Transfer Entropy in MDPs with Temporal Logic Specifications

Suda Bharadwaj Mohamadreza Ahmadi Takashi Tanaka Ufuk Topcu

***Abstract*— Emerging applications in autonomy require the need for control techniques that take into account uncertain environments, communication and sensing constraints while satisfying high-level mission specifications. Motivated by this need, we consider a class of Markov decision processes (MDPs), along with a *transfer entropy* cost function. In this context, we study high-level mission specifications as co-safe linear temporal logic (LTL) formulae. We provide a method to synthesize a policy that minimizes the weighted sum of the transfer entropy and the probability of failure to satisfy the specification. We derive a set of coupled non-linear equations that an optimal policy must satisfy. We then use a modified Arimoto-Blahut algorithm to synthesize solve the non-linear algorithms. Finally, we demonstrated the proposed method on a navigation and path planning scenario of a Mars rover.**

I. Introduction

Autonomous systems are expected to deliver increasingly complex missions in dynamic and uncertain environments. In space applications, for example, these systems are in addition fettered by communication or sensing restrictions. For instance, in the upcoming Mars 2020 rover mission, a Mars rover is tasked to safely explore an uncertain environment and coordinate with a scouting helicopter [1]. Missions of such sophisticated nature will necessitate on-board autonomy [2], [3]. Nonetheless, tight sensing constraints, due to the power consumption of on-board sensors and transmitters, and bandwidth limitation on data sent from the Earth and orbiting satellites [4], [5] further complicates the navigation task. In these cases, it is necessary for autonomous agents to make decisions to complete their task with *limited information*.

Markov decision processes (MDPs) are one of the most widely studied models for decision-making under uncertainty in the fields of artificial intelligence, robotics, and optimal control [6]. We model the interaction between autonomous agent and an uncertain environment using a Markov decision process (MDP) with an additional *transfer entropy* cost that we refer to as a transfer entropy MDP [7]. We use the additional transfer entropy cost [8] to quantify the directional information flow from the state of an MDP (representing the uncertain environment or the location of the autonomous agent) to the control policy. Intuitively, minimizing the transfer entropy promotes policies that rely less on the knowledge of the current state of the system. In communication theory, a related quantity called *directed information* has been used to measure channel capacities in feedback systems [9], [10] as well as a proxy for feedback data rate to controllers [11].

There has been significant work on quantifying information requirements for low-level control requirements, such as stability [12]. However, quantifying information requirements for high-level decision-making scenarios that we are interested in are not as widely studied. There have been model reduction techniques for MDPs under temporal logic constraints studied where states and actions that are completely irrelevant to the mission are removed [13], [14], [15]. However, these approaches do not *quantify* the information flow to the controller from the state. [16] examines directed information in MDPs to quantify information and policies are penalized if they vary too much from a completely uninformed starting point, e.g, take any action with equal probability. We are, on the other hand, interested in studying the causality of information from the state to the controller, i.e, we seek to penalize sending information that is not relevant for the decision-making process. Hence, transfer entropy is a more suitable information-theoretic metric than directed information in our setting.

We formally describe high-level mission specifications that are defined in temporal logic. Temporal logic has been used as a formal way to allow the user to relatively intuitively specify high-level specifications, in for example, robotics and autonomy applications [17], [18]. Several tools exist to synthesize policies in MDPs with probabilistic temporal logic specifications [19]. We study the effect of information restriction on satisfying temporal logic objectives in MDPs with a transfer entropy cost.

Contributions: We develop a novel framework to formally connect information-theoretic techniques for policy synthesis in MDPs with techniques from formal methods and probabilistic model checking. Specifically, our contributions are as follows:

(1) We develop a framework based on MDPs with a transfer entropy cost which places a cost on state variables that are 'expensive to observe' in an information-theoretic sense.

(2) We incorporate a temporal logic constraint by optimizing the weighted sum of the probability of satisfying a mission specification and the transfer entropy cost.

(3) In contrast to standard MDP policy computation under temporal logic specifications, the transfer entropy cost leads to randomized optimal policies [20], [21], [7] necessitating policy search in an infinite state space. To solve this efficiently, we exploit a necessary optimality condition that the policy must satisfy.

(4) We solve these coupled non-linear equations using a modified version of an iterative algorithm from [22].

(5) While the proposed method builds on earlier results in [7], we generalize the setting to penalize subsets of state variables and incorporate temporal logic constraints.

(6) We apply our results in a case study involving path

All authors are with the University of Texas at Austin. E-mail: {suda.b, mrahmadi, ttanaka, utopcu}@utexas.edu

planning for a Mars rover.

II. Preliminaries

The sequence $(x_0, x_1 ... x_t)$ is denoted x^t and the subsequence $x_l, x_{k+1} ... x_k$ is denoted by x_l^k. We use upper-case letters to denote random variables and lower-case letters for the realizations of the corresponding random variable.
We denote by $\mathcal{D}(\mathcal{X})$ the set of all probability distributions on a finite set $\mathcal{X}$, *i.e.* all functions $f : \mathcal{X} \to [0, 1]$ such that $\sum_{x \in \mathcal{X}} f(x) = 1$. Finally, for a set $\mathcal{S}$, we define $2^{\mathcal{S}}$ as the set of all subsets of $\mathcal{S}$ and S^ω as the set of all infinite sequences of elements in $\mathcal{S}$

A. Markov Decision Processes

Labeled Markov decision process (MDP): Consider a set $\mathcal{AP}$ of *atomic propositions* which can be used, for example, to mark a state as being a "faulty configuration" (reaching it is, thus, undesirable), for example an obstacle. A *labeled MDP* is an MDP whose states are labeled with atomic propositions. More formally, it is a tuple $M = (\mathcal{X}, \mathcal{U}, p, \mathcal{AP}, L)$ where

- $\mathcal{X}$ is a finite set of *states*,
- $\mathcal{U}$ is a finite alphabet of *actions*,
- $p : \mathcal{X} \times \mathcal{U} \to \mathcal{D}(\mathcal{X})$ is a *probabilistic transition function* that assigns, to a state $x \in \mathcal{X}$ and an action $u \in \mathcal{U}$, a probability distribution over the successor states. We abbreviate $p(x_t, u)(x_{t+1})$ by $p(x_{t+1}|x_t, u_t)$.
- $L : \mathcal{X} \to 2^{\mathcal{AP}}$ is the *labeling function* which indicates the set of atomic propositions which are true in each state of the MDP.

Runs and policies: A *run* from state x_0 with time horizon T is a sequence $\rho = x_0 u_0 x_1 u_1 \ldots, x_{T-1}, u_{T-1}, x_T$ of states and actions such that for all $0 \leq t \leq T$ we have $p(x_{t+1}|x_t, u_t) > 0$. A *policy* corresponds to a way of selecting actions based on the history of states and actions. While *deterministic stationary* policies are known to be sufficient for certain classes of problems, such as pure reachability [23], policies in general can be non-deterministic and history dependent. In this paper, we consider the general form and formally represent a policy as a conditional probability distribution $q_t(u_t|x^t, u^{t-1})$.

A run ρ is *consistent* with a policy q if it can be obtained by extending its prefixes using q. Formally, $\rho = x_0 u_0 x_1 u_1 \ldots$ is consistent with q if for all $t \geq 0$ we have that $u_t \in \{u | q_t(u|x^t, u^{t-1} > 0)\}$ and $p(x_{t+1}|x_t, u_t) > 0$

Markov chain: A Markov chain is a tuple $(\mathcal{X}, x_I, p)$ where $\mathcal{X}$ is (in our case) a finite set of states, $x_I \in \mathcal{X}$ is the initial state, and $p : \mathcal{X} \to \mathcal{D}(\mathcal{X})$ is a probabilistic transition function. An MDP M together with a policy q induces a *Markov chain* M^q. Notions of runs in a Markov chain are the same as those defined earlier.

Given a Markov chain $M^q = (\mathcal{X}, x_I, p)$, the state visited at the step t is a random variable. We denote by $h^k(x, \mathcal{B})$ the probability that a run starting from state x visits the set $\mathcal{B}$ in exactly k steps. By definition $h^{\leq i}(x, \mathcal{B}) = \sum_{k=0}^{i} h^k(x, \mathcal{B})$ denotes the probability that run from x reaches the set $\mathcal{B}$ in *at most* i steps where $h^0(x, \mathcal{B})$ is 0 if $x \notin \mathcal{B}$ and 1 otherwise.

B. Temporal Logic

Co-safe linear temporal logic: We utilize linear temporal logic (LTL) to specify the objectives of the system. For example, we can specify that an agent infinitely often patrols a certain set of states (liveness) while not entering undesirable states (safety). For the formal semantics of LTL, see [24]. We are interested in minimizing the expected information cost over a finite time horizon. However, this is not well defined for general LTL formulas as the cost can, in general, diverge. We will thus look at a class of formulas that can be satisfied in finite time called co-safe formulas. These are commonly used in optimal control of MDPs [25]. It was shown in [26] that any LTL formula in which the negation is only applied directly to the atomic propositions called *positive normal form* and which only uses the connectives $\Diamond$ (eventually), $\bigcirc$ (next), and $\mathcal{U}$ (until) are co-safe.

Deterministic finite automaton (DFA): Any co-safe LTL formula φ can be translated to a DFA [26]. A DFA is a tuple $\mathcal{A}_\varphi = (\mathcal{S}, s_I, 2^{\mathcal{AP}}, \delta, \text{Acc})$ where $\mathcal{S}$ is a finite set of states, $\mathcal{AP}$ is a set of atomic propositions, $2^{\mathcal{AP}}$ is the alphabet of the automaton. $\delta : \mathcal{S} \times 2^{\mathcal{AP}} \to \mathcal{S}$ is the transition function and $s_I \in \mathcal{S}$ is the initial state. The acceptance condition Acc is an accepting set of states Acc $\subseteq S$. Since φ is co-safe, it is known that all infinite sequences that satisfy φ have a finite *good prefix*. Let $w = w_0 w_1 \cdots \in (2^{\mathcal{AP}})^\omega$ be an infinite word in the language of the automaton such that $w \vDash \varphi$, then there exists $n \in \mathbb{N}$ such that $w_0, w_1, \ldots w_n \vDash \varphi$. Hence, after reaching an accepting state $s \in$ Acc, we can 'complete' the prefix by setting $\delta(s, \alpha) = s$ for all $\alpha \in 2^{\mathcal{AP}}$

Product MDP: Given an MDP $M = (\mathcal{X}, \mathcal{U}, p, \mathcal{AP}, L)$ and a specification DFA $\mathcal{A}_\varphi = (\mathcal{S}, s_I, 2^{\mathcal{AP}}, \delta, \text{Acc})$, we can define a *product MDP*, $\mathcal{M} := M \times \mathcal{A}_\varphi$, as $\mathcal{M} := (\mathcal{V}, \mathcal{U}, \Delta, v_0, L_\varphi, \text{Acc}_\mathcal{M})$ where

- $\mathcal{V} = \mathcal{X} \times \mathcal{S}$;
- $\Delta : \mathcal{V} \times \mathcal{U} \to \mathcal{D}(\mathcal{V})$ is a probabilistic function such that $\Delta((x_{t+1}, s_{t+1})|(x_t, s_t)) = p(x_{t+1}|x_t, u_t)$ if $\delta(s_t, L(x_{t+1})) = s_{t+1}$;
- $v_0 = (x_0, s_I)$; is the initial state;
- $L_\varphi = L(s) \cup \{\text{acc}_\varphi\}$ if $s \in$ Acc and $L(s)$ otherwise; and
- $\text{Acc}_\mathcal{M}$ is the set of all states where the new atomic proposition acc_φ is true.

Simply, once a run ρ in $\mathcal{M}_\varphi$ reaches a state labeled with the atomic proposition acc_φ, it satisfies the formula φ. We denote a run ρ as satisfying φ by $\rho \vDash \varphi$. Hence, the problem of finding a policy q that maximizes the probability of satisfying a given co-safe LTL specification becomes a matter of synthesizing a strategy to reach a state in $\text{Acc}_\mathcal{M}$. This is a reachability problem in an MDP and can be solved using value iteration. This results in a *memoryless* policy in $\mathcal{M}_\varphi$. Intuitively, the DFA component states of the product MDP can be thought of a *memory state*. From this policy we can construct a *finite-memory* policy in M. For more details on this construction, we refer the reader to [27].

III. Problem Statement

In this section, we present the class of MDPs we consider and we formulate the problem under study.

Let $M = (\mathcal{X}, \mathcal{U}, p, \mathcal{AP}, L)$ be a finite labeled MDP. Let $\mu_t(x^t, u^{t-1})$ be the joint distribution defined recursively by the state transition probability $p(x_{t+1}|x_t, u_t)$ and a policy $q_t(u_t|x^t, u^{t-1})$ as

$$\begin{aligned}&\mu_{t+1}(x^{t+1}, u^t)\\ &= p_t(x_{t+1}|x_t, u_t) q_t(u_t|x^t, u^{t-1}) \mu_t(x^t, u^{t-1}).\end{aligned} \quad (1)$$

Let $\nu_t(u_t|\tilde{x}^t, u^{t-1})$ be the conditional distribution obtained by conditioning and marginalizing the joint distribution $\mu_t(x^t, u^{t-1})$. More specifically,

$$\nu_t(u_t|\tilde{x}^t, u^{t-1}) = \sum_{\bar{\mathcal{X}}^t} \mu_t(\bar{x}^t|\tilde{x}^t, u^{t-1}) q_t(u_t|x^t, u^{t-1}). \quad (2)$$

The *conditional mutual information* $I(\bar{X}^t; U_t|U^{t-1}, \tilde{X}^t)$ [28] can be explicitly written as

$$I(\bar{X}^t; U_t|U^{t-1}, \tilde{X}^t) := \sum_{\mathcal{X}^t, \mathcal{U}^t} \mu_t(x^t, u^{t-1}) \log \frac{q_t(u_t|x^t, u^{t-1})}{\nu_t(u_t|\tilde{x}^t, u^{t-1})}.$$

We assume that the cost of information transfer over the time horizon $0 \leq t \leq T-1$ is proportional to the (causally conditioned) *transfer entropy*.

$$I(\bar{X}^{T-1} \to U^{T-1} \| \tilde{X}^{T-1}) = \sum_{t=0}^{T-1} I(\bar{X}^t; U_t|U^{t-1}, \tilde{X}^t). \quad (3)$$

We note that the notion of *directed information* is introduced by [9] based on [29], and its generalization with causal conditioning by [30]. Intuitively, (3) can be understood as the information flow from a random process $\{\bar{X}_t\}$ to $\{U_t\}$ given $\{\tilde{X}_t\}$ as side information.

A transfer entropy MDP is an MDP with a split state space $\mathcal{X} = \bar{\mathcal{X}} \times \tilde{\mathcal{X}}$ and an associated transfer entropy cost as in equation (3) from $\{\bar{X}_t\}$ to $\{U_t\}$ given $\{\tilde{X}_t\}$. Formally, transfer entropy MDP is a tuple $M = (\bar{\mathcal{X}}, \tilde{\mathcal{X}}, \mathcal{U}, p, \mathcal{AP}, L)$, where $\bar{\mathcal{X}}$ denotes the *expensive* state variables, whereas $\tilde{\mathcal{X}}$ denotes the *free* state variables. To motivate this formulation, we present an example in which such a construction is natural.

Consider a Mars rover for the upcoming Mars 2020 mission [1]. Mars rovers have to complete their tasks in mostly unknown environments. Limited a priori knowledge of the terrain and possible obstacles can be provided from low-resolution satellite imagery. This information, however, is often not enough for decision-making as was evidenced by the Curiosity rover which suffered punctures, due to the unexpected presence of jagged, immobile, rocks embedded in the terrain. For the Mars 2020 mission, a helicopter has been proposed to act as a scout [1] to assist with planning. Figure 1 shows an artists' rendering of the helicopter flying ahead to scout. The helicopter can then transmit information of the terrain back to the rover which is used for planning to satisfy the mission specification.

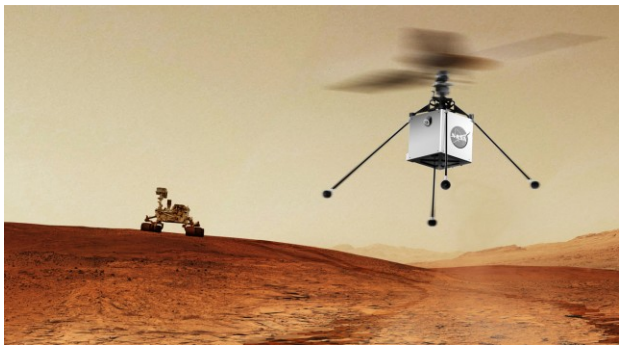


Fig. 1: Artist's rendering of the proposed helicopter to scout for the Mars rover. The helicopter can fly ahead and send information back to the rover about the presence of any obstacles. [1].

We model the dynamics of the rover in the Martian environment as an MDP with split state space $\mathcal{X} = \bar{\mathcal{X}} \times \tilde{\mathcal{X}}$. At every time step t, the component $\tilde{X}_t$ of the state vector is immediately available to the autonomous agent, e.g, from onboard sensors of the rover, while the component $\bar{X}_t$ is only available from a remote sensor, e.g, the scouting helicopter. We are thus interested in finding a policy $q_t(u_t|x^t, u^{t-1})$ that minimizes the information transfer from $\bar{X}$ to U. This information flow is captured by the transfer entropy cost. We can represent this system using a feedback control architecture shown in Figure 2.

Additionally, the rover has to satisfy specification φ, given by a co-safe LTL formula is to a given threshold $0 \leq D \leq 1$ in the probability. Let $\mathbb{P}^T_{q_t}(x_0 \vDash \varphi)$ be the probability of satisfaction of φ by policy q_t in finite time horizon T from initial state x_0. We define $J(X^T, U^{T-1}) := 1 - \mathbb{P}^T_{q_t}(x_0 \vDash \varphi)$ to be the probability of failure.

The main problem we study in this paper can be described as

$$\begin{aligned}\min_{\{q_t(u_t|x^t, u^{t-1})\}_{t=0}^{T-1}} & I(\bar{X}^{T-1} \to U^{T-1} \| \tilde{X}^{T-1})\\ \text{s.t } & J(X^T, U^{T-1}) \leq 1 - D.\end{aligned} \quad (4)$$

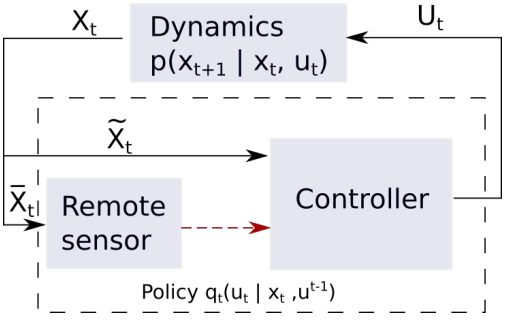


Fig. 2: Example of a feedback control architecture with part of the statespace $\bar{X}_t$ being measured remotely. The red arrow indicates a band limited communications channel so transmissions are restricted.

IV. Incorporating Temporal Logic Constraints

In this section, we demonstrate how to take into account high-level mission specifications in terms of a co-safe LTL formula and cast the constrained control problem into the form of optimization problem (4).

Consider a finite labeled MDP with transfer entropy cost $M = (\hat{\mathcal{X}}, \mathcal{U}, p, \mathcal{AP}, L)$ where, as before, the state space of M is split into expensive and cheap to measure state variables $\hat{\mathcal{X}} = \bar{\mathcal{X}}_e \times \tilde{\mathcal{X}}_f$. We are additionally given a specification DFA $\mathcal{A}_\varphi = (\mathcal{S}, s_I, 2^{\mathcal{AP}}, \delta, \text{Acc})$, and finite time horizon T. The product transfer entropy MDP is $\mathcal{M} := (\mathcal{V}, \mathcal{U}, \Delta, v_0, L_\varphi, \text{Acc}_\mathcal{M})$. Hence, we will have the state space $\mathcal{V} = (\bar{\mathcal{X}}_e \times \tilde{\mathcal{X}}_f) \times S$. Now, for notational simplicity, we set $\mathcal{X} = \mathcal{V}$, the free to measure state $\tilde{\mathcal{X}} = (\tilde{\mathcal{X}}_f, \mathcal{S})$ (we assume without loss of generality that the state in the automaton is freely known), and the expensive to measure state $\bar{\mathcal{X}} = \bar{\mathcal{X}}_e$. Let $X = (\bar{X}_e, \tilde{X}_f, S)$ and $x = (\bar{x}_f, \tilde{x}_s, s)$ be defined similarly. Thus, our state space is now $\mathcal{X} = \bar{\mathcal{X}} \times \tilde{\mathcal{X}}$ with random variable $X = (\bar{X}, \tilde{X})$.

We define a state-action cost in the product MDP in the following way. We define a function $c_t(x_t, u_t, x_{t+1})$, such that for every transition from x_t to x_{t+1}, the cost is 0 if neither x_t or x_{t+1} are in $\text{Acc}_\mathcal{M}$. The cost is -1 if $x_t \notin \text{Acc}_\mathcal{M}$ and $x_{t+1} \in \text{Acc}_\mathcal{M}$ and no state in $\text{Acc}_\mathcal{M}$ has been visited prior to reaching x_t. Intuitively, minimizing this quantity will result in a policy q that maximizes the probability of reaching $\text{Acc}_\mathcal{M}$ and hence, equivalently will maximize the probability of satisfying the temporal logic specification in M. The expected accumulated reward from state x_0 given by $\sum_{t=0}^{T-1} \mathbb{E}\{c_t(x_t, u_t, x_{t+1})\}$ will equal the *negative* of the reachability probability to the target set C in $T-$steps *i.e.* we have

$$\sum_{t=0}^{T-1} \mathbb{E}\{c_t(x_t, u_t, x_{t+1})\} = -h^{\leq T}(x, \text{Acc}_\mathcal{M}) \tag{5}$$

· Setting $J(X^T, U^{T-1}) = \sum_{t=0}^{T-1} \mathbb{E}\{c_t(x_t, u_t, x_{t+1})\}$, we obtain an equivalent formulation of (4) with cost function c_t as defined earlier.

Remark: The constrained optimization problem in equation (4) can be written as a *Lagrangian relaxation* in the following way

$$\min_{\{q_t\}_{t=1}^T} J(X^T, U^{T-1}) + \beta I(\bar{X}^T \to U^{T-1} || \bar{X}^T) \tag{6}$$

where β is a positive constant

Intuitively, this means that we want to minimize the information flow from the state variables in $\bar{\mathcal{X}}$ subject to the constraint on the accumulated cost J. Using the cost function defined in (5), this constrains the probability of not satisfying the specification. The rest of the paper will deal with (6)

V. Optimality Conditions

In this section, we derive a necessary optimality condition for (6). The result in this section generalizes [31], [32] to conditional directed information. In the following derivation, we assume $\beta = 1$ for simplicity. First, we rewrite the objective function in (6) explicitly as a function of q and ν. Using the definition of the causally conditioned directed information (3), the objective function can be written in a stage-additive form as $f(q_0, ..., q_{T-1}; \nu_0, ..., \nu_{T-1}) = \sum_{t=0}^{T-1} \ell_t$ with

$$\begin{aligned}\ell_t = &\sum\nolimits_{\mathcal{X}^t} \sum\nolimits_{\mathcal{U}^t} \mu_t(x^t, u^{t-1}) q_t(u_t | x^t, u^{t-1}) \\ &\times \Bigg(\sum\nolimits_{\mathcal{X}_{t+1}} p(x_{t+1} | x_t, u_t) c_t(x_t, u_t, x_{t+1}) \\ &\qquad + \log \frac{q_t(u_t | x^t, u^{t-1})}{\nu_t(u_t | \tilde{x}^t, u^{t-1})} \Bigg)\end{aligned}$$

where μ_t is recursively defined by q_t via (1). To analyze our cost function $f(q; \nu)$ to be minimized, we note the following simple lemma, which is a straightforward generalization of [33, Theorem 4(b)].

Lemma 5.1: For fixed q, $f(q; \nu)$ is minimized by (2).

This lemma implies that, although q and ν must satisfy (2) (we write $\nu(q)$ to emphasize that ν_t for $0 \leq t \leq T-1$ is a function of q_t for $0 \leq t \leq T-1$), the constraint (2) will be automatically satisfied by solving $\min_{q,\nu} f(q; \nu)$. In particular, if q^* is an optimal solution to (6), and if $\nu^* = \nu(q^*)$, then (q^*, ν^*) is an optimal solution to $\min_{q,\nu} f(q; \nu)$. Since optimality of (q^*, ν^*) implies coordinate-wise optimality of q^*, this implies

$$q^* \in \arg\min_q f(q; \nu^*). \tag{7}$$

Thus, if q^* is an optimal solution to (6), it necessarily satisfies $\nu^* = \nu(q^*)$ and (7) simultaneously. The next lemma shows that the optimal solution to the right hand side of (7) can be obtained analytically.

Lemma 5.2: For fixed ν^*, define ρ_t^* and ϕ_t^* backward in time by

$$\begin{aligned}\phi_t^*(x^t, u^{t-1}) &= \sum\nolimits_{\mathcal{U}_t} \nu_t^*(u_t | \tilde{x}^t, u^{t-1}) \exp\{-\rho_t^*(x^t, u^t)\} \\ \rho_t^*(x^t, u^t) &= \sum\nolimits_{\mathcal{X}_{t+1}} p(x_{t+1} | x_t, u_t) \\ &\qquad \times \{c_t(x_t, u_t, x_{t+1}) - \log \phi_{t+1}(x^{t+1}, u^t)\}\end{aligned}$$

with terminal condition $\phi_T^*(x^T, u^{T-1}) = 1$. Then, the optimal solution to $\min_q f(q; \nu^*)$ satisfies

$$q_t^*(u_t | x^t, u^{t-1}) = \frac{\nu_t^*(u_t | \tilde{x}^t, u^{t-1}) \exp\{-\rho_t^*(x^t, u^t)\}}{\phi_t^*(x^t, u^{t-1})} \tag{8}$$

μ_t-almost everywhere for each $0 \leq t \leq T-1$.

Proof: See Appendix I ∎

The main result of this section is thus summarized as follows.

Theorem 5.3: An optimal solution q^* to (6) necessarily

satisfies the following set of nonlinear equations

$$\mu^*_{t+1}(x^{t+1}, u^t) = p(x_{t+1}|x_t, u_t) q^*_t(u_t|x^t, u^{t-1}) \times \mu^*_t(x^t, u^{t-1}) \quad (9a)$$

$$\nu^*_t(u_t|\tilde{x}^t, u^{t-1}) = \sum_{\bar{\mathcal{X}}^t} \mu^*_t(\bar{x}^t|\tilde{x}^t, u^{t-1}) q^*_t(u_t|x^t, u^{t-1}) \quad (9b)$$

$$\rho^*_t(x^t, u^t) = \sum_{\mathcal{X}_{t+1}} p(x_{t+1}|x_t, u_t)\{c_t(x_t, u_t, x_{t+1}) - \log \phi^*_{t+1}(x^{t+1}, u^t)\} \quad (9c)$$

$$\phi^*_t(x^t, u^{t-1}) = \sum_{\mathcal{U}_t} \nu^*_t(u_t|\tilde{x}^t, u^{t-1}) \times \exp\{-\rho^*_t(x_t, u^t)\} \quad (9d)$$

$$q^*_t(u_t|x^t, u^{t-1}) = \frac{\nu^*_t(u_t|\tilde{x}^t, u^{t-1}) \exp\{-\rho^*_t(x^t, u^t)\}}{\phi^*_t(x^t, u^{t-1})} \quad (9e)$$

for each $0 \leq t \leq T-1$ with the given initial condition μ^*_0 and the terminal condition $\phi^*_T(x^T, u^{T-1}) = 1$.

A. Forward-backward algorithm

The optimality condition (9) is a set of coupled non-linear equations with respect to the variables $\mu^*, \nu^*, \rho^*, \phi^*, q^*$. In order to solve these we propose a numeric forward-backward algorithm. Firstly, note that if ρ^*, ϕ^*, q^* are known, μ^*, ν^* can be solved forwards in time. Similarly, if μ^*, ν^* are known then the others can be solved backwards in time.

To solve this, we do the following. First we make a guess for each of the variables. We then solve the forward-time equations for μ^*, ν^*. We use these values to then solve for ρ^*, ϕ^*, q^* backwards in time. This process is repeated until convergence. This can be viewed as a generalization of the Arimoto-Blahut algorithm [22].

Remark: We note that the problem formulation and derived equations are infinite-history, *i.e.* they depend on the state and control actions from $t = 0$ to $t = T-1$. In order to make this computationally tractable to solve, we modify the algorithm to search for the best policy of the form $q^*_t(u_t|x_t, u^{t-1}_{t-n})$ with some finite n. We refer the reader to [7] for more details on the similar algorithm and its convergence results.

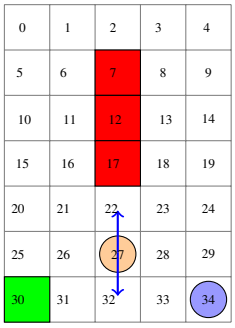
VI. Numerical Results

We consider a scenario where the rover is tasked with collecting samples from a specific region. The environment is modeled as an MDP as motion can be stochastic, i.e, slippage can occur. The mission is specified as a co-safe LTL specification.

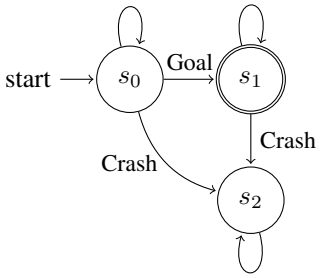
We analyze two different case studies. In the first experiment the rover has to plan around a moving obstacle, but the knowledge of the location of the moving obstacle is penalized. In the second experiment, the rover has some a priori knowledge of the terrain, but there is a cost to using any additional information.

A. Moving obstacle

We solve the motion planning problem under sensing constraints in a gridworld as shown in 3(a). Consider a scenario where rover is tasked with reaching the goal state in green whilst avoiding collisions with the red static obstacles and an orange moving obstacle that moves in the area shown. For example, the helicopter can be completing a separate mission and we do not want the rover and helicopter to collide, but we also want to limit their communication to conserve power. Hence, we treat the helicopter as a moving obstacle and add an information cost to its position.



(a) Gridworld with moving obstacle



(b) Specification DFA

Fig. 3: Gridworld and DFA with $\text{Acc}_{\mathcal{M}} = (s_1)$ depicting a scenario where agent in blue has to reach green target cell without crashing into the red static obstacles or the orange moving obstacle.

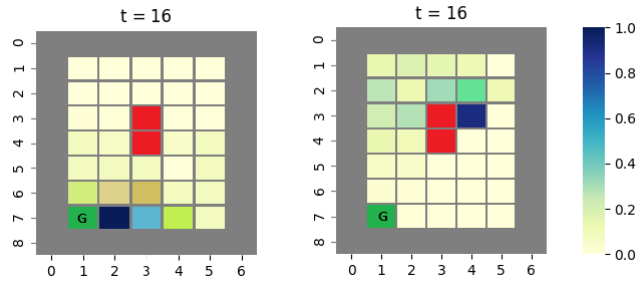
We express this in LTL as $\neg$'crash' $\mathcal{U}$ 'goal'. The atomic proposition 'crash' is true in the red static obstacles and when the state of the rover is the same as the state of the moving obstacle. The atomic proposition 'goal' is true in the green cell. The DFA representation is shown in Figure 3(b).

The rover has the choice of moving in 4 directions - North, South, East, and West or staying still. The motion is stochastic, i.e., it has a probability of slip. For example, if it chooses to move north, it has a probability to 'slip' (due to terrain effects like running sand) and move to a state north east or north west.

The state space of the MDP is (x, y, x_{obs}, y_{obs}) where (x, y) is the position of the rover and (x_{obs}, y_{obs}) is the position of the moving obstacle. We assume that the state of the moving obstacle to be expensive to observe. Formally, we let $\tilde{\mathcal{X}} = (x, y)$ and $\bar{\mathcal{X}} = (x_{obs}, y_{obs})$.

Since there is a probability to slip, the agent has a non-zero probability of crashing and not satisfying the specification if it goes the long way around the wall. If the agent knows the position of the moving obstacle at all times, it can plan to avoid collision, and hence the shorter path will have the higher probability of satisfaction. Intuitively, we expect to see if that we set the β parameter high, *i.e.* if the cost of information is high, the agent will go the long way around the wall as it will be too expensive to observe the moving obstacle. We use a time horizon $T = 25$ and test for $\beta = 0.5$ and $\beta = 5$

Figure 4 shows the probability distributions of the agent at a specific time $t = 16$. Clearly, in the case where $\beta = 0.5$, the agent is able to go through the region where the moving



(a) $\beta = 0.5$ (b) $\beta = 5$

Fig. 4: Probability distribution of the agent after 16/25 timesteps. a) has a low β which means information cost is low while b) has a high β and hence high cost on information

obstacle operates. However, when we increase the cost of information, the agent moves around the static obstacles.

B. Static obstacles

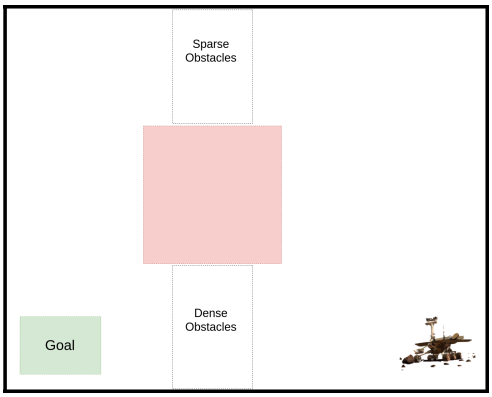
Now, we present the example of the Mars rover navigating in the presence of static obstacles. Figure 5a shows an example of a simple map of the environment that can obtained from a satellite image. This gives us a rough knowledge of the environment. We know the red region is impassable terrain, e.g. a jagged boulder. We also know that there is a region with a high density of obstacles and one with a low density of obstacles. All other regions are assumed to be obstacle free.

The helicopter can send information on the exact locations of obstacles to the rover to assist in path planning, however, we assign a cost to this information.

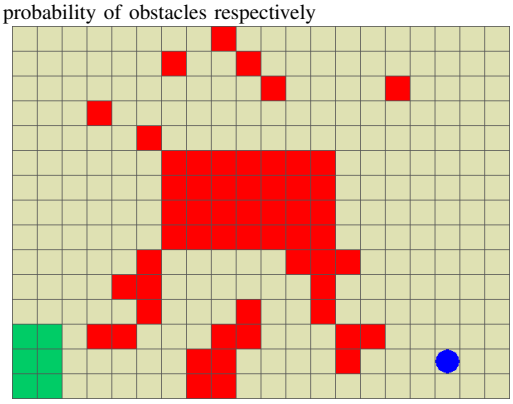
The LTL specification is again $\neg$'crash' $\mathcal{U}$ 'goal' where the goal is the green region. We use a time horizon $T = 40$. This time the state in the MDP is given by $(x, y, (o_1, \ldots, o_n))$ where $o_i \in [0, 1]$ are probability values indicating the likelihood there is an obstacle in state (x_i, y_i). We assign discrete values to o_i by constraining it to values in the set $o_i \in [0, 0.2, 0.4, 0.6, 0.8, 1]$.

We model the helicopter flying ahead and scouting by allowing o_i to transition to 0 or 1 with probability given by the value of o_1, if the rover is within distance d of the obstacle. More explicitly, the state $(x_i, y_i, (o_1, \ldots, o_j, \ldots, o_n))$ will transition to $(x_i, y_i, (o_1, \ldots, 1, \ldots, o_n))$ with probability o_j and transition to $(x_i, y_i, (o_1, \ldots, 0, \ldots, o_n))$ with probability $1 - o_j$. This will only happen if distance between the states (x_i, y_i) and (x_j, y_j) is less than or equal to a given range $d = 2$. $o_i = 1$ indicates there is an obstacle present in (x_i, y_i).

The region with sparse obstacle distribution has mostly $o_i = 0$ while the dense obstacle region has many more states with $o_i > 0$. This means that the rover will need the helicopter to scout ahead more often in the route with more obstacles. We assign the transfer entropy cost to states $(o_1, \ldots o_n)$ which will penalize using these states in the policy synthesis.

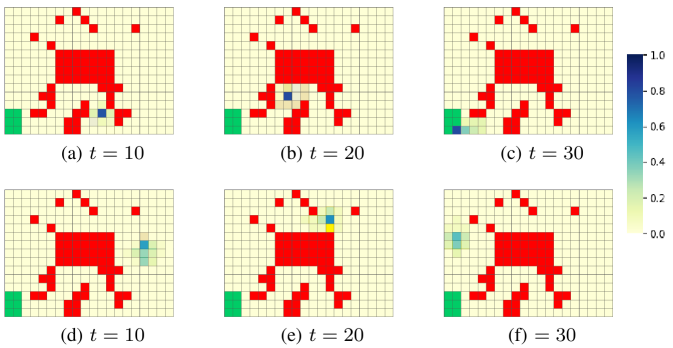


(a) Simple martian environment for case study. The red region is already known as impassable terrain. Additionally, two areas are identified as having a high and low probability of obstacles respectively



(b) True obstacle distribution of the scenario from 5a. Red cells are obstacles, and the green cells represent the target region the rover shown in blue is trying to reach.

Fig. 5: We represent the environment in 5a as a gridworld in 5b which we model as an MDP.



(a) $t = 10$ (b) $t = 20$ (c) $t = 30$

(d) $t = 10$ (e) $t = 20$ (f) $= 30$

Fig. 6: Probability distribution of the rover location over time for two cases: (a) - (c) : $\beta = 0$, (d) - (f): $\beta = 10$.

Figure 6 shows the evolution of the probability distribution of the agent when the information is free (i.e β is small) and when information is expensive (β is large). We see that when information is free, the rover takes the path through the dense obstacle distribution. Also note that since there is no information cost, the problem reduces to solving pure reachability and the policy is deterministic. When we set $\beta = 10$, the rover takes the path through the sparse obstacle region. Since there are fewer cells with non-zero probability of rocks, there is less need to sense for rocks and send the helicopter to scout d states ahead The transfer entropy cost from $(o_1, \dots o_n)$ to u is thus lower along the sparse obstacle path.

VII. Conclusion and Future Work

In this paper, we presented a formal way to integrate co-safe LTL constraints into a minimal-information MDP problem. This is the first step in analyzing temporal logic constraints in communication constrained problems. For future work, we aim to relax the co-safe requirement to allow more general classes of LTL formulas by analyzing the mean information cost over an infinite run. Furthermore, we aim to extend this work to a multiple coordinating agent formulation as this problem setting naturally lends itself to minimizing communication between agents who are trying to satisfy a joint specification.

References

[1] E. Landau, "Helicopter could be scout for mars rovers," *NASA/JPL News Release*, vol. 31, 2015.

[2] R. Francis, D. Gaines, and G. Osinski, "Advanced rover science autonomy experiments in preparation for the mars 2020 mission: Results from the 2016 canmars analogue mission," in *Lunar and Planetary Science Conference*, vol. 48, 2017.

[3] T. Estlin, D. Gaines, C. Chouinard, R. Castano, B. Bornstein, M. Judd, I. Nesnas, and R. Anderson, "Increased mars rover autonomy using ai planning, scheduling and execution," in *IEEE International Conference on Robotics and Automation*. IEEE, 2007, pp. 4911–4918.

[4] R. Sherwood, A. Mishkin, S. Chien, T. Estlin, P. Backes, B. Cooper, G. Rabideau, and B. Engelhardt, "An integrated planning and scheduling prototype for automated mars rover command generation," in *Sixth European Conference on Planning*, 2014.

[5] P. G. Backes, G. Rabideau, K. S. Tso, and S. Chien, "Automated planning and scheduling for planetary rover distributed operations," in *Proceedings in IEEE International Conference on Robotics and Automation*, vol. 2, 1999, pp. 984–991.

[6] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes," *Mathematics of Operations Research*, vol. 12, no. 3, pp. 441–450, 1987.

[7] T. Tanaka, H. Sandberg, and M. Skoglund, "Finite state markov decision processes with transfer entropy costs," *arXiv preprint arXiv:1708.09096*, 2017.

[8] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[9] J. Massey, "Causality, feedback and directed information," in *Proceedings in the International Symposium on Information Theory Applications*, 1990, pp. 303–305.

[10] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, 2009.

[11] E. I. Silva, M. S. Derpich, and J. Ostergaard, "An achievable data-rate region subject to a stationary performance constraint for lti plants," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1968–1973, 2011.

[12] G. N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans, "Feedback control under data rate constraints: An overview," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 108–137, Jan 2007.

[13] S. Bharadwaj, S. L. Roux, G. Perez, and U. Topcu, "Reduction techniques for model checking and learning in MDPs," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017, pp. 4273–4279.

[14] T. Brázdil, K. Chatterjee, M. Chmelík, V. Forejt, J. Křetínskỳ, M. Kwiatkowska, D. Parker, and M. Ujma, "Verification of markov decision processes using learning algorithms," in *International Symposium on Automated Technology for Verification and Analysis*. Springer, 2014, pp. 98–114.

[15] F. Ciesinski, C. Baier, M. Groesser, and J. Klein, "Reduction techniques for model checking markov decision processes," in *International Conference on Quantitative Evaluation of Systems*. IEEE, 2008.

[16] N. Tishby and D. Polani, *Information Theory of Decisions and Actions*. New York, NY: Springer New York, 2011, pp. 601–636.

[17] M. Svorenová, I. Cerna, and C. Belta, "Optimal control of MDPs with temporal logic constraints," in *CDC*, 2013, pp. 3938–3943. [Online]. Available: http://dx.doi.org/10.1109/CDC.2013.6760491

[18] B. Lacerda, D. Parker, and N. Hawes, "Optimal policy generation for partially satisfiable co-safe LTL specifications," in *IJCAI*, 2015, pp. 1587–1593. [Online]. Available: http://ijcai.org/Abstract/15/227

[19] J. Fu, S. Han, and U. Topcu, "Optimal control in Markov decision processes via distributed optimization," in *2015 54th IEEE Conference on Decision and Control*, Dec 2015, pp. 7462–7469.

[20] T. Tanaka, P. M. Esfahani, and S. K. Mitter, "LQG control with minimum directed information: Semidefinite programming approach," *IEEE Transactions on Automatic Control*, 2017.

[21] E. Todorov, "Efficient computation of optimal actions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 11 478–83, August 2009.

[22] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul 1972.

[23] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

[24] C. Baier and J. Katoen, *Principles of model checking*. MIT Press, 2008.

[25] B. Lacerda, D. Parker, and N. Hawes, "Optimal and dynamic planning for Markov decision processes with co-safe LTL specifications," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2014, pp. 1511–1516.

[26] O. Kupferman and M. Y. Vardi, "Model checking of safety properties," *Formal Methods in System Design*, vol. 19, no. 3, pp. 291–314, 2001.

[27] V. Forejt, M. Z. Kwiatkowska, G. Norman, and D. Parker, "Automated verification techniques for probabilistic systems." in *SFM*, vol. 11. Springer, 2011, pp. 53–113.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[29] H. Marko, "The bidirectional communication theory–a generalization of information theory," *IEEE Transactions on communications*, vol. 21, no. 12, pp. 1345–1351, 1973.

[30] G. Kramer, "Causal conditioning, directed information and the multiple-access channel with feedback," in *In Proceedings of the IEEE International Symposium on Information Theory*. IEEE, 1998, p. 189.

[31] C. D. Charalambous and P. A. Stavrou, "Optimization of directed information and relations to filtering theory," in *European Control Conference (ECC)*. IEEE, 2014, pp. 1385–1390.

[32] P. A. Stavrou, C. K. Kourtellaris, and C. D. Charalambous, "Information nonanticipative rate distortion function and its applications," in *Coordination Control of Distributed Systems*. Springer, 2015, pp. 317–324.

[33] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

Appendix I
Proof of Lemma 5.2

We will use the following basic result repeatedly.

Lemma 1.1: [33, Theorem 4(c)] For fixed $\mu(x)$ and $\nu(u)$, the optimal solution to

$$\min_{q(u|x)} \sum_{\mathcal{X}} \sum_{\mathcal{U}} \mu(x) q(u|x) \left(\log \frac{q(u|x)}{\nu(u)} + c(x,u) \right)$$

satisfies

$$q(u|x) = \frac{\nu(u)\exp\{-c(x,u)\}}{\sum_{\mathcal{U}} \nu(u)\exp\{-c(x,u)\}}$$

$\mu(x)$-almost everywhere.

To prove Lemma 5.2, it is sufficient to show the following statements hold for each $0 \le t \le T-1$.

(a) For fixed $q_0, ..., q_{t-1}$, the optimal solution to

$$\min_{q_t} f(q_0, ..., q_{t-1}, q_t, q^*_{t+1}, ..., q^*_{T-1}; \nu^*)$$

satisfies (8).

(b) For fixed $q^*_{t+1}, ..., q^*_{T-1}$ satisfying (8), we have

$$\sum_{k=t}^{T-1} \ell_k = -\sum_{\mathcal{X}^t} \sum_{\mathcal{U}^{t-1}} \mu_t(x^t, u^{t-1}) \log \phi^*_t(x^t, u^{t-1}).$$

We prove these statements by backward induction. For the time step $T-1$, notice that

$$\begin{aligned} f(q;\nu^*) &= \text{(constant)} + \\ &\sum_{\mathcal{X}^{T-1}} \sum_{\mathcal{U}^{T-1}} \mu_{T-1}(x^{T-1}, u^{T-2}) q_{T-1}(u_{T-1}|x^{T-1}, u^{T-2}) \\ &\times \left(\log \frac{q_{T-1}(u_{T-1}|x^{T-1}, u^{T-2}}{\nu^*_{T-1}(u_{T-1}|\tilde{x}^{T-1}, u^{T-2})} + \rho^*_{T-1}(x^{T-1}, u^{T-1}) \right) \end{aligned}$$

where "constant" is the term that does not depend on q_{T-1}. Lemma 1.1 is applicable to show that the minimizer q^*_{T-1} satisfies (8). Statement (b) can be shown directly by substituting $q_{T-1} = q^*_{T-1}$ as

$$\begin{aligned} \ell_{T-1} &= \sum_{\mathcal{X}^{T-1}} \sum_{\mathcal{U}^{T-1}} \mu_{T-1}(x^{T-1}, u^{T-2}) q^*_{T-1}(u_{T-1}|x^{T-1}, u^{T-2}) \\ &\quad \times \left(\log \frac{q^*_{T-1}(u_{T-1}|x^{T-1}, u^{T-2}}{\nu^*_{T-1}(u_{T-1}|\tilde{x}^{T-1}, u^{T-2})} + \rho^*_{T-1}(x^{T-1}, u^{T-1}) \right) \\ &= \sum_{\mathcal{X}^{T-1}} \sum_{\mathcal{U}^{T-1}} \mu_{T-1}(x^{T-1}, u^{T-2}) q^*_{T-1}(u_{T-1}|x^{T-1}, u^{T-2}) \\ &\quad \times \left(-\log \phi^*_{T-1}(x^{T-1}, u^{T-2}) \right) \\ &= -\sum_{\mathcal{X}^{T-1}} \sum_{\mathcal{U}^{T-1}} \mu_{T-1}(x^{T-1}, u^{T-2}) \log \phi^*_{T-1}(x^{T-1}, u^{T-2}) \\ &\quad \times \underbrace{\sum\nolimits_{\mathcal{U}_{T-1}} q^*_{T-1}(u_{T-1}|x^{T-1}, u^{T-1})}_{=1}. \end{aligned} \tag{10}$$

To complete the proof, we show that if (a) and (b) hold for the time step $t+1$, then they also hold for the time step t. Since (b) is hypothesized for $t+1$, using ρ^*_t, it is possible to write

$$\begin{aligned} &f(q_0, ..., q_t, q^*_{t+1}, ..., q^*_{T-1}; \nu^*) \\ &= \text{(constant)} + \sum_{\mathcal{X}^t} \sum_{\mathcal{U}^t} \mu_t(x^t, u^{t-1}) q_t(u_t|x^t, u^{t-1}) \\ &\qquad \times \left(\log \frac{q_t(u_t|x^t, u^{t-1}}{\nu^*_t(u_t|\tilde{x}^t, u^{t-1})} + \rho^*_t(x^t, u^t) \right) \end{aligned}$$

where "constant" is the term that does not depend on q_t. Lemma 1.1 is applicable once again to show that the minimizer q^*_t satisfies (8). Statement (b) for the time step t can be shown by the direct substitution. Details are similar to (10).