Negative Utilitarian Optimal Safety

by Sven Nilsen, 2018

Local Optimal Safety Theory (LOST) gives the local optimal decision theory relative to unsafe states assuming zero self-confidence. In this paper I generalize LOST to include utility maximization regarding safety with extra penalty for unsafe states. This extension is called “Negative Utilitarian Optimal Safety” or abbrivated NUOS. The difference in behavior is that NUOS is capable of rationally avoiding obstacles in the environment leading to greater safety. NUOS is Instrumentally Rational on its own, but can be composed with LOST under Zen Rationality for on/off self-confidence.

Local Optimal Safety Theory (LOST) is a simple decision theory that maximizes the unsafe minimum distance when unsafe, and minimizes the safe minimum distance when safe. Distance is measured in the number of steps required to reach a state. LOST is useful analytically because it can be used to extract exact solutions to some simple environments that require Zen Rationality (extended Instrumental Rationality with higher order goal reasoning). These environment assumes that the future decisions of the agent can not be trusted, but it still has to maximize the probability of success.

However, in complex environments, LOST is no longer rational, because it can not plan ahead and gets stuck in states of local maximum safety. For example, when a person runs from a fire, it is stupid to enter a closed labyrinth if there is a way around it. A LOST agent will run straight into the labyrinth.

The problem is that a LOST agent is not utilitarian. It has no sense of whether a given choice is stupid or smart in the long run, it only cares about whether that choice is further away from an unsafe state.

To fix this, one simply uses the distance as utility instead of the expected reward:

$E(x : [S]) = \sum i \{ U(x_i) \}$ The expected reward from a list of states

$$U(s : S) = \text{if } f(s) \{\text{steps}(f, c, n, s)\} \text{ else } \{P \cdot \text{steps}(\backslash(x) = !f(x), c, n, s)\}$$

Where `P` is an extra penalty for unsafe states (usually a big negative number).

The function `f` returns `true` for unsafe states and `false` otherwise, `c` gives a list of new states per state (the available choices), and `n` is a number that controls the bounded search for minimum distance through the function `steps`.

The probability associated with the reward is 100%, because there is no uncertainty about the distance.

LOST can solve problems requiring Zen Rationality but NUOS can not (it is Instrumentally Rational), because it assumes perfect self-control when cutting corners close to unsafe states. The more shallow the search is, the more the algorithm approximates LOST, but the deeper the search is, the more closeness to unsafe behavior is possible, which can be dangerous if the agent loses self-control.

An optimal safe agent who keeps self-control up to some bounded search and expects to lose all self-control, the zen rational solution is to swap NUOS with LOST. Since LOST can be swapped with any algorithm, one can swap back to NUOS. This means on/off self-confidence is solved by NUOS/LOST.