Absolute Safety in Light Partially Observable Markov Decision Processes

May 1, 2019

1 Problem Description

The general problem that we seek to solve is a multi-stage decision problem where an action $a \in \mathcal{A}$ is chosen at each step. The system dynamics are determined by a function F, which maps a fully observable component of the state, $x \in \mathcal{X}$, the action, and a non-deterministic input, $d \in \mathcal{D}(x)$, to a new state[1]. That is, at each stage, k,

$$x_{k+1} = F(x_k, a_k, d_k). \tag{1}$$

The term *disturbance* will often be used for d in a broad sense that represents any process that is not determined from the state and action.

This problem is approached from two vantage points. From the safety standpoint discussed in Section 1.1, we seek to choose actions so that safety constraints will not be violated for any sequence of d's. From an optimization standpoint, we seek to maximize the expected reward if the d's are random variables with a known joint distributions. This optimization problem is formulated as a partially observable Markov decision process in Section 1.2.

We consider a special case where there is a static, but partially observable state component, θ, that affects the disturbance set. Specifically, $\mathcal{D}(x, \theta) \subseteq \mathcal{D}(x)$, is a set valued function of this partially observable state.

1.1 Safety (Discrete Time)

In this document, safety will be defined with an *instantaneous safety function*, $l(x)$, which maps every state in the fully observable state space, $\mathcal{X}$, to a real number. If $l(x) \geq 0$, then x is considered instantaneously safe. If $l(x) < 0$, then x is unsafe.

To guarantee future safety, a control policy that will keep x in an instantaneously safe region at all time in the future must exist. To reason about

[1] The term "state" is used loosely here. It should be noted that x is not the complete Markov state for the probabilistic problem. In most cases the context should resolve ambiguity, but, when necessary, the terms "fully observable state component", which is always denoted x, and "Markov state", which is always denoted s, will be used for clarity

this concept, we introduce the *safety value function* or simply *safety function*, L, which will allow us to classify states and actions as safe or unsafe. Intuitively, L represents the minimum safety margin for all times in the future. If $L(x, k, \tilde{\Theta}) \geq 0$, then it is possible to avoid instantaneously unsafe states in the future given that $\theta \in \tilde{\Theta} \subseteq \Theta$. In this case x is said to be a *safe state*.

Let the action be determined by a function $a_k = a(x_k, k)$ and the disturbance by $d_k = d(x_k, a_k, k) \in \mathcal{D}(x_k, \theta)$. The state-action safety function is defined as follows:

$$L\left(x_k, a_k, k, \tilde{\Theta}\right) = \sup_{a} \inf_{\theta \in \tilde{\Theta}, d} \min_{k \leq i \leq K} l(x_i) \tag{2}$$

where

$$x_{i+1} = F(x_i, a(x_i, i), d(x_i, a(x_i, i), i)) \quad \forall\, i \in \{k \,..\, K-1\}.$$

If $L\left(x_k, a_k, k, \tilde{\Theta}\right) \geq 0$, a_k is said to be a safe action at x_k. The state safety function is simply

$$L\left(x_k, k, \tilde{\Theta}\right) = \sup_{a_k \in \mathcal{A}} L\left(x_k, a_k, k, \tilde{\Theta}\right). \tag{3}$$

Calculating $L\left(x_k, a_k, k, \tilde{\Theta}\right)$ involves solving a Stackelberg game. Fortunately, in some cases, sufficient approximations may be calculated offline[2].

Is this general enough to encompass unknown obstacles?

correct? Can we say anything about the computational complexity of this game? Is it any easier than a POMDP?

1.2 Mixed Observability Markov Decision Process

From an optimization perspective, we assume that $\{d_k\}_{k=1}^{K}$ are random variables with a known joint distribution. Specifically, the nondeterministic dynamics are governed by a function G, so that

$$d_k = G(s_k, v_k), \tag{4}$$

where $s_k \in \mathcal{S}$ is the Markov state of the system (see Eq. (5)), and v_k is a random variable that is independent at each time step. The Markov state consists of the fully-observable state, x_k, a partially observable state $\eta_k \in \mathcal{H}$, and the partially observable static parameter $\theta_k \in \Theta$,

$$s_k = (x_k, \eta_k, \theta_k) \tag{5}$$

Crucially, the presence of the hidden state, (η, θ), which links d_k between timesteps, allows for *online learning* and the opportunity to make better decisions by gathering information.

[2] Unfortunately, it is not true in general that

$$L\left(x_k, a_k, k, \tilde{\Theta}\right) \geq \inf_{\theta \in \tilde{\Theta}} L\left(x_k, a_k, k, \{\theta\}\right).$$

I think...

The optimization objective is to maximize the expected value of a reward function, $\mathcal{R}$.

$$\text{maximize}\, \mathrm{E}\left[\sum_{k=1}^{K} \mathcal{R}(s_k, a_k)\right] \tag{6}$$

This problem is a partially observable Markov decision process (POMDP). More specifically, it belongs to the class of mixed observability Markov decision processes (MOMDPs). This MOMDP is defined as follows:

- The **state space** is $\mathcal{S} = \mathcal{X} \times \mathcal{H} \times \Theta$.
- The **action space** is $\mathcal{A}$.
- The **observation space** is $\mathcal{O}$. We assume that an observation $o_k \in \mathcal{O}$ contains perfect information about x_k, but may also contain information about η or θ.
- The **reward function** is $\mathcal{R}$.
- The **transition distribution** is implicitly defined by F, G, and a hidden state transition function, F_η, defined so that $\eta_{k+1} = F_\eta(s_k, a_k, w_k)$, where w_k is a random variable that is independent at each time step. The combination of these functions will be denoted with $\mathcal{T} = (F, G, F_\eta)$.
- The **observation distribution** is defined by $o_k = H(s_{k-1}, a_{k-1}, s_k, v_k)$. As stated above, we assume that o_k contains perfect information about x_k.

This POMDP will will be compactly represented as.

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{R}, \mathcal{T}, \mathcal{Z}). \tag{7}$$

The optimal solution to $\mathcal{M}$ is a policy, π^* that maps the history of previous actions and observations to the next action to take to optimize Eq. (6). There has been a great deal of research into solving POMDPs approximately both online and offline, and the optimal solution will automatically learn about the hidden state (η_k, θ). However, if the problem is solved as formulated above, the policy may enter unsafe states.

2 Safe MOMDPs

To guarantee safety, we can prune the action space of $\mathcal{M}$ to only include safe actions. At each state, x, let the safe action space be defined as

$$\mathcal{A}_{\text{safe}}(x, k, \tilde{\Theta}) = \left\{a \in \mathcal{A} : L\left(x, a, k, \tilde{\Theta}\right) \geq 0\right\}. \tag{8}$$

A new safe MOMDP, can be defined as

$$\mathcal{M}_{\text{safe}} = (\mathcal{S}, \mathcal{A}_{\text{safe}}, \mathcal{O}, \mathcal{R}, \mathcal{T}, \mathcal{Z}). \tag{9}$$

This problem has the safety property defined in Proposition 1.

Proposition 1. *If $L(x_0, 0) \geq 0$, then, for any policy for $\mathcal{M}_{safe}$, every state that can be reached is safe, i.e. $l(x_k) \geq 0 \, \forall \, k \in \{1..K\}$.*

TODO: write out proof. It is conceptually simple: No action can lead to a state where unsafe states can't be avoided in the future.

3 Learning about Safety

A key property of this formulation is that, by eliminating possible θ from Θ, the size of $\tilde{\Theta}$ can be reduced and $\mathcal{A}_{\text{safe}}$ enlarged. The optimal solution of $\mathcal{M}_{\text{safe}}$ will include active learning about this hidden state so that goals can be accomplished safely.

4 Continuous-time safety formulation

TODO

5 Potential objections

5.1 The optimal solution to the POMDP will automatically be safe if there is infinite cost assigned to instantaneously unsafe states.

- Yes, but it is hard to find optimal solutions to POMDPs.
- We don't usually know exact probability distributions, we may want to be safe with respect to disturbances that have zero probability.
- If the probability distributions are trained on data, we often will not see the worst-case d_k.