Value Functions for Temporal Logic: Optimal Policies and Safety Filters

Oswin So*,1, William Sharpless*,2, Sylvia Herbert[2], Chuchu Fan[1]

***Abstract*— While Bellman equations for basic reach, avoid, and reach-avoid problems are well studied, the relationship between value optimality and policy optimality becomes subtle in the undiscounted infinite-horizon setting, particularly for more complicated tasks. Greedily maximizing the Q-function can produce policies that indefinitely defer task completion for reach-avoid problems, or equivalently, Until specifications, even when the value function is optimal. Building upon recent results decomposing the value function for temporal logic (TL) into a graph of constituent value functions, we construct non-Markovian policies based on state history that avoid this pathology and prove their optimality with respect to the quantitative robustness score for nested Until, Globally, and Globally-Until specifications. We further show how the Q function can serve as a safety filter for complex TL specifications, extending prior results beyond simple avoid or reach-avoid tasks.**

I. Introduction and Related Work

The value function is central to optimal control and reinforcement learning (RL), encoding the optimal objective over action sequences. While it can be computed via dynamic programming (DP), this quickly becomes intractable in high dimensions. Instead, RL leverages the Bellman equation (BE), a recursive characterization of the value that enables approximation without grid-based DP. The associated Q-function further allows optimal policies to be obtained via single-step maximization. This framework is typically applied when objectives are defined by temporal sums of rewards (or costs), as in Markov Decision Processes (MDPs).

While this approach has been successful, the additive combination of rewards and costs in RL can be difficult; the optimal trajectory may incur significant cost so long as it obtains high reward, leading practitioners to iteratively tune hyperparameters until desirable performance is observed. To overcome this limitation, values based on temporal maxima and minima have been introduced from the field of Hamilton-Jacobi reachability (HJR) [1, 2]. In recent years, several works have integrated the max-min BE of HJR into RL frameworks to design safe and performant algorithms [3–6].

HJR characterizes values for basic safety and liveness tasks, called the reach, avoid and reach-avoid problems [7]. Notably, the objectives of these values are equivalent to the quantitative semantics of basic predicates in Temporal Logic (TL) [8–11], namely eventually (F), always (G), and until (U) [12]. TL itself is merely a specification language but offers an attractive way to automatically translate and reduce complex task specifications into values for RL.

Fig. 1: Hardware rollouts from a Crazyflie drone using an unsafe task policy, the task policy filtered for safety, and the task policy filtered for the specification. Filtering for safety maintains safety but violates the specification. Our proposed safety filter for TL specifications guarantees the specification is satisfied.

Recently, it was demonstrated that HJR and the HJ-RL algorithms can be extended to broader classes of TL via algebraic decomposition of the value into a graph of values [13, 14]. This was first considered in [13], where the dual-objective values, called the reach-reach (F ∧ F) and reach-always-avoid (F ∧ G) problems, proved to decompose into coupled sets of BEs, and later extended to a broad class of TL specifications in [14]. These methods scale the approximation of the value in high-dimension and model free settings to complex task specifications.

However, a novel challenge arises in these settings: simple single-step maximization of the corresponding Q function does not necessarily yield the optimal action sequence for infinite-horizon trajectories (see Ex. 1), a fundamentally non-Markovian problem [13]. While decomposition allows one to solve the value for infinite trajectories, it remains to show when the optimal inputs must "switch" from single-step maximization of the current value to a constituent value in the graph, corresponding to the remaining unsatisfied and infinite-horizon specifications.

In [13], a Markovian policy is derived with an efficient state-augmentation that tracks relevant quantitative semantics and yields the optimal switching condition. However, this approach is derived in the dual-objectives and is difficult to generalize to complex formulae, particularly those that include reach-avoid-loop specifications (GU). In this work, we propose a generalization of the value and Q-function to state histories to compute non-Markovian policies optimal

[1]Oswin So and Chuchu Fan are with the Massachusetts Institute of Technology, Cambridge, MA 02139, USA. {oswinso, chuchu}@mit.edu

[2]William Sharpless and Sylvia Herbert are with the University of California, San Diego, CA 92093, USA. {wsharpless, sherbert}@ucsd.edu

with respect to any state history. This approach extends to the broad class of TL specifications for which the value may be decomposed [14]. Moreover, this approach offers the novel ability to synthesize safety filters for a user-defined nominal policy that (1) guarantee specification satisfaction in a least-restrictive manner and (2) do so in a manner that is optimal given the prior sub-optimal path of the nominal policy.

In this work, the optimality of the policy over histories and the guarantee of satisfaction are made possible by augmenting the system with a temporal state (Def. 4). For completeness, we re-derive the policies in [13] for the Until predicate via the novel framework and we show the approach proposed in this work extends to the larger class of formulae defined in Def. 3. Ultimately, this work offers a practical approach to generate an optimal policy for the value of a broad class of TL formulae, and demonstrates the result may be used to filter a nominal policy to ensure complex task satisfaction.

The contributions of this work are as follows:

1) We generalize the optimal value function and Q function (Def. 1) and optimal policy (Def. 2) for state histories to handle the non-Markovian nature of the problem.
2) For a broad class of TL formulae (Def.3), we use this framework to construct a compositional optimal policy for the associated robustness metric (Thms. 1-8).
3) Using the Q function over (sub-optimal) history yields a least-restrictive safety filter for a nominal policy to ensure satisfaction (Thms. 9,10).

II. Preliminaries

A. Problem Setting and Notation

We consider a discrete-time system $x_{t+1} = f(x_t, a_t)$ with state $x_t \in \mathcal{X} \subseteq \mathbb{R}^n$ and action $a_t \in \mathcal{A} \subseteq \mathbb{R}^m$. Let $x_{0:t} \in \mathcal{X}^{t+1}$ and $a_{0:t} \in \mathcal{A}^{t+1}$ denote state and action trajectories of length $t+1$, and use $x_t = x_{t:t} \in \mathcal{X}$ and $a_t = a_{t:t} \in \mathcal{A}$ for brevity. Let $\mathcal{X}^+ \coloneqq \bigcup_{t=1}^{\infty} \mathcal{X}^t$ be the set of all finite-length state trajectories. For a (non-Markovian) policy $\pi : \mathcal{X}^+ \to \mathcal{A}$, we say that the infinite trajectory $x^\pi_{0:\infty}$ is generated by π from $x_{0:t} \in \mathcal{X}^+$ if $x^\pi_{0:t} = x_{0:t}$ and $x^\pi_{k+1} = f(x^\pi_k, \pi(x^\pi_{0:k}))$ for all $k \geq t$. In general, we use the word "policy" to refer to a non-Markovian policy (dependent on history), and specify when it is Markovian [1].

We use Linear Temporal Logic (LTL) [15] specifications augmented with quantitative semantics via the robustness score [9] from Signal Temporal Logic (STL) [16]. LTL formulae comprise atomic propositions p (Boolean variables that depend on the current state), logical operators ($\wedge$, $\vee$, $\neg$) and temporal operators Until (U) and Next (X), given by the following grammar in Backus-Naur Form:

$$\phi \coloneqq \top \mid p \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid \phi_1 \mathsf{U} \phi_2 \mid \mathsf{X}\phi, \tag{1}$$

where $\top$ is the Boolean constant true. Other temporal operators can be derived from these, e.g., "Finally" F and "Globally" G. Let Ψ denote the set of all LTL formulae, and let Prop denote the set of *propositional formulae*, i.e., formulae that do not contain temporal operators. We write propositional formulae in italics (e.g., p), while general formulae are in serif (e.g., r).

We additionally associate every TL formula $\psi \in \Psi$ with a quantitative semantic called the *robustness score* [9] $\rho_{[\psi]} : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ that quantifies the satisfaction of ψ by an infinite length state trajectory $x_{0:\infty} \in \mathcal{X}^{\mathbb{N}}$, with $\rho_{[\psi]}(x_{0:\infty}) \geq 0$ if and only if $x_{0:\infty}$ satisfies ψ. Specifically, we associate every atomic proposition p with a function $p : \mathcal{X} \to \mathbb{R}$ such that $p(x_0) = \rho_{[p]}(x_{0:\infty})$ for any trajectory $x_{0:\infty}$. For brevity, **we abuse notation to refer to a formula and its robustness score interchangeably**, e.g., $\psi(x_{0:\infty}) \coloneqq \rho_{[\psi]}(x_{0:\infty})$, with the meaning clear from context. ρ is defined recursively as follows [9]:

$$\begin{aligned}
&\rho_{[\neg\psi]}(x_{0:\infty}) \coloneqq -\psi(x_{0:\infty}) && \rho_{[\psi_1 \wedge \psi_2]}(x_{0:\infty}) \coloneqq \psi_1(x_{0:\infty}) \wedge \psi_2(x_{0:\infty}) \\
&\rho_{[\mathsf{X}\psi]}(x_{0:\infty}) \coloneqq \psi(x_{1:\infty}) && \rho_{[\psi_1 \vee \psi_2]}(x_{0:\infty}) \coloneqq \psi_1(x_{0:\infty}) \vee \psi_2(x_{0:\infty}) \\
&\rho_{[\mathsf{G}\psi]}(x_{0:\infty}) \coloneqq \min_{t \geq 0} \psi(x_{t:\infty}) && \rho_{[\psi_1 \mathsf{U} \psi_2]}(x_{0:\infty}) \coloneqq \max_{t \geq 0} \psi_2(x_{t:\infty}) \wedge \min_{0 \leq s < t} \psi_1(x_{s:\infty}) \\
&\rho_{[\mathsf{F}\psi]}(x_{0:\infty}) \coloneqq \max_{t \geq 0} \psi(x_{t:\infty})
\end{aligned} \tag{2}$$

where we have abused notation and used $\wedge$, $\vee$ to denote $\min$, $\max$ respectively.

B. Optimal Control for Temporal Logic

We now consider the problem of maximizing the robustness score of a given LTL formula $\psi \in \Psi$, i.e.,

$$\max_{a_{0:\infty}} \rho_{[\psi]}(x_{0:\infty}), \quad x_{t+1} = f(x_t, a_t),\ \forall t \geq 0. \tag{3}$$

Following standard optimal control theory [17], we can define an optimal value function V and Q function Q for (3):

Definition 1. *For a TL formula ψ and $t \geq 0$, define the value function $\mathbf{V}^*_{[\psi]} : \mathcal{X}^+ \to \mathbb{R}$ (on histories) as*

$$\mathbf{V}^*_{[\psi]}(x_{0:t}) \coloneqq \max_{a_{t:\infty}} \rho_{[\psi]}(x_{0:\infty}), \tag{4}$$

and the Q function $\mathbf{Q}_{[\psi]} : \mathcal{X}^+ \times \mathcal{A} \to \mathbb{R}$ (on histories) as

$$\mathbf{Q}_{[\psi]}(x_{0:t}, a_t) \coloneqq \max_{a_{t+1:\infty}} \rho_{[\psi]}(x_{0:\infty}), \tag{5}$$

where $x_{k+1} = f(x_k, a_k)$ for $k \geq t$. This generalizes the conventional value function and Q function on states. In particular, for any $t \geq 0$, the value and Q function on states is recovered when the history has a length 1, i.e.,

$$V^*_{[\psi]}(x_t) \coloneqq \mathbf{V}^*_{[\psi]}(x_{t:t}), \quad Q_{[\psi]}(x_t, a_t) \coloneqq \mathbf{Q}_{[\psi]}(x_{t:t}, a_t). \tag{6}$$

Lemma 1. *For any $\psi \in \Psi$, fix $\bar{x}_{0:k} \in \mathcal{X}^+$ for $k \geq 0$, and $\bar{a}_k \in \mathcal{A}$. Let $\bar{x}_{k+1} = f(\bar{x}_k, \bar{a}_k)$. Then,*

$$\mathbf{Q}_{[\psi]}(\bar{x}_{0:k}, \bar{a}_k) = \mathbf{V}^*_{[\psi]}(\bar{x}_{0:k+1}). \tag{7}$$

To simplify the results, we assume the $\max$ in the definitions of $\mathbf{V}^*_{[\psi]}$, $\mathbf{Q}_{[\psi]}$, and robustness scores (2) are attainable. This can be guaranteed with the following assumption:[2]

[1] One can always augment the state with the (infinite-dimensional) history to make the policy Markovian. dHowever, this would be inefficient and we avoid this to distinguish the proposed Markovian approach

[2] The assumptions can be relaxed by using $\sup$ instead of $\max$ in (4) and (5), but this yields ϵ-optimal policies instead of optimal policies.

Assumption 1. *The robustness score of all atomic propositions takes values in a finite set.*

This generalizes the finite state space assumption in [13] to additionally enable policy construction in infinite state spaces. Crucially, the use of indicator functions for atomic propositions satisfies Asm. 1.

We next define optimality for our non-Markovian problem.

Definition 2. *A policy $\pi : \mathcal{X}^+ \to \mathcal{A}$ is **optimal** (over histories) for LTL formula $\psi \in \Psi$ if, for all histories $x_{0:t} \in \mathcal{X}^+$ for $t \geq 0$, the trajectory $x^\pi_{0:\infty}$ generated by π from $x_{0:t}$ achieves the maximal robustness score, i.e.,*

$$\rho_{[\psi]}(x^\pi_{0:\infty}) = \mathbf{V}^*_{[\psi]}(x_{0:t}). \tag{8}$$

Remark 1 (Notions of Optimality)**.** The notion of optimality (over histories) in Def. 2 is stronger than the conventional notion in MDPs, which only requires optimality over initial states as used in [13]. The two notions of optimality coincide for Markovian problems [17], but can differ for non-Markovian problems.

In this work, we construct optimal policies for the following fragment of TL formulas defined recursively.

Definition 3. *Define the set of solvable formulas $\mathcal{S}$ as the smallest set of TL formulas satisfying:*

- ***(S1) Propositional.*** *If $p \in \mathrm{Prop}$, then $p \in \mathcal{S}$.*
- ***(S2) Until.*** *If $p \in \mathrm{Prop}$ and $\varphi \in \mathcal{S}$, then $p \mathsf{U} \varphi \in \mathcal{S}$.*
- ***(S3) Next.*** *If $\varphi \in \mathcal{S}$, then $\mathsf{X}\varphi \in \mathcal{S}$.*
- ***(S4) Globally.*** *If $p \in \mathrm{Prop}$, then $\mathsf{G}p \in \mathcal{S}$.*
- ***(S5) Globally-Until.*** *If $p_i, r_i \in \mathrm{Prop}$ for $i = 1, \dots, N$ with $N \geq 1$, then $\mathsf{G}\big(\bigwedge_{i=1}^N p_i \mathsf{U} r_i\big) \in \mathcal{S}$.*
- ***(S6) Disjunction.*** *If $\varphi_1, \varphi_2 \in \mathcal{S}$, then $\varphi_1 \vee \varphi_2 \in \mathcal{S}$.*
- ***(S7) Propositional conjunction.*** *If $p \in \mathrm{Prop}$ and $\varphi \in \mathcal{S}$, then $p \wedge \varphi \in \mathcal{S}$.*

We will construct an optimal policy for each item in Def. 3. The case for Def. 3-(S1) is straightforward, since $p \in \mathrm{Prop}$ does not depend on any actions. In the following sections, we construct an optimal policy for remaining items in Def. 3 before showing that our proposed policy optimally solves the entire fragment $\mathcal{S}$ (Thm. 8).

III. Constructing an Optimal Policy for Until

We first construct an optimal policy for the Until operator, U, as in Def. 3-(S2). Until is a fundamental building block of LTL: every LTL formula can be rewritten using only the logical operators $\neg$, $\vee$ and the temporal operators U and X [15]. Yet, constructing an optimal policy for an Until specification is not straightforward, as we show next.

In discounted MDPs, the policy that greedily maximizes the Q function is optimal [17, Corollary 6.2.8]. However, as shown next, this is not true for Until specifications, as the greedy policy may indefinitely defer task completion, an issue also identified in recent works [18].

Counterexample 1. Consider a finite state space $\mathcal{X} = \{0, 1\}$, and a finite action space $\mathcal{A} = \{0, 1\}$ with dynamics $f(x, a) = a$. Consider the formula $\psi = \mathsf{F}\mathbb{1}_1 \equiv \top \mathsf{U} \mathbb{1}_1$, which requires the system to eventually reach state 1. Since state 1 can be reached in one step from any state, the Q function is equal to 1 for all history-action pairs, i.e., for any $x_{0:t} \in \mathcal{X}^+$ and a_t, $Q_{[\psi]}(x_{0:t}, a_t) = 1$. In particular, for any history $x_{0:t}$,

$$0 \in \operatorname*{argmax}_{a_t \in \mathcal{A}} Q_{[\psi]}(x_{0:t}, a_t). \tag{9}$$

However, the policy $\pi(x_{0:k}) = 0$ for all $x_{0:k}$ does not satisfy the specification ψ, since it results in the trajectory $x_k = 0$ for all $k \geq 0$, which never reaches 1 and thus has robustness score 0 despite $\mathbf{V}^*_{[\psi]}(x_{0:k}) = 1$ for all $x_{0:k}$.

To prevent indefinitely delaying score maximization, we seek a time-optimal policy that achieves the maximal score in minimum time, which we next explore. For the rest of Sec. III, we consider $\psi \coloneqq q \mathsf{U} \mathsf{r}$ for $q \in \mathrm{Prop}$, $\mathsf{r} \in \Psi$.

A. Finite Horizon Until

Consider the finite-horizon versions of the Until operator (as in STL [16] or metric temporal logic [19]), which does not suffer from this issue due to the finite horizon. Choosing the shortest horizon that still recovers the maximal score forces completion as soon as possible.

Definition 4 (Time Augmented System)**.** *Define the augmented state space $\tilde{\mathcal{X}} \coloneqq \mathcal{X} \times \mathbb{Z}$ with state $\tilde{x} = [x, \mathtt{t}] \in \tilde{\mathcal{X}}$, where $\mathtt{t} \in \mathbb{Z}$ is a timer that decreases every step, and the dynamics function $\tilde{f} : \tilde{\mathcal{X}} \times \mathcal{A} \to \tilde{\mathcal{X}}$ is given by*

$$\tilde{f}([x, \mathtt{t}], a) = [f(x, a), \mathtt{t} - 1]. \tag{10}$$

For conciseness, we write the trajectory $\tilde{x}_{0:\infty}$ as $[x_{0:\infty}, \mathtt{t}_0]$ for $\tilde{x}_0 = [x_0, \mathtt{t}_0]$.

We use this timer to enforce the time limit within which a TL formula must be satisfied. The timer is initialized at $\mathtt{t}_0$, and satisfaction must occur by $\mathtt{t} = 0$. We formalize this as $\tilde{\psi} \coloneqq q \mathsf{U} (\mathsf{r} \wedge r_{\mathtt{t} \geq 0})$, where $r_{\mathtt{t} \geq 0}(x, \mathtt{t}) = \infty$ for $\mathtt{t} \geq 0$ and $-\infty$ otherwise. Note that $\rho_{[\tilde{\psi}]}([x_{0:\infty}, \mathtt{t}_0]) = -\infty$ if $\mathtt{t}_0 < 0$. In other words, r must be satisfied within the time limit specified by the timer to achieve a positive robustness score, which resembles a finite-horizon Until specification. We now make this connection precise.

Lemma 2. *$\tilde{\psi}$ has the same semantics as the timed Until operator $\mathsf{U}_{[0, \mathtt{t}_0]}$:*

$$\rho_{[\tilde{\psi}]}([x_{0:\infty}, \mathtt{t}_0]) = \max_{0 \leq t \leq \mathtt{t}_0} \min\{\mathsf{r}(x_{t:\infty}), \min_{0 \leq k < t} q(x_k)\}. \tag{11}$$

where the maximization over t only considers the interval $[0, \mathtt{t}_0]$ instead of $[0, \infty)$.

Proof. We prove this by induction on $\mathtt{t}_0$. For $\mathtt{t}_0 = 0$,

$$\begin{aligned} \rho_{[\tilde{\psi}]}([x_{0:\infty}, 0]) &= \mathsf{r}(x_{0:\infty}) \vee \big(q(x_0) \wedge -\infty\big), \\ &= \mathsf{r}(x_{0:\infty}) = \max_{0 \leq t \leq 0} \min\{\mathsf{r}(x_{t:\infty}), \min_{0 \leq k < t} q(x_k)\}. \end{aligned}$$

Now, assume the statement holds for $\mathtt{t}_0 = n$. Then,

$$\begin{aligned}
&\rho_{[\tilde{\psi}]}([x_{0:\infty}, n+1]) \\
&= \mathsf{r}(x_{0:\infty}) \vee \big(q(x_0) \wedge \rho_{[\tilde{\psi}]}([x_{1:\infty}, n])\big), \\
&= \mathsf{r}(x_{0:\infty}) \vee \big(q(x_0) \wedge \max_{0\le t\le n} \min\{\mathsf{r}(x_{t+1:\infty}), \min_{1\le k<t+1} q(x_k)\}\big), \\
&= \max_{0\le t\le n+1} \min\{\mathsf{r}(x_{t:\infty}), \min_{0\le k<t} q(x_k)\}.
\end{aligned}$$

Therefore, by induction, it holds for all $\mathtt{t}_0 \ge 0$. □

Thus, for initial state $[x_0, \mathtt{t}_0]$, we use the timed until operator $\mathsf{U}_{[0,\mathtt{t}_0]}$ as a shorthand such that $q\mathsf{U}_{[0,\mathtt{t}_0]}\mathsf{r}$ refers to the formula $q\mathsf{U}(\mathsf{r} \wedge r_{\mathtt{t}\ge 0})$. Note that for any $x_{0:\infty}$ and $\mathtt{t}_0$,

$$\rho_{[q\mathsf{U}_{[0,\mathtt{t}_0]}\mathsf{r}]}([x_{0:\infty}, \mathtt{t}_0]) \le \rho_{[q\mathsf{U}\mathsf{r}]}(x_{0:\infty}). \tag{12}$$

We next look at when the finite-horizon Until recovers the infinite-horizon Until. Specifically, we identify a "witness time" when the supremum in the definition of U (2) is attained.

Lemma 3. *Under Asm. 1, for any $x_{0:\infty}$, there exists a smallest witness time $\tau = \sigma_{[\psi]}(x_{0:\infty}) \ge 0$ where*

$$\rho_{[\psi]}(x_{0:\infty}) \coloneqq \sup_{t\ge 0} \min\{\mathsf{r}(x_{t:\infty}), \min_{0\le k<t} q(x_k)\}, \tag{13}$$

$$= \min\{\mathsf{r}(x_{\tau:\infty}), \min_{0\le k<\tau} q(x_k)\}, \tag{14}$$

In other words, the supremum in the definition of U is first attained at some finite time $\sigma_{[\psi]}(x_{0:\infty})$.

Proof. By Asm. 1, the robustness score takes values in a finite set. This is preserved under min. Thus, the blue term in (13) also takes values in a finite set and attains its supremum at a finite time τ. □

The above witness time holds for any trajectory. We now define an optimal witness time, which is the smallest among all trajectories that achieve the optimal value.

Definition 5. *For $x_{0:t} \in \mathcal{X}^+$, define the optimal witness time $\sigma^*_{[\psi]} : \mathcal{X}^+ \to \mathbb{Z}_{\ge 0}$ as*

$$\sigma^*_{[\psi]}(x_{0:t}) = \min\{\sigma_{[\psi]}(x_{0:\infty}) \mid \rho(x_{0:\infty}) = \mathbf{V}^*_{[\psi]}(x_{0:t})\}. \tag{15}$$

In other words, $\sigma^*_{[\psi]}(x_{0:t})$ is the smallest witness time among all trajectories starting at $x_{0:t}$ that achieve the optimal value $\mathbf{V}^*_{[\psi]}(x_{0:t})$.

Importantly, $\tilde{\psi}$ has the same value as ψ at $\mathtt{t}_0 = \sigma^*_{[\psi]}$.

Lemma 4. *Let $\mathtt{t}_s = \sigma^*_{[\psi]}(x_{0:s}) - s$ be the optimal witness time for $x_{0:s}$, and $\mathtt{t}_0 = \mathtt{t}_s + s = \sigma^*_{[\psi]}(x_{0:s})$. Then,*

$$\max_{a_t} Q_{[\tilde{\psi}]}([x_{0:s}, \mathtt{t}_0], a_t) = \mathbf{V}^*_{[\tilde{\psi}]}([x_{0:s}, \mathtt{t}_0]) = \mathbf{V}^*_{[\psi]}(x_{0:s}). \tag{16}$$

We will show how relaxing the minimum-time requirement enables construction of a *Markovian* optimal policy that achieves the maximal score without requiring keeping track of the timer state $\mathtt{t}$ (Sec. III-C).

In the next subsections, we construct an optimal policy for ψ. We first consider a simplified formula by replacing $\mathsf{r} \in \Psi$ with $V^*_{[\mathsf{r}]} \in \mathrm{Prop}$. We then construct a policy for ψ by combining the prefix of the simplified policy with a suffix that handles the remaining r.

B. Policy for Until with AP by replacing r with $V^*_{[\mathsf{r}]}$

A challenge with ψ is that $\mathsf{r} \in \Psi$ is arbitrary. Thus, we first consider a simple but related formula φ. For the same $q \in \mathrm{Prop}$ and $\mathsf{r} \in \Psi$, consider $\varphi \coloneqq q\mathsf{U}V^*_{[\mathsf{r}]}$ and its finite-horizon counterpart $\tilde{\varphi} \coloneqq q\mathsf{U}_{[0,\mathtt{t}_0]}V^*_{[\mathsf{r}]}$, which are defined similarly to ψ and $\tilde{\psi}$. We first show the optimal value is unchanged.

Lemma 5. *For any $n \ge 0$, and any $(x_{0:t}, \mathtt{t}_0) \in \mathcal{X}^+ \times \mathbb{Z}_{\ge 0}$,*

$$\mathbf{V}^*_{[\tilde{\psi}]}([x_{0:t}, \mathtt{t}_0]) = \mathbf{V}^*_{[\tilde{\varphi}]}([x_{0:t}, \mathtt{t}_0]) \tag{17}$$

In particular, this implies both $\mathbf{V}^*_{[\psi]}(x_{0:t}) = \mathbf{V}^*_{[\varphi]}(x_{0:t})$ and $\sigma^*_{[\psi]}(x_{0:t}) = \sigma^*_{[\varphi]}(x_{0:t})$, for all $x_{0:t} \in \mathcal{X}^+$.

We now construct a policy for φ and show that it is optimal.

Definition 6. *Define the Markovian policy $\pi_{[\tilde{\varphi}]} : \tilde{\mathcal{X}} \to \mathcal{A}$ as*

$$\pi_{[\tilde{\varphi}]}([x, \mathtt{t}]) \coloneqq \arg\max_a Q_{[\tilde{\varphi}]}([x, \mathtt{t}], a). \tag{18}$$

Now, define the non-Markovian policy $\pi_{[\varphi]} : \mathcal{X}^+ \to \mathcal{A}$ as

$$\pi_{[\varphi]}(x_{0:k}) \coloneqq \pi_{[\tilde{\varphi}]}([x_k, \sigma^*_{[\varphi]}(x_{0:k}) - k]). \tag{19}$$

Lemma 6. *For any trajectory $x_{0:\infty}$ and any $\mathtt{t}_0$, let $s \in [0, \mathtt{t}_0]$, and denote $c_\tau \coloneqq \min_{0\le k<\tau} q(x_k)$. Then,*

$$\begin{aligned}
&\rho_{[\tilde{\psi}]}([x_{0:\infty}, \mathtt{t}_0]) \\
&= \max_{0\le\tau<s} \min\{\mathsf{r}(x_{\tau:\infty}), c_\tau\} \vee \min\{c_\tau, \rho_{[\tilde{\varphi}]}([x_{s:\infty}, \mathtt{t}_s])\},
\end{aligned} \tag{20}$$

and

$$\begin{aligned}
&\rho_{[\tilde{\varphi}]}([x_{0:\infty}, \mathtt{t}_0]) \\
&= \max_{0\le\tau<s} \min\{V^*_{[\mathsf{r}]}(x_\tau), c_\tau\} \vee \min\{c_\tau, \rho_{[\tilde{\varphi}]}([x_{s:\infty}, \mathtt{t}_s])\}.
\end{aligned} \tag{21}$$

Moreover, by maximizing over $a_{s:\infty}$, we get

$$\begin{aligned}
&\mathbf{V}^*_{[\tilde{\psi}]}([x_{0:s}, \mathtt{t}_0]) = \mathbf{V}^*_{[\tilde{\varphi}]}([x_{0:s}, \mathtt{t}_0]) \\
&= \max_{0\le\tau<s} \min\{V^*_{[\mathsf{r}]}(x_\tau), c_\tau\} \vee \min\{c_\tau, V^*_{[\tilde{\varphi}]}([x_s, \mathtt{t}_s])\}.
\end{aligned} \tag{22}$$

A challenge to proving the optimality of $\pi_{[\varphi]}$ is that $\sigma^*_{[\varphi]}(x_{0:k}) - k$ may not follow the dynamics of the timer state $\mathtt{t}$ in the augmented system for the optimality of $\pi_{[\tilde{\varphi}]}$ to transfer to $\pi_{[\varphi]}$. We show next that this does not happen.

Lemma 7. *Let $\bar{x}_{0:\infty}$ be generated by $\pi_{[\varphi]}$ from $\bar{x}_{0:s}$. Then,*

$$\sigma^*_{[\varphi]}(\bar{x}_{0:k}) = \sigma^*_{[\varphi]}(\bar{x}_{0:s}), \quad \forall k \ge s. \tag{23}$$

We are now ready to prove the optimality of $\pi_{[\varphi]}$.

Theorem 1. *$\pi_{[\varphi]}$ is optimal (Def. 2) for TL formula φ.*

Proof Sketch. Fix $\bar{x}_{0:s} \in \mathcal{X}^+$, and let $\bar{x}_{0:\infty}$ be the trajectory generated by $\pi_{[\varphi]}$ from $\bar{x}_{0:s}$. Let $\mathtt{t}_s \coloneqq \sigma^*_{[\varphi]}(\bar{x}_{0:s}) - s$, and $\mathtt{t}_0 = \sigma^*_{[\varphi]}(\bar{x}_{0:s})$. By (12) and Lem. 4,

$$\rho_{[\tilde{\varphi}]}([\bar{x}_{0:\infty}, \mathtt{t}_0]) \le \rho_{[\varphi]}(\bar{x}_{0:\infty}) \le \mathbf{V}^*_{[\varphi]}(\bar{x}_{0:s}) = \mathbf{V}^*_{[\tilde{\varphi}]}([\bar{x}_{0:s}, \mathtt{t}_0]).$$

It suffices to show the first and last term are equal. By Lem. 6, it is enough to prove $\rho_{[\tilde{\varphi}]}([\bar{x}_{s:\infty}, \mathtt{t}_s]) = V^*_{[\tilde{\varphi}]}([\bar{x}_s, \mathtt{t}_s])$. By Lem. 7, the quantity $\sigma^*_{[\varphi]}(\bar{x}_{0:k}) - k$ simulates the timer in the augmented system. Hence the

suffix follows the greedy finite-horizon policy $\pi_{[\bar{\varphi}]}$.

A backward induction on the timer therefore yields $\rho_{[\bar{\varphi}]}([\bar{x}_{k:\infty}, \mathtt{t}_k]) = V^*_{[\bar{\varphi}]}([\bar{x}_k, \mathtt{t}_k])$ for all $k \in [s, \sigma^*_{[\varphi]}(\bar{x}_{0:s})]$, and in particular at $k = s$. Therefore $\rho_{[\bar{\varphi}]}([\bar{x}_{s:\infty}, \mathtt{t}_s]) = V^*_{[\bar{\varphi}]}([\bar{x}_s, \mathtt{t}_s])$, and the claim follows from the initial inequality chain. □

C. A Markovian Policy for φ

While $\pi_{[\varphi]}$ (19) in Def. 6 is optimal, evaluating the control at state x_k requires keeping track of the timer state in the augmented state space, resulting in a non-Markovian policy on the original state space $\mathcal{X}$. Similar to [13], we can construct an alternative *Markovian* optimal policy $\hat{\pi}_{[\varphi]}$ that depends only on the current state x_k.

Definition 7. *Define the Markovian policy $\hat{\pi}_{[\varphi]}$ as*

$$\hat{\pi}_{[\varphi]}(x) \coloneqq \pi_{[\bar{\varphi}]}([x, \sigma^*_{[\varphi]}(x)]), \tag{24}$$

where $\pi_{[\varphi]}$ is defined as in (18)*.*

Theorem 2. *$\hat{\pi}_{[\varphi]}$ is also optimal for φ.*

Remark 2. By Lem. 4, the infinite-horizon value function can be solved by taking a limit of the values of finite-horizon problems. Consequently, this process yields the value functions and optimal actions for all states for all horizon lengths. $\hat{\pi}_{[\varphi]}$ can be computed during this process by storing the action that maximizes the Q function for the first step n where $V^*_{[\bar{\varphi}]}([x, n])$ converges.

D. Policy for Until with General r

We now tackle the original problem of finding an optimal policy for general $\mathsf{r} \in \Psi$. We propose two candidate policies corresponding to the two optimal policies for φ defined in Def. 6 and Def. 7, respectively. For both, we make the following assumption, which we later show is satisfied for a large class of TL formula (Sec. VI).

Assumption 2. *There exists a known (potentially non-Markovian) optimal policy $\pi_{[\mathsf{r}]}$ for r, i.e., for $x_{0:\infty}$ generated by $\pi_{[\mathsf{r}]}$ from $x_{0:s}$,*

$$\rho_{[\mathsf{r}]}(x_{0:\infty}) = V^*_{[\mathsf{r}]}(x_{0:s}). \tag{25}$$

Note that Asm. 2 holds for $\mathsf{r} \in \mathrm{Prop}$ (Def. 3-(S1)).

Definition 8. *Define the policy $\pi_{[\psi]}$ as*

$$\pi_{[\psi]}(x_{0:k}) \coloneqq \begin{cases} \pi_{[\varphi]}(x_{0:k}), & n_k - k > 0, \\ \pi_{[\mathsf{r}]}(x_{n_k:k}), & n_k - k \leq 0, \end{cases} \tag{26}$$

where $n_k \coloneqq \sigma^*_{[\psi]}(x_{0:k})$ and $\pi_{[\varphi]}$ is defined in (19)*.*

Definition 9. *Define the policy $\hat{\pi}_{[\psi]}$ as*

$$\hat{\pi}_{[\psi]}(x_{0:k}) \coloneqq \begin{cases} \hat{\pi}_{[\varphi]}(x_k), & k < n^*, \\ \pi_{[\mathsf{r}]}(x_{n^*:k}), & k \geq n^*, \end{cases} \tag{27}$$

where $n^* = \min\{n : \sigma^*_{[\varphi]}(x_n) = 0\}$ is the smallest time such that $\sigma^*_{[\varphi]}(x_{n^*}) = 0$, and $\hat{\pi}_{[\varphi]}$ is defined in (24)*.*

Remark 3. *$\hat{\pi}_{[\psi]}$ is Markovian on the original state space $\mathcal{X}$ if $\pi_{[\mathsf{r}]}$ is Markovian on $\mathcal{X}$.*

Theorem 3. *$\pi_{[\psi]}$ is an optimal policy for ψ.*

Proof Sketch. Fix a history $x_{0:s}$ and let $n^* \coloneqq \sigma^*_{[\psi]}(x_{0:s})$. The proof splits into two cases.

($n^* \geq s$): The optimal witness has not yet occurred. From time s onward, the problem is equivalent to a finite-horizon Until problem with remaining horizon $n^* - s$. Since $\pi_{[\psi]}$ matches $\pi_{[\varphi]}$ until the witness is reached, we use the optimality of $\pi_{[\varphi]}$ (Thm. 1) and the equivalence in value of ψ and φ (Lem. 5) to conclude that $\pi_{[\psi]}$ also attains the optimal value for this finite-horizon problem. When the witness time n^* is reached, the policy switches to the suffix controller $\pi_{[\mathsf{r}]}$ which is optimal for r. After showing that $n_k - k$ correctly simulates the timer state, we conclude that the suffix after time n^* is exactly the trajectory generated by $\pi_{[\mathsf{r}]}$ from x_{n^*}, and therefore attains the optimal r-value. Finally, the full trajectory attains the optimal value $\mathbf{V}^*_{[\psi]}(x_{0:s})$ by the history-level decomposition of Until.

($n^* < s$): The optimal witness already lies in the fixed prefix. Thus the Until objective has already switched to the suffix problem for r at time n^*. By construction, $\pi_{[\psi]}$ follows the optimal policy for r from then on, so the resulting suffix achieves the optimal r-value. Combining this with the fixed prefix contribution $\min_{0 \leq i < n^*} q(x_i)$ shows again that the full trajectory attains $\mathbf{V}^*_{[\psi]}(x_{0:s})$.

Combining both results yields the desired result. □

We omit the proof for $\hat{\pi}_{[\psi]}$ but it follows similarly.

Theorem 4. *$\hat{\pi}_{[\psi]}$ is an optimal policy for ψ.*

Remark 4 (Computing the policy without witness times)**.** The optimal witness time $n^* = \sigma^*_{[\psi]}(x_0)$ (Def. 5) need not be computed explicitly, since it is defined as the first time τ such that $\min\{V^*_{[\mathsf{r}]}(x_\tau), \min_{0 \leq k < \tau} q(x_k)\} \geq \min\{V^*_{[\psi]}(x_{\tau+1}), \min_{0 \leq k < \tau+1} q(x_k)\}$ by (14). Thus, it suffices to track $\min_{0 \leq k < \tau} q(x_k)$ and check the above condition at each time step τ. Similarly, $\sigma^*_{[\varphi]}(x_0) = 0$ if and only if $V^*_{[\mathsf{r}]}(x_0) \geq \max_a q(x_0) \wedge V^*_{[\varphi]}(f(x_0, a))$. Thus, it suffices to check the above condition at each time to determine when to "switch" to $\pi_{[\mathsf{r}]}$.

We compare $\pi_{[\psi]}$ (26) and $\hat{\pi}_{[\psi]}$ (27) in Figure 2. $\pi_{[\psi]}$ requires tracking either the timer $\mathtt{t}$ or the cumulative minimum $\min_{0 \leq k < \tau} q(x_k)$ per Rmk. 4, but switches to $\pi_{[\mathsf{r}]}$ in the fewest steps. In contrast, $\hat{\pi}_{[\psi]}$ only requires storing a single boolean to track whether $\sigma^*_{[\varphi]}(x_k) = 0$ has been reached, but potentially takes more steps to switch to $\pi_{[\mathsf{r}]}$.

IV. Policies for Next, Disjunction and Conjunction

Next, we construct optimal policies for the X, $\vee$ and $\wedge$ operators to tackle Def. 3-(S3), 3-(S6) and 3-(S7). Fortunately, these operators do not suffer from the deferral issue of U as in Counterexample 1 and thus have more straightforward optimal policies when the operands are propositional. However, in

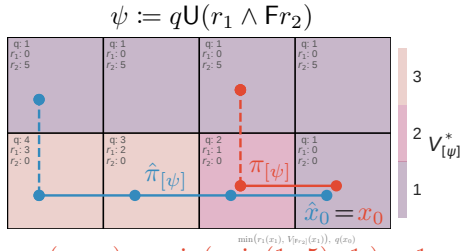

Fig. 2: **Comparing $\pi_{[\psi]}$ and $\hat{\pi}_{[\psi]}$.** Both $\pi_{[\psi]}$ and $\hat{\pi}_{[\psi]}$ achieve the optimal robustness at $x_0 = \hat{x}_0$. However, $\pi_{[\psi]}$ achieves this in 2 steps, while $\hat{\pi}_{[\psi]}$ reaches a higher $r_1(\hat{x}_3) = 3$, albeit taking 4 steps.

the general case where the operands are TL formulas, the optimal policies can be more complex. To handle this, we use the same technique as in Sec. III-D and assume the existence of optimal policies for the operand formulas (see Asm. 2).

Definition 10. *Let $\pi_{[\mathsf{r}]}$ be an optimal policy for r. Then, define $\pi_{[\mathsf{X}\mathsf{r}]}$ as*

$$\pi_{[\mathsf{X}\mathsf{r}]}(x_{0:k}) := \begin{cases} \operatorname{argmax}_{a\in\mathcal{A}} V^*_{[\mathsf{r}]}(f(x_0, a)), & k = 0, \\ \pi_{[\mathsf{r}]}(x_{1:k}), & k > 0. \end{cases} \tag{28}$$

Definition 11. *For set I, let $\pi_{[\mathsf{r}_i]}$ be an optimal policy for r_i for each $i \in I$. Define the policy $\pi_{[\bigvee_{i\in I}\mathsf{r}_i]}$ as*

$$\pi_{[\bigvee_{i\in I}\mathsf{r}_i]}(x_{0:k}) := \pi_{[\mathsf{r}_{i^*}]}(x_{0:k}), \tag{29}$$

where $i^* \in \operatorname{argmax}_{i\in I} V^*_{[\mathsf{r}_i]}(x_0)$ is any maximizer.

Definition 12. *Let $\pi_{[\mathsf{r}]}$ be an optimal policy for r. Define the policy $\pi_{[q\wedge\mathsf{r}]}$ as*

$$\pi_{[q\wedge\mathsf{r}]} = \pi_{[\mathsf{r}]}. \tag{30}$$

Theorem 5. *The policies defined in Def. 10–12 are optimal for $\mathsf{X}\mathsf{r}$, $\bigvee_{i\in I}\mathsf{r}_i$ and $q \wedge \mathsf{r}$, respectively.*

V. Policy for Globally

Finally, we construct an optimal policy for the G operator, another important operator in TL specifications due to its ability to express safety and invariance properties [15].

A. Globally with Propositional Operand

When the operand is propositional (Def. 3-(S4)), the greedy Markovian policy is optimal.

Definition 13. *Define the Markovian policy $\pi_{[\mathsf{G}q]} : \mathcal{X} \to \mathcal{A}$ as*

$$\pi_{[\mathsf{G}q]}(x) := \operatorname*{argmax}_{a\in\mathcal{A}} Q_{[\mathsf{G}q]}(x, a). \tag{31}$$

Theorem 6. *$\pi_{[\mathsf{G}q]}$ is an optimal policy for $\mathsf{G}q$.*

B. Globally with Conjunctions of Until operands

Finally, we tackle the case of GU (Def. 3-(S5)). When the operand contains the U operator, this encodes a Büchi acceptance criterion that requires infinitely many visits to a set of states [20], which requires a more complex policy to solve the deferral issue as in Counterexample 1. Using the recursive relationships for G, we have that

$$\mathsf{G}(q\mathsf{U}r) \equiv (q\mathsf{U}r) \wedge \mathsf{X}\mathsf{G}(q\mathsf{U}r). \tag{32}$$

Thus, GU can be viewed as U with an infinite tail of GU. Consequently, a naively constructed policy for GU by maximizing the Q function can also result in a policy that infinitely defers satisfying the U component, and thus fails to satisfy the GU formula.

Let I be a nonempty finite set. We now consider $\psi := \mathsf{G}(\bigwedge_{i\in I} q_i\mathsf{U}r_i)$. Without loss of generality, we take $I = \{1, 2, \ldots, N\}$. The following lemma, taken from [14], makes precise this recursive structure of GU.

Lemma 8 ([14, Theorem 3])**.** *For any $i \in I$,*

$$\psi \equiv \tilde{q}_i\mathsf{U}(\tilde{r}_i \wedge \mathsf{X}\psi), \quad \tilde{q}_i := q_i \wedge w_i, \quad \tilde{r}_i := r_i \wedge w_i \tag{33}$$

where $w_i := \bigwedge_{j\in I\setminus\{i\}}(q_j \vee r_j)$.

Using Lem. 8, we can also show the following result.

Corollary 1. *Let $\chi_i[\cdot]$ denote the RHS of (33) but with ψ replaced by an arbitrary formula, i.e.,*

$$\chi_i[\theta] := \tilde{q}_i\mathsf{U}(\tilde{r}_i \wedge \mathsf{X}\theta). \tag{34}$$

Then, $\chi_i[\psi] \equiv \psi$ for all $i \in I$, and

$$\psi \equiv \chi_1[\psi] \equiv \chi_1[\chi_2[\cdots\chi_N[\psi]\cdots]]. \tag{35}$$

We now construct an optimal policy $\pi_{[\psi]}$ for ψ by making use of this infinite nested structure in Cor. 1. Specifically, we will iteratively visit each U component in sequence, then repeat this process infinitely many times, which ensures that we satisfy the Büchi acceptance criteria of ψ.

Note. The order that the U components are visited can be arbitrary (see [14]). We use ascending order for simplicity.

Definition 14. *For $i \in I$, let the **until** formula ϕ_i denote the i-th nested U formula in Cor. 1, i.e.,*

$$\phi_i := \chi_i[\chi_{i+1}[\cdots\chi_N[\psi]\cdots]] = (q_i\wedge w_i)\mathsf{U}(r_i\wedge w_i\wedge\phi_{i+1}). \tag{36}$$

Now, define the policy $\pi[\psi]$ as

$$\pi_{[\psi]}(x_{0:k}) := \pi_{[\phi_1]}(x_{0:k}), \tag{37}$$

where we use the optimal policy for U ((26) or (27)), since ϕ_1 is a (deeply nested) U formula.

Theorem 7. *$\pi_{[\psi]}$ from Def. 14 is an optimal policy.*

VI. Optimal Policies for Compositions

In the previous sections, we have established optimal policies for U (Thm. 3 and 4), X, disjunctions and conjunctions with a single temporal operand (Thm. 5), for G (Thm. 6), and GU (Thm. 7). These correspond exactly to the items in the definition of the fragment $\mathcal{S}$ in Def. 3. Consequently, combining these results enables us to construct optimal policies for any formula in $\mathcal{S}$ via structural induction:

Theorem 8 (Optimal policies for $\mathcal{S}$)**.** *For $\varphi \in \mathcal{S}$, the policy $\pi_{[\varphi]}$ constructed with Thm. 3–7 is optimal (Def. 2) and satisfies* (8)*.*

A. Expanding $\mathcal{S}$ by Rewriting Conjunctions

A notable gap in the fragment $\mathcal{S}$ is the absence of conjunctions of multiple temporal operators, e.g., $\mathsf{F} r_1 \wedge \mathsf{F} r_2$. However, we can rewrite many such conjunctions into an equivalent single U formula, which is in $\mathcal{S}$ by **(S2)** [14]. Namely, [14, Lemma 7] implies that conjunctions of GU, U, and G formulas with propositional operands are equivalent to a formula in $\mathcal{S}$.

Corollary 2. *For finite sets $\mathcal{I}$ and $\mathcal{J}$, and $q_i, \mathsf{r}_i, q_j, r_j, q \in$ Prop, the formula*

$$\mathsf{p}_{\mathcal{I},\mathcal{J}} := \bigwedge_{i\in\mathcal{I}} \mathsf{G}(q_i \,\mathsf{U}\, \mathsf{r}_i) \;\wedge\; \bigwedge_{j\in\mathcal{J}} (q_j \,\mathsf{U}\, r_j) \;\wedge\; \mathsf{G}\, q \tag{38}$$

is equivalent to a formula that is in $\mathcal{S}$.

By Cor. 2, $\mathcal{S}$ contains all conjunctions of GU, U, and G formulas with propositional operands. Moreover, the recursive structure of **(S2)** and **(S3)** allows nesting, so $\mathcal{S}$ contains a wide variety of formulas beyond those covered by Cor. 2.

Remark 5. There are two limitations. (1) The absence of a general conjunction rule for two formulae. Formulas such as $\mathsf{FG}\, p_1 \wedge \mathsf{GF}\, p_2$ are not *syntactically* in $\mathcal{S}$. However, this is equivalent to single (albeit nested) F, i.e., $\mathsf{FG}\, p_1 \wedge \mathsf{GF}\, p_2 \;\equiv\; \mathsf{F}\big(\mathsf{G}(p_1 \mathsf{U} (p_2 \wedge p_1))\big)$, and it can be verified that the right-hand side belongs to $\mathcal{S}$. (2) Nesting of temporal operators within G is limited. For example, $\mathsf{G}\big(p_1 \vee \mathsf{F}(p_2 \wedge \mathsf{F} p_3)\big)$ cannot be rewritten into an equivalent formula in $\mathcal{S}$.

VII. Safety Filter

Define the failure set $\mathcal{F} \subset \mathcal{X}^+$ as the set of state prefixes from which the specification ψ cannot be satisfied with non-negative robustness score, i.e.,

$$\mathcal{F} := \{x_{0:k} \in \mathcal{X}^+ : \max_{a_{k:\infty}} \rho_{[\psi]}(x_{0:\infty}) < 0\}. \tag{39}$$

We now show that Q constitutes a safety monitor [21] for the TL formula ψ under the fallback policy $\pi_{[\psi]}$.

Lemma 9. *Under the fallback policy $\pi_{[\psi]}$, $Q_{[\psi]}$ is a safety monitor for the specification ψ, i.e.,*

$$Q_{[\psi]}(x_{0:k}, a_k) \geq 0 \implies x_{0:k} \notin \mathcal{F}. \tag{40}$$

Proof. This follows by definition of Q (5) and $\mathcal{F}$ (39). □

Consequently, we can construct a least-restrictive safety filter [21] which guarantees future satisfiability of ψ under any task policy $\tilde{\pi}$ by only intervening when $\tilde{\pi}$ would violate the specification:

Theorem 9. *Consider the following least-restrictive intervention scheme ϕ:*

$$\phi(x_{0:k}, a_k) = \begin{cases} a_k, & Q_{[\psi]}(x_{0:k}, a_k) \geq 0, \\ \arg\max_a Q_{[\psi]}(x_{0:k}, a), & \textit{otherwise}. \end{cases} \tag{41}$$

Then, ϕ is a safety filter [21], i.e.,

$$\begin{aligned} &Q_{[\psi]}(x_{0:k}, \pi_{[\psi]}(x_{0:k})) \geq 0 \\ \implies\; &Q_{[\psi]}(x_{0:k}, \phi(x_{0:k}, a)) \geq 0, \quad \forall a \in \mathcal{A}. \end{aligned} \tag{42}$$

This does not guarantee the system under the filtered controls satisfies ψ for the same reasons as Counterexample 1. To guarantee ψ, we need additional assumptions on ψ that prevent it from indefinitely delegating to the future.

Theorem 10. *Let ψ be a TL formula such that, along any trajectory $x_{0:\infty}$, $V^*_{[\psi]}(x_t) \geq 0$ for all $t \geq 0$ implies that $\rho_{[\psi]}(x_{0:\infty}) \geq 0$. Then, any trajectory $\bar{x}_{0:\infty}$ generated by the system under the intervention scheme ϕ will satisfy the specification ψ with non-negative robustness score, i.e., for any task policy $\tilde{\pi}$, we have $\rho_{[\psi]}(\bar{x}_{0:\infty}) \geq 0$, where*

$$x_{t+1} = f(x_t, \phi(x_{0:t}, \tilde{\pi}(x_{0:t}))), \quad \forall t \geq 0. \tag{43}$$

Proof. By the universal safety filter theorem [21], the safety filter ϕ guarantees that $V^*_{[\psi]}(x_t) \geq 0$ for all $t \geq 0$. Applying the assumption then completes the proof. □

Corollary 3. *The assumption in Thm. 10 is guaranteed to be satisfied if ψ does not contain U, or if all U operators in ψ are bounded, e.g., $\mathsf{U}_{[0,N]}$ for some $N < \infty$.*

VIII. Demonstrations

A. Two-Agent `GridWorld`: Filtering Collaboration

To showcase the previous results, we solve a two-agent `GridWorld` problem (4-dim.) for a complex specification. Namely, we solve the optimal policy $\hat{\pi}$, dubbed $\pi_{\textbf{spec}}$, associated with the decomposed value [14] for the specification below, and demonstrate (1) the ability of $\pi_{\textbf{spec}}$ to optimally guide the system towards satisfaction and (2) the ability of $\pi_{\textbf{spec}}$ to filter unsatisfactory nominal policies such that satisfaction is achieved regardless. In all scenarios, each agent is defined by a 2-dim. position and has actions $\{\leftarrow, \rightarrow, \uparrow, \downarrow, \emptyset\}$ which move it one space (or not at all). The value and policy are solved in the joint state-action space.

The `GridWorld` problem is defined by the specification

$$\begin{aligned} \psi_{\textbf{spec}} := \mathsf{F}_{[0,60]} r_{ws} \wedge (\neg r_{ws} \mathsf{U} r_g) \wedge \mathsf{FG} r_{ws} \wedge \\ \mathsf{G}(\mathsf{F} r_{wd} \wedge \mathsf{F} r_{sw}) \wedge \neg r_d \mathsf{U} r_k \wedge \mathsf{G} r_{safe}. \end{aligned} \tag{44}$$

This specification reads: Be in the worksite (r_{ws}) in **60** time units ($\mathsf{F}_{[0,60]} \equiv \top \mathsf{U}_{[0,60]}$ as in Sec. III-A) and eventually always be in the worksite but don't enter without gear (r_g). Eventually, loop wood (r_{wd}) and saw (r_{sw}). One agent gets the key (r_k) to open the doors (r_d), don't hit walls, and keep exactly 2 units between agents to avoid collision and losing communication ($r_{safe} := \neg r_{walls} \wedge r_{\|\cdot\|_1 = 2}$). We show a trajectory guided by $\pi_{\textbf{spec}}$ in the left of Fig. 3.

Next, we demonstrate Thm. 10, showing that $\pi_{\textbf{spec}}$ may serve as a safety filter for unsatisfactory user-defined policies, and do so in two cases: "**coffee**" and "**tea**". In each case, we generate two nominal policies (via the same approach), in which agent 1 takes action for $\psi_1 := \mathsf{F} r_{\text{coffee}} \wedge \mathsf{G} \neg r_{walls}$ or $\psi_2 := \mathsf{F} r_{\text{tea}} \wedge \mathsf{G} \neg r_{walls}$, and agent 2 takes action for $\psi_{\textbf{spec}}$, and the joint actions are filtered by $\psi_{\textbf{spec}}$. The rollouts for

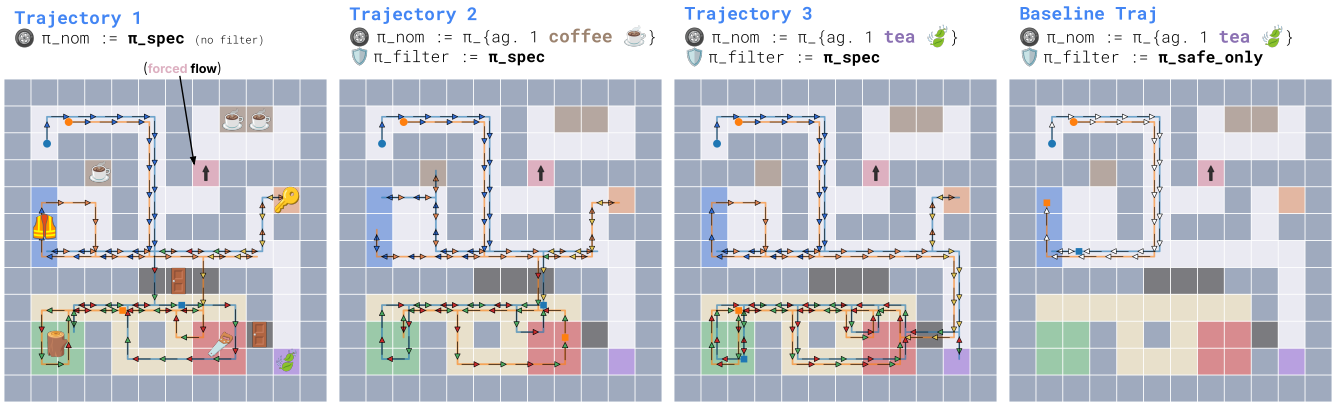

Fig. 3: **Demonstration of optimal policy $\hat{\pi}$, independently and as a filter:** Here, we demonstrate the optimal policy for the value for a given specification (top) in a two-agent `GridWorld` system. The trajectory for the policy alone is given in the left plot, while the two center plots demonstrate its usage as a filter for two nominal policies to get **coffee** and **tea** respectively. The plot on the far right demonstrates the problem with a "safety-only" filter here, which although safe is unable to yield a satisfactory trajectory.

the trajectory in either case can be seen in the center grids of Fig. 3. Lastly, we compare this approach with the standard HJ/CBF approach, in which the filter is only defined by the safety spec. $\psi_{\texttt{safe\_only}} := \mathsf{G} r_{safe}$. In this "baseline" case, despite agent 2 being defined by $\psi_{\texttt{spec}}$, without a safety filter for the joint system, deadlock arises between the two-agent policies and the system trajectory fails to satisfy $\psi_{\texttt{spec}}$, shown on the right side of Fig. 3.

B. Single-Agent Drone Safety Filter

We demonstrate the execution of the policy $\hat{\pi}$ in a single-agent scenario with a Crazyflie drone, navigating obstacles towards a worksite (Figure 1). The problem is modeled in the `GridWorld` dynamics, sending waypoints to the low-level controller that correspond to cell centers. Here, a reduced form of spec. (44) is used, namely $\psi_{\texttt{spec}} := \mathsf{F}_{[0,50]} r_{ws} \wedge \mathsf{G}(\mathsf{F} r_{wd} \wedge \mathsf{F} r_{sw}) \wedge \neg r_d \mathsf{U} r_k \wedge \mathsf{G} r_{safe}$. We use this to filter a nominal policy which solely tries to reach the worksite $\psi := \mathsf{F} r_{ws}$, and compare it with a filter with only the safety spec. $\psi_{\texttt{safe\_only}}$ and the unfiltered nominal policy (Fig. 1). Ultimately, the latter two get trapped or crash; only the trajectory filtered with $\psi_{\texttt{spec}}$ satisfies the spec.

IX. Conclusion

This work constructs optimal policies for a large class of TL formulas by recursively applying the optimal policies for simpler subformulas. We also show the optimal policy can be used as a safety filter to guarantee satisfaction of TL specifications under certain assumptions. Future work includes expanding the solvable fragment to include more general conjunctions and nested temporal operators and investigating game-theoretic extensions.

References

[1] Mitchell et al. "A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games". *IEEE TAC* (2005).

[2] Fisac et al. "Reach-avoid problems with time-varying dynamics, targets and constraints". *HSCC*. ACM. 2015.

[3] Fisac et al. "Bridging hamilton-jacobi safety analysis and reinforcement learning". *ICRA*. 2019.

[4] Hsu et al. "Safety and Liveness Guarantees through Reach-Avoid Reinforcement Learning". *RSS*. 2021.

[5] Ganai et al. "Learning Stabilization Control from Observations by Learning Lyapunov-like Proxy Models". *ICRA* (2023).

[6] So et al. "Solving Minimum-Cost Reach Avoid using Reinforcement Learning". *NeurIPS*. 2024.

[7] Bansal et al. "Hamilton-jacobi reachability: A brief overview and recent advances". *CDC*. 2017.

[8] Pnueli. "The temporal logic of programs". *SFCS*. IEEE. 1977.

[9] Donzé et al. "Robust satisfaction of temporal logic over real-valued signals". *FORMATS*. 2010.

[10] Fainekos et al. "Robustness of temporal logic specifications for continuous-time signals". *Theoretical Computer Science* (2009).

[11] Camacho et al. "LTL and beyond: Formal languages for reward function specification in Reinforcement Learning". *IJCAI*. 2019.

[12] Chen et al. "Signal temporal logic meets reachability: Connections and applications". *WAFR*. Springer. 2018.

[13] Sharpless et al. "Dual-Objective Reinforcement Learning with Novel Hamilton-Jacobi-Bellman Formulations". *ICLR*. 2026.

[14] Sharpless et al. "Bellman Value Decomposition for Task Logic ..." *arXiv preprint* (2026).

[15] Baier et al. *Principles of model ...* MIT press, 2008.

[16] Maler et al. "Monitoring temporal properties of continuous signals". *FTRTFT*. 2004.

[17] Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[18] Li et al. "Solving Reach-and Stabilize-Avoid Problems Using Discounted Reachability". *arXiv preprint* (2025).

[19] Alur et al. "Real-time logics: complexity and expressiveness". *Information and Computation* (1993).

[20] Vardi et al. "An automata-theoretic approach to automatic program verification". *LICS*. 1986.

[21] Hsu et al. "The safety filter: A unified view of safety-critical control in autonomous systems". *Annual Review of Control, Robotics, and Autonomous Systems* (2023).