in *approximation in policy space* [2]. This ADP framework seeks suboptimal solutions by optimizing over a tractable, parameterized subset of the policy space. Although [15] did not explicitly involve control or MDPs, the DP techniques applied therein constitute a principled and widely adopted methodology for tackling information-theoretic MDPs and POMDPs, also known as information-state MDPs [2].

Another major ADP approach, instead of directly targeting policies, is *approximation in value space*, which focuses on approximating the cost function. This strategy is particularly advantageous for MDPs with prohibitively large or continuous state spaces, as it enables more scalable methods that mitigate computational intractability (see, e.g., [2]). A notable class of methods within this paradigm is the rollout, which begins with an offline base-policy (heuristic) approximation and then applies online cost improvement. Rollout has demonstrated strong empirical effectiveness in reducing complexity while preserving performance guarantees [2], and can also be interpreted as a form of model predictive control (MPC) that leverages a base policy [16].

Contributions: In this work, we consider a discrete-time

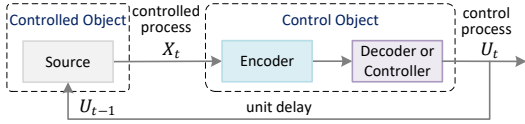


Fig. 1: Controlled discrete-time dynamic system formulated as a finite-horizon MDP

finite-horizon directed information-constrained MDP (DI-MDP) problem in a feedback-controlled system (see Fig. 1), where the goal is to identify the optimal control policy that minimizes DI payoff subject to stage-wise state-dependent cost constraints. We first provide a new DP reformulation for the studied problem using a Q -factor representation in terms of state and control policy pair, which can facilitate offline approximation for the base policy. We then propose a novel truncated-rollout-based backward-forward ADP framework that mitigates the computational burden of the studied problem by decoupling the offline and online phases. The offline training phase involves backward functional base policy and associated Q-factor approximation using a rolling or short horizon (see Algorithm 1). The online phase conducts forward rollout optimization with lookahead minimization followed by instantaneous Q-factor evaluations, to realize policy and cost improvement over any finite horizon (see Algorithm 2 and Theorem 4). Both phases admit stage-wise convergence guarantees (see Lemmas 2, 3). We corroborate our theoretical results with numerical simulations, demonstrating the cost improvement of the rollout method compared to a previously proposed policy approximation method [15] (see Fig. 2), and the computational complexity observed in executing the offline and online phases for the two methods (see Table II), where the guaranteed performance shows the scalability with increased computational resources.

Notation: We denote the set of real numbers by \mathbb{R} . $\mathbb{N} \triangleq \{1, 2, \dots\}$, $\mathbb{N}_0 \triangleq \{0, 1, \dots\}$, and $\mathbb{N}_j^N \triangleq \{j, \dots, N\}$, $j \leq N$, $N \in \mathbb{N}$. We denote a sequence of random variables (RVs) by $X^t = \{X_0, X_1, \dots, X_t\}$, $t \in \mathbb{N}_0^N$ and their real values by $x^t \in \mathcal{X}^t = \{\mathcal{X}_0, \dots, \mathcal{X}_t\}$, where \mathcal{X}_t denotes the finite set and hence \mathcal{X}^t the set sequence. A truncated sequence of RVs is denoted by $X_j^t = \{X_j, \dots, X_t\}$, $j \in \mathbb{N}_0^t$, $t \geq j$, and its realizations by $x_j^t = \{x_j, \dots, x_t\} \in \mathcal{X}_j^t = \{\mathcal{X}_j, \dots, \mathcal{X}_t\}$, $t \geq j$. The distribution of a RV X on \mathcal{X} is denoted by $P(x)$ and the conditional distribution of a RV U given $X = x$ is denoted by $P(u|x)$. Functional dependencies are indicated with square brackets, e.g. $P[Q](x)$ and $P[u](x)$ represent a distribution $P(x)$ that depends on another distribution Q or another realization u , respectively. The expectation operator is denoted by $\mathbb{E}\{\cdot\}$, and $\mathbb{E}^\mu\{\cdot\}$ denotes expectation with respect to a given distribution $\mu(\cdot)$.

II. PROBLEM STATEMENT AND PRELIMINARIES

We consider the controlled discrete-time dynamic system over any finite horizon, illustrated in Fig. 1.

A. System Model

First, we describe each component of the system model in Fig. 1. Consider a controlled source process $X^t \in \mathcal{X}^t$ defined by a time index $t \in \mathbb{N}_0^N$ with given initial distribution $X_0 \sim$

$P_0(x_0)$ and a discrete control process $U^t \in \mathcal{U}^t$ in the system. We assume that the elements of the sets \mathcal{X}^t and \mathcal{U}^t are finite. At each time t , the controlled process $\{X_t : t = 0, \dots, N\}$ is Markov conditional on the past controlled object X_{t-1} and the past control $U_{t-1} = u_{t-1}$. The controlled process is defined by a transition kernel given by the conditional probability distribution

$$P_t(x_t|x_{t-1}, u_{t-1}). \quad (1)$$

The control object or control process $\{U_t : t = 0, \dots, N\}$ is generated according to a randomized history-dependent control policy, hereinafter denoted by μ_t , i.e.,

$$\mu_t = P_t(u_t|u^{t-1}, x_t), \quad (2)$$

with the initial control policy μ_0 to represent $P_0(u_0|x_0)$. Note that (2) depends only on the current controlled object X_t since such a structural property was identified in [17, Section IV], [18, Theorem 4.1]. On the other hand, for each $t = 0, 1, \dots, N$, the controlled object X_t is “partially observed” through the control U_t after applying the policy μ_t by the control object. The information of the controlled object filtered from the control process is summarized via an information state or belief given by the following recursion

$$\begin{aligned} P_{t+1}(x_t|u^t) &= \frac{P_t(x_t, u^t)}{P_{t+1}(u^t)} \\ &= \frac{\sum_{x_{t-1} \in \mathcal{X}_{t-1}} \mu_t P_t(x_t|x_{t-1}, u_{t-1}) P_t(x_{t-1}|u^{t-1})}{\sum_{x_{t-1} \in \mathcal{X}_{t-1}} \mu_t P_t(x_t|x_{t-1}, u_{t-1}) P_t(x_{t-1}|u^{t-1})}, \end{aligned} \quad (3)$$

which is Markov, conditional on u^{t-1} , $P_t(x_{t-1}|u^{t-1})$. Finally, the control process $\{U_t : t = 0, 1, \dots, N\}$ induces the conditional probability distribution $P_t(u_t|u^{t-1})$, $\forall t$, hereafter denoted by ν_t , which is specified by the control object and information state via the following expression

$$\begin{aligned} \nu_t = P_t(u_t|u^{t-1}) &= \sum_{x_{t-1} \in \mathcal{X}_{t-1}} \mu_t P_t(x_t|x_{t-1}, u_{t-1}) \\ &\quad \times P_t(x_{t-1}|u^{t-1}). \end{aligned} \quad (4)$$

Here, by denoting the information state by b_t and $P_t(x_t|x_{t-1}, u_{t-1})$ by w_t , (3) can be expressed as a system transition equation of the form $b_{t+1} = f_t(b_t, \mu_t, w_t)$.

B. System Performance

We specify the system performance using an information-theoretic cost as the objective function, subject to state-dependent cost constraints. Define the decision policy π over the finite horizon N as a sequence of control policies $\pi = \{\mu_0, \dots, \mu_N\}$, and let the set of all admissible decision policies be Π such that $\pi \in \Pi$. At each time stage $t \in \mathbb{N}_0^N$, the information-theoretic cost of the system incurred by the control policy μ_t is quantified by the conditional mutual information [19]

$$\begin{aligned} I(X_t; U_t|U^{t-1}) &\triangleq \mathbb{E}^{\mu_t} \left\{ \log \left(\frac{\mu_t}{\nu_t} \right) \right\} \\ &= \sum_{\substack{x_{t-1} \in \mathcal{X}_{t-1} \\ u^t \in \mathcal{U}^t}} \log \left(\frac{\mu_t}{\nu_t} \right) \mu_t w_t b_t P_t(u^{t-1}). \end{aligned} \quad (5)$$

We define directed information [4] theoretic cost as the additive cost functional

$$C^\pi(X^N, U^N) \triangleq \sum_{t=0}^N I(X_t; U_t | U^{t-1}). \quad (6)$$

Given a state-dependent cost function $\rho_t(x_t, u_t)$ at each t , defined over the state-control pair as $\rho_t: \mathcal{X}_t \times \mathcal{U}_t \rightarrow \mathbb{R}$, and a sequence of nonnegative constraints thresholds $\{D_0, \dots, D_N\}$, the stage-wise state-dependent cost constraint associated with a control policy μ_t is specified by $\mathbb{E}^{\mu_t}\{\rho_t(X_t, U_t)\} \leq D_t$, where

$$\mathbb{E}^{\mu_t}\{\rho_t(X_t, U_t)\} = \sum_{\substack{x_{t-1} \in \mathcal{X}_{t-1} \\ u^t \in \mathcal{U}^t}} \rho_t(x_t, u_t) \mu_t w_t b_t P_t(u^{t-1}). \quad (7)$$

Our DI-MDP is formulated as a constrained optimization problem, which seeks a decision policy π that solves the following sequential stochastic optimization problem

$$\begin{aligned} \inf_{\pi \in \Pi} \quad & C^\pi(X^N, U^N) \\ \text{s.t.} \quad & \mathbb{E}^{\mu_t}\{\rho_t(X_t, U_t)\} \leq D_t, \forall t \in \mathbb{N}_0^N. \end{aligned} \quad (8)$$

We state some noteworthy remarks related to (8).

Remark 1: (i) Solving (8) is equivalent to solving the lower bound (nonanticipative rate-distortion function) of the finite horizon sequential source coding problem with stage-wise distortion constraints [18], where the information-theoretic payoff (6) interprets the expected compression rate, the control policy μ_t represents the test-channel (i.e., reconstruction probability distribution), and (7) as the stage-wise fidelity constraint. (ii) In contrast, the TERMDP problem in [13] adopts a different formulation by reversing the objective and the constraint in (8); the payoff is regularized by transfer entropy. This leads to a nonconvex optimization problem (as discussed in [13]), where the coupling between the information-regularization and the standard cost terms hinders the attainment of the global optimum.

C. Reformulation of (8)

We now present the reformulation of problem (8). We first leverage the key convexity properties of DI and II from [15, Theorem 2, 3] with given b_t obtained for fixed $U^{t-1} = u^{t-1}$ for any $t \in \mathbb{N}_0^N$. This allows to apply the Lagrange duality theorem [20] with Lagrange multipliers $\{s_t \leq 0, t \in \mathbb{N}_0^N\}$, converting (8) into an unconstrained convex optimization problem. We then reformulate it as a finite-horizon *information-state MDP* [2], where for any $t \in \mathbb{N}_0^N$, each element of *information state* b_t takes values in the continuous interval $(0, 1)$, resulting in a continuous information-state space encompassing an infinite number of possible distributions representing b_t . The *control policy* (action) μ_t , the *random disturbance* w_t , and the stage-wise *cost function* $g_t(b_t, \mu_t) = \log\left(\frac{\mu_t}{\nu_t}\right) - s_t(\rho_t(x_t, u_t) - D_t)$ are defined accordingly. Table I summarizes the correspondence between the components of this information-state MDP and those of the original problem (8).

TABLE I: Information-State MDP Corresponds to (8)

Information-state MDP	Connection to (8)
information state	$b_t = P_t(x_{t-1} U^{t-1} = u^{t-1})$
control policy (action)	$\mu_t = P_t(u_t U^{t-1} = u^{t-1}, x_t)$
transition	$P_t(b_{t+1} b_t, \mu_t)$
cost function	$g_t(b_t, \mu_t)$
(in contrast of reward)	$= \log\left(\frac{\mu_t}{\nu_t}\right) - s_t(\rho_t(x_t, u_t) - D_t)$

By invoking the *principle of optimality* [2], we obtain the following finite horizon stochastic DP recursion in Q-factor form, expressed explicitly in terms of any given information state b_t and corresponding control policy μ_t pair, and computed backward in time $t = N, \dots, 0$

$$\begin{aligned} Q_t^*(b_t, \mu_t) = & \sum_{\substack{x_{t-1} \in \mathcal{X}_{t-1}^t, u_t \in \mathcal{U}_t}} \left\{ \left(\log\left(\frac{\mu_t}{\nu_t}\right) - s_t \rho_t(x_t, u_t) \right) \right. \\ & \left. + \min_{\mu_{t+1}} Q_{t+1}^*(b_{t+1}, \mu_{t+1}) \right\} w_t \mu_t b_t \} + s_t D_t, \end{aligned} \quad (9)$$

with the terminal condition $Q_{N+1}^*(b_{N+1}, \mu_{N+1}) = 0$ for all b_{N+1} and μ_{N+1} . Note that the given information state b_t for fixed $U^{t-1} = u^{t-1}$ appearing in (9) can be determined at each time-stage t through the recursion update (3), and this update is Markovian conditioned on the current U^{t-1} and b_t . Furthermore, at $t = 0$, we adopt the initial assumptions of our system model by setting the initial control policy and marginal distribution as $\mu_0 = P_0(u_0 | x_0)$ and $P_0(u_0 | u_{-1}) = P_0(u_0)$ respectively. Hence, the recursion at the initial time stage in (9) is adjusted accordingly.

The stochastic DP recursion given by (9) can be solved using the backward-forward method from [15], which relies heavily on finely discretizing the continuous information-state space into a large finite information-state space \mathcal{B}_t per stage during the offline training phase. The offline computational complexity due to the curse of dimensionality of solving (9) motivates the usage of a novel, efficient ADP method introduced in the next section to obtain reachable results.

III. MAIN RESULTS

In this section, we present our main results. We propose a truncated rollout-based ADP Framework that admits *approximation in value space* to find the optimal solution.

The rollout approach is a sample-based heuristic method where an offline-approximated base policy is further improved in an online mechanism [2]. To obtain the offline-approximated base policy via sampling, we avoid computing $Q_t(b_t, \mu_t)$ backward over the entire horizon N . Instead, we operate over a truncated horizon, which naturally leads to a corresponding truncated rollout approach. We first introduce the approximate Q-factor $\tilde{Q}_t(b_t, \mu_t)$ at any $t \in \mathbb{N}_0^N$

$$\begin{aligned} \tilde{Q}_t(b_t, \mu_t) = & \tilde{Q}_t[g_t, \tilde{Q}_{t+1}](b_t, \mu_t) = \sum_{\substack{x_{t-1} \in \mathcal{X}_{t-1}^t \\ u_t \in \mathcal{U}_t}} \left\{ \left(\log\left(\frac{\mu_t}{\nu_t}\right) \right) \right. \\ & \left. - s_t \rho_t(x_t, u_t) + \tilde{Q}_{t+1}(b_{t+1}, \mu_{t+1}) \right\} w_t \mu_t b_t \} + s_t D_t. \end{aligned} \quad (10)$$

Herein, we adopt a receding control approach by approximating the future \tilde{Q}_{t+1} over a much shorter *rolling horizon*

N_s (also as known as a sampling or *receding horizon* [2]) compared with N ($N_s \ll N$). Thus, \tilde{Q}_t is rewritten as

$$\tilde{Q}_t(b_t, \mu_t) = \tilde{Q}_t[g_t, \tilde{Q}_{N_s}^{\bar{\pi}}](b_t, \mu_t) = \sum_{\substack{x_{t-1} \in \mathcal{X}_{t-1}^t \\ u_t \in \mathcal{U}_t}} \left\{ \left(\log \left(\frac{\mu_t}{\nu_t} \right) - s_t \rho_t(x_t, u_t) + Q_{N_s}^{\bar{\pi}}(b_{t+1}, \mu_{t+1}) \right) w_t \mu_t b_t \right\} + s_t D_t, \quad (11)$$

where $Q_{N_s}^{\bar{\pi}}$ is estimated backward over N_s , and $\bar{\pi}$ is the obtained base policy defined on a discretized information-state space denoted by $\tilde{\mathcal{B}}_t$ such that $b_t \in \tilde{\mathcal{B}}_t$ with its values selected from the interval $(0, 1)$. Thus, the backward computation admits an offline training procedure with analytical recursions of the base control policy. To facilitate the backward approximation of $Q_{N_s}^{\bar{\pi}}$, we employ the following lemmas adapted from [15].

Lemma 1: (Double minimization) For any $t \in \mathbb{N}_0^N$, let $s_t \leq 0$ and $D_t > 0$, then for a fixed w_t , and a given information state $\{b_t \in \tilde{\mathcal{B}}_t\}$ obtained for a fixed $U^{t-1} = u^{t-1}$, minimizing (9) admits a double minimum as

$$Q_t^*(b_t, \mu_t^*) = \min_{\mu_t} \min_{\nu_t} \sum_{\substack{x_{t-1} \in \mathcal{X}_{t-1}^t \\ u_t \in \mathcal{U}_t}} \left\{ \left(\log \left(\frac{\mu_t}{\nu_t} \right) - s_t \rho_t(x_t, u_t) + Q_{t+1}^*(b_{t+1}, \mu_{t+1}) \right) \mu_t w_t b_t \right\} + s_t D_t, \quad t = N, \dots, 0. \quad (12)$$

Moreover, for a fixed μ_t , the right hand side (RHS) of (12) is minimized by (4), whereas for fixed ν_t , the RHS of (12) is minimized by

$$\mu_t = \frac{\nu_t A_t[b_t](x_t, u_t, s_t)}{\sum_{u_t \in \mathcal{U}_t} \nu_t A_t[b_t](x_t, u_t, s_t)}, \quad (13)$$

where $A_t[b_t](x_t, u_t, s_t) = e^{s_t \rho_t(x_t, u_t) - Q_{t+1}^*(b_{t+1}, \mu_{t+1})}$.

Lemma 2: (Base policy approximation algorithm) For each $t \in \mathbb{N}_0^N$, consider a fixed w_t , and a given b_t obtained for a fixed $U^{t-1} = u^{t-1}$. Moreover, for any t , let $s_t \leq 0$ and $\nu_t^{(0)}$ be the initial output distribution with non-zero components, and let $\nu_t^{(k+1)} = \mu_t[\nu_t^{(k)}]$ and $\mu_t^{(k+1)} = \nu_t[\mu_t^{(k)}]$ be expressed wrt the previous k^{th} iteration as follows

$$\mu_t^{(k+1)} = \frac{\nu_t^{(k)} A_t[b_t](x_t, u_t, s_t)}{\sum_{u_t \in \mathcal{U}_t} \nu_t^{(k)} A_t[b_t](x_t, u_t, s_t)}, \quad (14)$$

$$\nu_t^{(k+1)} = \nu_t^{(k)} \sum_{x_t \in \mathcal{X}_t} \frac{b_t w_t A_t[b_t](x_t, u_t, s_t)}{\sum_{u_t \in \mathcal{U}_t} \nu_t^{(k)} A_t[b_t](x_t, u_t, s_t)}. \quad (15)$$

Then as $k \rightarrow \infty$, we obtain for any t that

$$Q_t(b_t, \mu_t^{(k)}) \rightarrow Q_t^*(b_t, \mu_t^*),$$

whereas $Q_t(b_t, \mu_t^{(k)})$ is expressed as

$$Q_t(b_t, \mu_t^{(k)}) = \sum_{x_t \in \mathcal{X}_t, u_t \in \mathcal{U}_t} b_t w_t \mu_t^{(k)} \left(\log \left(\frac{\mu_t^{(k)}}{\nu_t^{(k)}} \right) + Q_{t+1}^*(b_{t+1}, \mu_{t+1}) \right). \quad (16)$$

Therefore, to approximate $Q_{N_s}^{\bar{\pi}}$ over the rolling horizon N_s , the results provided by Lemma 2 can be implemented

Algorithm 1 Offline Base Control Policy Approximation

Input: given $\{w_t : t \in \mathbb{N}_{N-N_s+1}^N\}$,
given base information state $b_t \in \tilde{\mathcal{B}}_t$, Lagrange multipliers $\{s_t \leq 0 : t \in \mathbb{N}_{N_s}^N\}$, error tolerance $\epsilon > 0$

- 1: **Initialize** $\{\nu_t^{(0)} : t \in \mathbb{N}_{N-N_s+1}^N\}$
- 2: **for** $t = N : N - N_s + 1$ **do**
- 3: $k \leftarrow 0$
- 4: **while** $T_{U_t}[u^{t-1}, b_t] - T_{L_t}[u^{t-1}, b_t] > \epsilon$ **do**
- 5: $\mu_t^{(k)} \leftarrow (14)$
- 6: $\nu_t^{(k+1)} \leftarrow (15)$
- 7: $Q_t(b_t, \mu_t^{(k)}) \leftarrow (16)$
- 8: $k \leftarrow k + 1$
- 9: **end while**
- 10: **end for**
- 11: $Q_{N_s}^{\bar{\pi}}(b_t, \mu_t) \leftarrow Q_{N-N_s+1}^*[g_t, Q_{N-N_s+2}^*](b_t, \mu_t^*)$

Output: $\{\mu_t^*(b_t) : t \in \mathbb{N}_{N-N_s+1}^N, b_t \in \tilde{\mathcal{B}}_t\}$,
 $\{\nu_t^*[b_t] : t \in \mathbb{N}_{N-N_s+1}^N, b_t \in \tilde{\mathcal{B}}_t\}$,
 $\{Q_{N_s}^{\bar{\pi}}(b_t, \mu_t) : b_t \in \tilde{\mathcal{B}}_t, \mu_t \in \mu_t^*(b_t)\}$.

via backward computation from the terminal stage $t = N$ to stage $t = N - N_s + 1$, as detailed in Algorithm 1. The following lemma complements Algorithm 1 by providing a stopping criterion to terminate the procedure after a finite number of iterations.

Lemma 3: (Stopping criterion of Algorithm 1) For each $t \in \mathbb{N}_0^N$, the quantity D_{s_t} given by

$$D_{s_t} = \sum_{x_t \in \mathcal{X}_t, u_t \in \mathcal{U}_t} \mu_t^* w_t b_t \rho_t(x_t, u_t), \quad (17)$$

admits the following bounds

$$Q_t^*(b_t, \mu_t) \leq - \sum_{x_t \in \mathcal{X}_t} (b_t w_t \log \left(\sum_{u_t \in \mathcal{U}_t} \nu_t A_t[b_t](x_t, u_t, s_t) \right)) - \sum_{u_t \in \mathcal{U}_t} \nu_t c_t[u^{t-1}](u_t) \log c_t[u^{t-1}](u_t) + s_t D_{s_t}, \quad (18)$$

$$Q_t^*(b_t, \mu_t) \geq - \sum_{x_t \in \mathcal{X}_t} (b_t w_t \log \left(\sum_{u_t \in \mathcal{U}_t} \nu_t A_t[b_t](x_t, u_t, s_t) \right)) - \max_{u_t \in \mathcal{U}_t} \log c_t[u^{t-1}](u_t) + s_t D_{s_t}, \quad (19)$$

where $c_t[u^{t-1}](u_t)$ is expressed as a function of fixed u^{t-1}

$$c_t[u^{t-1}](u_t) = \sum_{x_t \in \mathcal{X}_t} \frac{b_t w_t A_t[b_t](x_t, u_t, s_t)}{\sum_{u_t \in \mathcal{U}_t} \nu_t A_t[b_t](x_t, u_t, s_t)}.$$

Lemma 3 generates a stopping criterion for Algorithm 1 at the k -th iteration by setting the estimation error ϵ per time stage, i.e., $T_{U_t}[u^{t-1}, b_t] - T_{L_t}[u^{t-1}, b_t]$, where

$$T_{U_t}[u^{t-1}, b_t] = - \sum_{u_t \in \mathcal{U}_t} \nu_t c_t[u^{t-1}](u_t) \log c_t[u^{t-1}](u_t),$$

$$T_{L_t}[u^{t-1}, b_t] = - \max_{u_t \in \mathcal{U}_t} \log c_t[u^{t-1}](u_t).$$

Comments on Algorithms 1: The control policy $\{\mu_t^*(b_t) : t \in \mathbb{N}_{N-N_s+1}^N, b_t \in \tilde{\mathcal{B}}_t\}$, the output distribution $\{\nu_t^*[b_t] : t \in \mathbb{N}_{N-N_s+1}^N, b_t \in \tilde{\mathcal{B}}_t\}$, and the cost-to-go functions $\{Q_t^* : t \in \mathbb{N}_{N-N_s+1}^N\}$ are approximated over the base information-state space $\tilde{\mathcal{B}}_t$ discretized under the fixed $U^{t-1} = u^{t-1}$. After computing these quantities backward

over the rolling horizon N_s , we obtain $Q_{N_s}^{\bar{\pi}}(b_t, \mu_t)$ and proceed with online one-step truncated-rollout lookahead minimization starting from the initial information state $\tilde{b}_1 = P_0(x_0|u_0)$ and initial output distribution $\nu_0 = P_0(u_0)$. The online procedure is detailed in Algorithm 2.

Comments on Algorithms 2: In the online truncated-rollout approximation, for each t with an associated information state b_t , the rollout control policy $\tilde{\mu}_t$ is determined by minimizing the approximate Q-factor $\tilde{Q}_t^{\bar{\pi}}(b_t, \mu_t)$ averaged over $P_t(u^{t-1})$

$$\tilde{\mu}_t = \arg \min_{\mu_t \in \bar{\pi}} \sum_{u^{t-1} \in \mathcal{U}^{t-1}} \tilde{Q}_t^{\bar{\pi}}(b_t, \mu_t) P_t(u^{t-1}), \quad (20)$$

where $\tilde{Q}_t^{\bar{\pi}}(b_t, \mu_t)$ is obtained by one-step truncated-rollout lookahead minimizing approximate Q-factor (11), expressed as

$$\tilde{Q}_t^{\bar{\pi}}(b_t, \mu_t) = \min_{\mu_t} \tilde{Q}_t[g_t, Q_{N_s}^{\bar{\pi}}](b_t, \mu_t). \quad (21)$$

Hence, the online rollout procedure at each t involves first a functional minimization followed by an instantaneous policy evaluation. Specifically, the minimization in (21) corresponds to steps 3–9 of Algorithm 1, with Q_t and Q_{t+1} replaced by \tilde{Q}_t and $Q_{N_s}^{\bar{\pi}}$ respectively, where convergence is guaranteed by Lemma 2. After determining the rollout policy $\tilde{\mu}_t$, the subsequent information state \tilde{b}_{t+1} is updated via recursion in (3). Consequently, our rollout approach yields an approximate solution to (8), generating an information-state trajectory $\{\tilde{b}_t, t \in \mathbb{N}_1^N\}$. The resulting rollout policy sequence, $\hat{\pi} = \{\mu_0, \tilde{\mu}_1, \dots, \tilde{\mu}_N\}$, achieves the minimum defined by (20) for each $t \in \mathbb{N}_1^N$. We next show that the proposed rollout policy performs no worse than the base policy.

Theorem 4: For any $t \in \mathbb{N}_1^N$, denote the by $\tilde{J}_t^{\hat{\pi}}(b_t)$ and $\tilde{J}_t^{\bar{\pi}}(b_t)$ the cost-to-go averaged over $P_t(u^{t-1})$ corresponding to the rollout policy $\hat{\pi}$ and base policy $\bar{\pi}$, respectively, starting at information state b_t , which are expressed as

$$\tilde{J}_t^{\hat{\pi}}(b_t) = \sum_{u^{t-1} \in \mathcal{U}^{t-1}} \tilde{Q}_t^{\hat{\pi}}[g_t, Q_{N_s}^{\bar{\pi}}](b_t, \tilde{\mu}_t) P_t(u^{t-1}), \quad (22)$$

$$\tilde{J}_t^{\bar{\pi}}(b_t) = \sum_{u^{t-1} \in \mathcal{U}^{t-1}} \tilde{Q}_t^{\bar{\pi}}[g_t, Q_{N_s}^{\bar{\pi}}](b_t, \mu_t) P_t(u^{t-1}). \quad (23)$$

Then $\tilde{J}_t^{\hat{\pi}}(b_t) \leq \tilde{J}_t^{\bar{\pi}}(b_t)$ for all b_t .

Proof: We prove this by induction backward in time. Clearly it holds for $t = N$ since $\tilde{J}_N^{\hat{\pi}} = \tilde{J}_N^{\bar{\pi}}$ according to (9). Assuming it holds for time stage $t + 1$, we have for all b_t

$$\begin{aligned} \tilde{J}_t^{\hat{\pi}}(b_t) &\stackrel{(a)}{=} \sum_{u^{t-1} \in \mathcal{U}^{t-1}} \tilde{Q}_t^{\hat{\pi}}[g_t(b_t, \tilde{\mu}_t), Q_{N_s}^{\bar{\pi}}(\tilde{b}_{t+1})](b_t, \tilde{\mu}_t) P_t(u^{t-1}) \\ &\stackrel{(b)}{\leq} \sum_{u^{t-1} \in \mathcal{U}^{t-1}} \tilde{Q}_t^{\bar{\pi}}[g_t(b_t, \tilde{\mu}_t), Q_{N_s}^{\bar{\pi}}(\tilde{b}_{t+1})](b_t, \tilde{\mu}_t) P_t(u^{t-1}) \\ &\stackrel{(c)}{=} \min_{\mu_t \in \bar{\pi}} \sum_{u^{t-1} \in \mathcal{U}^{t-1}} \tilde{Q}_t^{\bar{\pi}}[g_t(b_t, \mu_t), Q_{N_s}^{\bar{\pi}}(b_{t+1})](b_t, \mu_t) P_t(u^{t-1}) \end{aligned}$$

Algorithm 2 Online Rollout Evaluation

Input: $\{\tilde{\mathcal{B}}_t : t \in \mathbb{N}_{N-N_s+1}^N\}$ of given $\{b_t : t \in \mathbb{N}_{N-N_s+1}^N\}$,
 $\{\mu_t^*(b_t) : t \in \mathbb{N}_{N-N_s+1}^N, b_t \in \tilde{\mathcal{B}}_t\}$,
 $\{\nu_t^*[b_t] : t \in \mathbb{N}_{N-N_s+1}^N, b_t \in \tilde{\mathcal{B}}_t\}$,
 $\{Q_{N_s}^{\bar{\pi}} : b_t \in \tilde{\mathcal{B}}_t\}$.

1: **Initialize** $\mu_0 = P_0(u_0|x_0)$, $P_1(u^0)$, $\tilde{b}_1 = P(x_0|u_0)$

2: **for** $t = 1 : N$ **do**

3: $\tilde{Q}_t^{\bar{\pi}}(\tilde{b}_t, \mu_t) \leftarrow$ step 3-9 in Algorithm 1

4: $\tilde{\mu}_t \leftarrow$ (20)

5: $\tilde{b}_{t+1} \leftarrow$ (3)

6: **end for**

Output: $\hat{\pi} = \{\mu_0, \tilde{\mu}_1, \dots, \tilde{\mu}_N\}$, $\{\tilde{b}_t, t \in \mathbb{N}_1^N\}$,
 $\{\tilde{\nu}_t : t \in \mathbb{N}_0^N\}$, $C^{\hat{\pi}}(X^N, U^N)$.

$$\begin{aligned} &\stackrel{(d)}{\leq} \sum_{u^{t-1} \in \mathcal{U}^{t-1}} \tilde{Q}_t^{\bar{\pi}}[g_t(b_t, \mu_t), Q_{N_s}^{\bar{\pi}}(b_{t+1})](b_t, \mu_t) P_t(u^{t-1}) \\ &= \tilde{J}_t^{\bar{\pi}}(b_t), \end{aligned}$$

where $\stackrel{(a)}{=}$ is the DP equation for rollout policy (22), $\stackrel{(b)}{\leq}$ holds because of the inductive assumption, $\stackrel{(c)}{=}$ holds by the definition of the online rollout minimization according to (20), and $\stackrel{(d)}{\leq}$ holds by the DP equation for the base policy (23). This completes the proof. ■

Theorem 4 shows the cost improvement property and guarantees that the rollout policy does not degrade the performance compared with the base policy. Furthermore, repeated rollout can be performed by constructing an improved discretized information-state space $\tilde{\mathcal{B}}_t$ based on empirical observations from the previous rollout trajectory. The updated base policy $\tilde{\pi}$ is re-approximated via Algorithm 1 and subsequently enhanced via Algorithm 2. Hence, repeated rollout can progressively refine the approximate optimal solution to problem (8).

IV. NUMERICAL EXAMPLE

This section provides numerical simulations to support our theoretical findings that led to Algorithms 1, 2. We consider binary alphabets, i.e., $\{\mathcal{X}_t = \mathcal{U}_t = \{0, 1\} : t \in \mathbb{N}_0^N\}$, with Hamming distance function given by

$$\rho_t(x_t, u_t) \equiv \rho(x_t, u_t) = \begin{cases} 0, & \text{if } x_t = u_t, \\ 1, & \text{if } x_t \neq u_t, \end{cases} \quad \forall t. \quad (24)$$

We consider finite previous control memory (memory-1 $u^{t-1} \approx u_{t-1}$) applied for information state $b_t = P_t(x_{t-1}|u_{t-1})$, control policy $\mu_t = P_t(u_t|u_{t-1}, x_t)$, and output distribution $\nu_t = P_t(u_t|u_{t-1})$. Under this assumption, we consider the base information-state space $\tilde{\mathcal{B}}_t$, which consists of a matrix comprising two “quantized” binary probability distributions drawn from the original continuous state space. We denote with n_t each quantization level per t , which results in an information-state space $\tilde{\mathcal{B}}_t$ with size $|\tilde{\mathcal{B}}_t| = n_t^{|\mathcal{U}_{t-1}|} = n_t^2$, representing combinations of 2 out of n_t quantized binary distributions.

Example 1: (Time-invariant binary symmetric controlled Markov chain) The source distribution w_t at each $t \in \mathbb{N}_1^N$

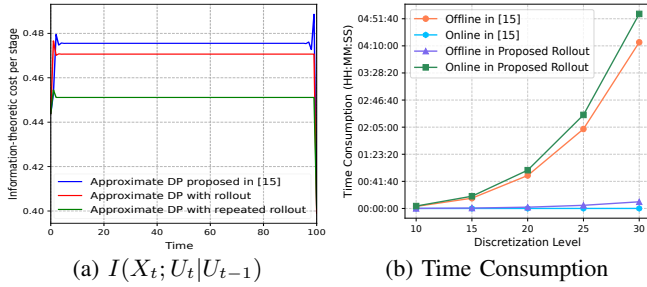


Fig. 2: Illustration of the stage-wise information-theoretic cost and time consumption comparison

	Offline Phase	Online Phase
Proposed scheme	$\mathcal{O}(N_s n^2 \frac{1}{\epsilon})$	$\mathcal{O}(N n \frac{1}{\epsilon})$
[15]	$\mathcal{O}(N n^2 \frac{1}{\epsilon})$	$\mathcal{O}(N n)$

TABLE II: Computation Complexity

is chosen such that for each t , we have

$$w_t = \begin{pmatrix} 1 - \alpha_0 & \alpha_0 & 1 - \alpha_1 & \alpha_1 \\ \alpha_0 & 1 - \alpha_0 & \alpha_1 & 1 - \alpha_1 \end{pmatrix}, \quad (25)$$

where the first two columns are under the condition of $u_{t-1} = 0$ and the rest for $u_{t-1} = 1$, where $\alpha_0, \alpha_1 \in (0, 1)$. Moreover, we choose the quantization levels $\{n_t = n : t \in \mathbb{N}_1^N\}$ and the stage-wise Lagrangian $\{s_t = s : t \in \mathbb{N}_0^N\}$. We demonstrate some results applying the proposed rollout with Algorithms 1, 2 by multi-core processing in Fig. 2a for $n = 20$, $\alpha_0 = 0.4$, $\alpha_1 = 0.8$, $s = -2$, $N = 100$, $N_S = 5$.

We then illustrate the computational complexity of our proposed scheme compared with the policy approximation approach in [15]. As shown in Table II, the proposed rollout-based method significantly reduces offline computation time by operating over a much shorter rolling horizon ($N_s \ll N$). Although the online rollout lookahead minimization of our proposed approach is a bit time-consuming, it offers a clear cost improvement over the prior approach in [15] under the same information-state space discretization level (See Fig. 2a, 2b). We also demonstrate the cost performance of repeated rollout in Fig. 2a where the updated information-state space $\tilde{\mathcal{B}}_t$ with $|\tilde{\mathcal{B}}_t| = n_t^2$, is constructed based on the empirical value range. Using this refined $\tilde{\mathcal{B}}_t$ for a new round of offline base policy approximation, we observe a further cost improvement after the online rollout evaluation.

We also observe that, with the comparable computational time, the proposed method requires indeed less random access memory (RAM) to realize approximation compared to the prior approach in [15]. This memory efficiency enables the offline base policy approximation to explore a larger discretized information-state space, thereby leading to improved online performance. Moreover, this efficiency enhances the proposed method's scalability, making it well-suited for extension to more complex settings as additional computational resources become available.

V. CONCLUSION

We proposed a DI-MDP problem over a finite horizon, aiming to derive the optimal control policy by minimizing the DI subject to stage-wise state-dependent cost constraints.

We first reformulated the problem as an unconstrained information-state MDP and derived DP recursion in Q-factor form. To address the intractability caused by the continuous information-state space in the offline training phase, we proposed a new rollout-based ADP framework that involves offline base policy approximation followed by online approximate policy improvement. Our theoretical results are supplemented with simulation studies that showed the cost performance improvement and demonstrated the scalability.

REFERENCES

- [1] E. C. Strinati et al., "Goal-oriented and semantic communication in 6G AI-native networks: The 6G-GOALS approach," in *Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2024, pp. 1–6.
- [2] D. Bertsekas, *Reinforcement learning and optimal control*. Athena Scientific, 2019.
- [3] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2005.
- [4] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl.*, Nov. 27–30 1990, pp. 303–305.
- [5] C. D. Charalambous, C. K. Kourtellis, and I. Tzortzis, "Information transfer of control strategies: Dualities of stochastic optimal control theory and feedback capacity of information theory," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 5010–5025, 2017.
- [6] T. Tanaka, P. M. Esfahani, and S. K. Mitter, "LQG control with minimum directed information: Semidefinite programming approach," *IEEE Trans. Autom. Control*, vol. 63, no. 1, pp. 37–52, Jan 2018.
- [7] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4525–4540, Nov 2019.
- [8] P. A. Stavrou, M. Skoglund, and T. Tanaka, "Sequential source coding for stochastic systems subject to finite rate constraints," *IEEE Trans. Autom. Control*, vol. 67, no. 8, pp. 3822–3835, 2022.
- [9] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.
- [10] Y. Savas, M. Ornik, M. Cubuktepe, M. O. Karabag, and U. Topcu, "Entropy maximization for markov decision processes under temporal logic constraints," *IEEE Transactions on Automatic Control*, vol. 65, no. 4, pp. 1552–1567, 2019.
- [11] Y. Savas, M. Hibbard, B. Wu, T. Tanaka, and U. Topcu, "Entropy maximization for partially observable markov decision processes," *IEEE transactions on automatic control*, vol. 67, no. 12, pp. 6948–6955, 2022.
- [12] T. L. Molloy and G. N. Nair, "Entropy-regularized partially observed markov decision processes," *IEEE Transactions on Automatic Control*, vol. 69, no. 1, pp. 379–386, 2023.
- [13] T. Tanaka, H. Sandberg, and M. Skoglund, "Transfer-entropy-regularized markov decision processes," *IEEE Transactions on Automatic Control*, vol. 67, no. 4, pp. 1944–1951, 2021.
- [14] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, 1972.
- [15] Z. He, C. D. Charalambous, and P. A. Stavrou, "A new finite-horizon dynamic programming analysis of nonanticipative rate-distortion function for markov sources," in *Proc. European Control Conf.*, 2025, (to appear). [Online]. Available: <https://arxiv.org/abs/2411.11698>
- [16] D. Bertsekas, *Rollout, policy iteration, and distributed reinforcement learning*. Athena Scientific, 2021.
- [17] C. D. Charalambous and P. A. Stavrou, "Optimization of directed information and relations to filtering theory," in *Proc. European Control Conf.*, 2014, pp. 1385–1390.
- [18] P. A. Stavrou, T. Charalambous, C. D. Charalambous, and S. Loyka, "Optimal estimation via nonanticipative rate distortion function and applications to time-varying Gauss-Markov processes," *SIAM J. on Control Optim.*, vol. 56, no. 5, pp. 3731–3765, 2018.
- [19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
- [20] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.